Michael Scharkow
Automatische Inhaltsanalyse und maschinelles Lernen

Automatische Inhaltsanalyse und maschinelles Lernen

Michael Scharkow

Dissertation, Universität der Künste Berlin, 2011

Cover-Illustrationen: *Robots* by Will Strong

Druck und Verlag: epubli GmbH, Oranienstraße 183, 10999 Berlin
www.epubli.de

ISBN 978-3-8442-1670-7

Danksagung

An erster Stelle möchte ich mich bei Monika Suckfüll und Hans-Jürgen Weiß bedanken, die sich nicht nur bereit erklärten, dieses Dissertationsprojekt zu betreuen, sondern mich auch in der Folge ermutigt und unterstützt haben, gleichermaßen kritisch wie entschlossen dieses Thema zu verfolgen.

Ganz herzlich danke ich zudem meinen Codierern Johanna Frey, Elisabeth Günther, David Maurer, Julia Piontek, Nora Riecker und Benjamin Scharkow. Dank gebührt auch Benjamin Fretwurst, Frank Marcinkowski, Torsten Maurer und Joachim Trebbe, die mir entweder persönlich Zugang zu ihren Codeplänen und Codieranweisungen gewährt oder dieselben sogar ganz offiziell publiziert haben, so dass ich in dieser Arbeit auf ihre reichhaltigen Kenntnisse zum Thema Inhaltsanalyse zurückgreifen konnte. Ich danke ebenso Christian Siefkes, der mich zuerst auf die Möglichkeiten des maschinellen Lernens für die quantitative Textanalyse hingewiesen hat.

Thorsten Quandt hat mich nicht nur ermuntert, diese Arbeit zügig fertigzustellen, sondern mir auch Zeit und Ressourcen dafür zu Verfügung gestellt, wofür ich mich herzlich bedanke. Carina Kordes und Alexander Ort haben die Arbeit schnell und gründlich korrekturgelesen. Ihnen danke ich ebenso wie meinen Kolleginnen und Kollegen an der Universität der Künste Berlin und der Universität Hohenheim, die mich in diesem Dissertationsvorhaben unterstützt haben.

Besonderer Dank gilt meinem wissenschaftlichen Sparringspartner Jens Vogelgesang, der diese Arbeit durch seine unermüdliche Hilfs- und Diskussionsbereitschaft nachhaltig geprägt hat. Einen besseren Mitstreiter in den Höhen und Tiefen der Forschungsarbeit kann man sich nicht wünschen.

Ohne die Geduld und Unterstützung von Antje Bretzmann würde es diese Arbeit nicht geben – danke für alles.

Stuttgart, im Dezember 2011

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

1.1 Problemstellung

Der Ausgangspunkt dieser Arbeit besteht nicht so sehr in einer wissenschaftlichen Frage- als vielmehr einer Problemstellung, die seit Jahrzehnten fast jede quantitative Inhaltsanalyse begleitet: Man kann als Forscher nie so viele Dokumente nach so vielen Kategorien codieren – bzw. codieren lassen – wie man eigentlich möchte. Der hohe Aufwand, der nicht nur mit der Entwicklung geeigneter Messinstrumente, sondern auch mit der eigentlichen Messung verbunden ist, führt vielfach zu einer forschungsökonomisch motivierten Reduktion des Stichproben- oder Codebuchumfangs. Gleichsam ist die quantitative Inhaltsanalyse, wie sie von Berelson (1952) und anderen gerade auch in Abgrenzung zu hermeneutischen oder anderen qualitativen Verfahren der Textanalyse (vgl. Ritsert, 1972) definiert wurde, auf die Quantität der realisierten Codierungen angewiesen. So basiert der Inferenzschluss, der sich auf den Entstehungs- oder Wirkungskontext einer Mitteilung bezieht (Krippendorff, 2004a), auf einer Aggregation von vielen Einzelmessungen, die für sich genommen kaum von Interesse sind (Früh, 2007, 63).

Methodologisch lässt sich die Bedeutung der Codierquantität in mindestens zweierlei Hinsicht begründen: Einerseits steht der realisierte Stichprobenumfang in direktem Verhältnis zur Generalisierbarkeit und Genauigkeit statistischer Inferenzschlüsse (Casella & Berger, 2002). Andererseits lassen sich komplexe theoretische Konstrukte umso zuverlässiger und valider messen, je umfangreicher das Messinventar ist bzw. je mehr Messgelegenheiten wahrgenommen wurden (Moosbrugger, 2007). Die Inhaltsanalyse profitiert folglich sowohl von einer Erhöhung des

Stichproben- wie des Variablenumfangs.[1] Die Frage ist nun, auf welche Weise sich der Umfang der Codierung steigern lässt, und eine mögliche Antwort lautet: Automatisierung.

Seitdem Computer auch für sozialwissenschaftliche Anwendungen verfügbar waren, haben sich immer wieder Forschergruppen mit den Möglichkeiten der Automatisierung von Inhaltsanalysen beschäftigt (Stone, 1997). Dabei stand zuerst nicht nur das Ziel im Mittelpunkt, möglichst viele Dokumente in möglichst kurzer Zeit zu codieren, sondern auch die Überlegung, dass eine automatische Analyse stets vollständig reliabel und damit replizierbar ist. Wenn es gelänge, substanzielle inhaltsanalytische Fragestellungen in maschinenlesbare Regelsätze zu überführen, würden sich gänzlich neue Möglichkeiten der Analyse ergeben.

Obwohl es in den folgenden Jahrzehnten sowohl bei der Verfügbarkeit digitaler Inhalte als auch in der Entwicklung von Hard- und Software enorme Fortschritte gegeben hat, ist die automatische Inhaltsanalyse eine Randerscheinung in der Kommunikationswissenschaft geblieben. Die Methodenentwicklung auf diesem Gebiet scheint mit dem Bedarf, der sich durch neue Fragestellungen und eine Vervielfachung des Untersuchungsmaterials ergibt, nicht Schritt zu halten. Selbst bei genauerer Lektüre der – zumeist sehr kurzen – Lehrbuchkapitel oder entsprechender Überblicksartikel, etwa von Züll & Alexa (2001), gewinnt man den Eindruck, dass sich im Großen und Ganzen seit den 1960er Jahren konzeptionell nicht viel auf dem Gebiet automatischer Inhaltsanalyse getan hat. Es dominieren schlagwortbasierte oder Co-Occurrence-Ansätze, die auf die Arbeiten von Stone et al. (1966) bzw. Iker & Harway (1969) zurückgehen, und an die in den letzten Jahrzehnten immer wieder – und zu Recht – dieselben Kritikpunkte gerichtet wurden.[2] Insbesondere ha-

1 Diese Feststellung gilt in gleichem Maße für andere standardisierte Datenerhebungsverfahren wie Befragung und Beobachtung, mit denen die Inhaltsanalyse nicht nur zahlreiche methodologische Grundannahmen teilt, sondern auch konkrete Strategien und Prozeduren im Forschungsprozess (Brosius et al., 2009).

2 Dies gilt im Prinzip auch für die stärker computerlinguistisch ausgerichteten Ansätze, die seit den 1980er Jahren von eine Forschergruppe in Amsterdam entwickelt werden (van Cuilenburg et al., 1988). Obwohl dort in den letzen Jahren große Fortschritte auf dem Gebiet der Aussagenanalyse gemacht wurden (Atteveldt, 2008), haben diese bislang recht wenig Resonanz im Fach erfahren.

ben die klassischen computergestützten Analyseverfahren den Nachteil, entweder nur mit großem Aufwand oder gar nicht mit konventionellen Operationalisierungs- und Codierstrategien vereinbar zu sein. Dies führt wiederum zu einer Auseinanderentwicklung statt einer Konvergenz verschiedener inhaltsanalytischer *Schulen*, was in einem kleinen Fach wie der Kommunikationswissenschaft letztendlich die Weiterentwicklung ihrer zentralen Methode (Scheufele & Engelmann, 2009, 146) eher schwächt denn stärkt (Früh, 2007, 8).

Zwei Überlegungen standen am Beginn dieser Arbeit: Erstens sind die Möglichkeiten der Automatisierung bei quantitativen Inhaltsanalysen bislang nicht ausreichend evaluiert worden, sowohl bei der eigentlichen Codierung als auch in den übrigen Teilen des Forschungsprozesses. Zweitens muss jede methodische Neu- und Weiterentwicklung anschlussfähig an die bisherige Forschungspraxis sein, um überhaupt die Chance zu haben, diese substanziell zu verbessern. Die Methodenentwicklung in anderen sozialwissenschaftlichen Disziplinen zeigt, dass die gezielte Automatisierung einzelner Forschungsschritte die Qualität und Effektivität der Erhebung und Analyse erheblich steigern kann. Als Beispiel sei nur der Einsatz von Computern bei adaptiven Tests (Van Der Linden & Glas, 2000) in der Erziehungswissenschaft und Psychologie genannt. Ein erstes Ziel dieser Arbeit ist es daher, die Frage zu diskutieren, welche Schritte im inhaltsanalytischen Forschungsprozess sich besonders für die Automatisierung eignen bzw. unter welchen Bedingungen sich überhaupt methodische Vorteile aus der Automatisierung ergeben.

In den folgenden Kapiteln geht es nicht nur um die grundsätzlichen Möglichkeiten der automatischen Inhaltsanalyse, sondern konkret um die Anwendung von Verfahren aus dem maschinellen Lernen für die Textcodierung. Dabei handelt es sich allgemein um statistische Algorithmen und deren Implementation in Computerprogrammen, die anhand von Beispieldaten komplexe Problemlösungen generieren und diese für die Verarbeitung neuer Daten nutzen. Nach Alpaydin (2008, xiii) ist maschinelles Lernen vor allem dann sinnvoll, „wenn keine menschliche Expertise verfügbar ist oder wenn Menschen nicht in der Lage sind, ihre Expertise zu erklären.“ Genau dies ist bei Inhaltsanalysen häufig der Fall: Es fällt den meisten Codierern nach dem Training leicht, einen Satz oder

ganzen Artikel der Kategorie *Umweltschutz* zu zuzuordnen, aber die Entwicklung diktionärbasierter Verfahren (vgl. Abschnitt 3.3.1) zeigt, dass es selbst erfahrenen Wissenschaftlern schwer fällt, dafür eine geeignete Wortliste zu erstellen, die gleichermaßen trennscharf und umfassend ist (Schönbach, 1982). Hier spielen induktive Klassifikationsverfahren ihre Stärken aus, die statt festen Regeln nur Beispieldokumente für alle Kategorien benötigen. Maschinelles Lernen hat bislang nur wenig Eingang in die Kommunikationswissenschaft gefunden (vgl. aber Atteveldt, 2008), so dass hier eine sozialwissenschaftliche Einordnung von Konzepten und empirischen Studien aus der Informatik unumgänglich ist.

Grundsätzlich soll im Folgenden die Frage untersucht werden, ob durch die Nutzung von Verfahren aus dem maschinellen Lernen zugleich valide, reliable *und* umfangreiche Analysen von Texten möglich werden. Anders formuliert: Kann Automatisierung das Problem der geringen Codier*quantität* lösen helfen und dadurch die *Qualität* inhaltsanalytischer Inferenzen verbessern? Um diese Frage zu beantworten, ist erstens eine konzeptionelle Auseinandersetzung mit dem analytischen Potenzial und den Grenzen automatischer Verfahren nötig. Zweitens muss empirisch geprüft werden, ob sich die Verfahren tatsächlich für den kommunikationswissenschaftlichen Forschungsalltag eignen. Diese Evaluation ist Kern des zweiten Teils dieser Arbeit, in dem neben der reinen Machbarkeit auch untersucht wird, von welchen Einflussgrößen die Qualität und Effektivität der automatischen Codierung abhängt.

Da eine solche Evaluationsstudie zwangsläufig in einem konkreten Anwendungskontext verankert sein muss, liegt der Schwerpunkt der Argumentation auf der thematischen Analyse von Online-Nachrichten. Dies ist vor allem der Tatsache geschuldet, dass Online-Inhaltsanalysen im Fach stetig an Relevanz gewinnen (Rössler, 2010), und dass viele Automatisierungsmöglichkeiten in diesem Kontext klarer zu Tage treten als etwa bei der Analyse von Printmedien. Da für die empirische Evaluation zuerst konzeptionelle und technische Voraussetzungen geschaffen werden müssen, wurde im Rahmen dieser Arbeit ein Softwarepaket entwickelt, mit dem sich manuelle und automatische Analysen von Online-Inhalten durchführen lassen (vgl. die Dokumentation in Anhang A).

Mit der Verknüpfung von methodologisch-konzeptionellen Überlegungen zur Automatisierung der Inhaltsanalyse und deren empirischer Evaluation möchte ich in dieser Arbeit der Aufforderung von Früh (2007, 293) folgen, nicht nur „die formalen Computerroutinen zur Bedeutungserkennung in Texten zu evaluieren, sondern auch die optimalen Einsatzgebiete von CUI [computerunterstützte Inhaltsanalyse, M.S.] und konventioneller Inhaltsanalyse gegeneinander abzugrenzen". Wenn dies gelingt, wäre nicht nur der methodologische Diskurs zu diesem Thema auf eine breitere empirische Basis gestellt, es ließen sich auch konkrete Handlungsoptionen für den inhaltsanalytischen Forschungsprozess ableiten.

1.2 Aufbau der Arbeit

Im folgenden Kapitel 2 werde ich den methodologischen Begründungszusammenhang für eine Automatisierung der Inhaltsanalyse diskutieren. Dabei geht es in einem ersten Schritt um den Stellenwert der *Quantität* für die quantitative Analyse von Mitteilungen. Aus diesem lässt sich, wie ich zeigen werde, ein nicht nur forschungsökonomisch begründbarer Bedarf an Automatisierung der Arbeitsschritte herleiten, der sich in den Begriffen *Skalierbarkeit* und *Reproduzierbarkeit* zusammenfassen lässt (vgl. Franzosi, 1995). Kapitel 3 bietet einen historischen und systematischen Überblick über computergestützte Verfahren der Inhaltsanalyse. Ein Vergleich der Vor- und Nachteile dieser Verfahren führt schließlich zur Forschungsfrage, inwieweit die überwachte Textklassifikation als Best-Practice-Verfahren für die Verknüpfung manueller und automatischer Ansätze bei der Analyse sehr umfangreicher Stichproben und der Erprobung umfangreicher Codebücher gelten kann.

In Kapitel 4 geht es um die Herausforderungen und Problemfelder der Automatisierung von Inhaltsanalysen mit Verfahren aus dem maschinellen Lernen. Dabei werde ich neben der computergestützten Datenerhebung von Online- und klassischen Medieninhalten sowie deren Archivierung und Bereinigung auch auf die spezifischen Anforderungen der überwachten Textklassifikation eingehen, die sich aus der Verwendung manueller Codierungen ergeben. Dies betrifft einerseits den Prozess

des maschinellen Lernens selbst als auch die Evaluation der Codier- bzw. Klassifikationsqualität.

Die Frage, ob und wie sich Verfahren aus dem maschinellen Lernen tatsächlich für die sozialwissenschaftliche Inhaltsanalyse eignen, wird in einer umfangreichen experimentellen Evaluationsstudie untersucht, die in den Kapiteln 5 bis 7 dargestellt ist. Um das Potential der überwachten Textklassifikation zu untersuchen, wird anhand einer umfangreichen Stichprobe von Online-Nachrichten eine Inhaltsanalyse durchgeführt, die sich auf dokumentierte Codebücher aus der Programm- und Nachrichtenwertforschung stützt. Um das Verfahren zu evaluieren, werden zunächst alle Dokumente manuell codiert, um daraus anschließend Trainings- und Testdaten für die automatische Klassifikation zu generieren. Die Evaluation besteht dabei aus zwei Teilstudien, die sich unterschiedlichen Forschungsfragen bzw. Hypothesen widmen: In Teilstudie 1 geht es um die grundsätzliche Bestimmung der automatischen *Klassifikationsqualität* sowie um die Frage, von welchen Faktoren diese Qualität abhängt. Teilstudie 2 zielt dagegen auf die Frage der *Klassifikationseffektivität* ab, d.h. wie viele manuelle Codierungen nötig sind, um eine bestimmte automatische Klassifikationsqualität zu erreichen, und wie sich dieser Trainingsprozess optimieren lässt. Die Ergebnisse beider Teilstudien werden abschließend zusammengefasst und hinsichtlich ihrer Konsequenzen für die Frage der Automatisierung von Textanalysen durch maschinelles Lernen diskutiert.

Im letzten Kapitel der Arbeit werden die konzeptionellen Grundlagen und die Ergebnisse der empirischen Evaluation nochmals reflektiert. Aus diesen Überlegungen ergibt sich nicht nur die Antwort auf die Frage dieser Arbeit, wie und ob sich Inhaltsanalysen durch maschinelles Lernen sinnvoll automatisieren lassen, sondern auch konkrete Empfehlungen für den Umgang mit automatischen Verfahren und deren Weiterentwicklung. Nicht zuletzt soll es dabei auch um die Reichweite der empirischen Ergebnisse gehen, die sich aus den inhaltlichen und methodischen Einschränkungen der Arbeit ergeben.

2 Methodologische Herausforderungen quantitativer Inhaltsanalysen

2.1 Grundlagen

Die sozialwissenschaftliche Inhaltsanalyse, wie sie in dieser Arbeit verstanden wird, basiert auf der Prämisse, dass sich durch die systematische Untersuchung von Kommunikation Rückschlüsse auf individuelle und gesellschaftliche Phänomene ziehen lassen, die jenseits des konkreten Kommunikats liegen (Merten, 1995; Mayntz et al., 1974, 151). Die Beschreibung von verbalen und anderen Mitteilungen in der öffentlichen und privaten Sphäre ist daher kein Selbstzweck, sondern dient letztlich der Analyse von individuellen Einstellungen und sozialen Strukturen und Prozessen. Nicht zuletzt weil viele inhaltsanalytisch erfassbare Phänomene den zentralen Gegenstand der Kommunikationswissenschaft darstellen, hat sich spätestens seit den 1970er Jahren die Inhaltsanalyse als deren zentrales Erhebungsverfahren bewährt (Brosius et al., 2009, 139). Dabei ist die methodologische Weiterentwicklung zwangsläufig eng verbunden mit der Theorieentwicklung des Faches sowie der Veränderung von Medienangeboten und Kommunikationskanälen. Die Inhaltsanalyse als zentrale Methode der Kommunikationswissenschaft kann aus zwei Blickwinkeln kritisiert und letztlich weiterentwickelt werden, die sich zwar gegenseitig bedingen, im Folgenden jedoch zuerst einzeln diskutiert werden: eine methodologische und eine eher fachwissenschaftliche Perspektive.

Um die Folgen dieser Interdependenz von Gegenstand, Theorie und Methode besser zu verstehen, ist zuerst ein Blick auf die Definition inhaltsanalytischer Forschung notwendig. Krippendorff (2004a) widmet dieser Definitionsarbeit, die auch Jahrzehnte nach den Arbeiten von Lasswell et al. (1952) und Berelson (1952) nicht abgeschlossen ist, ein umfangreiches Kapitel und schlägt selbst folgende kurze Definition der

Inhaltsanalyse vor, die einige höchst relevante Anknüpfungspunkte für methodologische Überlegungen bietet:

> Content analysis is a research technique for making replicable and valid inferences from text (or other meaningful matter) to the contexts of their use. (Krippendorff, 2004a, 18)

Aus methodologischer Perspektive ist es demnach Aufgabe der inhaltsanalytischen Methodenentwicklung, das Verfahren daraufhin zu untersuchen, ob und wie sich für beliebige Fragestellungen und Inhalte gewährleisten lässt, dass die Inferenzschlüsse auf Basis der Codierung von Mitteilungen größtmögliche Zuverlässigkeit und Gültigkeit besitzen. Ein zentrales Kriterium für jede empirische Studie ist dabei die Transparenz der Messung und Replizierbarkeit der Ergebnisse. Auch wenn sozialwissenschaftliche Inhaltsanalysen bislang selten so folgenreich sind wie standardisierte Leistungstests oder klinische Studien, für die es längst eigene Fachzeitschriften fest definierte Qualitätskriterien (American Educational Research Association, 1985; ICH, 1996; Deutsche Forschungsgemeinschaft, 1999) gibt, müssen doch die Ansprüche an die Methodenentwicklung dieselben sein. In dieser Hinsicht ist die methodologische Diskussion zum Einsatz quantitativer Inhaltsanalysen erst einmal fach- und gegenstandsunabhängig. Nicht von ungefähr wurden und werden wichtige Impulse zur inhaltsanalytischen Forschung zu großen Teilen aus Nachbardisziplinen wie der Soziologie (Popping, 2000), Politikwissenschaft (Schrodt & Donald, 1990) oder Psychologie (Gottschalk, 2000) gesetzt.

Trotzdem gilt es, zumal in dieser Arbeit, die Fachperspektive der Kommunikationswissenschaft auf die Inhaltsanalyse nicht zu übergehen. Wichtige Anstöße für Methodeninnovationen sind nicht aus methodologischem Selbstzweck heraus entstanden, sondern weil sich die Rahmenbedingungen – neue Fragestellungen und Theorien, neue Forschungsgegenstände, neue Kommunikationskanäle – geändert haben, unter denen inhaltsanalytische Studien konzipiert und durchgeführt werden. Noch vor 25 Jahren bestand die Medienstichprobe einer Fernsehprogrammanalyse aus einer Handvoll Sendern, während die Email als

Kommunikationsmedium noch nicht einmal auf dem Radar der Kommunikationswissenschaft erschienen war. Doch nicht nur der Gegenstand beeinflusst Methodenwahl und -entwicklung, auch neue Fragestellungen und Theorien. So hat etwa die Framing-Forschung im letzten Jahrzehnt nicht nur dafür gesorgt, dass viele klassische Rezeptionsforscher nun umfangreiche Inhaltsanalysen durchführen, sondern auch zu einem lebhaften methodischen Diskurs über die Messung von Medienframes geführt (Scheufele, 2003; Matthes, 2007; Matthes & Kohring, 2008). Um auf die o.g. Definition von Krippendorff zurückzukommen: Es wandeln sich die Phänomene, auf die inhaltsanalytische Referenzen abzielen, weil sich die untersuchten Mitteilungen und ihre Entstehungs- oder Nutzungskontexte wandeln. Dies erzeugt eine Nachfrage nach methodischer Weiterentwicklung und Innovation der Inhaltsanalyse.

In dieser Arbeit geht es nun um einen spezifischen Aspekt der quantitativen Inhaltsanalyse – die Relevanz der *Quantität* der Codierung, deren Implikationen und Konsequenzen. Der Quantitätsbegriff kommt in der methodologischen Diskussion zur Inhaltsanalyse zumeist nur am Rande vor, und dann häufig nur aus forschungsökonomischer Sichtweise.[1] Seine Implikationen sind jedoch so weitreichend, dass ich diese im Folgenden ausführlicher diskutieren will, sowohl aus methodologischer als auch aus kommunikationswissenschaftlicher Perspektive. Die zentrale Argumentation dieses Kapitels lässt sich wie folgt kurz zusammenfassen: Die bedeutendste aktuelle Herausforderung inhaltsanalytischer Methodenentwicklung liegt in der Skalierbarkeit (Verhalten des inhaltsanalytischen Verfahrens bei wachsender Menge analysierbarer Codierdaten). Dies ist sowohl aus methodologischer Perspektive, im Sinne der Steigerung der Zuverlässigkeit und Validität der Inferenzen, wünschbar, als auch notwendig, um viele aktuelle Fragen der Kommunikationswissenschaft, etwa im Bereich der Online-Kommunikation, überhaupt sinnvoll beantworten zu können. In den meisten Fällen lässt sich die Quantität der Codierdaten nur durch den Einsatz automatischer Verfahren steigern und gleichzeitig die Reproduzierbarkeit der Ergebnisse sichern. Automatisierung ist daher eine denkbare und plausible Lösung für viele forschungsprakti-

1 So betont Rössler (2005) in seinem Lehrbuch, dass es bei der Inhaltsanalyse um eine *große Zahl von Botschaften* [Hervorhebung im Original] geht.

sche und methodische Probleme umfangreicher Inhaltsanalysen, auch wenn dadurch ggf. andere Probleme entstehen. In den nächsten beiden Abschnitten möchte ich diese Argumente ausführen, um anschließend für eine neuerliche Auseinandersetzung mit automatischen Verfahren der Inhaltsanalyse zu plädieren.

2.2 Relevanz der Codierquantität für die Inhaltsanalyse

2.2.1 Methodenperspektive: Qualitität der Inferenzen

Einer der wichtigsten Gründe, warum jedes inhaltsanalytische Forschungsprojekt nicht nur von einer Steigerung der Codierqualität, sondern auch der Codierquantität profitiert, liegt in der simplen Tatsache, dass Inferenzen, die auf mehr Informationen aufbauen, zuverlässiger und valider sind als solche, die auf weniger Informationen aufbauen. Konkret lässt sich dies aus aus der modernen Messtheorie und den Annahmen statistischer Inferenz ableiten. Zunächst lege ich der Inhaltsanalyse das einfache Kommunikationsmodell nach Früh (2007) zugrunde, nach dem ein Kommunikator eine Mitteilung konstruiert, aus der dann der Rezipient (und damit auch der Codierer) die Aussageabsicht des Kommunikators rekonstruiert (vgl. Abbildung 2.1).

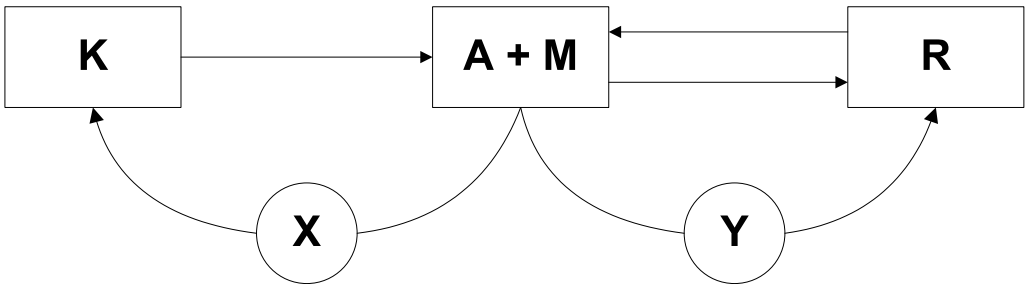

Abbildung 2.1: Einfaches Kommunikationsmodell nach Früh (2007, 43)

Die erste mögliche Inferenz jeder Inhaltsanalyse bezieht sich nun auf den mit X bezeichneten Pfad, d.h. von der vorliegenden Mitteilung (A+M) auf die Kommunikationsintention des Urhebers K.[2] Die zweite mögliche Inferenz, die im Modell mit Y bezeichnet ist, bezieht sich auf den Rezipienten. Da dieser Inferenzschluss nur auf Basis einer Inhaltsanalyse m.E. deutlich schwieriger realisierbar ist, konzentriere ich mich hier zunächst nur auf den ersten Schluss auf den Kommunikator. Das einfache Modell von Früh lässt sich nicht nur theoretisch erweitern, etwa um Kontextvariablen, sondern auch konzeptionell an die Grundannahmen quantitativer Datenerhebung und -analyse rückbinden. Hierfür bietet sich das Prozessmodell stochastischer Textgenese und -codierung von Benoit et al. (2009b) an, das sich detailliert auf das Verhältnis von Kommunikator, Aussage und „gelenkter Rezeption" (Wirth, 2001) in der Inhaltsanalyse bezieht. Obwohl das in Abbildung 2.2 dargestellte Modell primär auf die Analyse von Wahlprogrammen in der Politikwissenschaft abzielt, lässt es sich problemlos für jede sozialwissenschaftliche Fragestellung generalisieren.

Das Prozessmodell geht von einer latenten, d.h. unbeobachtbaren Einstellung μ des Kommunikators aus, die mit Hilfe einer Inhaltsanalyse rekonstruiert werden soll. Dies ist jedoch nur eine mögliche Art von Inferenz, die sich aus einem diagnostischen Erkenntnisinteresse (Früh, 2007, 44) ergeben kann. Ebenso ist es möglich, gleich auf der zweiten Stufe des Prozessmodells anzusetzen, in der der Kommunikator eine latente Aussage π entwickelt, die er mitteilen möchte. Die intendierte Aussage wird dann in einem *stochastischen* Prozess in eine konkrete Mitteilung transformiert. Dieser Schritt stellt nach Benoit et al. (2009b) die erste methodologische Herausforderung für die Inhaltsanalyse dar: Jede vorliegende Mitteilung, ob sprachlich oder mit anderen Zeichensystemen generiert, ist das Produkt eines nicht-deterministischen Prozesses und hätte bei gleicher Aussageintention auch eine andere konkrete Gestalt annehmen können. Einfach ausgedrückt gibt es unendlich viele Arten,

2 Der Begriff Intention ist allerdings insofern nicht ganz zutreffend, als dass auch unbeabsichtigte Aspekte des Kommunikatorverhaltens Ziel der Inferenz sein können, d.h. die rekonstruierte Aussage nicht zwangsläufig bewusst getätigt worden ist. Dies ist u.a. bei der Anwendung von Inhaltsanalysen in der psychologischen Diagnostik der Fall (Gottschalk, 2000).

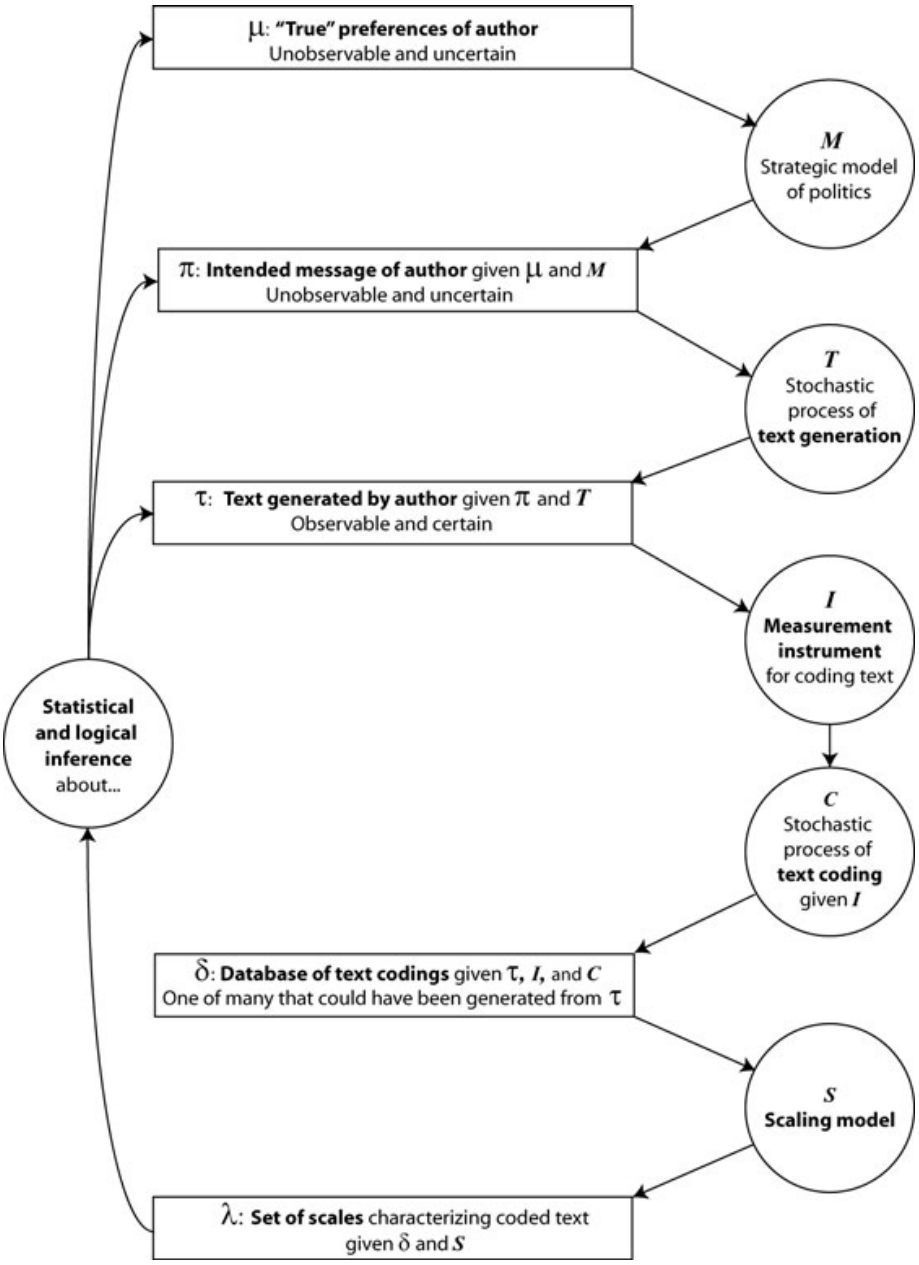

Abbildung 2.2: Prozess der stochastischen Textgenese und -codierung nach Benoit et al. (2009b, 498)

dieselbe Mitteilung zu kommunizieren, so dass jede Inferenz die Unsicherheit und Ungenauigkeit der Text*genese* berücksichtigen muss. Die einzelne manifeste Mitteilung τ ist also nur eine Realisation aus einer Grundgesamtheit möglicher Mitteilungen.

Die zweite Ursache von Unsicherheit und Ungenauigkeit liegt im Codierprozess selbst, konkret in der Auswahl des Messinstruments I und dem Akt der eigentlichen Codierung C. Um zu verstehen, worin die Vorteile einer hohen Zahl von Codierungen liegen, ist ein Blick auf die klassische Testtheorie hilfreich: Danach steigt erstens die Reliabilität eines Tests mit der Testlänge (Moosbrugger, 2007), zweitens können mehrere unterschiedliche Messungen ein komplexes latentes Konstrukt wie eine Aussage vielfach besser abbilden als ein einzelnes Item (Nunally & Bernstein, 1978; Lewis-Beck et al., 2004, 673). Weber (1983) hat erstmals auf das Potential von Messmodellen für die Inhaltsanalyse hingewiesen. Allerdings konzentriert er seine Argumentation auf die Wortebene. Konzeptionell spricht jedoch nichts außer Kostengründen dagegen, auch bei konventionellen Analysen auf Textebene mit mehreren Indikatoren pro Konstrukt zu arbeiten.

Ebenso wie die Verwendung von mehreren Operationalisierungstrategien bei der Codierung desselben latenten Inhalts ist auch die Mehrfachcodierung ein erprobtes Mittel zur Qualitätsverbesserung von Inhaltsanalysen. Einerseits ist die Messung valider, weil mehrere Codierer ihr Vorwissen und ihre Rezeptionskompetenz zur Anwendung bringen können, andererseits ist auch die Reliabilität höher, weil sich die zufälligen Messfehler einzelner Codierer insgesamt aufheben. Schließlich hat die Mehrfachcodierung auch den methodischen Vorteil, Reliabilitätswerte anhand der Normalcodierung berechnen sowie den Prozess der Codierung statistisch modellieren zu können (vgl. Scharkow, 2010a). So kann Carpenter (2008) zeigen, dass mit Hilfe mehrerer parallel arbeitender Codierer entweder per einfacher Mehrheitsentscheidung oder unter expliziter Berücksichtigung individueller Codiererfähigkeiten und -präferenzen die Gesamtreliabilität der Codierung gesteigert werden kann.

Zusammenfassend lässt sich festhalten, dass jede Vergrößerung des bei Benoit et al. (2009b) mit δ bezeichneten Datensatzes mit Codierungen die Unsicherheit und Ungenauigkeit der Inferenzen auf die latente

Kommunikationsintention zu reduzieren hilft. Wer aus einer Stichprobe codierter Mitteilungen allgemein gültige Schlüsse auf ihren Entstehungs- oder Nutzungskontext ziehen will, profitiert daher gleich in dreifacher Hinsicht von einer höheren Quantität an Codierungen, die sich aus dem Produkt von Mitteilungen $\times$ Variablen im Codebuch ergibt.

Erhöhung der Reichweite der Inferenzen
Die offensichtlichste Konsequenz einer geringen Anzahl von Codierungen besteht in der Reichweite der Ergebnisse. Obwohl die Stichprobenziehung bei Inhaltsanalysen ohnehin ein Problem darstellt, da zumeist keine repräsentative Auswahl möglich oder umsetzbar ist (vgl. Gehrau et al., 2005), bleibt bei einer kleinen Stichprobe oft unklar, inwiefern etwa beobachtete Differenzen wenigstens für die untersuchten Medien gelten. Dies gilt z.B. bei längsschnittliche Fragestellungen, für die oft nur sehr wenige Messzeitpunkte zur Verfügung stehen, aus denen dann aber Trends über einen längeren Zeitraum abgeleitet werden sollen. So muss sich die umfangreiche Studie von Bruns & Marcinkowski (1997) bzw. Marcinkowski et al. (2001) auf vier Messzeitpunkte beschränken, die allerdings einen Zeitraum von über 20 Jahren abbilden sollen. Ähnliches gilt für die Studie von Wessler (2008) zur Berichterstattung über die Europäische Union.

Reduktion des Stichprobenfehlers (Sicherheit)
Eng mit dem oben genannten Argument verbunden ist die zweite Konsequenz umfangreicher Codierungen: Der Stichprobenfehler wird mit zunehmender Anzahl der Codierungen kleiner, die Inferenzen damit genauer. Gerade in den Sozialwissenschaften, in denen eher kleine statistische Differenzen und Zusammenhänge die Regel sind, können vielfach Hypothesen nur mit ausreichender statistischer Power getestet werden, die sich wiederum direkt aus dem Stichprobenumfang ergibt.

Reduktion des Messfehlers (Genauigkeit) und Validität
Selbst bei gegebenem Stichprobenumfang lässt sich mit einer Mehrfachcodierung mit unterschiedlichen oder gleichen Instrumenten die statistische Inferenz verbessern, weil sich dadurch die Reliabilität der Messung erhöht. Da bei korrelativen Analysen die Zusammenhänge zwischen

messfehlerbehafteten Variablen verzerrt sind, profitiert jede Inhaltsanalyse von einer höheren Reliabilität der Codierung (vgl. Abschnitt 4.4). Die Verwendung von mehreren Variablen zur Messung eines komplexen Konstrukts ist zudem häufig valider als eine Einfachmessung.

Die oben genannten Gründe sprechen dafür, bei jeder Inhaltsanalyse nicht nur in die Qualität, sondern auch in die Quantität der Codierung zu investieren. Aus dieser Perspektive ist das konventionelle Vorgehen, bei dem Beiträge aus *wenigen* Medienangeboten und *wenigen* Messzeitpunkten von genau *einem* Codierer anhand genau *eines* Indikators, d.h. einer Codieranweisung pro theoretischem Konstrukt, codiert werden, sicher nicht optimal. Allerdings werden die meisten Forschungsleiter zu Recht darauf hinweisen, dass eine Inhaltsanalyse selbst zu diesen Bedingungen mit hohen Kosten verbunden ist, und eine suboptimale Messung besser als keine Messung ist. Trotzdem bleibt es eine zentrale methodologische Herausforderung für die Inhaltsanalyse, Strategien für eine quantitative Ausweitung der Datenbasis bei gleichbleibendem oder sogar geringerem Aufwand zu entwickeln.

2.2.2 Fachperspektive: Forschungsfragen- und gegenstände

Neben den oben vorgestellten methodologischen Argumenten gibt es in der Kommunikationswissenschaft zahlreiche fachliche Gründe, die gesteigerte Anforderungen an den Umfang von Inhaltsanalysen stellen. Exemplarisch möchte ich nur auf zwei aktuelle Entwicklungen im Fach eingehen, die das Problem illustrieren: Erstens das zunehmende Interesse an (halb-)öffentlicher interpersonaler Kommunikation im Internet, die einen nahezu unerschöpflichen Strom an leicht erheb- und schwer analysierbaren Daten produziert (Rössler, 2010). Zweitens die Verknüpfung von inhaltsanalytischen und Befragungs- oder Beobachtungsdaten auf der Ebene einzelner Rezipienten (Wolling, 2002). Im ersten Fall wird die Weiterentwicklung der Methode durch neue Formen und Quantitäten von Mitteilungen stimuliert, in letzterem durch immer spezifischere Fragestellungen auf dem Gebiet der Nutzungs- und Rezeptionsforschung. Die methodischen Implikationen dieser Entwicklungen möchte ich hier nur kurz skizzieren.

Herausforderungen durch Online-Inhalte

Spätestens seit dem Aufsatz von Morris & Ogan (1996) ist die Bedeutung der Analyse von Online-Kommunikation für das Fach praktisch unumstritten. Im Internet lassen sich unzählige Formen synchroner und asynchroner, interpersonaler und öffentlicher Kommunikation beobachten, die inhaltsanalytische Zugänge für die Überprüfung alter und neuer Theorien geradezu herausfordern. Trotzdem kann die Methodenentwicklung nur schwer mit dem technischen, kulturellen und sozialen Wandel durch und mit Hilfe von Online-Kommunikation mithalten. Dies liegt nicht nur, aber sicher auch an der Tatsache, dass die methodologische Reflexion und Weiterentwicklung der Inhaltsanalyse vor allem auf die Untersuchung klassischer Print- und Rundfunkmedien ausgerichtet war und ist. Hier haben sich durch die Erfahrungen vieler Forschergruppen, etwa in der Programmforschung (Weiß, 1998) oder der politischen Kommunikationsforschung (Pfetsch, 2004; Wessler, 2008), Standards und vielfach bewährte Verfahren in der Stichprobenziehung, Codebuchentwicklung und Feldarbeit etabliert. Nicht nur angesichts der Multimedialität, Flüchtigkeit und Dynamik von Online-Kommunikaten, sondern aufgrund ihres bloßen Umfangs sind nur wenige bewährte Forschungspraktiken ohne weiteres einsetzbar. Selbst umfangreiche Analysen der üblichen Handvoll Qualitätszeitungen oder Fernsehvollprogramme erfordern nur einen Bruchteil der Standardisierung und des Projektmanagements, das für die Analyse auch nur eines kleinen Ausschnitts der Blogosphäre, von Twitter-Mitteilungen oder Webforen notwendig ist.

Selbst wenn man sich nur auf die journalistischen Angebote im World Wide Web konzentriert, stellt allein der Umfang der Medienstichprobe eine Herausforderung für die Forschung da. In einer aktuellen Studie zählen Neuberger et al. (2009) über 500 überregionale deutschsprachige journalistische Angebote im Internet, wobei ein vergleichsweise strenges Kriterium verwendet wurde. Hinzu kommt eine mindestens sechsstellige Zahl an deutschprachigen Blogs, die zumindest quasi-journalistischen Charakter haben, d.h. ein Massenpublikum ansprechen (wollen) (vgl. Busemann & Gscheidle, 2010). Selbst wenn man, wie dies Rössler (2010) für alle neuen Kommunikationsformen konstatiert, zunächst nur an einfachen, strukturbeschreibenden Inhaltsanalysen im Rahmen einer öffent-

lichkeitsbasierten Online-Programmforschung (Zeller & Wolling, 2010) interessiert ist, erfordert dies selbst unter Verwendung von Stichproben eine große Zahl an Untersuchungseinheiten. Verlässt man jedoch das klassische Feld öffentlicher, journalistisch orientierter Medienangebote, stößt man erst recht auf neue Herausforderungen: Zu Beginn des Jahres 2011 wurden auf der größten Social-Network-Plattform Facebook täglich eine Millarde neuer Inhalte, d.h. Bilder, Postings, Kommentare oder Mitteilungen erstellt. Ein durchschnittlicher Nutzer produziert im Monat über 90 solcher Mitteilungen.[3] Auf der Mikroblogging-Plattform Twitter werden täglich 140 Millionen Kurzmitteilungen erstellt, die Nutzerzahlen wachsen um 500.000 User pro Tag.[4] Ähnliche Größenordnungen an privaten oder (halb-)öffentlichen Mitteilungen werden auf vielen anderen Onlineangeboten erreicht, egal ob es um Kommentare zu Youtube-Videos oder Produktempfehlungen bei Amazon geht. Natürlich lassen sich auch hier Stichproben ziehen, so dass nicht zwangsläufig riesige Datenmengen analysiert werden müssen. Allerdings ist gerade der *long tail* dieser Mitteilungen, d.h. randständige und selten diskutierte Themen, für die Forschung interessant, weil sich so erstmals öffentliche und vor allem interpersonale Kommunikation zu Spezialthemen analysieren lässt. Dies wiederum erfordert häufig das Sammeln und (Vor-)Codieren großer Mengen von Mitteilungen.

Die bloße Quantität analysierbarer Online-Inhalte ist jedoch nicht das einzige Forschungsproblem der Online-Inhaltsanalyse. Angesichts des vergleichsweise jungen Mediums und der Tatsache, dass bisherige kommunikationswissenschaftliche Inhaltsanalysen zumeist an öffentlichen, d.h. publizistischen Mitteilungen interessiert waren, sind sowohl die theoretische als auch die methodisch-konzeptionelle Basis für die Analysen solcher Inhalte eher unterentwickelt (Rössler, 2010). Selbst wenn man repräsentative und große Stichproben von Facebook- oder Twitter-Mitteilungen erheben könnte, gibt es bislang schlicht kaum erprobte Instrumente, die die immense thematische und sprachliche Breite sowohl öffentlicher als auch interpersonaler Kommunikate abbilden können. Dementsprechend orientieren sich bisherige Analysen vor allem an tradi-

3 Angaben des Anbieters unter http://www.facebook.com/press/info.php?statistics.
4 Angaben des Anbieters unter http://blog.twitter.com/2011/03/numbers.html.

tionell publizistischen Ansätzen, etwa die Untersuchung deutschsprachiger Twitter-Inhalte von Neuberger et al. (2010). Um die inhaltliche Vielfalt dieser Online-Kommunikate zuverlässig und valide zu messen, ist eine entsprechend breite Operationalisierung notwendig. Wenn schon die Codebücher klassischer Programmanalysen zum Teil dutzende Themenvariablen bzw. -kategorien enthalten, wird eine entsprechende Themenanalyse von nutzergenerierten Inhalten nochmals deutlich umfangreichere Codierungen erfordern.

Schließlich ist wiederum im Anschluss an Rössler (2010) festzuhalten, dass es in der Kommunikationswissenschaft an systematischen Methodenexperimenten mangelt, die die Wirksamkeit etwa bestimmter Stichprobenverfahren, Operationalisierungs- oder Codierungsstrategien für die Analyse von Online-Kommunikation untersuchen (vgl. Scharkow, 2010a). Wenn es gelingt, die Codierquantität bei gleichem Aufwand zu erhöhen, erlaubt dies dem Forscher, auch einmal beim Ausprobieren neuer Codepläne oder Codierstrategien zu scheitern. Dies führt zumeist – nicht zwangsläufig und nicht als einziger Weg – zu einer Steigerung der Qualität der Forschung. Betrachtet man inhaltsanalytische Arbeit als kumulativ, kann man in jedem Fall aus eigenen und fremden Erfahrungen mit neuen Forschungsgegenständen und Verfahren lernen.

Inhaltsanalysen in der Rezeptionsforschung

Der Bedarf an umfangreichen Inhaltsanalysen ergibt sich nicht nur aus dem rasch wachsenden medialen Angebot, sondern auch aus der Nachfrage der Kommunikator- und Rezeptionsforschung nach inhaltsanalytische Daten. Spätestens seit der Agenda-Setting-Studie von Erbring et al. (1980) kann die Verknüpfung von Befragungsdaten mit individuell zugewiesenen inhaltsanalytischen Daten als etabliertes Verfahren der Rezeptionsforschung angesehen werden. Dieses Vorgehen wurde in den letzten Jahren nicht nur methodisch verfeinert (Wolling, 2002), sondern in einer Vielzahl unterschiedlicher Themenstellungen angewandt, von der Kultivierungs- (Lücke, 2007) über die Nachrichtenwerttheorie (Fretwurst, 2008) bis hin zum Framing (Matthes, 2007). Ein zentrales Problem dieser Verknüpfung ist die Vielfalt relevanter, d.h. von den Befragten rezipierter Medienangebote, die es ggf. zu analysieren gilt, um differenzierte Aussa-

gen über Medieneffekte treffen zu können. Während bei TV-Nachrichten auch nach Einführung des privaten Fernsehens ein relativ übersichtliches Angebot von Vollprogrammen zu codieren ist, wird die Situation bei Hörfunk- und Printmedien deutlich schwieriger. Schon bei Erbring et al. (1980) konnten nicht alle genutzten Tageszeitungen analysiert werden, und auch in neueren Studien wird zumeist mit stark reduzierten Angebotsdaten gearbeitet (Arlt et al., 2010). Angesichts der vielbeschworenen „Fragmentierung des Publikums“ (Goertz, 2009) wird der Bedarf an umfangreichen Inhaltsanalysen noch zunehmen, da der Anteil an Befragten, die dieselben Inhalte rezipiert haben, kontinuierlich sinkt. Dies wird umso deutlicher, wenn man den Umfang an Online-Angeboten betrachtet, die ein Internetnutzer tagtäglich besucht. Obwohl durch die Verknüpfung von Logfile- und Online-Inhaltsanalysen eine Rekonstruktion von Rezeptionsprozessen in beispielloser Breite und Tiefe möglich wäre, scheitert dies bislang nicht nur an der Verfügbarkeit von Nutzungsdaten, sondern vor allem auch an der Quantität an Codierungen, die für eine solche Studie notwendig ist. Angesichts der im vorangegangenen Abschnitt referierten Zahlen zu nutzergenerierten Inhalten im Internet stellt ggf. schon die Codierung der während eines Untersuchungszeitraums von einigen Tagen oder Wochen von wenigen Nutzern *produzierten* Inhalte eine Herausforderung dar. Gerade weil online-basierte Anschlusskommunikation eines der wichtigsten neuen Themen der Agenda-Setting-Forschung ist (Haas et al., 2010; Vu & Gehrau, 2010), ist auf diesem Gebiet zukünftig eine enorme Nachfrage nach Inhaltsanalysen zu erwarten.

Ein großer Bedarf an umfangreichen Analysen medialer Inhalte ist in den letzten Jahren auf dem Gebiet der klassischen Rezeptionsforschung zu verzeichnen, die bislang eher experimentell orientiert war. In dem Maße, in dem sich die Forschung auf die Wirkung natürlicher, d.h. nicht experimentell veränderter oder generierter Stimuli konzentriert, ist es notwendig diese zu analysieren. Angesichts der Vielzahl an potentiell wirkungsrelevanten Simuluseigenschaften, vor allem bei audiovisuellem Material, ist der Umfang entsprechender Codebücher zumeist forschungsökonomisch begrenzt: Man konzentriert sich zumeist auf die reine Text- bzw. Inhaltsebene *oder* visuelle *oder* akustische Merkmale (Holicki & Brosius, 1988; Suckfüll, 1997). Um zuverlässig bestimmen zu können, welche

dieser Merkmale einzeln oder gemeinsam Wirkungen beim Rezipienten entfalten, müssen möglichst viele Stimuluseigenschaften codiert werden. Gerade wenn theoretisch noch nicht geklärt ist, welche Merkmale einer Botschaft genau welche Effekte haben, profitiert die Rezeptionsforschung von einer möglichst großen Quantität der Inhaltsanalyse.

2.3 Automatisierung als Lösung?

Im vorangegangenen Kapitel habe ich argumentiert, dass die Zukunft inhaltsanalytischer Forschung vor allem mit dem Problem konfrontiert sein wird, eine ausreichende Codierquantität zu gewährleisten, auf deren Basis sich zuverlässige und valide Inferenzschlüsse ziehen lassen. Dieses Problem lässt sich am ehesten – und vielleicht auch ausschließlich – durch Automatisierung lösen, d.h. die Nutzung von Computerprogrammen für bestimmte Aufgaben im Forschungsprozess. Dies ist keineswegs eine neue Erkenntnis (Stone, 1969a; Diefenbach, 2001), aber angesichts der oben formulierten methodologischen und fachwissenschaftlichen Herausforderungen ist die Relevanz automatischer Verfahren zweifellos gewachsen. Krippendorff (2004b, XXI) bemerkt etwa dazu: „[C]omputer aids participate in content analysis much as human analysts do. They become part of its methodology, with transparency being a major issue." Im folgenden Abschnitt sollen daher die Vor- und Nachteile der Automatisierung inhaltsanalytischer Forschung sowohl aus methodologischer als auch kommunikationswissenschaftlicher Anwendungsperspektive diskutiert werden. Dabei geht es um die grundsätzliche Frage, ob und wie Computerunterstützung die Inhaltsanalyse bereichern kann, wenn es um die Erhöhung der Codierquantität bei gleichzeitiger Sicherung von Reliabilität und Validität der Analyse geht.

2.3.1 Methodenperspektive: Skalierbarkeit und Qualität

Aus forschungsökonomischer Perspektive hängt der Einsatz eines bestimmten manuellen oder automatischen Verfahrens vor allem von dessen Skalierbarkeit ab. Unter Skalierbarkeit wird in der Informatik und anderen Disziplinen, etwa der Betriebswirtschaft, die Fähigkeit einer Software,

einer Methode oder einer Organisation verstanden, mit der Anzahl zu verarbeitender Aufgaben zu wachsen oder zu schrumpfen, ohne übermäßig an Effektivität oder Effizienz einzubüßen (Abbott & Fisher, 2010).

Diese Anforderung lässt sich auch an sozialwissenschaftliche Methoden stellen: Ein Beispiel hierfür ist die Online-Befragung, mit der sich sowohl kleine Ad-hoc-Befragungen im Rahmen studentischer Projekte als auch umfangreiche Online-Access-Panel mit hunderttausenden Teilnehmern durchführen lassen. Der Aufwand für die Erhebung bleibt dabei fast gleich, wenn man von der Rekrutierung der Befragten absieht, die ohnehin das schwerwiegendste Problem dieses Erhebungsverfahrens ist.

Angesichts der oben erläuterten Wünschbarkeit von möglichst umfangreichen inhaltsanalytischen Daten ist die Aufwärtsskalierbarkeit ein wichtiges Kriterium, d.h. die Frage, wie aufwändig das Hinzufügen (a) neuer Untersuchungseinheiten oder (b) neuer Indikatoren, d.h. Variablen im Codebuch, ist. Bei der klassischen manuellen Inhaltsanalyse ist dies ein linearer Zusammenhang: Die Zahl der notwendigen Codierungen entspricht genau dem Produkt aus Variablen und Untersuchungseinheiten (vgl. Abbildung 2.3). Dies führt dazu, dass ab einem bestimmten Stichproben- oder Codebuch-Umfang der Gesamtaufwand der Analyse fast nur aus den Kosten der Codierung besteht, während alle anderen Arbeitsschritte der Inhaltsanalyse kaum noch ins Gewicht fallen. Aufgrund des linearen Zusammenhangs von Aufwand und Quantität der Codierung kann man festhalten, dass klassische manuelle Inhaltsanalysen nur schlecht aufwärts skalieren.

Da automatische Verfahren in ihrer technischen Leistungsfähigkeit zumeist nur von der Hardwareausstattung begrenzt sind, ist hier nur ein geringerer Zusammenhang von Codierquantität und -kosten zu erwarten (vgl. die gepunktete Linie in Abbildung 2.3). Dies gilt in besonderem Maße für die Skalierung der Stichprobengröße, die bei vollautomatischen Verfahren mit minimalen Kosten verbunden ist (Monroe & Schrodt, 2008, 352).Dieser Vorteil der Verarbeitung großer Mengen an Dokumenten war in der Vergangenheit das häufigste Argument, das für die Verwendung automatischer Verfahren vorgebracht wurde (Schrodt & Donald, 1990; King & Lowe, 2003; Früh, 2007). Dies gilt im Übrigen nicht nur für den Arbeitsschritt der Codierung, sondern auch für die Datenerhebung

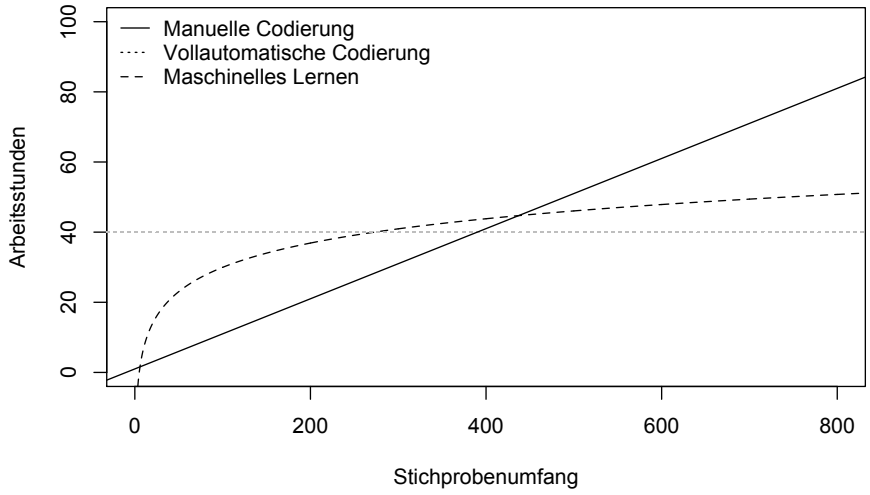

Abbildung 2.3: Skalierbarkeit bei manueller und automatischer Inhaltsanalyse

oder Auswertung, d.h. alle Aufgaben, in denen die Zahl der Untersuchungseinheiten den Aufwand beeinflusst. In dieser Hinsicht skalieren automatische Verfahren hervorragend mit dem Stichprobenumfang. Erwartungsgemäß sollte sich das maschinelle Lernen als halbautomatisches Verfahren mit sinkendem Aufwand pro zusätzlicher Untersuchungsheit zwischen manueller und vollautomatischer Codierung einordnen lassen (vgl. die gestrichelte Linie in Abbildung 2.3)

Ein anderes Bild ergibt sich, wenn man statt der Stichprobe das Codebuch erweitern möchte. Der Aufwand für die Entwicklung eines validen Instruments für die automatische Codierung ist zumeist deutlich höher als bei der klassischen Codebuchentwicklung, weil jede Regel explizit und deterministisch ausgeführt und in Maschinensprache überführt werden muss. So entstehen schon vor der Codierung des ersten Dokuments hohe Anlaufkosten, die sich ggf. durch die Skalierung des Stichprobenumfangs nicht wieder auffangen lassen. Dies ist gerade deshalb problematisch, weil für viele neue Fragestellungen und Medienangebote erst kleine, eher

explorative Studien angebracht sind, um etwa verschiedene Operationalisierungstrategien zu testen.[5] Ob nun ein automatisches Verfahren im Aufwand der Operationalisierung ähnlich wie ein klassisches Codebuch verhält, sich folglich besser oder schlechter mit der Zahl der Variablen skalieren lässt, hängt sehr von den eingesetzten automatischen Verfahren ab, die in Kapitel 3 noch ausführlich vorgestellt werden. Zudem hängt dies auch von der – oft fehlenden – Erfahrung mit computerbasierten Ansätzen ab. Grundsätzlich lässt sich aber sagen, dass für einen klassisch geschulten Forscher der Aufwand automatischer Verfahren sehr hoch sein kann, so dass man gerade bei kleineren Studien von einer schlechteren Abwärtsskalierbarkeit ausgehen muss, wie dies in Abbildung 2.3 deutlich wird.

Um eine Auf- und Abwärtsskalierbarkeit sowohl hinsichtlich der Zahl der Variablen als auch der Untersuchungseinheiten sicherzustellen, sind Verfahren notwendig, die die Vorteile manueller Inhaltsanalyse (Flexibilität der Operationalisierung, Validität, Anschlussfähigkeit im Fach) und automatischer Codierung (Effektivität und Effizienz) vereinen. Auf diese Weise könnte ein Forschungsteam ein inhaltsanalytisches Instrument anhand einer kleinen Stichprobe entwickeln, das anschließend möglichst nahtlos auf eine beliebig große Zahl an Untersuchungseinheiten angewendet werden kann. Ob das maschinelle Lernen dieses Versprechen einhalten kann, soll in dieser Arbeit geklärt werden.

Auch wenn Skalierbarkeit auf den ersten Blick primär forschungsökonomisch motiviert scheint, ist sie eine zentrale Voraussetzung für die methodologische Entwicklung der Inhaltsanalyse: Wenn der Aufwand für die Codierung niedrig ist, ist man bei begrenzten Mitteln eher bereit, sowohl inhaltlich als auch methodisch Neuland zu betreten, alternative Instrumente und Forschungsdesigns systematisch auszuprobieren und auf neue Angebote zu reagieren. Wiederum zeigen die Erfahrungen mit Online-Befragungen, dass niedrige Anwendungskosten die methodologische Diskussion positiv stimulieren (vgl. Taddicken, 2008; Zerback et al., 2008; Kaczmirek, 2009).

5 Erschwerend kommt hinzu, dass die Anwendung automatischer Verfahren weit weniger gut dokumentiert ist als die konventionelle Inhaltsanalyse, für die nicht nur zahlreiche Lehrbücher sondern auch entsprechende Lehrveranstaltungen angeboten werden.

Reliabilität und Reproduzierbarkeit
Ein zweiter wesentlicher Vorteil der Automatisierung inhaltsanalytischer Forschungsschritte liegt in der Reproduzierbarkeit der Ergebnisse. Dies betrifft nicht nur, aber in besonderem Maße den eigentlichen Codierprozess, der durch den Einsatz menschlicher Helfer zwangsläufig nur unvollständig replizierbar ist. Hier liegt der offensichtlichste Vorteil automatischer Verfahren, die stets vollständige Reproduzierbarkeit versprechen, da es sich bei Computerprogrammen um deterministische Prozesse handelt (Krippendorff, 2004a). Selbst wenn die Codierung oder Stichprobenziehung stochastische Komponenten enthält, ist auch dies potentiell dokumentier- und damit reproduzierbar.

Über die reine Codierung hinaus gewährleistet die Nutzung von Computersoftware die Replikation inhaltsanalytischer Forschungsergebnisse, weil sie einerseits den Spielraum für mögliche Fehler und Missverständnisse einschränkt, andererseits eine effiziente Art der Dokumentation einzelner Entscheidungen und Forschungsschritte bietet, und damit eine weitere Forderung von Krippendorff (2004a) erfüllt, nämlich die Zusammenarbeit von einzelnen Forschern bzw. Forschungsteams zu erleichtern. Nutzt man für viele routinemäßig anfallenden Aufgaben – die Stichprobenziehung, die Verteilung der Untersuchungseinheiten auf die Codierer, das Datenmanagement oder die Durchführung von Reliabilitätstests – eine Softwarelösung, müssen lediglich die relevanten Input-Parameter dokumentiert werden, um anderen die Replikation der Studie zu ermöglichen. Angesichts der noch immer höchst unterschiedlichen Dokumentationspraxis für Inhaltsanalysen (Lauf, 2001) bleibt zu hoffen, dass die Verfügbarkeit leicht bedienbarer Software nicht nur zu einer Standardisierung der Dokumentation, sondern auch zu einer sachgerechten Durchführung von Studien motiviert. Einfacher formuliert: Wenn ein Verfahren keinen zusätzlichen Aufwand verursacht, wird es eher durchgeführt und dokumentiert.

Ein wichtiges Ziel der methodologischen Diskussion zur Inhaltsanalyse sollte daher sein, nicht nur Standards für die Durchführung und Dokumentation zu definieren, wie dies im Bereich der Befragung und Testentwicklung seit Jahrzehnten der Fall ist (American Educational Research Association, 1985; Deutsche Forschungsgemeinschaft, 1999),

sondern auch die Möglichkeiten der Qualitätssicherung durch Automatisierung im Blick haben. Dies gilt selbstverständlich nicht nur für die Erhebung der Daten, sondern auch für deren statistische Auswertung und Archivierung. Auch hier scheint die Survey-Forschung methodisch fortgeschrittener zu sein, wie ein Blick in den Bestand des Zentralarchivs für Empirische Sozialforschung zeigt. In jüngster Zeit wurden auch für die dezentrale Archivierung von Forschungsdaten Softwarelösungen entwickelt, die letztlich keinem anderen Ziel als der Reproduzierbarkeit empirischer Forschung dienen (King, 1995, 2003).

Schließlich lässt sich der methodologische Anspruch an wissenschaftliches Arbeiten, reproduzierbare Ergebnisse zu erzeugen, auch forschungsökonomisch begründen: Angesichts der Quantität zu analysierender Daten ist es schlicht ineffektiv und ineffizient, stets von neuem zu codieren und dabei absichtlich oder unabsichtlich das Rad neu zu erfinden. Viele wesentliche Bestandteile des inhaltsanalytischen Forschungsprozesses sind noch immer unzureichend dokumentiert und erschweren eine Replikation der Ergebnisse. Eine Standardisierung und mögliche Automatisierung dieser Schritte würde m.E. dazu führen, dass mehr Raum für eine methodische und inhaltliche Entwicklung der Forschung durch Replikation und Erweiterung entsteht.

Validität

Eine letzte methodologische Frage der Automatisierung ist die Validität der Verfahren. Hier muss man differenziert nach der Reichweite der Inferenzen und der Angemessenheit eines Ansatzes für eine spezifische Fragestellung unterscheiden: Grundsätzlich müssen automatische Verfahren nicht weniger valide sein als manuelle, insbesondere wenn der Computer eher handwerkliche Aufgaben übernimmt. Es liegt auf der Hand, dass die Anzahl gezählter Wörter ein valider Indikator für die Länge eines Textes ist, und dass dies mit automatischen Verfahren mindestens genauso gut zu messen ist wie manuell. Meist geht jedoch das Interesse an inhaltsanalytischen Inferenzen über die reine Textbeschreibung hinaus, und hier kann man zumindest über die Angemessenheit der Operationalisierung streiten, wenn etwa Textlänge als Indikator für den Nachrichtenwert (Schulz, 1976) verwendet wird. Dies ist aber kein

verfahrenstechnisches Problem, sondern eines des Inferenzschlusses auf Merkmale außerhalb der konkreten Botschaft.

Automatische Verfahren können nun aus zwei Gründen eingesetzt werden, die unterschiedliche Folgen für die Bewertung von deren Validität haben: 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 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 (Rössler, 2005; Früh, 2007). 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 für die Validität. 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. So banal es klingen mag: Manuelle und automatische Verfahren der Inhalsanalyse sind nicht an sich (in-)valide, sondern nur im Bezug auf die Inferenzschlüsse, die aus ihnen gezogen werden. Generell kann man jedoch nicht sagen, ob automatische Ansätze weniger valide sind als manuelle oder umgekehrt.

2.3.2 Fachperspektive: Anwendungsfelder und Nutzen

Neben der grundlegenden methodologischen Einordnung automatischer Verfahren stellt sich auch die Frage, für welche Forschungsansätze oder Themengebiete der Kommunikationswissenschaft diese nun besonders geeignet oder ungeeignet sind. Dies lässt sich nicht ohne weiteres beantworten, weil die Anwendungsmöglichkeiten und auch die tatsächlichen Anwendungen enorm vielfältig sind. Obwohl noch immer randständig, wurden und werden automatische Verfahren für fast alle Themengebiete und Forschungsfragen eingesetzt. Ein entsprechender Überblick würde

den Rahmen dieser Arbeit sprengen, jedoch sind im folgenden Kapitel 3 zu einzelnen Ansätzen automatischer Textanalyse beispielhaft Studien genannt, in denen diese Verwendung fanden. An dieser Stelle möchte ich daher nur einige ausgewählte kommunikationswissenschaftliche Anwendungskontexte hinsichtlich ihrer Automatisierbarkeit diskutieren.

Ein erstes Kriterium, an dem sich die Eignung automatischer Verfahren illustrieren lässt, ist das zu untersuchende Medium: So sind digitale Texte sicher am ehesten für eine automatische Verarbeitung geeignet, weil sie erstens bereits maschinenlesbar vorliegen und zweitens die meisten Softwarepakete zur Inhaltsanalyse nur für Textanalysen konzipiert sind. Für die Verarbeitung audiovisueller Inhalte sind bereits die technischen Möglichkeiten deutlich beschränkter. Die stiefmütterliche Behandlung non-verbaler Stimuli in der Inhaltsanalyse ist jedoch keineswegs nur auf automatische Verfahren beschränkt. Rössler (2010) merkt kritisch an, dass die Kommunikationswissenschaft insgesamt die Analyse von multimedialen Inhalten bislang weitgehend ignoriert hat. Wenn schon keine etablierten Instrumente für die quantitative manuelle Analyse von Fotografien, Musikstücken oder Filmen zur Verfügung stehen, ist es umso schwieriger, entsprechende automatische Verfahren zu entwickeln und dann auch vergleichend zu evaluieren. Wer audiovisuelle Mitteilungen hinsichtlich nonverbaler Merkmale analysieren will, muss dies bislang manuell tun. Dies muss nicht zwingend bedeuten, dass automatische Verfahren dafür nicht geeignet sind, jedoch ist die Verfügbarkeit von Know-How und entsprechender Software bislang nicht gegeben.

Neben der Beschaffenheit des Untersuchungsmaterials ist auch der Untersuchungsrahmen einer Inhaltsanalyse entscheidend für die Wahl automatischer oder manueller Verfahren. Letztere sind in den meisten Fällen deutlich flexibler zu gestalten, weil es ggf. einfacher ist, den Codierern eine entsprechende Klassifikationslogik beizubringen als einem Computer. In dieser Hinsicht eignen sich eher offene, qualitativ orientierte Analysen nicht für automatische Verfahren, die eine starke Strukturierung der Operationalisierung erfordern. Ist jedoch die konsistente Codierung über lange Untersuchungszeiträume von übergeordnetem Interesse, wie etwa in der kontinuierlichen Themenstruktur- oder Medienresonanzanalyse, spielen automatische Verfahren ihre Vorteile

Skalierbarkeit und Reproduzierbarkeit aus. Gerade wenn es um Langzeitanalysen und Prognosemodelle geht, ist die Konsistenz der Codierung zumeist wichtiger als deren Genauigkeit. Fasst man die bisher genannten zwei Punkte zusammen, ergibt sich als ein sehr erfolgversprechendes Anwendungsgebiet automatischer Verfahren das kontinuierliche Themenmonitoring von Online-Inhalten, und genau hierfür werden diese in der kommerziellen Forschung auch intensiv genutzt (Gürtler & Kronewald, 2010). Umgekehrt gibt es auch Anwendungsfelder, für die sich automatische Verfahren bislang gar nicht eignen, etwa die tiefgehende Analyse komplexer audiovisueller Inhalte, bei der nicht selten entsprechendes Expertenwissen über Narration, Schnitt oder Musik vorausgesetzt wird.

Am Beispiel der Filmanalyse (Suckfüll, 1997) lässt sich auch eine weitere Vergleichsperspektive zwischen manuellen und automatischen Verfahren illustrieren: das Ziel der Inferenz. Grundsätzlich eignen sich automatische Ansätze vor allem für deskriptive, d.h. mitteilungszentrierte Analysen, bei denen die Inferenzschlüsse relativ nahe am analysierten Gegenstand liegen. Ein Beispiel ist auch hier die Programmstrukturforschung, ob bei Rundfunk- oder Online-Medien (Weiß, 1998; Zeller & Wolling, 2010). Soll von den Mitteilungen jedoch direkt auf deren Auswirkungen auf den Rezipienten geschlossen werden, ist der Einsatz menschlicher Codierer praktisch alternativlos. In dem Maße, in dem Codierer in ihrer Eigenschaft als Mediennutzer die Untersuchungseinheiten eher bewerten oder ihre Wirkung beurteilen sollen, verbietet sich der Einsatz von Computern. Einschränkend möchte ich jedoch betonen, dass diese Art von manueller Bewertung eher mit dem Begriff der *Rezeptionsanalyse* (Kepplinger, 2009) zu beschreiben ist, weil es gerade nicht um den Ausschluss subjektiver Erfahrung des einzelnen Codierers geht (vgl. Scharkow, 2010a). Trotzdem ist festzuhalten, dass sich automatische Verfahren gerade nicht dazu eignen, die menschliche Wahrnehmung und Verarbeitung bei der Medienrezeption zu ersetzen.

Zusammenfassend kann man sagen, dass sich automatische Verfahren prinzipbedingt vor allem für die beschreibende Analyse großer Mengen digitaler Texte eignen, für andere Anwendungsfelder der Inhaltsanalyse, etwa Bild- und Filmanalyse, jedoch bislang ungeeignet sind. Trotzdem ist aber gerade die Themenanalyse digitaler (oder digitalisierter) Texte

ein so großes Anwendungsgebiet, dass es sich lohnt, über die Möglichkeiten und Grenzen automatischer Verfahren für diese Anwendungsfälle nachzudenken.

2.4 Automatische Verfahren als Alternative oder Ergänzung?

Angesichts der unterschiedlichen Stärken und Schwächen automatischer und manueller Verfahren der Inhaltsanalyse könnte eine Zusammenfassung lauten, dass man beides doch am besten als Ergänzung zueinander verstehen sollte. Dies ist sicher grundsätzlich zutreffend und doch nicht die ganze Wahrheit. Schon jetzt werden viele automatische Verfahren ganz selbstverständlich für Arbeitsschritte eingesetzt, die zuvor manuell erledigt wurden, vor allem im Bereich der Datenerhebung und Auswertung. Selten werden heute noch Papierausgaben von Printmedien gesammelt und an die Codierer verteilt, wenn es E-Paper oder Online-Datenbanken wie LexisNexis gibt (vgl. Abschnitt 4.1). Noch seltener werden Stichprobenpläne oder Codierbögen ohne Computerunterstützung erstellt, von der statistischen Auswertung der Daten ganz abgesehen. Dies ist nicht immer unproblematisch, vor allem wenn man sich dabei auf Fremdanbieter oder intransparente Software verlassen muss, aber häufig forschungsökonomisch legitimierbar. Wenn sich nachweisen lässt, dass ein automatisches Verfahren mit einiger Sicherheit ähnlich zuverlässig und valide funktioniert wie ein manuelles, wird 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. So gibt es kaum einen Grund, warum ein Forscher von Hand zu bestimmten Zeiten bestimmte Websites abrufen und abspeichern sollte, wenn dies der Computer schneller und zuverlässiger kann (vgl. Rüdiger & Welker, 2010). Auch das Wörterzählen wird zumeist halbautomatisch, d.h. unter Verwendung von Textverarbeitungsprogrammen durch die Codierer, durchgeführt, ohne dass dies als problematisch angesehen wird.

Es stellt sich hinsichtlich dieser Beobachtungen der Forschungspraxis die Frage, ob die Zusammenfassung dieses Kapitels nicht lauten müsste: Das Ziel der Methodenentwicklung ist es, jeden Arbeitsschritt der Inhaltsanalyse zu automatisieren, solange dies – und das ist die entscheidende Einschränkung – methodisch vertretbar ist, d.h. zu ausreichend zuverlässigen und gültigen Ergebnissen führt. Mit anderen Worten, es gibt keinen Grund, die Möglichkeiten der Automatisierung nicht zumindest zu prüfen und sich ggf. für ein automatisches Verfahren zu entscheiden. Angesichts der Vorteile automatischer Verfahren hinsichtlich Dokumentation und Reproduzierbarkeit wäre es m.E. sogar wünschenswert, automatische Verfahren zu verwenden, selbst wenn diese nicht mit geringerem Aufwand oder Ressourcenverbrauch verbunden wären. Dies ist vor allem der Fall, wenn die Anlaufinvestitionen – das Beschaffen und Lernen entsprechender Software, die Entwicklung passender Verfahren – hoch sind, und sich auf den ersten Blick nicht lohnen. Gerade wenn man an der Skalierbarkeit der Codierquantität, der Reliabilität und Reproduzierbarkeit der eigenen Analysen interessiert ist, führt an automatischen Verfahren kaum ein Weg vorbei.

Angesichts der vielen Einschränkungen, denen automatische Verfahren unterliegen, ist es selbstverständlich, dass weiterhin viele Schritte im Forschungsprozess manuell erledigt werden müssen. Kein Computer kann Forschungsfragen formulieren und in entsprechende Analysestrategien umsetzen. Kein Computer kann von sich aus verbale oder nonverbale Mitteilungen verstehen. Zudem gibt es schlicht für die meisten inhaltsanalytischen Fragestellungen keine fertigen Lösungen, die man so flexibel einsetzen kann wie menschliche Helfer. Hier gilt dann doch, dass automatische Verfahren die klassisch manuelle Arbeitsweise eher ergänzen, etwa bei der Entwicklung des Codebuchs oder der Durchführung von Reliabilitätstests.[6] Damit automatische Verfahren, wie von Krippendorff (2004a, XI) gefordert, zum selbstverständlichen Repertoire jedes inhaltsanalytisch arbeitenden Forschers werden können, müssen diese erst einmal bekannt sein. Nur wenn man die konzeptionellen Stärken und Schwächen sowie die Funktionsweise der einzelnen Ansätze kennt, ist man in der Lage,

6 Einige Vorschläge, wie sich dies praktisch durchführen lässt, sind in der Beschreibung des Forschungsinstruments im Anhang A zu finden.

entweder auf Basis theoretischer Überlegungen oder empirischer Evaluationsstudien und Methodenexperimenten Aussagen darüber zu treffen, wie zuverlässig und valide sie bezogen auf die eigene Forschungsfrage sind.

In fast 60 Jahren Forschungspraxis sind automatische Verfahren bislang in vielen Disziplinen und für viele Fragestellungen eingesetzt worden. Allerdings sind die Möglichkeiten der Automatisierung gerade angesichts der rasanten Entwicklung entsprechender Softwarealgorithmen bislang nur wenig im Kontext sozialwissenschaftlicher Anforderungsszenarien diskutiert worden. Im nächsten Kapitel werden daher klassische und neue Ansätze computergestützter Codierung dargestellt und daraufhin untersucht, inwiefern sie die Skalier- und Reproduzierbarkeit von Inhaltsanalysen verbessern können. Dabei konzentriere ich mich auf die Automatisierung der Codierung als zentralem Arbeitsschritt der Inhaltsanalyse, während die Datenerhebung und anderere Aufgaben im Kontext automatischer Inhaltsanalysen im darauf folgenden Kapitel 4 diskutiert werden.

3 Automatische Inhaltsanalyse in den Sozialwissenschaften

3.1 Grundlagen computergestützter Verfahren

3.1.1 Eine kurze Geschichte automatisierter Inhaltsanalyse

In diesem Abschnitt soll die Entwicklung automatischer Verfahren der Inhaltsanalyse anhand der zentralen Problemstellungen und deren (versuchter) Lösungen skizziert werden. Es geht dabei eher um die Systematisierung der Methodenentwicklung auf diesem Gebiet als um eine historische Betrachtung. Aus diesem Grunde wird auf eine Darstellung der Geschichte der Inhaltsanalyse (vgl. u.a. Berelson, 1952; Holsti, 1969; Lisch & Kriz, 1978; Krippendorff, 2004a) ebenso verzichtet wie auf eine Systematisierung der Methodenentwicklung nach Fächern (Stone et al., 1966; Diefenbach, 2001). Schließlich soll an dieser Stelle auch kein Überblick über die Anwendung automatischer Verfahren in veröffentlichten empirischen Studien gegeben werden, da diese von Züll & Landmann (2002) umfassend dokumentiert wurden.

Vor der eigentlichen Systematisierung des Forschungsgebiet automatischer Inhaltsanalyse ist eine kurze Begriffsdefinition angebracht. Angesichts der Vielzahl von Möglichkeiten, Computer in verschiedenen Phasen des inhaltsanalytischen Forschungsprozesses einzusetzen, ist der vielfach verwendete Begriff computerunterstützte Inhaltsanalyse (CUI) streng genommen obsolet, da heute so gut wie jede Inhaltsanalyse – wie fast sämtliche empirische Forschung – in irgendeiner Weise durch Computer unterstützt wird, sei es bei der Berechnung von Stichprobenplänen, der Eingabe von Daten durch die Codierer oder der Auswertung und grafischen Aufbereitung der Ergebnisse. Luzar (2004) schlägt daher in Anlehnung an Stuckardt vor, zwischen computer*unter*stützter und computer*ge*stützter Inhaltsanalyse zu unterscheiden. Nur bei letzterer wird

die Codierung durch den Computeralgorithmus und ohne Eingriff des Forschers getroffen, während unter den ersten Begriff auch Annotationssysteme wie CETA, qualitative Textanalyse-Software wie Atlas.Ti oder Datenbanken mit Schlagwortsuche wie LexisNexis fallen. Diese Unterscheidung ist zwar inhaltlich angebracht, aber begrifflich vorbelastet, da beides in der Vergangenheit synonym verwendet wurde. In dieser Arbeit soll in Anlehnung an Monroe & Schrodt (2008) der Begriff *automatische* Inhaltsanalyse verwendet werden, der alltagssprachlich verankert ist und das Vorgehen treffender beschreibt. Zudem ist die Effizienz der Verfahren bezogen auf große Textmengen in der Definition verankert, obwohl diese Effizienz streng genommen erst eine Folge der Funktionsweise von automatischen Verfahren und kein Merkmal derselben ist.

> As a rule-of-thumb, we consider a system fully automated if the marginal cost of analyzing additional texts goes to zero as the size of the corpus being analyzed increases, and the coding is completely replicable given a set of software, dictionaries, and so forth. (Monroe & Schrodt, 2008, 352)

In diesem Sinne ist von automatischer Inhaltsanalyse genau dann zu sprechen, wenn tatsächlich ein Computeralgorithmus zur Codierung verwendet wird, d.h. die einzelne Codierentscheidung bezogen auf die relevante Untersuchungseinheit *nicht* vom Forscher getroffen wird. Dieser ist wiederum für die Entwicklung von Codierregeln, deren Umsetzung in maschinenlesbare Form (Software) und die Interpretation der Ergebnisse verantwortlich. In den folgenden Abschnitten geht es also zunächst nur um die Automatisierung der Codierung als wichtigstem Schritt jeder Inhaltsanalyse.

Die Geschichte und Entwicklung der automatischen Inhaltsanalyse lässt sich m.E. nach auf drei zentrale Herausforderungen zurückführen: (1) die Konzeption inhaltsanalytischer Fragestellungen und deren automatischer Durchführung, (2) die Entwicklung von Software für die automatische Datenverarbeitung und -analyse und (3) die Bereitstellung und Analyse digitaler bzw. maschinenlesbarer Dokumente.

Die erste Entwicklungsphase computerunterstützter und automatischer Inhaltsanalyse seit Ende der 50er Jahre ist vor allem gekennzeich-

net durch das Experimentieren mit dem neuen Instrument, das auch an großen Universitäten nur in wenigen Großrechnern zur Verfügung stand, den sich die wenigen mutigen Sozialwissenschaftler auch noch mit allen anderen Disziplinen teilen mussten.[1] Konzeptionell standen diese ersten Studien fast ausschließlich in der Tradition der Textstatistik, d.h. des Wörterzählens, die seit den 20er Jahren in vielen Disziplinen wie der Politik- oder auch Literaturwissenschaft angewandt wurde (Stone et al., 1966; Holsti, 1969). Sie war damit einerseits hinter der methodologischen Entwicklung der Inhaltsanalyse zurückgeblieben, die sich mit den innovativen Ansätzen von Lasswell et al. (1952) oder Osgood (1959) und der Allerton House Conference 1955 zeigte (Pool, 1959). Andererseits waren gerade die Pioniere der Inhaltsanalyse zunehmend von deren Aufwand abgeschreckt und setzten große Erwartungen in die computergestützten Verfahren, deren Weiterentwicklung als zentral für den Erfolg der Methode angesehen wurde (Stone, 1997).

Zu diesem Zeitpunkt war die computergestützte Inhaltsanalyse noch Hochrisikoforschung, die mit zahlreichen Problemen konfrontiert war: Einerseits waren Hard- und Software so limitiert, dass nur kleine Textkorpora mit wenigen Variablen analysiert werden konnten (Iker & Harway, 1969). Anderseits gab es so gut wie keine maschinenlesbaren Dokumente, so dass alle Untersuchungseinheiten in einem aufwändigen und fehleranfälligen Prozess auf Lochkarten übertragen werden mussten. Im Gegensatz zu den heutigen Möglichkeiten war die computergestützte Inhaltsanalyse zu Beginn der 60er Jahre weder kostengünstiger – eine halbe Stunde Rechenzeit kostete so viel wie das Monatsgehalt einer Sekretärin – noch mit weniger Aufwand verbunden als die manuelle Codierung (Stone, 1997, 42).

Mit der Entwicklung des General Inquirer (Stone et al., 1966) und Words (Iker & Harway, 1969) waren nicht nur die ersten relativ leicht benutzbaren Software-Pakete verfügbar, sondern auch die grundlegenden

1 Eine allgemein anerkannte Pionierstudie ist auf dem Gebiet der automatischen Textanalyse nicht auszumachen, auch und gerade weil die ersten Studien damals von vielen Wissenschaftlern aus unterschiedlichen Disziplinen unabhängig voneinander durchgeführt wurden. Dies änderte sich schlagartig mit der Annenberg-Konferenz 1967 (Gerbner et al., 1969).

konzeptionellen Auseinandersetzungen mit der automatischen Inhaltsanalyse weitgehend abgeschlossen. Die auf der Annenberg-Konferenz (Gerbner et al., 1969) vorgestellten Studien setzten Maßstäbe, an denen sich für die nächsten Jahrzehnte die meisten Arbeiten auf dem Gebiet orientieren sollten. Fast alle Ansätze, die in diesem Kapitel diskutiert werden, etwa diktionärbasierte oder Co-Occurrence-Analysen, waren zu diesem Zeitpunkt bereits entwickelt, auch wenn die technische Umsetzung in den Folgejahren natürlich leichter wurde.

In den 70er und 80er Jahren wurde das Forschungsprogramm, das mit dem General Inquirer begonnen wurde, durch die Entwicklung von Diktionären (vgl. Abschnitt 3.3.1) in vielen Disziplinen und Anwendungsfeldern fortgeschrieben. Während in der angelsächsischen Sozialforschung das methodologische Interesse an der automatischen Inhaltsanalyse nachließ (Weber, 1984, 127), gab es in Deutschland zu dieser Zeit vielfältige Entwicklungen auf diesem Gebiet (Deichsel, 1975; Lisch & Kriz, 1978). Unter anderem wurde das Programmpaket Textpack entwickelt, das ursprünglich für die Codierung offener Fragen gedacht war, in der Folge aber auch für die Analyse von Dokumenten eingesetzt wurde (Schönbach, 1978, 1982; Klingemann et al., 1984).[2]

Das Problem der Archivierung und Distribution großer Dokumentenmengen in maschinenlesbarer Form war bis Ende der 70er Jahre ungelöst, so dass sich viele Studien nur auf bereits existierende und wenig aktuelle Dokumente stützen konnten. DeWeese (1977) konnte als erster Forscher tagesaktuelle Medieninhalte automatisch digital archivieren, indem er auf die zunehmend verbreiteten Satzgeräte der Verlage direkt zugriff. Dieser Ansatz wurde auch von einem deutschen Forschungsteam weiterverfolgt, da er sich vor allem für genuin kommunikationswissenschaftliche Fragestellungen sehr gut eignete (Bröker, 1984). Im Jahr 1979 stellte der Dienstleister LexisNexis erstmals digitale Ausgaben von amerikanischen Tageszeitungen per Datenfernübertragung zur Verfügung.

2 Bis in die 80er Jahre war die Portierung der hochoptimierten Spezialsoftware General Inquirer oder Textpack von Großrechnern auf andere Rechnerarchitekturen mit hohen Kosten verbunden. Erst der Einsatz von Hochsprachen wie Fortran oder Pascal erleichterte die Weiterentwicklung und Verbreitung der Software, die dann auch auf konventionellen Desktop-PC lief.

Mit der Entwicklung der Personal Computer und der weiteren Verfügbarkeit digitaler Medieninhalte setzte eine zweite Welle groß angelegter automatischer Inhaltsanalysen ein. Fan (1988) demonstrierte, dass mit Computerunterstützung umfangreiche Längsschnittstudien, etwa im Agenda Setting und Framing, durchgeführt werden können (Fan & McAvoy, 1989; Fan, 1997). Das Forschungsteam um Schrodt entwickelte eine Software, um internationale Ereignisse aus Tickermeldungen zu extrahieren (Schrodt & Donald, 1990).

Mit den Fortschritten der Künstlichen Intelligenz-Forschung wurden auch Analysemethoden jenseits diktionärbasierter und textstatistischer Verfahren wiederentdeckt. Auf den großen Enthusiasmus von Weber (1984, 142) und anderen folgte jedoch bald die Ernüchterung, dass Computer in absehbarer Zeit keine Texte lesen und verstehen könnten (van Cuilenburg et al., 1988). Nichtsdestotrotz begannen verschiedene Sozialwissenschaftler, syntaktisch-semantische Inhaltsanalysen mit Computerunterstützung zu realisieren, obwohl deren Möglichkeiten mangels generalisierbarer Software-Algorithmen ähnlich begrenzt waren (und zum Teil noch sind) wie die wortbasierten Verfahren in den 60er Jahren.

Seit den 90er Jahren und der Verbreitung des Internet kann das Problem der Verfügbarkeit von digitalen Textinhalten als gelöst gelten. Vielmehr ist es wieder eine Herausforderung, dieser Informationsmenge überhaupt begegnen zu können. Mit zunehmender Rechenkapazität und der Entwicklung statistischer Algorithmen gibt es zumindest in der Informatik und einzelnen sozialwissenschaftlichen Forschungsfeldern auch wieder eine Besinnung auf einfache Bag-of-Words-Ansätze, bei der die syntaktische Struktur von Texten weitgehend außer Acht gelassen wird (Hillard et al., 2007).

Durch die zunehmende Verfügbarkeit (halb-)öffentlicher interpersonaler Kommunikation in Emails, Online-Foren, Chats, Newsgroups oder Social Network Sites wie Facebook oder Myspace sind nicht nur methodisch, sondern vor allem auch inhaltlich neue Impulse für die Inhaltsanalyse festzustellen. Klassische Fragestellungen nach dem Inhalt der Kommunikate werden zusätzlich durch relationale Analysen ergänzt, die umfangreiche und komplexe Kommunikationsprozesse rekonstruieren helfen (vgl. Carley, 1997; Diesner et al., 2005). Die Grenzen hin zur

automatisierten Online-Beobachtung sozialer Interaktionen sind dabei fließend.

Trotz aller Entwicklungen ist die automatische Inhaltsanalyse jedoch noch immer nicht im Mainstream sozialwissenschaftlicher Forschung angekommen. In den aktuellen Auflagen einschlägiger Lehrbücher wird dem Thema meist nur eine Handvoll Seiten gewidmet, in den deutschsprachigen (Merten, 1995; Früh, 2007; Rössler, 2005; Maurer & Reinemann, 2006) noch weniger als in den amerikanischen (Neuendorf, 2002; Riffe et al., 2005). Einzig bei Krippendorff (2004a) findet sich eine umfangreiche Darstellung des Themas. Hinzu kommen zwei Lehrbücher, die sich entweder explizit im Titel (Popping, 2000) oder faktisch (Weber, 1990) mit automatischen Verfahren befassen. Dieser recht spärliche Bestand an Literatur ist auch die Motivation dafür, die Methode in den folgenden Abschnitten ausführlicher darzustellen.[3]

3.1.2 Begriffe und Forschungsprozess

Quantitative Inhaltsanalysen folgen in der Regel einem mehr oder minder standardisierten Ablauf, der mit der Formulierung der Forschungsfrage beginnt und bei der Darstellung und Interpretation der Ergebnisse endet. Dieser Forschungsprozess ist vielfach dargestellt worden, etwa bei Früh (2007, 96), Riffe et al. (2005, 55) oder Krippendorff (2004a, 86), und soll daher hier nicht näher erläutert werden. Automatische und manuelle Analysen haben weitgehend dieselben Anforderungen an Hypothesenbildung, Definition der Untersuchungseinheiten, Stichprobendesign und -ziehung, Qualitätskriterien (Validität und Reliabilität) und Darstellung der Ergebnisse. Sie unterscheiden sich vor allem in zwei Punkten: Erstens orientiert sich die Kategorienbildung und Operationalisierung bei automatischen Verfahren stärker an den technischen Möglichkeiten, zweitens erfolgt die eigentliche Messung oder Codierung, auch als *Text Processing* bezeichnet, durch einen Computeralgorithmus anstelle von menschlichen Codierern. Folgt man dem Phasenmodell von Früh (2007, 96), unterscheiden sich manuelle und automatische Inhaltsanalyse vor allem in der

3 Die folgenden Abschnitte stellen eine Erweiterung eines früher veröffentlichten Beitrags (Scharkow, 2010b) dar, in dem auch entsprechende Software-Pakete vorgestellt werden.

Entwicklungs- und Anwendungsphase, auch wenn viele im Folgenden vorgestellten Ansätze mit spezifischen Forschungsfragen und Auswertungsstrategien verbunden sind. Die zentrale Aufgabe, Inferenzschlüsse von Texten auf deren Anwendungskontext zu ziehen (Krippendorff, 2004a, 18), liegt dabei immer in den Händen des Forschers. Insofern sind immer nur einzelne Analyseschritte mehr oder minder automatisierbar, nicht aber die Forschungsmethode an sich.

Auch wenn sich seit den 60er Jahren die Rahmenbedingungen, unter denen die einzelnen Schritte der Datenerhebung und -codierung ablaufen, gewandelt haben, blieb doch der grundlegende Forschungsprozess bei automatischen Inhaltsanalysen weitgehend unverändert. In Abbildung 3.1 ist der Kern der eigentlichen Analyseschritte zusammengefasst – Hypothesenbildung, Stichprobenziehung und Interpretation werden dabei mitgedacht, müssen an dieser Stelle aber nicht ausführlich erläutert werden, da sie nicht spezifisch für automatische Verfahren sind. Die Reihenfolge der Schritte ist nicht streng linear zu verstehen, da beispielsweise Regeldefinitionen und Datenerhebung unabhängig voneinander stattfinden können. Es scheint allerdings forschungspraktisch sinnvoll, die Software-Codierregeln anhand bereits erhobener und ggf. vorbehandelter Daten zu entwickeln, um verschiedene Fehlerquellen separieren zu können. Bei manueller Codierung (Annotation) müssen die Daten ohnehin bereits bereinigt vorliegen.

Wenn Fragestellung, Untersuchungsanlage und ggf. Hypothesen festgelegt sind, beginnt die eigentliche Feldphase automatischer Inhaltsanalysen mit der Bereitstellung von maschinenlesbaren Inhalten. 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-Archive 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 der (Online-)Stichprobenziehung, des Datenmanagements und des Zugriffs auf relevante Informationen aus großen Datenmengen geworden. Dieses Thema wird ausführlicher in Abschnitt 4.1 behandelt.

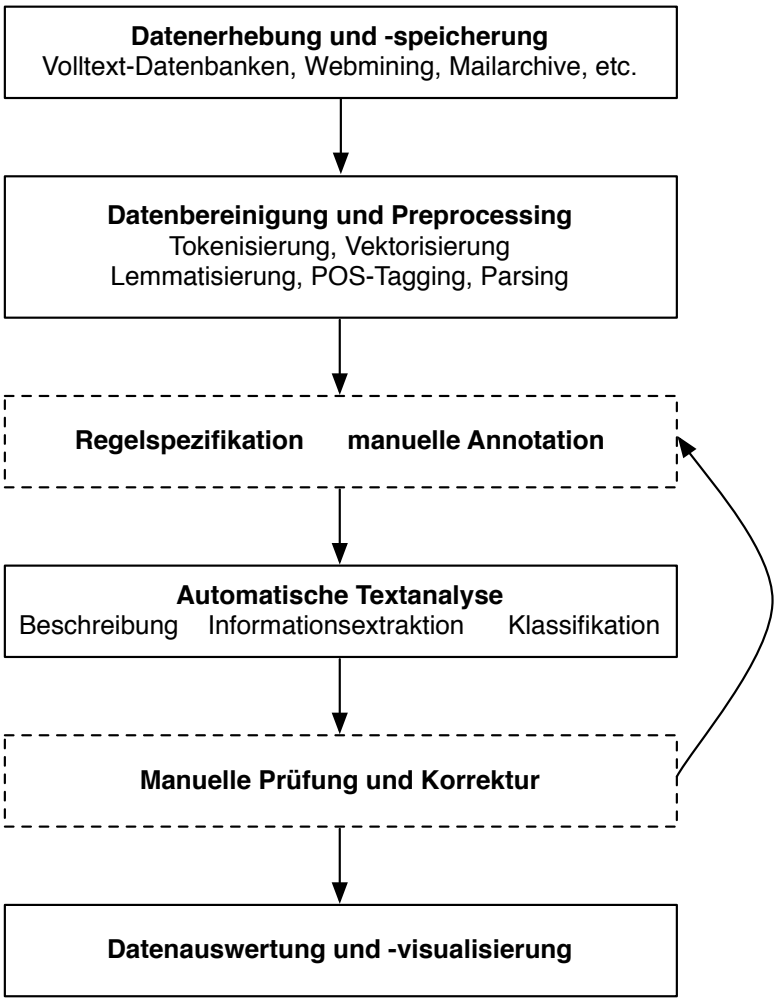

Abbildung 3.1: Typischer Ablauf automatischer Inhaltsanalysen
Die gestrichelten Kästchen stellen Phasen dar, die nicht automatisch ablaufen.

Nachdem ein Text aus dem Internet oder anderen Quellen maschinenlesbar vorliegt, muss er zunächst bereinigt werden, etwa durch Entfernung von irrelevanten oder nicht-textuellen Inhalten.[4] Zudem sollte der Text in ein standardisiertes Format, etwa ASCII-Text, HTML oder XML, umgewandelt werden (vgl. Feinerer et al., 2008), um die Weitergabe und mittelfristige Archivierbarkeit zu gewährleisten.

Anschließend werden die für die Analyse relevanten Merkmale, die so genannten *Features*, aus dem Text extrahiert. In den meisten Fällen sind dies Wörter (Unigramme) bzw. Wortgruppen definierter Länge (N-Gramme). Text ist in dieser Form nichts weiter als eine geordnete Abfolge von Wörtern bzw. ein ungeordneter Worthaufen (*Bag-of-Words*). Analysen auf Zeichen- oder Buchstabenebene sind zwar technisch genauso leicht realisierbar, mangels sinnvoller Fragestellungen und Hypothesen jedoch eher selten anzutreffen.

Vor der eigentlichen Analyse werden die Textdaten häufig mit statistischen oder linguistischen Verfahren vorbehandelt, um die nachfolgenden Analysen zu vereinfachen. Häufig werden etwa besonders häufig oder selten vorkommende Wörter entfernt, gebeugte Wortformen durch Stammformen ersetzt oder Synonyme aufgelöst. Der Nutzen dieses Vorbehandelns (*Preprocessing*) der Dokumente ist aber nicht unumstritten, da hierdurch die inhaltliche und sprachliche Vielfalt eines Textes unter Umständen erheblich verringert wird und darunter ggf. die Validität der automatischen Codierung leiden kann. In Abschnitt 4.2 werden die Möglichkeiten und Konsequenzen des Preprocessing daher ausführlicher diskutiert, da sie auch in der Methodenentwicklung eine wichtige Rolle spielen. Außerdem wird im empirischen Teil der Arbeit der Einfluss zweier häufig verwendeter Preprocessing-Verfahren bei der Verwendung induktiver Klassifikationsalgorithmen überprüft.

Nach der mehr oder minder aufwändigen Vorbehandlung der Texte werden die Datensätze je nach Fragestellung in Untersuchungseinheiten umtransformiert. Am häufigsten ist dabei die Beitrags- oder Dokumen-

4 Im Folgenden wird sich die Diskussion vollständig auf die automatische Textanalyse konzentrieren. Gerade statistische Analysemethoden eignen sich jedoch auch für nicht-textuelle Inhalte. Voraussetzung ist lediglich, dass die Extraktion relevanter Features (z.B. Töne, Farben, Schnitte) automatisierbar ist und diese als Variablen quantifizierbar sind.

Tabelle 3.1: Beispiel einer Term-Dokument-Matrix

	a	fun	is	mining	of	sequence	text	words
Doc 1	0	1	1	1	0	0	1	0
Doc 2	2	0	1	0	1	1	1	1

Quelle: Feinerer et al. (2008, 10)

tenebene. Während die Segmentierung von Dokumenten in kleinere Einheiten, d.h. Absätze oder Sätze, in der Kommunikationswissenschaft häufig zu finden ist, etwa bei der Analyse von Frames, ist die Aggregation vieler Beiträge in größere Korpora eher selten, da diese außerhalb der Korpuslinguistik kaum von Interesse sind.

Nach der Feature-Extraktion und dem Preprocessing liegen die Textdaten zumeist als Dokument-Term-Matrix (vgl. Tabelle 3.1) vor, d.h. einem Datensatz, in dem die Zeilen Dokumenten (ggf. auch Absätzen oder Sätzen) entsprechen und die Spalten den extrahierten Features, d.h. zumeist Wörtern oder Wortgruppen. Die einzelnen Zellen enthalten Informationen zum Vorkommen und ggf. der Häufigkeit eines Terms in einer Untersuchungseinheit. Dieses Datenformat erlaubt mit geringem Aufwand eine Vielzahl statistischer Analysen, die sich teilweise mit gängigen Programmen wie SPSS oder R durchführen lassen, auch wenn spezielle Software ggf. besser mit den umfangreichen Matrizen umgehen kann (vgl. Manning & Schütze, 1999; Feinerer et al., 2008).

Für stärker sprachorientierte Textanalysen, in denen die Syntax und Semantik der Aussagen im Vordergrund steht, eignet sich das Datenformat der Term-Dokument-Matrix nicht, da viele relevante Textmerkmale, etwa die Wortstellung, verloren gehen. Hier wird vielfach auf andere Datenformate zurückgegriffen, die mehr Informationen enthalten und dann anwendungsspezifisch nochmals transformiert werden müssen (Atteveldt, 2008).

In der Untersuchungsphase jeder Inhaltsanalyse werden verschiedene Textcodierungen durchgeführt. Die hierfür verwendeten Verfahren werden im folgenden Abschnitt klassifiziert und anschließend ausführ-

lich dargestellt. Dabei ist sowohl ein deduktives als auch ein induktives Vorgehen möglich (Früh, 2007, 72-74), d.h. es können Kategorien aus dem Text entwickelt und/oder auf diese angewandt werden. Die meisten Verfahren eignen sich jedoch nur für eine der beiden Strategien, so dass ggf. verschiedene automatische Analysen kombiniert werden sollten. Die Entwicklung und Überprüfung verschiedener Kategorien oder Codierschemata ist ein iterativer Prozess, wie durch den Feedback-Pfeil in Abbildung 3.1 angedeutet wird. Dabei profitiert die heutige Forschung von der wachsenden Leistungsfähigkeit der Computer, die dafür sorgt, dass auch bei großen Dokumentenzahlen die eigentliche Codierung so schnell verläuft, dass man von einem *Instant Feedback* sprechen kann.

Auch die bei jeder Inhaltsanalyse notwendigen Reliabilitäts- und Validitätsprüfungen können größtenteils automatisiert werden, so dass die eigentliche Herausforderung in der Interpretation der Testerergebnisse und deren Integration in die statistische Analyse liegt, die in Abschnitt 4.4 diskutiert wird. Bei der Beurteilung der Reliabilität liegt eine weitere Schwierigkeit in der Verknüpfung unterschiedlicher Maße für manuelle und maschinelle Codierung.

Die Frage der Ergebnisdarstellung und -interpretation steht weniger im Zentrum dieser Arbeit, da sich diese bei konventionellen Inhaltsanalysen in gleicher Weise stellt wie bei computergestützten Verfahren. Bei automatischen Analysen von großen Textdatensätzen entsteht dabei leicht die Schwierigkeit, eine Vielzahl von Variablen und Ausprägungen sinnvoll zu visualisieren. Dies gilt jedoch auch für umfangreichere manuelle Analysen (vgl. Adam, 2008).

3.1.3 Typologien der Verfahren automatischer Textanalyse

In diesem Abschnitt sollen Kriterien diskutiert werden, nach denen Verfahren automatischer Inhaltsanalyse systematisiert werden können. Dies geschieht einerseits mit dem Ziel, die Einordnung der einzelnen Ansätze und Studien zu erleichtern, auf die ich im folgenden Kapitel eingehen werde. Andererseits ist es im Rahmen dieser Arbeit notwendig, die bisherigen Typologien um eine wichtige Dimension – die der Spezifikation von Codierregeln – zu erweitern. Es soll *nicht* darum gehen, allgemein

inhaltsanalytische Ansätze und Fragestellungen zu klassifizieren, wie dies von vielen Autoren – von Berelson (1952) und North et al. (1963) bis Krippendorff (2004a) und insbesondere von Merten (1995) – bereits ausführlich getan wurde.

Der Versuch, automatisierte inhaltsanalytische Verfahren zu klassifizieren, ist keineswegs neu. Bereits im Rahmen der Annenberg-Konferenz Ende der 60er Jahre wurden zwei unterschiedliche Ansätze einander gegenübergestellt: Die weitgehend theoriefreie Exploration von Worthäufigkeiten und Konkordanzen auf der einen, die hypothesengeleitete Kategorisierung von Dokumenten nach bestimmten Schlagwörtern auf der anderen Seite (Stone, 1997, 41). Beide Ansätze gehen dabei von der Prämisse aus, dass sich allein auf lexikalischer Ebene, nämlich durch Vorkommen bzw. Häufigkeit einzelner Wörter, der Inhalt eines Textes erschließt. Roberts (1997a) spricht in diesem Zusammenhang von *thematischer* Analyse.

Zwei neuere Typologien von West (2001) bzw. Roberts (1997a, 2000) basieren auf der Unterscheidung von Osgood (1959) in *instrumentelle* und *repräsentationale* Inhaltsanalysen. Hier stehen sich vor allem zwei Fragen im Vordergrund: 1. Welche Bestandteile oder Attribute von Texten sind als Basis für Inferenzschlüsse hinsichtlich des Kommunikators oder der Mitteilung von Nutzen? 2. Wie stark beziehen sich Analysen auf die konkrete Kommunikationsintention des Urhebers bzw. die Interpretation des Forschers (vgl. Shapiro, 1997)?

West (2001) ordnet verschiedene Ansätze in einem Kontinuum an, das von der einfachen Aussagenanalyse (*Overt Message Analysis*) bis hin zur generischen grammatischen Analyse von Kommunikaten reicht. Erstere versucht, die vom Urheber intendierte Nachricht zu analysieren und zu verstehen, sie ist daher stark von der zielgerichteten Kommunikationsstrategie desselben abhängig. Letztere ist sehr viel allgemeiner und nach West (2001, 83) völlig unabhängig von der Kommunikationsintention des Urhebers. Auf der einen Seite des Kontinuums (vgl. die Typologie in Tabelle 3.2) steht die Analyse von Texteigenschaften, die fast vollständig der Kontrolle des Urhebers unterliegt, auf der anderen Seite die linguistische Analyse, die autoren- und kontextunabhängig sein sollte.

Tabelle 3.2: Typologie inhaltsanalytischer Ansätze nach West

Ansatz	Analysestrategie	Problem
Overt Message Analysis	Extraktion von Informationen aus Aussagen	Implentation von echtem algorithmischem Textverständnis bislang nicht gelungen
Representational Analysis	Zuordnung von Wörtern zu Kategorien, Klassifikation von Texten	Verbindung von Wörtern zu inhaltlichen Kategorien begründungsbedürftig
Associational Analysis	Auszählung von Zusammenhängen zwischen Wörtern	Zusammenhang von sprachlichen und kognitiven Konkordanzen begründungsbedürftig
Grammatical Analysis	Analyse von Zeichen-, Wort- und Textstatistiken	Inferenzschluss auf Textentstehung oder -wirkung begründungsbedürftig

Quelle: West (2001)

Ebenfalls in Anlehnung an Osgood unterscheidet Roberts (1997a, 2000) zwei Dimensionen, anhand derer inhaltsanalytische Verfahren klassifiziert werden können. Auf der einen Seite steht die epistemologische Differenzierung zwischen instrumenteller und repräsentationaler Textanalyse und -interpretation. Als instrumentell bezeichnet Roberts (2000, 262) die Untersuchung von Mitteilungen, die sich auf konzeptionelle Vorgaben des Forschers stützt und daher „valide Inferenzen trotz der [kommunikativen] Strategien des Urhebers“(Osgood, 1959, 75, zitiert nach Roberts, 2000) ermögliche. Repräsentationale Analysen versuchen hingegen, „intendierte Aussagen des Autors zu finden und klassifizieren“ (Shapiro, 1997, 228, zitiert nach Roberts, 2000).

Obwohl der Versuch einer Klassifikation jenseits operationaler und statistischer Überlegungen begrüßenswert ist, halte ich diese Unterscheidung ebenso wie die Typologie von West (2001) für problematisch: Erstens werden Erkenntnisinteresse, Analysestrategie und Inferenzschluss konfundiert, was letztendlich dazu führen kann, dass sich repräsenta-

tionale und instrumentelle Verfahren lediglich in der Interpretation der Ergebnisse, nicht aber der eigentlichen Datenerhebung und -analyse unterscheiden. Zweitens ist der Gegensatz zwischen Autorenintention und Interpretationshoheit des Forschers ebenso wenig haltbar wie die Überzeugung, mit repräsentationalen Verfahren könnte der manifeste Inhalt eines Kommunikats objektiv gemessen werden (West, 2001, 82). Eine ausführliche Diskussion und Kritik dieses Problems findet sich bei Krippendorff (2004a, Kap. 2), der auch darauf hinweist, dass eine Vielzahl von Inferenzen bezüglich Quelle, Wirkung oder Kontext von Kommunikaten inhaltsanalytisch begründbar sind. Drittens sind beide Klassifikationen nicht spezifisch auf die methodologischen Problemstellungen *automatischer* Inhaltsanalysen ausgerichtet, obwohl durchaus beispielhaft computergestützte Ansätze in die Typologie eingeordnet werden.

In letzterer Hinsicht ist die zweite Dimension der Klassifikation von Roberts (2000) deutlich besser geeignet, da sie sich auf die Quantifizierung von sprachlichen Symbolen und deren algorithmischer Verarbeitung bezieht: Roberts unterscheidet zwischen thematischen, semantischen und netzwerkbasierten Analysen von Texten und illustriert diese anhand der Datenmatrizen, die spezifisch für den jeweiligen Ansatz sind. Die klassische thematische Analyse ordnet jede Untersuchungseinheit einem oder mehreren Themen zu, während die semantische Analyse pro Aussage Subjekt, Handlung und Objekt festhält. In netzwerkanalytischen Ansätzen werden Referenzen zwischen semantischen Einheiten festgehalten.[5] Die Unterscheidung zwischen thematischen und semantischen Ansätzen wurde von verschiedenen Autoren aufgenommen (Popping, 2000; Krippendorff, 2004a) und wird auch in der im Folgenden vorgestellten Typologie in anderer Form wieder auftauchen.

In dieser Arbeit soll eine weitere Typologie vorgeschlagen werden, die in Teilen die Unterscheidung zwischen semantischen und thematischen Analysen enthält, aber vor allem die Anforderungen und Einflussmög-

5 Streng genommen ist das netzwerkanalytische Verfahren nicht auf derselben Abstraktionsebene angeordnet wie die beiden anderen Ansätze. So kann man netzwerkanalytische Analysemethoden auf bereits erhobene thematische oder semantische Daten anwenden. Die netzwerkanalytische Verarbeitung semantischer Daten ist dabei häufiger der Fall, weshalb Krippendorff (2004a) und Atteveldt (2008) beides als *Semantic Network Analysis* zusammenfassen.

lichkeiten des Forschers und des Computers im Forschungsprozess als Maßstab nimmt. Die Typologie ist daher eher methodisch-praktisch als epistemologisch motiviert.

Grundsätzlich lassen sich unüberwachte explorative und überwachte hypothesengeleitete Verfahren automatischer Textcodierung unterscheiden. Für erstere ist keinerlei aufwändige und kostenintensive Regelspezifikation bzw. manuelle Codierung der Texte (vgl. Abbildung 3.1 auf Seite 51) notwendig – es kann sofort mit der automatischen Analyse begonnen werden. Unüberwachte Textanalysen sind dementsprechend mit dem geringsten Aufwand verbunden, da sich der gesamte Analyseprozess vollautomatisch durchführen lässt. Der Nachteil solcher Verfahren liegt jedoch auf der Hand: Der Einfluss des Forschers auf die Codierung ist begrenzt, da die Analyse nur indirekt beeinflusst werden kann, etwa durch die Wahl geeigneter Analyseverfahren und Algorithmen. Da die Ergebnisse unüberwachter Analysemethoden lediglich statistisch determiniert sind, ist die Interpretation der Ergebnisse durch den Forscher umso wichtiger. Die Validität der Inferenzen ist gerade bei diesen Verfahren häufig Gegenstand heftiger Kritik.

Für die sozialwissenschaftliche Praxis ist die Vorgabe von Codierschemata durch den Forscher in den meisten Fällen essentiell, da es bei der Inhaltsanalyse immer um *spezifische* Dimensionen eines Textinhaltes geht und diese im Voraus für den Codierprozess operationalisiert werden müssen. Hier sind überwachte Verfahren notwendig, bei denen der Forscher der Software Regeln oder Beispiele vorgibt, 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 manuellen Codierung von Texten ab. Es muss daher stets ein Kompromiss zwischen dem Operationalisierungsaufwand 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 computergestützter Analyse verorten. Daher lohnt es sich, die vorhandenen Verfahren überwachter Codierung feiner zu systematisieren. Hierbei

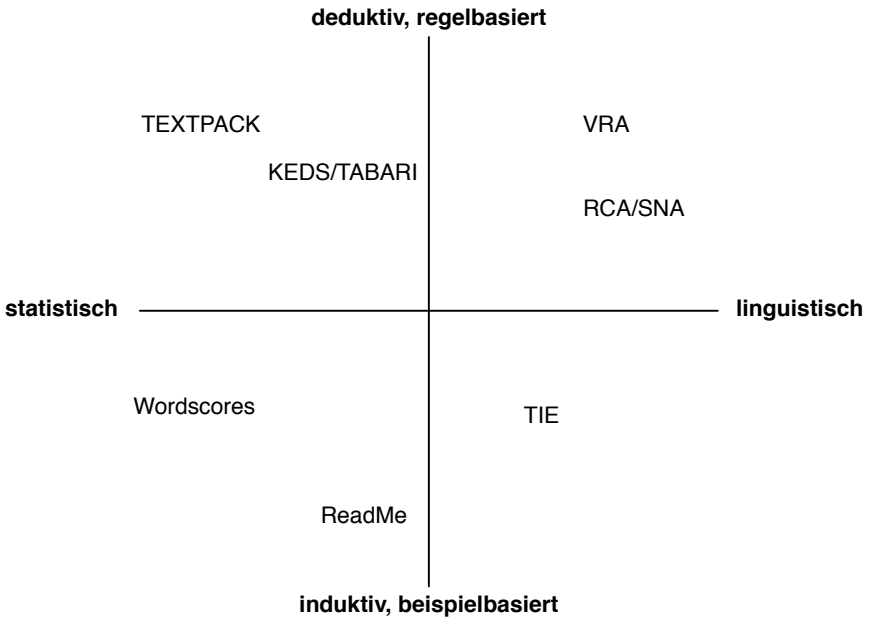

Abbildung 3.2: Klassifikation hypothesengeleiteter Textanalyse-Software

kann man sich an zwei Dimensionen orientieren: Der Grad sprachlicher Analysetiefe und die Art der Regelspezifikation. Legt man diese beiden Dimensionen zu Grunde, kann man statistische und linguistische sowie deduktive und induktive Verfahren unterscheiden.

Die erste Dimension in Abbildung 3.2 stellt statistische und linguistische Ansätze einander gegenüber (vgl. Monroe & Schrodt, 2008): Wird bei rein wortbasierten Verfahren jedes Wort oder N-Gramm als eine einfache, isolierte Variable aufgefasst, versuchen syntaktisch-semantische 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 gerichtete Beziehung zwischen „Peter“ (Subjekt) – „sehen“ (Verb) – „Haus“ (Objekt). Traditionell versucht man mit statistischen Verfahren, Dokumente thematisch zu klassifizieren (ein Dokument handelt von Peter, wenn das Wort „Peter“ vorkommt), während linguistische Ver-

fahren auf Aussageebene Beziehungen analysieren (die Entität „Peter“ vollzieht die Handlung „sehen“ im Bezug auf die Entität „Haus“) und damit Antworten auf offene Fragen zum Text geben sollen. Hinsichtlich des Erkenntnisinteresses entspricht diese Unterscheidung derjenigen von Roberts (2000) in thematische und semantische Analysen. Aus der Perspektive der konkreten Softwareimplementation sind thematische Analyse häufig leichter umzusetzen, weil sie auf klassischen statistischen Verfahren beruhen, die nichts von der sprachlichen Struktur der Inhalte wissen müssen. Linguistische Analysen bedürfen hingegen spezialisierter Software (z.B. Syntax-Parser), die die sprachliche Struktur der Aussagen extrahieren können. 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).

Die zweite Dimension der Typologie in Abbildung 3.2 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 verbreitet Praxis bei der automatischen Textanalyse. Dies führt allerdings dazu, dass manuelle und automatische Verfahren stark voneinander abgekoppelt sind. Eine manuelle Codiererschulung ist in den meisten Fällen beispielbasiert, da viele komplexe Konstrukte nach wiederholtem Üben richtig codiert werden, aber selten explizite und umfassende Regeln (etwa, wann ein Presseartikel negativ über einen Wahlkandidaten berichtet) vom Forscher formuliert werden. Grundsätzlich erfordern also deduktive Verfahren mehr konzeptionelle Vorarbeit vom Forscher, während induktive Verfahren vor allem auf viele und zuverlässig codierte Beispieltexte angewiesen sind. Die Entwicklung inhaltsanalytischer Instrumente profitiert dabei von der gewachsenen Rechenleistung moderner Computer: Nach jeder Änderung der Regeln oder zusätzlichen Annotationen kann das Instrument getestet werden, weil die automatische Codierung selbst mit

großen Mengen an Texten und/oder Kategorien nur Minuten dauert. Der Feedback-Prozess in Abbildung 3.1 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).

Insgesamt lassen sich automatische Verfahren der Inhaltsanalyse in der hier vorgestellten Typologie in zwei bzw. drei Dimensionen gegenüberstellen: Grundlegend unterscheide ich in unüberwachte (deskriptive bzw. explorative) vs. überwachte Verfahren. Innerhalb der überwachten Verfahren in statistische vs. linguistische und deduktive vs. induktive Ansätze. Diese Kategorien bilden auch die Grundlage für die nächsten Abschnitte.

3.2 Deskriptive und explorative Verfahren

3.2.1 Textstatistik

Die Berechnung von Text-, Satz- und Wortstatistiken ist eines der ältesten und einfachsten Verfahren automatischer Textanalyse. Es beruht auf der Auszählung von bestimmten Zeichen bzw. Zeichenkombinationen in Dokumenten, der anschließenden beschreibenden Darstellung der Ergebnisse und nicht zuletzt der Überprüfung von Zusammenhangs- und Unterschiedshypothesen, bezogen auf unterschiedliche Texte oder Textkorpora. Der Vorteil von automatischen Verfahren beruht dabei nicht nur auf der Tatsache, dass Computer solcherlei Aufgaben extrem schnell und vollständig reliabel erledigen, sondern auch in der besonderen Schwierigkeit, die Menschen bei dieser Tätigkeit, im Gegensatz zu syntaktischen oder semantischen Analysen, haben. Der Zugewinn an Zeit *und* Genauigkeit gegenüber manuellem Auszählen ist daher gerade auf diesem Gebiet enorm. Beim Vergleich computergestützter Auszählungen mit den eigenen früheren Versuchen kommen Mosteller & Wallace (1964, 7) zu der „wichtigen empirischen Erkenntnis“: „people cannot count, at least not very high“.

Obwohl die Beschreibung von Texten durch Häufigkeiten und Mittelwerte auf den ersten Blick trivial erscheint, können doch verschiedene interessante und wissenschaftlich relevante Konzepte mit textstatisti-

schen Maßen operationalisiert werden. In Anlehnung an die Typologie von Holsti (1969, 26) lassen sich textstatistische Verfahren sowohl für die Beschreibung der Kommunikate als auch zur Inferenz auf deren Entstehungsbedingungen und Rezeptionskontexte einsetzen.

Beschreibung der Kommunikate
Ein prominentes Beispiel für die Relevanz textbeschreibender Analysen ist die Disziplin der Korpuslinguistik, in der Sprachtheorien – etwa auf lexikographischer und syntaktischer Ebene – aus der statistischen Analyse von umfangreichen Dokumentenkorpora gewonnen und überprüft werden (Lemnitzer & Zinsmeister, 2006). Die bekanntesten Theorien auf diesem Gebiet wurden zwar ohne Computerhilfe formuliert, aber erst mit diesen empirisch auf eine breite Basis gestellt. Das gilt insbesondere für das Zipfsche Gesetz, nach dem die Häufigkeit (bzw. Wahrscheinlichkeit) eines Wortes umgekehrt proportional zur Position desselben in einer geordneten Rangreihe aller Wörter ist: $p(x) \sim \frac{1}{n}$ (Zipf, 1965).

Auf Basis dieser Auszählung lassen sich besonders häufig vorkommende Wörter (so genannte Stopwörter) einer Sprache dokumentieren (vgl. Quasthoff, 1998). Da es sich bei diesen Stopwörtern zumeist um Präpositionen, Artikel, Pronomen oder Konjunktionen handeln, die in fast allen Textsorten und Themenkontexten vorkommen, werden diese häufig aus der eigentlichen Wortfrequenz-Analyse ausgeschlossen, um den Blick auf die thematisch relevanten Wörter zu erleichtern. Eine um Stopwörter bereinigte Häufigkeitstabelle für Wörter und Wortgruppen kann in vielen Fällen zur Veranschaulichung und Zusammenfassung großer Textmengen eingesetzt werden. Dies lässt sich auch grafisch visualisieren, etwa indem die Schriftgröße proportional zur Worthäufigkeit gewählt wird, wie dies etwa beim automatischen Verschlagworten (*Tagging*) von Blog-Einträgen geschieht (vgl. Brooks & Montanez, 2006). Die in Abbildung 3.3 dargestellte *Word Cloud* basiert auf den bereinigten Worthäufigkeiten einer Stichprobe von 1000 Schlagzeilen und Leads aus der Evaluationsstudie (vgl. Kapitel 5ff.) und wurde mit dem Online-Tool Wordle[6] generiert. Hierbei wird schlicht die Schriftgröße proportional zur Worthäufigkeit gewählt.

6 http://wordle.net

Abbildung 3.3: Worthäufigkeiten aus einer Stichprobe von 1000 Nachrichtenmeldungen von Mai 2008 bis Mai 2009
Quelle: 12 deutsche Nachrichtenwebsites (vgl. Abschnitt 6.1)

Über die Differenzbildung von generell und beitragsspezifisch häufig vorkommenden Wörter können, wie Landmann & Züll (2008) zeigen, auch außergewöhnliche Themen oder besondere Ereignisse identifiziert werden. Die dabei herausragenden Begriffe können gleichzeitig zur Beschreibung dieser Ereignisse verwendet werden.

Schlüsse auf die Entstehung von Texten
Textstatistische Verfahren haben besonders in der literaturwissenschaftlichen Forschung einen hohen Stellenwert. Dabei kommt der Stilometrie und der damit eng verbundenen Autorenschaftsforschung eine zentrale Bedeutung zu. Die Grundidee bei diesen Verfahren ist die Spezifizität sprachlicher Textgestaltung, die dabei hilft, Autoren oder Epochen zu identifizieren. Obwohl die Idee zu diesem Vorgehen schon aus der Mitte des 19. Jahrhunderts stammt, wurden entscheidende Fortschritte

auf diesem Gebiet erst mit der Möglichkeit automatischer Verfahren erzielt (vgl. Holmes, 1998; Tankard, 2001). In ihrer bahnbrechenden Studie zu den Federalist-Papers konnten Mosteller & Wallace (1964) die nicht namentlich gekennzeichneten Artikel aufgrund der Häufigkeiten von Füllwörtern wie „any“, „while“ und „upon“ den Autoren Alexander Hamilton und James Madison zuordnen. In den folgenden Jahrzehnten wurden unzählige textstatistische Indikatoren wie durchschnittliche Wort- und Satzlänge, Anzahl von Satz- und Sonderzeichen sowie die Häufigkeiten bestimmter Wendungen und sogar Buchstabenkombinationen auf eine ebenso große Zahl von Dokumenten angewandt (vgl. Grieve, 2007). Von der Zuordnung von Sonetten zum Shakespeare-Kanon über die Identifikation des Autors von *Primary Colors* bis hin zu juristischen Analysen von Täterschaft (Holmes, 1998; Adams, 2004), derer sich in jüngster Zeit auch Sicherheitsbehörden bei der Überwachung von Online-Kommunikation bedienen (Abbasi & Chen, 2005; Pennebaker & Chung, 2009), gibt es eine große Bandbreite von Einsatzmöglichkeiten.

Obwohl der grundsätzliche Nutzen stilometrischer Verfahren heute kaum bestritten wird, konnte doch bislang kein Patentrezept für alle Anwendungsfälle entwickelt werden. Der recht vollmundigen Ankündigung von Morton (1963), mit Hilfe von sieben Indikatoren (u.a. Satzlänge und Häufigkeit des Verbs *sein*) die Autorenschaft der Paulus-Episteln geklärt und damit die Bibelforschung revolutioniert zu haben, setzte Ellison (1965) eine Replik entgegen, dass dann der Joyce'sche *Ulysses* von mindestens fünf Autoren geschrieben worden sei, von denen allerdings keiner auch *Portrait of the Artist as a Young Man* verfasst haben könne (Holsti, 1969, 87). Hoover (2003) kommt in seiner Zusammenfassung zum Ansatz der literarischen Wortschatzanalyse sogar zu der Schlussfolgerung, insgesamt sei das Verfahren nicht reliabel und valide genug, um sinnvoll in der Autorenschaftsforschung eingesetzt zu werden.

Textstatistische Analysen spielen jedoch nicht nur in den Geisteswissenschaften eine wichtige Rolle. Auch in der Psychologie, insbesondere der Individualdiagnostik, gibt es eine lange Tradition, Patientengespräche und andere Aussagen zu transkribieren und quantitativ auszuwerten (Gottschalk & Gleser, 1969; Gottschalk, 2000; Pennebaker et al., 2003). Ein in diesem Zusammenhang wichtiges Maß ist die *Type-Token-Relation*

(TTR), also das Verhältnis von Gesamtzahl an Wörtern zur Menge unterschiedlicher Wörter im Text. Dieser Indikator für den Vokabularreichtum korreliert mit verschiedenen Persönlichkeits- und Entwicklungsmerkmalen (Holsti, 1969, 75f.). Die TTR und das Verhältnis von Adjektiven und Verben unterscheiden sich beispielsweise signifikant bei gesunden und schizophrenen Menschen (Hammer & Salzinger, 1964).

Auch in der Kommunikationswissenschaft finden textstatistische Maße häufig Verwendung. So lässt die durchschnittliche Länge (Wörterzahl) eines Artikels zum Beispiel Schlüsse auf Genre, Medium oder journalistische Arbeitspraxis zu. Die TTR könnte ebenso als Indikator journalistischen Stils verwendet werden, etwa zur Unterscheidung von Qualitäts- und Boulevardmedien. Zudem ist die Beitragslänge ein entscheidender Indikator für den Nachrichtenwert einer Meldung (Schulz, 1976).

Schlüsse auf die Wirkung von Texten

Auch die Komplexität und Lesbarkeit eines Textes lässt sich durch Wort- und Satzlänge oder Umfang des benutzten Vokabulars erschließen (DuBay, 2004; Best, 2006; Krippendorff, 2009). Lesbarkeitsmaße, die sich aus mehreren solcher Indikatoren zusammensetzen, werden häufig in der Bildungs- und Usability-Forschung eingesetzt. Die dabei verwendeten Formeln, etwa von Flesch (1948), Coleman & Liau (1975), Bjornsson (1983) oder Bamberger & Vanecek (1984), enthalten oft dieselben oder ähnliche Indikatoren (Satz- und Wortlänge, Anzahl langer Wörter) mit unterschiedlicher Gewichtung. Für die automatische Analyse eignen sich allerdings nicht alle Indikatoren gleich gut: Silbenbasierte oder syntaktische Indikatoren lassen sich nicht so problemlos umsetzen wie zeichen- oder wortbasierte Maße, die ohne phonetische oder syntaktische Regelsätze auskommen, die zuerst spezifiziert werden müssten.

Obwohl die Lesbarkeitsanalyse in vielen älteren Standardwerken mehr oder minder ausführlich beschrieben ist (Holsti, 1969; Lisch & Kriz, 1978), werden Maße der Textkomplexität und Lesbarkeit vergleichsweise selten angewandt. Eine aktuelle Ausnahme ist die Arbeit von Kercher (2010), in der sich auch eine Übersicht zum Forschungsfeld und eine Liste von Lesbarkeitsformeln findet.

Ein zentrales Problem textstatistischer Verfahren liegt in der Tatsache, dass sie zumeist induktiv zur Hypothesengenerierung eingesetzt und dann häufig ohne bzw. mit ex-post-Erklärungen generalisiert werden. Korrelate textstatistischer Eigenschaften werden selten theoretisch abgeleitet, sondern durch den Vergleich gemessener Text- und Urhebereigenschaften entdeckt (vgl. Pennebaker & Chung, 2009, 454). Krippendorff (2004a, 344) äußert sich daher kritisch zu dieser operationalistischen Arbeitsweise:

> Unfortunately, the history of content analysis is full of examples of researchers who have merely declared counts to be indices of phenomena, usually of social or political significance, without spelling out how their claims could be validated.

Sowohl bei den Ursachen (z.B. Eigenschaften des Urhebers) als auch bei den Folgen (z.B. Lesbarkeit) ist es daher notwendig, textstatistische Maße mit anderen Daten zu validieren. Dies gilt selbstverständlich für jegliche Inhaltsanalyse, jedoch ist es besonders augenfällig, dass für sich genommen das Zählen von Zeichen und Wörtern kaum wissenschaftliche Relevanz hat. Der vergleichsweise leichten Umsetzung steht daher ein größerer Anspruch an die (Mess-)Theoriebildung gegenüber.

3.2.2 Co-Occurrence und Latente Semantische Analyse

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 bi- bzw. multivariate Erweiterung der einfachen Wortstatistik. Die Co-Occurrence-Analyse basiert auf der Annahme, dass kognitiv bzw. semantisch zusammenhängende Konstrukte auch räumlich nahe beieinander stehen (Krippendorff, 2004a, 206). Betrachtet man die Wörter innerhalb eines spezifizierten Rahmens, etwa in kompletten Sätzen oder Absätzen, lässt sich das gemeinsame Auftreten bestimmter Begriffe in eine Kontingenztabelle oder eine Ähnlichkeitsmatrix überführen (Landmann & Züll, 2004; Galliker & Herman, 2003). Da jedoch aus stilistischen Gründen synonyme Wörter selten innerhalb eines Satzes oder Absatzes

erscheinen, können und sollten auch Co-Occurrences höherer Ordnung und Komplexität in die Berechnung der Ähnlichkeitsmatrix einbezogen werden (vgl. Stone, 1969a, 527).[7]

Die Grundlagen dieser Assoziationsanalyse legte bereits Osgood (1959), der allerdings mit codierten Kategorien statt Rohdaten in Form von Wörtern und Wortgruppen arbeitete und dann mittels χ^2-Tests statistisch signifikante Zusammenhänge untersuchte. Abschließend wurden relevante Zusammenhänge durch Netzwerkgraphen visualisiert – ein Verfahren, dass auch heute noch bei Marken-Assoziationsanalysen gebräuchlich ist (Teichert & Schöntag, 2009).

Da nach dem Zipfschem Gesetz Worthäufigkeiten extrem schief verteilt sind und die Kontingenzmatrizen schon bei kurzen Dokumenten sehr umfangreich werden, bieten sich statt χ^2-Tests explorative Verfahren zur Datenverdichtung an, um bedeutsame Zusammenhänge überhaupt identifizieren zu können. So lassen sich mit einer Wort-Co-Occurrence-Matrix konventionelle hierarchische Clusteranalysen durchführen, die eine große Menge an Wörtern zu wenigen semantische Gruppen zusammenfassen. Mit einer anschließenden multidimensionalen Skalierung lassen sich die semantischen Cluster sowie deren Positionierung zueinander visualisieren (Salisbury, 2001). Abbildung 3.4 zeigt eine solche Darstellung zum Thema Irak-Krieg von Landmann & Züll (2004).

Die erste und bekannteste Anwendung dieses explorativen Ansatzes stammt von Iker & Harway (1969), die dafür auch die erste Software WORDS entwickelten. Aufgrund der Restriktionen in der verfügbaren Hardware war es nötig, selbst kleine Textkorpora aufwändig manuell vorzubehandeln, da letztendlich nur 200 Wörter in der Berechnung der Co-Occurrence-Matrix Platz fanden. Dieses Problem ist mittlerweile zumindest für gängige Textmengen gelöst, etwa durch die Verwendung von effizienten neuronalen Netzen wie im Programm CATPAC, das für die meisten automatischen Co-Occurrence-Analysen verwendet wird (Doerfel & Barnett, 1996; Salisbury, 2001). Da jedoch die Interpretation von Clusterlösungen und MDS-Grafiken mit vielen zehntausend Objekten

7 Eine Co-Occurrence zweiter Ordnung liegt z.B. vor, wenn X und Y nie zusammen auftauchen, da sie Synonyme sind, aber sowohl X als auch Y häufig mit dem Wort Z in Verbindung stehen (Lemaire & Denhière, 2006). Diese Logik lässt sich beliebig fortführen.

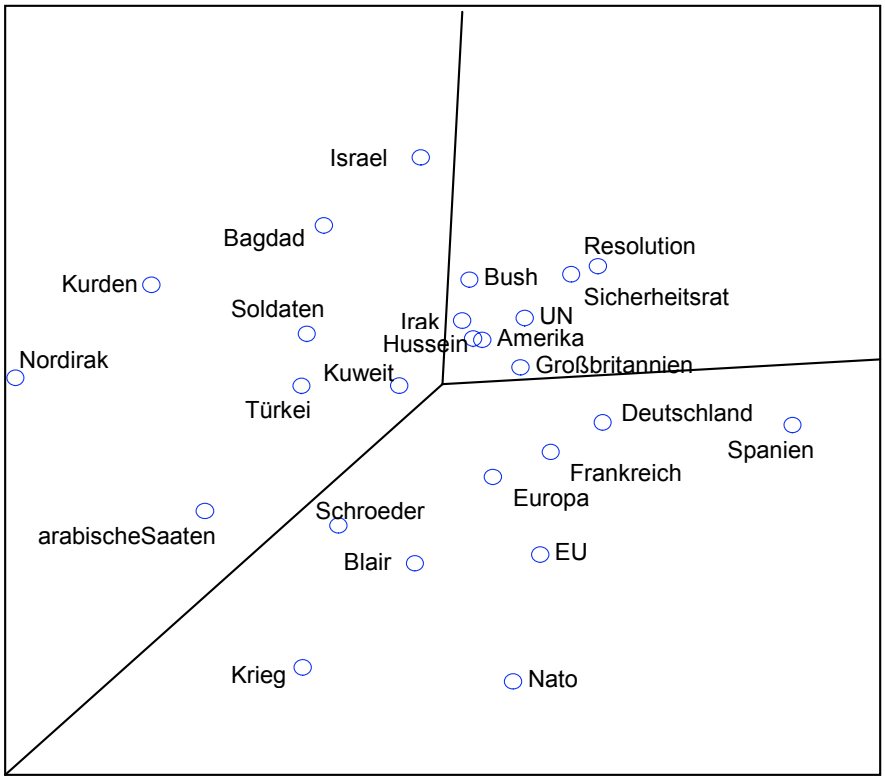

Abbildung 3.4: Multidimensionale Skalierung von Co-Occurrence-Daten zum Kriegsbeginn im Irak, Quelle: Landmann & Züll (2004, 135)

ohnehin schwierig ist, wird auch mit moderner Hard- und Software die Selektion von relevanten Wörtern zum Problem (Galliker & Herman, 2003, 100). In den meisten Studien wird daher zuerst eine Liste der am häufigsten vorkommenden Wörter erstellt, in deren Umfeld dann Co-Occurrences mit anderen Begriffen gezählt werden.

Eng mit der Co-Occurence-Analyse verwandt ist ein faktoranalytisches Verfahren, die Latente Semantische Analyse oder Indizierung (LSI), die ebenfalls der Verdichtung von großen Dokument-Text-Matrizen dient, um mit diesen reduzierten Komponenten analytisch weiterzuarbeiten. Sie basiert grundsätzlich auf der Logik der Einzelwertzerlegung von

Matrizen, die auch der Hauptkomponentenanalyse zugrunde liegt (Deerwester et al., 1990; Manning et al., 2008). Bereits bei Iker & Harway (1969) werden die Wörter als Variablen (und nicht als Fälle) behandelt und deren gemeinsames Auftreten korrelativ analysiert.[8] Best (1997) nutzt LSI, um thematisch verwandte Diskussionen in Online-Newsgroups zu gruppieren und deren Entwicklungen längsschnittlich zu analysieren.

Cluster- und faktoranalytische Co-Occurrence-Ansätze basieren grundsätzlich auf denselben theoretischen Überlegungen, unterscheiden sich jedoch hinsichtlich ihrer Zielsetzung: Wählt man Wörter als Fälle, können im Text vorhandene semantische Strukturen aufgedeckt werden, jedoch geht die Zuordnung zu bestimmten Dokumenten verloren. Im Vordergrund der Analyse steht dementsprechend die Informationsextraktion und -verdichtung, nicht die Klassifikation einzelner Texte. Ein Beispiel für diesen Ansatz ist die Analyse von Markenassoziationen, bei der die gesamte verbale Beschreibung eines Image und nicht die einzelne Äußerung den Untersuchungsgegenstand darstellt (Salisbury, 2001; Teichert & Schöntag, 2009). Stephen (1999) analysiert mit einer Co-Occurrence-Analyse die Titel von kommunikationswissenschaftlichen Zeitschriftenartikeln und deckt dabei verschiedene Themencluster in der Disziplin auf. Klebanov et al. (2008) analysieren politische Reden unter dem Begriff der *lexikalischen Kohäsion* und können so zentrale semantische Konzepte aus Äußerungen Margaret Thatchers extrahieren.

Auch in der psychologischen Forschung von Iker & Harway (1969) geht es um die Analyse von kognitiven Zusammenhängen und nicht primär um die Klassifikation der Probanden. Da sie jedoch faktoranalytisch an die Daten herangehen, ist die nachträgliche Zuordnung bzw. sogar

8 Sowohl bei der Cluster- als auch bei der Hauptkomponentenanalysen stellt sich die Frage, ob und ggf. welche Verfahren angesichts der Verteilung von Worthäufigkeiten überhaupt sinnvoll sind. Korrelative Verfahren gehen von mindestens intervallskalierten Daten aus – eine Annahme, die streng genommen nicht gegeben ist. Mosteller & Wallace (1964) legen ihren Analysen daher eine Poisson-Verteilung für Zähldaten zugrunde. Bei der Clusteranalyse ist hingegen die Wahl der passenden Agglomerationsmethode von zentraler Bedeutung. Diese hängt einerseits von Skalenniveau der Daten ab, andererseits von theoretischen Überlegungen, etwa ob das gemeinsame Auftreten von Wörtern genauso relevant ist wie das gemeinsame Nicht-Auftreten (Brosius, 2006, 625ff.). Gerade das gemeinsame Nicht-Auftreten von seltenen Wörtern führt bei korrelativen Verfahren zu künstlich hohen Zusammenhängen.

die Zuweisung von Factor-Scores zu individuellen Antworten möglich. Dieses Verfahren setzen beispielsweise auch Simon & Xenos (2004) für die Verdichtung und Interpretation verschiedener latenter Antwortkategorien bei offenen Fragen ein. Auch bei der Operationalisierung von Frames als Cluster von thematischen Elementen sind Co-Occurrence-Ansätze sinnvoll einzusetzen (Miller, 1997; Matthes & Kohring, 2008).

Der Nutzen von explorativen Co-Occurrence-Verfahren ist im Fach höchst umstritten. Schon in der Abschlussdebatte der Annenberg-Konferenz (Stone, 1969a) wurde deutlich, dass einige Forscher das Verfahren als theorielos und empiristisch ablehnen, während andere gerade in der Abwesenheit von Forschereinflüssen Chancen für maximale Objektivität und Reliabilität sehen (Salisbury, 2001, 69). Während Iker & Harway (1969) die Möglichkeit betonen, auf diese Weise Themen und Konzepte aus den Texten selbst hervortreten zu lassen, verweist etwa Krippendorff (2004a, 22) explizit darauf, dass Texte an sich keinerlei Bedeutung haben, sondern diese immer vom Leser bzw. Inhaltsanalytiker anhand des Textes konstruiert wird. Bezogen auf Co-Occurrence-Analysen heißt das, dass die Interpretation der gefundenen textstatistischen Zusammenhänge dem Forscher obliegt. 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 & Züll (2004) bei ihrer Evaluation des Programms CATPAC feststellen. Obwohl das Verfahren selbst vollautomatisch abläuft, sind in den meisten publizierten Co-Occurrence-Studien so viele manuelle Vorbereitungen nötig, z.B. in der Auswahl der relevanten Wörter, dass man kaum von einem vollautomatischen theoriefreien Ansatz sprechen kann. Dies lässt sich anhand der Studie von Galliker & Herman (2003) illustrieren, in der explizit nur die Co-Occurrence-Tabellen im Bezug auf die Wörter Mann/Frau berechnet und visualisiert werden. Hier ist bereits eine Verknüpfung mit diktionärbasierten Verfahren (Kap. 3.3.1) erkennbar.

Grundsätzlich eignen sich 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.

3.2.3 Automatische Dokumentklassifikation

Die automatische unüberwachte Klassifikation von Dokumenten ist eine klassische Anwendung im Bereich der statistischen Textanalyse (Jain et al., 1999; Hotho et al., 2005). Basierend auf einer – nach Bedarf reduzierten und semantisch verdichteten – Dokument-Term-Matrix, in der jedes Dokument einen Fall darstellt, werden wie bei der Co-Occurrence-Analyse clusteranalytische Verfahren angewandt, um einander ähnliche Dokumente zu gruppieren, man spricht daher in der Informatik auch von *Document Clustering*.

Fast alle Suchmaschinen und Web-Verzeichnisse bedienen sich clusteranalytischer Verfahren, um eine große Zahl an Web-Dokumenten sinnvoll zu strukturieren (Broder et al., 1997). Aus diesem Grund ist die Literatur zu diesem Thema sehr umfangreich und die Methodenentwicklung schreitet rasant voran (Manning et al., 2008). Im folgenden Abschnitt werden daher nur die Grundlagen und einige sozialwissenschaftliche Beispiele für unüberwachte Dokumentklassifikation vorgestellt.

Der vollautomatischen Klassifikation von Texten liegt die Annahme zugrunde, dass Dokumente, in denen die gleichen Wörter 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 bei unterschiedlichen Dokumentlängen euklidische Distanzen für die Clusterung weniger gut geeignet sind (Ghosh & Strehl, 2006; Manning et al., 2008). Die dabei entstehende Distanzmatrix der Dokumente kann anschließend als Ausgangspunkt für beliebige clusteranalytische Verfahren eingesetzt werden. Dabei wird grundsätzlich zwischen hierarchisch-agglomerativen und partitionierenden Verfahren unterschieden (Aldenderfer & Blashfield, 1984).

Hierarchisch-agglomerative Verfahren
Bei der hierarchischen Clusteranalyse werden schrittweise die Dokumente, die sich am ähnlichsten sind, in einem Cluster zusammengefasst. Dies ist bei der Klassifikation von Texten besonders sinnvoll, da die impliziten (und vom Forscher nicht erzwingbaren) Zielkategorien ohnehin häufig in Form eines einer hierarchischen Kategorienbaums vorliegen. Ausgehend von Einzeldokumenten über spezifische und allgemeinere Cluster kann ein ex post definierbarer Abstraktionsgrad der Kategorien festgelegt werden.

Abbildung 3.5 zeigt ein Dendrogramm für die hierarchische Clusterung von Reuters-Meldungen. Die Meldungen im unteren Teil der Grafik könnten z.B. zunächst unter der Kategorie *Zinsen*, auf höherer Ebene unter *Finanzen* oder sogar *Wirtschaft* zusammengefasst werden. Die Entscheidung, wie viele und welche Cluster nun sinnvoll sind, hängt letztlich immer von der nachträglichen Interpretation des Forschers ab. Auch wenn es verschiedene statistische Gütemaße für die optimale Clusterzahl gibt, sind inhaltliche Validitätskriterien häufig relevanter für die Annahme oder Ablehnung einer bestimmten Clusterlösung (vgl. Manning et al., 2008, 356-60).

Das bekannteste Beispiel für die Anwendung hierarchischer Textklassifikation ist der Online-Service *Google News*, in dem innerhalb eines Ressorts thematisch verwandte Artikel aus verschiedenen Quellen gruppiert dargestellt werden. Ein Vorbild für diesen Dienst war u.a. das an der Columbia University entwickelte NewsBlaster (McKeown et al., 2002), für das Nachrichtentexte und -bilder ausgewertet wurden. Eine andere Textgrundlage wählen Brooks & Montanez (2006), die eine Themenhierarchie aus Blogeinträgen nur aus deren Technorati-Tags (Schlagwörtern) anstelle der Volltexte ableiten.

Während Document-Clustering vor allem im Bereich von Online-Suchmaschinen häufig eingesetzt wird, sind die Anwendungen im sozialwissenschaftlichen Kontext noch recht selten. In einer aktuelle Studie von Quinn et al. (2006) werden die Reden von Kongressabgeordneten nach Themengebieten geclustert, zudem werden die Möglichkeiten längsschnittlicher Policy-Agenda-Analysen mit den Mitteln automatischer Dokumentverarbeitung aufgezeigt.

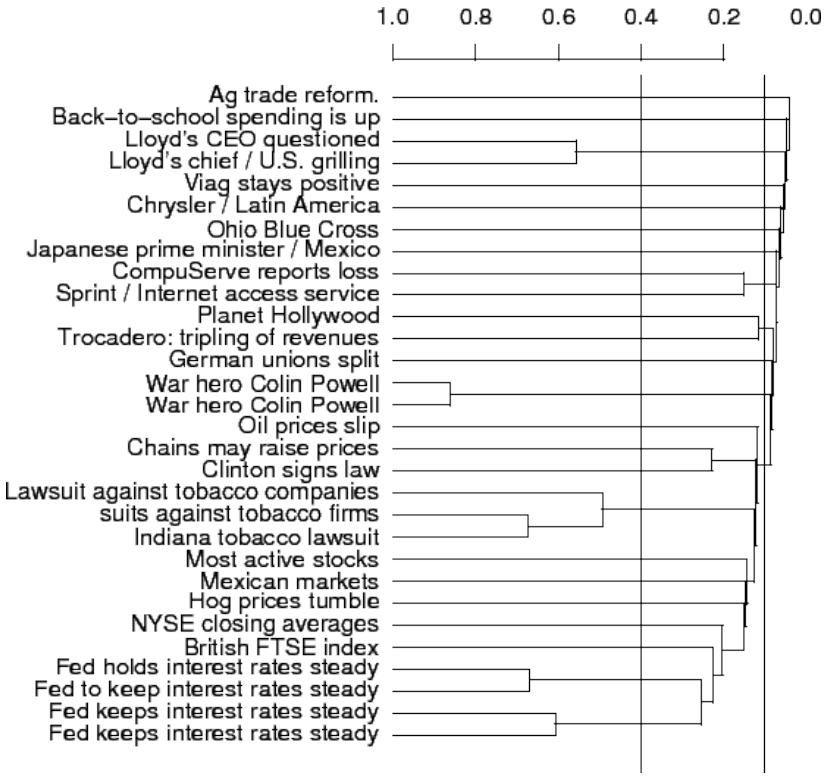

Abbildung 3.5: Dendrogramm einer automatischen Dokumentclusterung, Quelle: Manning et al. (2008, 379)

Partitionierende Verfahren

Da hierarchische Verfahren bei einer großen Zahl an Objekten sehr rechenintensiv sind – schließlich müssen für jeden Agglomerationsschritt alle Einheiten neu bewertet werden –, sind partitionierende Clusteranalysen für die automatische Klassifikation von Texten oft effizienter (Steinbach et al., 2000; Hotho et al., 2005). Hierbei wird zuerst die Zahl der Cluster festgelegt und dann von zufälligen Startwerten im Vektorraum die Distanz jedes Dokuments zu jedem Cluster berechnet. Dieser Schritt wird so lange wiederholt, bis die Clusterzuordnung stabil ist. Der am häufigsten verwendete Cluster-Algorithmus ist hierbei k-Means, wobei k die Anzahl der a priori festgelegten Cluster bezeichnet. Im Unterschied

zu hierarchischen Verfahren sind zudem die Clusterlösungen nicht ineinander geschachtelt, sondern alle Cluster sind auf einer Ebene, so dass Manning et al. (2008) von *flat clustering* sprechen.

Da die Clusteralgorithmen selbst vollautomatisch ablaufen, ist die wichtigste Aufgabe für den Forscher die Festlegung der Clusterzahl. Während bei hierarchischen Verfahren die 'richtige' Clusterzahl ex post bestimmt wird, muss sie bei einer partitionierenden Clusteranalyse bereits ex ante feststehen. Allerdings wird in den meisten Fällen schlicht ein Intervall mit plausiblen Clusterzahlen definiert und dann für jedes k ein eigener Durchlauf gestartet. Die Ergebnisse der einzelnen Durchläufe werden dann hinsichtlich ihrer Plausibilität verglichen (Quinn et al., 2006, 16).[9] Dabei ist eine visuelle Inspektion der Ergebnisse oft sinnvoll, denn ähnlich wie bei der Co-Occurrence-Analyse können auch Dokumentencluster zueinander positioniert und grafisch dargestellt werden (Fortuna et al., 2005; Di Giacomo et al., 2007).

Strategien für sinnvolle Dokumentcluster

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. Da es sowohl für die Zahl als auch die Ausprägung der Cluster viele verschiedene Möglichkeiten gibt, liegt eine wichtige Herausforderung für die angewandte Forschung in der systematischen und intersubjektiv überprüfbaren Auswahl der Clusterlösungen. Grimmer & King (2009, 4) weisen darauf hin, dass „selbst mit 100 Dokumenten die Zahl möglicher Klassifikationen dem 10^{28}-fachen aller Elementarteilchen des Universums entspricht." Zudem gäbe es kein objektiv bestes Verfahren, Texte zu klassifizieren, da potentiell jede Klassifikation gleich plausiblen Regeln folgt und kein Algorithmus über alle Klassifikationsprobleme hinweg gleich gut eingesetzt werden kann.

9 Alternativ kann natürlich auch einfach eine feste Clusteranzahl vorgegeben werden, etwa weil die Anzahl der Kategorien aus inhaltlichen Vorüberlegungen resultiert, die sich natürlich nicht in der Clusterung niederschlagen müssen.

Grimmer & King (2009) entwickeln ein Verfahren, bei dem nacheinander alle verfügbaren Clusteralgorithmen auf einen Textkorpus angewendet werden. Die entstehenden Clusterlösungen werden dann selbst in eine Distanzmatrix überführt und mithilfe einer Multidimensionalen Skalierung dargestellt. Anschließend kann durch eine visuelle Inspektion eine für die Forschungsfrage passende Clusterlösung ausgewählt bzw. aus mehreren Lösungen ein Clusterensemble generiert werden. Anhand mehrerer Textkorpora und Fragestellungen aus der politischen Kommunikationsforschung können die Autoren zeigen, dass die durch dieses Verfahren aufgedeckten Klassifikationen von unabhängigen *Ratern* als valide und relevant eingestuft werden. Die durch vollautomatische Klassifikation entdeckten Kategorien konnten anschließend auch durch überwachte Verfahren (vgl. Abschnitt 3.4.1) getestet werden (Grimmer & King, 2009, 17-19).[10]

3.3 Deduktive Ansätze

In der klassischen quantitativen Inhaltsanalyse werden Texte anhand zuvor festgelegter Regeln gelesen und codiert. Die Kriterien, die die Rezeption der Codierer steuern bzw. strukturieren, müssen daher vor der eigentlichen Codierung feststehen (vgl. Wirth, 2001). Bei den im Folgenden vorgestellten deduktiven Verfahren werden Texte anhand eines festen Regelsatzes codiert, der direkt in der Software festgelegt ist.[11] Das Zuordnen von Texten bzw. Textstellen zu geschlossenen oder offenen Kategorien ist hierbei ein vollständig deterministischer Vorgang, der hinsichtlich der vorgegebenen Regeln perfekte Reliabilität und einen enorm hohen Durchsatz verspricht. Bei einem vorhandenen Regelsatz steigt der Codieraufwand für eine große Anzahl an Dokumenten nur minimal, es bedarf lediglich leistungsfähigerer Hardware oder mehr Rechenzeit.

10 Das vorgestellte Verfahren wurde auch in einem Publikationsprojekt umgesetzt, für das 100 politikwissenschaftliche Essays in einem Sammelband durch Dokumentclusterung thematisch geordnet wurden (King et al., 2009).

11 Das bedeutet keinesfalls, dass die vom Forscher aufgestellten Kategorien ausschließlich theoriegeleitet und damit deduktiv entwickelt werden müssen (Früh, 2007, 72-74).

3.3.1 Diktionärbasierte Verfahren

Seit der Frühzeit computergestützter Inhaltsanalyse stellen diktionärbasierte Verfahren das wichtigste und lange Zeit auch einzige Mittel hypothesengeleiteter deduktiver Textcodierung dar (Stone et al., 1966; North et al., 1963; Holsti, 1969; Deichsel, 1975; Klingemann, 1984). Es verwundert daher nicht, dass dieser Ansatz den meisten empirischen Studien zugrunde liegt und bis heute, und vor allem in Lehrbüchern, fast synonym mit automatischer Textanalyse verwendet wird (vgl. Lisch & Kriz, 1978; Weber, 1990; Merten, 1995; Früh, 2007; Rössler, 2005; Maurer & Reinemann, 2006).[12]

Analyselogik und Vorgehen
Die Grundlogik diktionärbasierter Verfahren ist seit Jahrzehnten praktisch unverändert (vgl. North et al., 1963, 130): 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. Mit der Analysesoftware kann dann problemlos nach den Wortstämmen gesucht werden, die sie enthaltenden Dokumente bzw. Textabschnitte werden entsprechend klassifiziert. Da dieser letzte Schritt schon in den 60er Jahren vergleichsweise einfach zu implementieren war, konzentrierte sich der Großteil der Forschungsarbeit auf die vorbereitenden Schritte des Preprocessing (vgl. Abschnitt 4.2) und der Diktionärentwicklung. Die beiden vorherrschenden Softwarepakete für diktionärbasierte Inhaltsanalyse General Inquirer[13] (Stone et al., 1966) und Textpack[14] (Klingemann et al., 1984) sind dementsprechend aus einer Sammlung verschiedener Tools zur Textverarbeitung und -codierung entstanden, die nur lose gekoppelt sind.

12 Sowohl die Darstellung der Verfahren als auch die Kritik daran basieren dabei noch immer auf der sogenannten kontextfreien Einwort-Analyse (KOFEINA) (Deichsel, 1975, 43), obwohl schon seit den 60er Jahren der syntaktische Kontext der Wörter starke Berücksichtigung fand und problemlos Wortgruppen und komplexere N-Gramme codierbar sind.

13 http://www.wjh.harvard.edu/~inquirer/

14 http://www.gesis.org/dienstleistungen/methoden/software/textpack/

Bei der Codierung selbst sind zwei Kriterien festzulegen: die Codiereinheit und das Skalenniveau der Codierung. Traditionell werden bei diktionärbasierten Verfahren Kategorien auf der Ebene einzelner Sätze zugeordnet und anschließend auf Artikelebene aggregiert. Technisch ist es jedoch kein Problem, die Klassifikation gleich auf Dokumentebene vorzunehmen. Bezieht man bei der Skalierung die einfache Häufigkeit der Wörter einer Kategorie ein, unterscheiden sich satz- und dokumentbasierte Codierungen lediglich in der Prozentuierungsbasis. Alternativ kann auch nur dichotom codiert werden, ob eine Kategorie der Codiereinheit zugeordnet werden kann oder nicht. Dabei stellt sich jedoch die Frage, ob und wann das Vorkommen eines einzelnen Wortes aus einer Liste tatsächlich die thematische Ausrichtung eines Textes wiedergeben kann: Kann ein Dokument beispielsweise schon als Wirtschaftsnachricht codiert werden, wenn darin das Wort „Inflation“ vorkommt? Bei einem deterministischen Verfahren muss vom Forscher festgelegt werden, (a) wie hoch der Schwellenwert für eine binäre Codierung ist und (b) ob alle Wörter gleichgewichtig bei der Berechnung dieses Schwellenwertes sind.

Weber (1983) weist darauf hin, dass eine reliable und valide Messung unabhängig von der Codiereinheit und der Skalierung äquivalente Ergebnisse erzielen sollte. Die von ihm geforderte umfassende Evaluationsstudie steht bislang jedoch noch aus. Eine ältere Untersuchung von Saris-Gallhofer et al. (1978) am Beispiel von Osgoods *Evaluative Assertion Analysis* zeigt jedoch, dass etwa die von Holsti (1966) festgesetzten Skalierungs- und Aggregationsregeln deutliche Auswirkungen auf die Validität der Messung haben.

Diktionäre

Berelsons (1952, 147) Diktum, dass eine Inhaltsanalyse mit ihren Kategorien steht und fällt, gilt insbesondere bei der Arbeit mit Diktionären. Aufgrund des vollständig deterministischen Codiervorgangs ist kein Raum für Unschärfen, Doppeldeutigkeiten und Kontextfaktoren, die jedoch natürliche Sprache erst auszeichnen. Bei jeder Kategorie im Diktionär muss ausführlich überprüft werden, ob die darin verzeichneten Wörter (bzw. Wortstämme) gleichzeitig trennscharf und vollständig sind. Schon für relativ einfache Kategoriensysteme, wie etwa dem *Wörterbuch*

Umweltschutz von Schönbach (1982) mit 245 Wörtern, sind wiederholte Verfeinerungen notwendig. So mussten etwa die Wörter *Umwelt* und *Abfall* wegen mangelnder Trennschärfe wieder entfernt werden, um die Anzahl falsch positiv codierter Dokumente zu senken.[15]

Die Konstruktion von Diktionären geht dabei zuerst theoriegeleitet vor, da feststehen muss, welche Dimensionen des Inhalts überhaupt von Interesse sind. Anschließend wird iterativ durch deduktive Probecodierungen und induktive Textinspektion das Wörterbuch verfeinert (Stone et al., 1966; Bengston & Xu, 2009). Grundsätzlich haben sich dabei zwei Entwicklungsstrategien herausgebildet, um die Reliabilität und Validität von Diktionären zu sichern:

1. Durch Listen von Keyword-in-Context-Zeilen (KWIC) kann für jede Untersuchungseinheit geprüft werden, ob die Kategorie den Inhalt sinnvoll wiedergibt. Falls dies nicht der Fall ist, muss entweder ein Diktionäreintrag geändert oder weitere Preprocessing-Regeln definiert werden, z.B. bei der Disambiguierung von Homonymen (Fan & McAvoy, 1989).
2. Alternativ oder zusätzlich werden mit semantischen Differentialen oder anderen Ratings potentiell relevante Wörter von vielen Laien oder Experten beurteilt, anschließend skaliert und in Kategorien geordnet. Dieses Vorgehen bietet sich vor allem bei psychologischen Konstrukten an (Gottschalk & Gleser, 1969; Saris-Gallhofer et al., 1978; Rosenberg et al., 1990; McTavish, 1997).

Nicht nur wegen des größeren Entwicklungsaufwands sind die Diktionäre ungleich wichtiger als die Programme, die damit Texte codieren, nicht zuletzt weil letztere aufgrund des rasanten Fortschritts in der Computertechnologie deutlich schneller altern.[16] Die großen sozialpsychologischen und soziologischen Diktionäre Harvard IV und Lasswell Values Dictionary (Kelly & Stone, 1975; Lasswell & Namenwirth, 1968; Züll et al., 1989), die über 8500 Wörter bzw. Wortgruppen und fast 500

15 In diesem Beispiel wurde also das Wörterbuch auf das Validitätskriterium Präzision statt auf Sensitivität (*Recall*) optimiert (vgl. Abschnitt 4.4.2).

16 So steht bereits in einer der ersten deutschsprachigen Monografien zu diesem Thema nicht die Software, sondern das *Hamburger Kommunikationssoziologische Wörterbuch* (HKW) von Deichsel (1975) im Vordergrund.

Kategorien enthalten, sind das Ergebnis vieler Jahrzehnte Arbeit, an der viele Forscher beteiligt waren, die laufend neue Kategorien hinzufügten (Krippendorff, 2004a, 287). Diese allgemeinen Wörterbücher sind jedoch die Ausnahme, da die meisten Diktionäre ad hoc anhand einzelner Fragestellungen entwickelt werden (Stone et al., 1966). Dem Vorteil des geringeren Entwicklungsaufwands und der genauen Anpassung des Instruments steht bei spezifischen Diktionären der große Nachteil der fehlenden Anschlussfähigkeit gegenüber.

Welches nun die erfolgversprechendere Strategie ist, wird im Fach kontrovers diskutiert: Weber (1983) fordert explizit die Entwicklung großer allgemeingültiger Diktionäre, während Krippendorff (2004a, 287) sich für spezialisierte Wörterbücher ausspricht. Beide Autoren weisen jedoch darauf hin, dass das Verfahren erst durch die Möglichkeit bzw. die tatsächliche Wiederverwendung von Diktionären seine wahren Qualitäten – Standardisierung, Transparenz, Effizienz – ausspielen kann. Anders ausgedrückt: Der Aufwand für die Erstellung eines guten Diktionärs ist in vielen Fällen höher als bei der manuellen Codierung, so dass der Nutzen diktionärbasierter Textanalyse erst mit steigender Dokumentenzahl bzw. vielen Replikationen die Kosten übersteigt.

Es lässt sich nach fast 50 Jahren Forschungsgeschichte jedoch feststellen, dass gerade diese Erwartungen sich sowohl bei spezialisierten als auch allgemeinen Diktionären nicht erfüllt haben. So wird die Mehrzahl der Wörterbücher in Tabelle 3.3 nur von deren Entwicklern und Mitarbeitern eingesetzt. In der Geschichte der Inhaltsanalyse tauchen die meisten Diktionäre nur einmal im Rahmen der entsprechenden Publikation auf, übergreifende Projekte sind zumeist versandet (Bröker, 1984, 160). Es gibt zudem nur wenige Verzeichnisse von Diktionären, und Überblicksartikel zu diesem Thema sind ebenfalls rar (Züll & Landmann, 2002; Stone et al., 1966, 140-41). Obwohl diese fehlende Systematisierung der Instrumente auch bei der konventionellen Inhaltsanalyse problematisch ist, wiegt sie bei computergestützten Verfahren umso schwerer, da die Daten nachweislich bereits elektronisch aufbereitet vorliegen. Die Forschung mit Diktionären ist daher hochgradig fragmentiert und durch viele Einzellösungen dominiert. Der erwartete Effizienz- und Effektivitätsgewinn durch die Wiederverwendung bestehender Wörterbücher ist

Tabelle 3.3: Übersicht verbreiteter General-Purpose-Diktionäre

Name des Diktionärs	Kategorien	Referenz
Harvard IV Dictionary	105	Kelly & Stone (1975)
Lasswell Values Dictionary	392	Lasswell & Namenwirth (1968)
Regressive Imagery Dictionary	65	Martindale (1975)
Minnesota Contextual Category System	116	McTavish et al. (1997)
Linguistic Word Count Dictionary	39	Pennebaker et al. (2007)

Quelle: Krippendorff (2004a, 285-87), http://www.textanalysis.info/

bislang jedenfalls nicht absehbar. Trotz wiederholter Anläufe hat es weder die deutsche noch die internationale Kommunikationswissenschaft bislang geschafft, Codebücher, Diktionäre und inhaltsanalytische Daten in ähnlicher Qualität zu archivieren und wissenschaftsöffentlich zur Verfügung zu stellen wie dies seit Jahrzehnten mit Umfragedaten und Befragungsinstrumenten geschieht.[17]

Trotz dieser Entwicklung ist allein die Möglichkeit, die Codierung vollständig nachzuvollziehen, zu diskutieren und zu verbessern, ein gewichtiges Argument für diktionärbasierte Verfahren. Selbst wenn Codebücher gut dokumentiert sind, bleibt der eigentliche Codiervorgang bei manuellen Inhaltsanalysen weitgehend im Dunklen, etwa weil die Codiererschulung so gut wie nie dokumentiert wird. Die Validität von Wörterbüchern und die Funktionsweise der Software lässt sich hingegen sehr gut nachvollziehen.

Da die Codierung von Dokumenten nach Stichwörtern ein deterministischer Prozess ist, kann für diktionärbasierte Verfahren vollständige

17 Der nach der aktuellen GLES-Studie (GÖFAK Medienforschung, 2010) aktuellste inhaltsanalytische Datensatz im Zentralarchiv in Köln datiert aus dem Jahr 1999. Die erstmals von Pool (1959) ausgesprochene Befürchtung, es werde sich keine Standardisierung der inhaltsanalytischen Instrumente entwickeln, hat sich damit als erstaunlich zutreffend erwiesen.

Reliabilität auf der Ebene manifester Texteigenschaften angenommen werden. Der Nachteil einer solch simplen Herangehensweise liegt jedoch in der oftmals geringen Validität der Ergebnisse, wenn es um relevante theoretische Konstrukte geht, die mit der Inhaltsanalyse gemessen werden sollen (vgl. Früh, 2007). 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 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, 2000). Selbst die Entwicklung eines einfachen Diktionärs für Sportmeldungen ist aufgrund der Vielzahl möglicher Indikatorbegriffe mit großem Aufwand verbunden.

Ein zentraler Vorteil von diktionärbasierten Verfahren liegt demgegenüber in der großen Effizienz. Selbst mit umfangreichen Wortlisten lassen sich gigantische Textmengen codieren, zumal das Problem der Schlagwortzählung leicht parallelisierbar ist (Dean & Ghemawat, 2004). Analysen von tausenden Dokumenten lassen sich problemlos an jedem Standard-PC durchführen, Online-Dienste wie Google News oder Blogpulse überwachen und indizieren hunderte Millionen Dokumente praktisch in Echtzeit.

3.3.2 Freitextrecherche

Auf den oben genannten Diensten, die Dokumente effizient indizieren und durchsuchbar machen, basiert das Verfahren der Freitextrecherche in Datenbanken oder Online-Suchmaschinen, das mittlerweile in der Kommunikationswissenschaft recht weit verbreitet ist (Tankard et al.,

1994; Hagen, 2001; Hollanders & Vliegenthart, 2008). Hierbei werden spezielle Suchanfragen zu definierten Begriffen, oft mit Boolschen Operatoren wie UND, ODER bzw. NICHT verknüpft, an externe Dienste gestellt, die dann ein Ergebnisse mit passenden Dokumenten zurückliefern, etwa LexisNexis für digitalisierte Inhalte von Printmedien oder Suchmaschinen wie Google für das World Wide Web. In den meisten Fällen dient die Freitextrecherche lediglich der Stichprobenziehung von Artikeln zu bestimmten Themenkomplexen, etwa zur Berichterstattung über das Rauchen (Wenger et al., 2001), den Klimawandel (McComas & Shanahan, 1999) oder die Person Hillary Clinton (Scharrer, 2002). Die eigentliche Codierung erfolgt dann zumeist manuell mit einem klassischen Codeplan.

Die Freitextsuche hat vor allem den Vorteil, dass sie kostengünstig und schnell zu realisieren ist, allerdings mit dem Nachteil, dass die Qualität der Anbieter zumeist nur bedingt empirisch überprüf- und selten änderbar ist. Zwei Probleme stehen dabei im Vordergrund der Kritik: Zum einen ist nicht gewährleistet, dass alle Beiträge aus der Grundgesamtheit indiziert wurden, zum anderen, dass durch die Abfrage alle relevanten Artikel gefunden wurden. Das erste Problem, das den Input der Datenbank betrifft, wird ausführlicher in Abschnitt 4.1.1 diskutiert und betrifft jeden Anbieter unabhängig von der konkreten Verfahrensweise der Studie.

Der zweite Kritikpunkt hängt jedoch mit der Analyselogik des Verfahrens an sich zusammen: Da die Grundgesamtheit der Beiträge dem Forscher nicht vorliegt, kann die Validität der Abfrage nicht vollständig geprüft werden. Es ist zwar möglich, den Anteil der fälschlich in die Stichprobe aufgenommenen Artikel und damit die Präzision der Abfrage durch manuelle Überprüfung zu bestimmen. Der Anteil der fälschlich *nicht gefundenen* Artikel, und damit die Sensitivität der Suchanfrage, ist jedoch aufgrund mangelnder Informationen über die Grundgesamtheit nicht abschätzbar. Die Abfragevalidität des gesamten Instruments ist mithin unklar (Hagen, 2001; Welker et al., 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. Anspruchsvollere Abfragen lassen sich zumeist nur bei einem Vollzugriff auf die Daten realisieren, etwa mittels so genannter regulärer Ausdrücke (Friedl, 2006), die in den meisten Programmiersprachen und in vielen Texteditoren implementiert sind.

Insgesamt stellt die Freitextrecherche eine sinnvolle Ergänzung zu manuellen und automatischen Analyseverfahren dar, insbesondere wenn sie für die Eingrenzung des Untersuchungsmaterials hinsichtlich eines spezifischen Zugriffskriteriums genutzt wird. Wie für diktionärbasierte Verfahren gilt jedoch auch hier, dass der Aufwand bei komplexeren Themen oder Fragestellungen sehr hoch und die Validität eingeschränkt ist, zumal im Gegensatz zu Diktionären die Länge der Suchanfrage stark begrenzt ist.

3.3.3 Regelbasierte Ansätze

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. Roberts, 1997a; King & Lowe, 2003).

Konzeptionell hat das Verfahren seine Wurzeln bei der *Assertion-Analysis* von Osgood (1959), wobei die Codierung lange Zeit ausschließlich manuell geleistet wurde und lediglich die Auswertung computergestützt war (Holsti, 1969; van Cuilenburg et al., 1988). Die Forschergruppe um Cuilenburg et al. (1986) entwickelte in den 80er Jahren das Annotationssystem CETA (Computer Aided Evaluative Text Analysis), mit dem zahlreichen Analysen auf dem Gebiet der politischen und Wirtschaftskommunikation (Kleinnijenhuis et al., 1997; Berg & Veer, 2000) durchgeführt wurden. Erst in jüngster Zeit konnten auf Basis von CETA (fast) vollautomatische Inhaltsanalysen durchgeführt werden (Atteveldt et al., 2008).

Analyselogik
Die in diesem Abschnitt vorgestellten Ansätze unterscheiden sich von den diktionärbasierten Verfahren sowohl in der Fragestellung als auch der Umsetzung: Es geht nicht (nur) um die Zuordnung eines Textes in eine oder mehrere thematische Kategorien, sondern um die Codierung und Analyse der Beziehungen zwischen den aussagetragenden Begriffen (Roberts, 1997b).Vereinfacht ausgedrückt: Regelbasierte Verfahren geben Auskunft darüber, worum es in einer Aussage geht. Sie beantworten die offene Frage: Welches Subjekt handelt wie bezogen auf welches Objekt?

Für die Umsetzung der automatischen Codierung bedeutet dies, dass der Computer in der Lage ist (bzw. sein muss), zusätzlich zur lexikalischen Ebene auch syntaktische und – in begrenztem Umfang – auch semantische Dimensionen des Textes in numerische Codes zu transformieren. Im Rahmen des Preprocessing, und hier vor allem bei der Disambiguierung von Wörtern, bietet schon der General Inquirer eine Art syntaktische Analyse (*Parsing*) an, doch steht diese bei regelbasierten Verfahren im Zentrum der Methode. Dazu ein stark vereinfachendes Beispiel in Form einer Schlagzeile oder Tickermeldung: „Hertha BSC feuert Trainer."

Bei einer thematischen, diktionärbasierten Analyse könnte der Code *Sport* vergeben werden, da sich sowohl „Trainer" als auch „Hertha BSC" in der entsprechenden Wortliste finden. Mit der primitiven und sicher unzureichenden Parsing-Regel, dass das erste kleingeschriebene Wort im Satz das Prädikat (P), alles davor das Subjekt (S) und alles danach das Objekt (O) ist, könnte problemlos eine Liste mit Akteuren erstellt werden, deren Handlungen besonders oft berichtet werden, welche Handlungen dies sind, wen sie betreffen. Der syntaktische Parser würde dann folgende Codierung vornehmen: „Hertha BSC^{S} feuert^{P} Trainer^{O}".

Zusätzlich könnte ein semantischer Parser erkennen, dass das Verb im Satz nicht in die Kategorie *Gewalt*, sondern *Arbeitsverhältnis* gehört und entsprechend das Wort annotieren.[18] Dieses simple Schema kann um vielfältige Attribute erweitert werden, wie die Studien von Atteveldt (2008) zeigen.

18 In diesem Fall könnte die Regel lauten, dass keine Präposition wie „auf" oder „mit" gegeben ist, die die kriegerische Bedeutung indizieren würde.

Als Output einer solchen automatischen Analyse liegen dann geordnete Daten vor, die man nach zahlreichen Kriterien, vor allem aber relational auswerten kann: Welche Mannschaft feuert besonders oft, wer wird oft gefeuert, über welche Interaktionen zwischen Trainer und Verein wird häufig berichtet? Dafür bieten sich vor allem netzwerkanalytische Verfahren an, die bislang fast ausschließlich auf Basis manueller Claim- oder Assertion-Analysen erfolgen (vgl. Carley, 1997; Adam, 2008).

Anwendung der regelbasierten Codierung

Wie Atteveldt (2008) treffend feststellt, ist eine syntaktisch-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). Diese Skepsis ist trotz der Weiterentwicklungen auf dem Gebiet der *Computational Linguistics* weiterhin angebracht, da bislang kein Computerprogramm auch nur annähernd das leisten kann, was als Textverständnis gelten könnte. Erschwerend kommt hinzu, dass das Ziel dieser Forschung unscharf ist:

> It seems that we understand so little of what it means to „understand" a text that knowing when an AI [Artificial Intelligence, M.S.] approach to understanding texts works is quite difficult. (West, 2001, 164)

Trotz aller Schwierigkeiten wurde das Verfahren der Informationsextraktion in der Politikwissenschaft, und dort insbesondere bei der Analyse internationaler Ereignisse, seit Beginn der 90er Jahre erfolgreich angewandt und weiterentwickelt. So konnten mit dem diktionärs- und regelbasierten System KEDS/TABARI[19] (Schrodt et al., 1994) erfolgreich die Schlagzeilen des Reuters-Tickers hinsichtlich internationaler Konflikte und anderer Ereignisse codiert werden. Dabei spielt jedoch das Wörterbuch für Akteure und Handlungen eine wichtigere Rolle als der vergleichsweise einfache (*shallow*) Parser. Schrodt & Donald (1990) weisen

19 http://web.ku.edu/~keds/software.html

selbst darauf hin, dass sie sich bei der Entwicklung von KEDS eher am diktionärbasierten Ansatz orientierten als an linguistischen Verfahren.[20]

Das von King & Lowe (2003) positiv evaluierte Tool VRA[21] (Virtual Research Assistant) arbeitet auf Basis eines vollständigen Satz-Parsings, mit dem die syntaktische Struktur der Schlagzeilen zerlegt und durch zusätzliche Kontextinformationen auf Basis von 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 themenspezifisch (internationale Ereignisse) sind und nur für die Analyse einfacher und relativ stark strukturierter Texte (englische Schlagzeilen) eingesetzt werden. Zudem ist wie bei der Diktionärentwicklung Expertenwissen für die Definition der Parsing-Regeln und die Erkennung von Akteuren erforderlich (Schrodt et al., 1994). Im Gegensatz zu Textpack oder GI sind KEDS/TABARI und VRA spezifisch für eine einzelne Fragestellung programmiert. Sowohl Parsing-Regelsätze als auch Wortlisten sind damit nur für Forscher mit Programmierkenntnissen überprüf- und modifizierbar.

Zwei Entwicklungen haben allerdings in jüngster Zeit dazu beigetragen, dass automatische semantisch-relationale Textanalysen einfacher und damit auch häufiger eingesetzt werden: So werden im Bereich des *Natural Language Processing* (NLP) immer leistungsfähigere Parser und andere Algorithmen entwickelt, die syntaktisch-semantische Analysen ermöglichen. Diese sind mittlerweile auch für andere Sprachen als Englisch verfügbar und liefern hier sehr ermutigende Ergebnisse, selbst bei umfangreicheren Texten (Atteveldt, 2008). Jedoch werden auch in diesem Fall noch Diktionäre und manuelle Codierungen zur Hilfe genommen.

Gelänge eine rein regelbasierte Zerlegung von Aussagen, könnten im Gegensatz zu diktionärbasierten Verfahren zumindest potentiell auch a priori undefinierte Akteure oder Handlungen codiert werden, solange

20 Im Prinzip kann KEDS/TABARI auch als diktionärbasiertes Tool betrachtet werden, dass zwei bzw. drei Kategoriensysteme implementiert: Akteure (Subjekt und Objekt) und Handlungen (Shellman, 2008).

21 http://vranet.com/

diese aus der Syntax der Aussage ableitbar sind.[22] Prinzipiell entspräche so ein Vorgehen einer Exploration von Textinhalten jenseits der basalen lexikalischen Ebene. So ließe sich z.B. in längsschnittlichen Studien das Auftauchen neuer politischer Akteure oder Handlungsweisen messen. Bislang ist der Einsatz solcher generischen Parser im Forschungsalltag nicht abzusehen, da einerseits sowohl die Entwicklung der Parsing-Algorithmen als auch deren Anwendungsmöglichkeiten noch weit von den Anforderungen der Sozialwissenschaften entfernt sind.

Regelbasierte Analysen strukturierter Inhalte
Der zweite Grund für eine verstärkte Auseinandersetzung mit regelbasierter Codierung liegt in der Tatsache, dass Inhalte im Internet in vielen Fällen sehr viel strukturierter als in Offline-Medien vorliegen, da sie ja für die digitale Verarbeitung gedacht sind. So schwer natürliche Sprachen automatisiert zu verarbeiten sind, so einfach ist dies bei Computersprachen: Der Urheber einer Email ist durch einen festen Platz im Mail-Header ebenso problemlos identifizierbar wie die Nachricht, auf die sie sich bezieht.[23] Das Ziel eines Hyperlinks ist mit jedem HTML-Parser deutlich leichter zu extrahieren als das Objekt einer Aussage in einem Satz. Dieses *HTML-Scraping* genannte Verfahren eignet sich für die Analyse von Nutzungsstatistiken für Youtube-Videos (Scharkow, 2007) oder Bewertungen und Kommentare der Internet Movie Database (Mirza & Scharkow, 2009) ebenso wie für die gezielte Sammlung von veröffentlichten Umfragedaten (Jackman, 2006).

Die Strukturierung von Informationen in HTML/XML und anderen standardisierten Formaten erleichtert syntaktische Analysen und besonders die relationale Inhaltsanalyse von Kommunikationsnetzwerken enorm. Die Arbeit von Luzar (2004) widmet sich gänzlich einem Verfahren, in dem die Struktur von Webdokumenten selbst den Untersu-

22 Wie Hillard et al. (2007) kritisch feststellen, kann KEDS/TABARI jedoch gerade keine unbekannten Akteure codieren, da es im Kern noch immer diktionärbasiert ist. Für VRA liegen dazu keine Informationen vor.

23 Die gesamte Struktur einer Email ist im RFC 5322 (http://tools.ietf.org/html/rfc5322) definiert. So gut wie alle Internet-Protokolle sind in einem solchen RFC (*Request for Comment*) dokumentiert, an den sich jeder Softwareentwickler bei der Implementierung zu halten hat.

chungsgegenstand darstellt. In ihrer Studie zeigt sich, dass strukturelle Eigenschaften von HTML- (und potentiell auch anderen) Dokumenten einen zusätzlichen analytischen Nutzen haben – etwa um zuverlässig den Anteil von visuellen Darstellungen am Inhalt zu messen –, die eigentliche Inhaltsanalyse aber nicht ersetzen können. Die Analyse von HTML-Tags ist jedoch auch und gerade bei der Sammlung und Vorbehandlung von Daten aus dem Internet von Interesse (vgl. Anhang A).

Der Nutzen von regelbasierten Verfahren bei der Online-Inhaltsanalyse liegt jedoch nicht so sehr in der Analyse einzelner Web-Seiten oder Emails, sondern vor allem in der Verknüpfung von Inhalts- und Netzwerkanalyse. Die Forschungsliteratur zu sozialen Online-Netzwerken ist im letzten Jahrzehnt nicht zuletzt aufgrund der relativ unproblematischen automatisierten Datenerhebung und –analyse explosionsartig angewachsen. Dabei wurden nicht selten die Online-Beobachtung von Kommunikationshandlungen (vgl. Rice, 1994) sowie automatische und manuelle Inhaltsanalysen der Kommunikate miteinander verknüpft. Da die vorliegende Literatur zu diesem Forschungsfeld schon jetzt unüberschaubar ist und zudem rasant wächst, sei nur auf wenige Übersichtsartikel verwiesen, in denen Kommunikationsstrukturen und Inhalte im USENET (Smith, 2003), in Chats und anderen synchronen Online-Medien (Rosen & Corbit, 2009) oder im World Wide Web analysiert wurden, wobei in letzterem vor allem Hyperlink-Strukturen regelbasiert vermessen wurden (Henzinger, 2001; Park, 2003; Park & Thelwall, 2003). Bei der Analyse von interpersonaler Kommunikation per Email kann sich die Forschung zunehmend auf große Text-Korpora stützen, die bereits bereinigt und für automatische Inhaltsanalysen aufbereitet sind. Besonders gut erforscht ist eine Sammlung von Mails aus dem Enron-Konzern, die der Wissenschaft nach staatsanwaltlichen Ermittlungen zur Verfügung gestellt wurden (Klimt & Yang, 2004; Diesner et al., 2005; McCallum et al., 2007).

In der deutschsprachigen Forschung ist vor allem die Studiensammlung von Stegbauer & Rausch (2006) zu nennen, in der verschiedene Gegenstände der „strukturalistischen Internetforschung“ auf diese Art analyisiert werden. Berendt et al. (2008) benutzen für ihre Analyse der deutschsprachigen Blogosphäre sowohl diktionär- als auch regelbasierte Verfahren.

3.4 Induktive Ansätze

3.4.1 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 grobe Richtlinien und Beispiele vermittelt. 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. Zudem führen zu umfangreiche Codieranweisungen zu einer übermäßigen kognitiven Belastung der Codierer, die im Gegenzug die Texte nur heuristisch statt systematisch verarbeiten (Wirth, 2001). Hohe Reliabilität und inhaltliche Validität, verstanden als Übereinstimmung mit den Vorgaben des Forschungsleiters (Früh, 2007), sind auch und gerade bei vergleichsweise unscharfen, nicht-deterministischen Codieranweisungen möglich.

Überwachte Verfahren der Textklassifikation machen sich die Tatsache zunutze, dass fast immer die beispielhafte Codierung von ausgewählten Texten deutlich weniger aufwändig ist als die Formulierung von komplexen Regelsätzen oder Diktionären (Sebastiani, 2002). Im Gegensatz zu den zuvor dargestellten Verfahren wird dabei induktiv ein Klassifikator definiert, d.h. die Extraktion für die Klassifikation relevanter Codierregeln übernimmt der Computer. Konkret bedeutet dies, dass 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. Diese Wörterbuch unterscheidet sich nicht nur in der Genese von den klassischen Vorgängern, sondern vor allem in der Gewichtung einzelner Begriffe für die Klassifikation. Während bei manuell erstellten Diktionären zumeist jeder

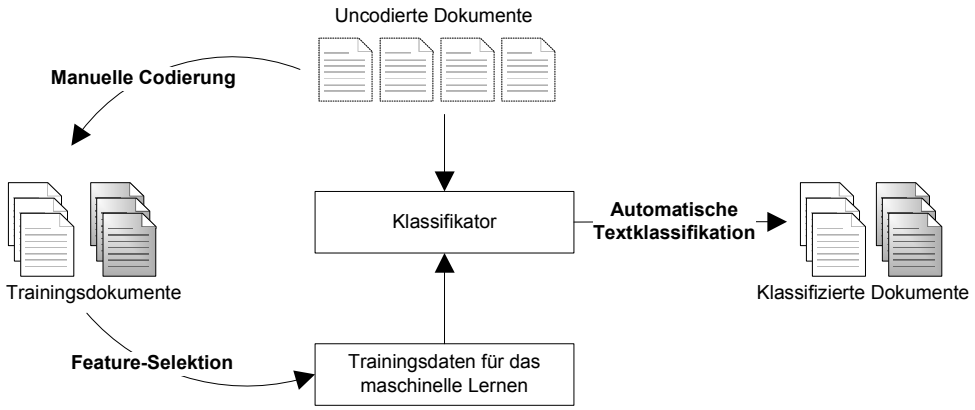

Abbildung 3.6: Funktionsweise überwachter Textklassifikation, Quelle: Evans et al. (2007, 1011)

Indikator einer Kategorie gleichgewichtig behandelt wird, z.B. „Fußball“ und „Schiedsrichter“ für eine Sportmeldung, wird das Gewicht jedes Wortes bei induktiven Verfahren empirisch aus der Wahrscheinlichkeit bestimmt, um damit erfolgreich zwischen den Kategorien unterscheiden zu können.

Ein sozialwissenschaftlich geläufiges Verfahren für einen induktiven binären Klassifikator wäre z.B. eine logistische Regression, bei der alle Wörter eines Textes als Prädiktoren (Features) 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 für jedes Dokument alle Features x und die *wahre* Klasse y bekannt sind, erhält jedes Feature ein Regressionsgewicht. Mit diesen lässt sich anschließend die Klassenzugehörigkeit eines uncodierten Dokuments, von dem alle Features x bekannt sind, leicht schätzen.

Das grundlegende Vorgehen bei einer induktiven Klassifikation lässt sich daher mit folgenden Schritten beschreiben (vgl. Abbildung 3.6):

1. Eine Anzahl Dokumente wird manuell nach einem gegebenen Codeplan codiert und in eine Kategorie (bzw. Klasse) eingeordnet. Daraus entsteht ein Trainingsdatensatz.

2. Ein Software-Klassifikator extrahiert aus dem Trainingsdatensatz Feature-Gewichte und ist damit einsatzbereit.
3. Der Klassifikator wird mit uncodierten Dokumenten konfrontiert, für die er die Wahrscheinlichkeiten der Klassenzugehörigkeit berechnet.

Für das oben genannte Beispiel zu Sportmeldungen liegt die Vermutung nahe, dass Wörter wie „Olympia“ oder „Doping“ ein hohes positives Gewicht 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. Da die Feature-Gewichte empirisch aus den vorgegebenen Daten geschätzt werden, ist es bei induktiven Verfahren möglich, dass diejenigen Wörter oder Wendungen am besten zwischen den Kategorien trennen, die dem Forscher selbst wenig bewusst sind.

Klassifikationsalgorithmen
Überwachte Klassifikation ist seit den 90er Jahren eines der meisterforschten Gebiete des maschinellen Lernens (Sebastiani, 2002). Dementsprechend stehen sehr viele Klassifikationsalgorithmen zur Verfügung, die statistisch unterschiedlich komplex, effektiv und effizient sind. Da es bereits eine große Menge vergleichender Literatur zu den verschiedenen Algorithmen gibt (z.B Joachims, 2002; Aas & Eikvil, 1999), soll hier auf eine umfangreiche Darstellung verzichtet werden. Stattdessen konzentriere ich mich auf die beiden Verfahren, die derzeit am häufigsten und erfolgreichsten eingesetzt werden (Felden et al., 2005): *Naive Bayes* (NB) und *Support Vector Machines* (SVM).

Naive Bayes-Klassifikatoren zeichnen sich durch eine große Effizienz, d.h. Geschwindigkeit bei Training und Klassifikation, und Effektivität, d.h. hohe Genauigkeit, aus. Ihr Funktionsprinzip ist gleichzeitig äußerst einfach, so dass auch für Forscher ohne statistisches Spezialwissen der Klassifikationsprozess keine *Black Box* ist. Für den generalisierbaren Fall eines binären Klassifikators muss für jedes Dokument d die Wahrscheinlichkeit berechnet werden, dass es zur Klasse c (oder zur Alternativklasse $\bar{c}$) gehört. Dieses Problem kann auf die Ebene einzelner Features (z.B. Wörter w) heruntergebrochen werden. Die Frage lautet dann: Wie

wahrscheinlich gehört Dokument d zur Klasse c, wenn Wort w darin vorkommt? Dies ist im Prinzip auch die Frage, die sich jeder Forscher bei der Konstruktion eines Diktionärs stellt. Mit Hilfe des Bayestheorems lässt sich diese Frage für jede Wort-Dokument-Kombination beantworten:

$$P(c|w) \propto P(c)P(w|c) \tag{3.1}$$

Dabei bezeichnet $P(c|w)$ die bedingte Wahrscheinlichkeit für Klasse c, gegeben Wort w. $P(c)$ ist die Priorwahrscheinlichkeit der Klasse, zumeist hergeleitet aus der relativen Häufigkeit der Klasse in den Trainingsdaten N_c/N. $P(w|c)$ ist schließlich die Wahrscheinlichkeit, dass Wort w in Texten der Klasse c vorkommt. Auch letztere lässt sich aus den relativen Häufigkeiten in den Trainingsdaten bestimmen. Verknüpft man nun die Wahrscheinlichkeiten pro Wort w zur Gesamtmenge an Wörtern n_d im Dokument ergibt sich:

$$P(c|d) \propto P(c) \prod_{k=1}^{n_d} P(w_k|c) \tag{3.2}$$

Die Wahrscheinlichkeit, dass ein Dokument d zur Klasse c gehört, lässt sich also aus der o.g. Priorwahrscheinlichkeit und dem Produkt der bedingten Wahrscheinlichkeiten, dass jedes enthaltene Wort w in Dokumenten der Klasse vorkommt, bestimmen. Aus Gründen der einfacheren Berechnung wird jedoch meist die Summe der logarithmierten Wahrscheinlichkeiten verwendet bzw. deren empirischer Schätzung $\hat{P}$ (Manning et al., 2008, 258):

$$\log P(c|d) = \log \hat{P}(c) + \sum_{k=1}^{n_d} \log \hat{P}(w_k|c) \tag{3.3}$$

Wie man aus Formel 3.2 erkennen kann, wird diese Bayesianische Klassifikation deshalb als naiv bezeichnet, weil sie von einer statistischen Unabhängigkeit der Wörter voneinander ausgeht. Obwohl diese Annahme bei natürlichsprachlichen Texten nicht zutrifft, sind Naive Bayes-Klassifikatoren erstaunlich leistungsfähig. Die Unabhängigkeit der Feature-Wahrscheinlichkeiten ist auch der Grund für die hohe Effizienz des Ver-

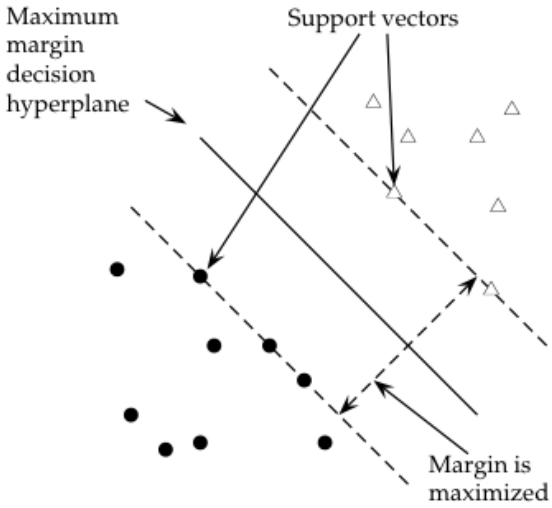

Abbildung 3.7: Struktur einer Support Vector Machine,
Quelle: Manning et al. 2008, 320

fahrens. Beim Training mit neuen Dokumenten müssen jeweils nur die einzelnen Wahrscheinlichkeiten für die vorkommenden Wörter $P(w|c)$ aktualisiert werden. Dies geschieht schneller als hochdimensionale Matrizenrechnungen (wie bei der logistischen Regression), was bei großen Mengen an Dokumenten von Vorteil ist. Insgesamt hat sich Naive Bayes als schnelles, robustes und leicht implementierbares Verfahren bei der Klassifikation von Texten bewährt, auch wenn diese in Studien von Joachims (2002) oder Hillard et al. (2008) leicht schlechter abschneidet als *Support Vector Machines*. Durant & Smith (2007) kommen jedoch zu genau umgekehrten Ergebnissen, so dass die Vermutung nahe liegt, dass die relative Leistung der Klassifikatoren je nach Problemstellung variiert. Eine mögliche Lösung dieses Problems liegt in der gleichzeitigen Verwendung mehrerer Klassifikatoren, dem sog. *Ensemble Learning*, wodurch nochmals die Qualität der Klassifikation erhöht werden kann (Hillard et al., 2008).

Support Vector Machines wurden erstmals von Joachims (1999, 2002) und Dumais et al. (1998) für die Klassifikation von Texten verwendet

(vgl. auch Leopold et al., 2007). Die Funktionsweise der SVM lässt sich am besten grafisch illustrieren, da die statistische Modellierung im Vergleich zur Naive Bayes oder Regressionsmodellen äußerst komplex ist. Anhand der Trainingsdaten wird eine sog. *Decision Hyperplane*, also eine Hyperebene für die Klassifikationsentscheidung, in den Vektorraum der Daten gelegt, die maximal weit von allen Datenpunkten entfernt ist. Betrachtet man wiederum ein Zwei-Klassen-Modell, wie es in Abbildung 3.7 dargestellt ist, besteht das Optimierungsproblem des Klassifikators darin, eine Hyperebene mit einem möglichst breiten Rand zwischen die Dokumente beider Klassen zu legen. Man spricht daher auch von einem *Large-Margin-Classifier*, der dafür sorgt, dass mit einer maximalen Wahrscheinlichkeit zwischen den Klassen unterschieden werden kann. Die Dokumente, die dann am nächsten zur Hyperebene liegen, konstituieren die *Support Vectors*, alle andere werden für Training und Klassifikation nicht berücksichtigt, da sie auf die Position der Entscheidungs-Hyperebene keinen Einfluss haben. SVM sind daher auch bei hochdimensionalen Feature-Räumen effizient, was sie besonders für Anwendungen in der Textklassifikation prädestiniert.

Anwendungen überwachter Textklassifikation
Obwohl die Grundidee für die überwachte Textklassifikation bis in die 60er Jahre zurückgeht, handelt es sich um ein vergleichsweise junges Forschungsfeld. Dies hat Konsequenzen für die Anwendbarkeit des Verfahrens: So wird auf dem Gebiet der Informatik Grundlagenforschung zu den Algorithmen und der statistischen Modellierung betrieben, die sich weitestgehend auf bereits vorhandene Textkorpora und Codierschemata stützt, etwa auf den Korpus von kategorisierten Reuters-Meldungen (Apté et al., 1994; Sebastiani, 2002; Debole & Sebastiani, 2005). Substanzielle Fragestellungen aus den Sozialwissenschaften spielen hier kaum eine Rolle. Eine Ausnahme stellt das große Anwendungsfeld der sog. *Sentiment Analysis* oder des *Opinion Mining* dar, in dem es um die Klassifikation von Meinungsäußerungen bzw. Bewertungen geht (vgl. Wiebe, 1994). In jüngster Zeit sind sowohl in der akademischen als auch der Marktforschung hunderte von Studien erschienen, bei denen zumindest teilweise auf überwachte Klassikationsverfahren zurückgegriffen wurde.

Einen hervorragenden Überblick zu diesem Thema bieten Pang & Lee (2008), die auch konzeptionelle Herausforderungen, etwa zur Frage, wie *Sentiment* sinnvoll operationalisiert werden kann, ausführlich diskutieren.

In den Sozialwissenschaften werden überwachte Klassifikatoren noch vergleichsweise selten angewandt bzw. mit anderen Begrifflichkeiten neu erfunden. Ein erstes Beispiel stellt die Klassifikation von Wahlprogrammen durch Laver et al. (2003) dar, die zuerst mit einem Kalibrations-Set die Feature-Gewichte errechnen und dann auf 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 wie bei NB- oder SVM-Klassifikatoren: Wenn in einem eher rechtskonservativen Wahlprogramm bestimmte Begriffe häufig vorkommen, ist ein anderes Programm mit diesen Wörtern mit hoher Wahrscheinlichkeit auch eher rechtskonservativ. Von den oben vorgestellten Klassifikationsverfahren unterscheidet sich das Wordscores genannte Programm auch darin, dass hier statt einer binären oder multinomialen Klassifikation ein metrischer Wert, nämlich ein Score auf einer Skala von politischen Positionen (z.B. Links-Rechts), generiert wird (Lowe, 2008). Auch wenn die Ergebnisse der vergleichsweise simplen Wordscores-Berechnung Probleme bei der Interpretation und Replikation aufweisen (Budge & Pennings, 2007a,b; Benoit & Laver, 2007; Martin & Vanberg, 2008), war und ist das Verfahren doch ein wichtiger erster Schritt bei der Verwendung induktiver Textklassifikation in der Politikwissenschaft und wird dort auch praktisch als Ergänzung oder gar Alternative zu manuellen Analysen von Wahlprogrammen oder Expertenbefragungen eingesetzt (Klemmensen et al., 2007; Bräuninger & Debus, 2008; Benoit et al., 2009a).

Die ersten Anwendungen der oben vorgestellten Naive Bayes bzw. SVM-Algorithmen sind ebenfalls in der Politikwissenschaft dokumentiert: Purpura & Hillard (2006) können zeigen, dass die thematische Zuordnung von Gesetzestexten aus der Datenbank des US-Kongresses mit überwachten Klassifikatoren genauso reliabel zu leisten ist wie mit menschlichen Codierern. Ein weiteres Anwendungsbeispiel, bei dem gleichzeitig mehrere Klassifikationsalgorithmen zur Codierung derselben Dokumente verwendet werden (*Classifier Ensemble*), bietet Stewards und

Zhukovs (2009) Analyse von Statements russischer Politiker und Militärs zur außen- und verteidigungspolitischen Agenda. Durant & Smith (2007) kommen bei ihrer Analyse von politischen Blogpostings zu ähnlich guten Ergebnissen wie Purpura & Hillard (2006). Sie können außerdem experimentell zeigen, dass nicht nur die Wahl des Klassifikationsalgorithmus, sondern auch die Feature-Selektion und das Preprocessing einen signifikanten Einfluss auf die Qualität der Klassifikation haben.

Hopkins & King (2010) verwenden ebenfalls statistische Maschinenlerntechniken, um große Mengen an Blog-Einträgen oder auch Emails aus dem Enron-Korpus zuverlässig zu klassifizieren. Das von ihnen entwickelte Verfahren hat allerdings den Nachteil (bzw. nach den Autoren den Vorteil), keine individuellen Dokumente zu klassifizieren, sondern nur die relativen Häufigkeiten der einzelnen Klassen inferenzstatistisch exakt schätzen zu können (vgl. auch King & Lowe, 2003). Zudem geben die Autoren an, dass ihre Methode nicht nur um Größenordnungen schneller ist als die individuelle Klassifikation von Dokumenten, sondern im Aggregat auch präziser sei als etwa Klassifikationen durch Support Vector Machines.[24] In jedem Fall eignet sich das Tool ReadMe nicht für Forschungsfragen, in denen Zusammenhänge auf Individualebene, etwa zwischen formalen und inhaltlichen Merkmalen der Texte, analysiert werden sollen.

3.4.2 Induktive 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. Für eine automatische Codierung von Textinhalten in eher offenen Forschungsansätzen ist die Informationsextraktion aus natürlichsprachlichen Texten, d.h. das „Verstehen“ eines Satzes durch den Computer, von großer Bedeutung. Wie in den vorangegangen Abschnitten gezeigt wurde, ist allein die manuelle Spezifikation syntaktisch-semantischer Analyseregeln mit erheblichem Aufwand und ungewissem Ausgang verbunden. Selbst Expertensysteme

24 Letzeres wird allerdings von Hillard et al. (2008) mit anderen Daten widerlegt.

zu themenspezifischen, vorstrukturierten 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 Mitteilung zu extrahieren ist. Bei den Verfahren, die in Abschnitt 3.3.3 vorgestellt wurden, ist das Vorgehen deduktiv, d.h. der Forscher muss einen vollständigen Regelsatz definieren, wie z.B. der Urheber einer Aussage zu erkennen ist (vgl. Atteveldt, 2008).

Praktischer 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. Dieses Vorgehen verhält sich gegenüber regelbasierten Ansätzen wie die überwachte Klassifikation zu Diktionären: Während bei diesen aus den Texten ein probabilistisches Wörterbuch aus den Beispielen gewonnen wird, werden bei der induktiven Informationsextraktion probabilistische Regelsätze algorithmisch generiert, die dann auf uncodierte Texte angewandt werden (Soderland, 1999). Obwohl auch dieses Forschungsproblem in der Informatik häufig erforscht wird, sind die Ergebnisse bislang weniger ermutigend als bei Klassifikationsaufgaben.

Während es schon bei menschlichen Codierern schwer fällt, ihnen die betreffenden Satzbestandteile zu markieren und semantische Objekte zu identifizieren, ist es noch schwieriger, dem Computer per Beispiel beizubringen, wo in einer Aussage die relevanten Informationen liegen, die dann im nächsten Satz scheinbar völlig anders angeordnet sind. Zwar ist eine Benutzerschnittstelle für eine digitale Markierung der Objekte bereits in bestehender Software für qualitative Inhaltsanalyse, z.B. MaxQDA (Kuckartz, 2007) oder Atlas.ti, vorhanden, allerdings werden die manuellen Annotationen nicht automatisch in Regeln überführt, da die Entwicklung passender Algorithmen noch auf sich warten lässt bzw. sozialwissenschaftlich bislang nicht rezipiert wurde.

Erste Fortschritte gibt es wie bei den anderen regelbasierten Verfahren vor allem dort, wo Texte relativ stark strukturiert und mit einem kleinen Wortschatz ausgestattet sind. Beispiele dafür sind Kleinanzeigen, Börsenticker oder Wettermeldungen (Soderland, 1999). Ein Überblick

zu verschiedenen Software-Algorithmen und kompletten Extraktionssystemen findet sich bei Muslea (1999). Der Trainable Information Extractor (TIE) von Siefkes (2007) konnte relativ gute Ergebnisse bei der Extraktion von Terminen und Sprechern aus Veranstaltungsankündigungen erzielen und liegt als gut dokumentierte Open Source Software vor. Im sozialwissenschaftlichen Kontext finden sich bislang fast keine Analysen mit induktiven Extraktionsverfahren. Eine Ausnahme ist die Studie von Atteveldt (2008). Diese erreicht eine mittlere Genauigkeit bei der Identifikation von Sprechern im niederländischen politischen Diskurs, muss aber noch immer Diktionäre für die Objektidentifikation zur Hilfe nehmen. Deshalb kann man nicht von einer vollständig induktiven Regelentwicklung sprechen. Noch scheinen syntaktisch-semantische Verfahren für ein induktives Vorgehen nicht weit genug entwickelt zu sein, obwohl die Lage weniger düster scheint, als van Cuilenburg et al. (1988) vor 20 Jahren 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.

3.5 Zwischenfazit – Überwachtes Lernen als Best Practice?

Zusammenfassend lässt sich festhalten, dass die technischen und infrastrukturellen Voraussetzungen für automatische Inhaltsanalysen heute so weit gegeben sind, dass auch umfangreichere Studien mit vertretbarem Aufwand durchgeführt werden können. Sowohl für vollautomatische unüberwachte Verfahren als auch diktionärbasierte Ansätze steht eine Vielzahl von kommerziellen und freien Softwarepaketen zur Verfügung (Alexa & Zuell, 2000; Scharkow, 2010a). Auch ist die Verfügbarkeit von digitalen Medieninhalten und anderen Textformen an sich kein Hindernis mehr (vgl. Abschnitt 4.1). Entscheidend für die Verbreitung automatischer Verfahren sind jedoch methodologische und auch forschungsökonomische Erwägungen: Welche Fragen können mit welcher Methode und welchem Aufwand valide und reliabel beantwortet werden? Bezogen auf den sozialwissenschaftlichen Forschungsalltag zeigt sich schnell, dass

nicht alle in diesem Kapitel vorgestellten Ansätze gleichermaßen erfolgversprechend sind: Vollautomatische Verfahren wie Textstatistik und *Document Clustering* sind wegen ihrer Funktionsweise nur eingeschränkt steuerbar und können daher nur für wenige spezifische Fragestellungen sinnvoll eingesetzt werden, vor allem in explorativen Studien sind sie aber durchaus von Nutzen. Diktionärbasierte Verfahren sind zwar verhältnismäßig gut dokumentiert und technisch leicht umzusetzen, die Entwicklung von eigenen Wörterbüchern – und diese braucht es in den meisten Fällen – erfordert jedoch so viel Expertise und Aufwand, dass eine manuelle Codierung häufig sowohl valider als auch kostengünstiger ist. Dies gilt auch für regelbasierte Ansätze, die zudem unter der schwierigen Automatisierbarkeit von syntaktischer und semantischer Decodierung der Aussagen leiden.

All diese Verfahren haben jedoch einen zentralen Nachteil: Sie sind weitestgehend von der konventionellen, gut dokumentierten und seit Jahrzehnten verfeinerten manuellen Codierpraxis abgekoppelt. Dies ist für die methodische Entwicklung der Inhaltsanalyse auf mehreren Ebenen problematisch. Erstens bedeutet die Entscheidung für automatische Verfahren zumeist auch eine Anpassung oder Neuausrichtung der Forschungsfrage, da sich der menschliche Codierer nicht einfach durch ein Wörterbuch oder einen Clusteralgorithmus ersetzen lässt. Zweitens erfordert auch heute der Umgang mit Textanalyse-Software und erst recht mit Tools des *Natural Language Processing* eine Expertise, die als Sozialwissenschaftler nur schwer zu erlangen ist. Zudem sind die Anforderungen oft gänzlich anders als bei konventionellen Inhaltsanalysen, so dass, nicht zuletzt durch die Spezialisierung der Verfahren auch auf personeller Ebene, der Spalt zwischen manuellen oder automatischen Ansätzen eher größer wird. Forscherinnen und Forscher, die beide Ansätze souverän beherrschen, waren und sind selten. Drittens sind vielfach automatische und konventionelle Inhaltsanalysen schlicht inkommensurabel, da ihnen eine gemeinsame Metrik für Reliabilität und Validität fehlt.[25]

25 Dies erschwert auch die Einordnung verschiedener Studien, in denen manuelle und (ggf. verschiedene) automatische Verfahren anhand einer Fragestellung miteinander verglichen werden (Rosenberg et al., 1990; Morris, 1994; Conway, 2006).

Eine Lösung für die skizzierten Probleme bieten meiner Meinung nach die hier vorgestellten induktiven Verfahren, insbesondere die überwachte Textklassifikation. Durch die Kombination von manueller Trainingscodierung und automatischer Normalcodierung vereint die Textklassifikation auf natürliche Weise die Vorteile – und natürlich auch Nachteile – beider Ansätze. Forscher und Codierer tun, was (nur) sie am besten können, nämlich theoretisch geleitete Interpretationen von Mitteilungen in numerische Codes zu überführen, während der Computer komplexe statistische Modelle daraus entwickelt und diese auf umfangreiche Dokumentkorpora anwendet. Fünf Argumente sprechen rein konzeptionell und methodologisch für den Einsatz überwachter Textklassifikation in der sozialwissenschaftlichen Inhaltsanalyse:

Anschlussfähigkeit Die automatische Klassifikation baut auf den beliebig theoretisch und operational begründeten Kategorien konventioneller Inhaltsanalysen auf. Ob diese sich für die Induktion von lexikalischen Klassifikationsregeln eignen, d.h. automatisierbar sind, ist eine empirische, aber keine methodologische Frage. Die Maßstäbe, nach denen eine automatische Codierung als reliabel und valide angesehen wird, sind grundsätzlich dieselben wie für die konventionelle Inhaltsanalyse.

Effizienz Lernende Algorithmen entlasten den Forscher von der aufwändigen Regeldefinition, sind aber in der Anwendung genauso effizient wie vollautomatische Verfahren. Der Umgang mit der Klassifikationssoftware ist zudem auf die Vergabe von Codes beschränkt, also das, was ohnehin in jeder konventionellen Inhaltsanalyse getan wird. Ein trainierter Klassifikator kann anschließend große Textmengen schnell und zuverlässig verarbeiten.

Sprach- und Materialunabhängigkeit Da überwachte Klassifikationsalgorithmen lediglich auf einer statistischen Modellierung und einem *Bag-of-Words*-Ansatz beruhen, sind sie ohne individuelle Anpassung mit unterschiedlichem Textmaterial in allen Sprachen anwendbar. Auf sprachspezifische Bereinigungs- und Analyseschritte kann daher grundsätzlich verzichtet werden. Sie können jedoch problemlos in den Analyseprozess integriert werden.

Reliabilität Lernende Klassifikatoren sind vollständig reliabel in dem Sinne, dass mit identischem Trainingsmaterial und gleichen Randbedingungen die Codierungen exakt reproduzierbar sind. Bezogen auf die inhaltliche Codierung sind sie aufgrund des – im Vergleich zu regel- oder diktionärbasierten Verfahren zusätzlichen – Induktionsschritts vom Trainingsmaterial zum statistischen Modell jedoch weniger zuverlässig. Der Klassifikator kann nie besser sein, als es die vorhandenen manuellen Codierungen erlauben, und wird zusätzlich eigene Codierfehler machen.

Validität Da die Validität der Codierung vor allem von der Validität des Kategorienschemas und dessen zuverlässiger Umsetzung abhängt, sind die Ergebnisse induktiver Klassifikationen prinzipiell genauso gültig wie die einer manuellen Codierung. Im Gegensatz zu deduktiven Verfahren ist die Validität nicht bereits a priori durch methodische Restriktionen begrenzt, allenfalls kann sich empirisch zeigen, dass sich eine Kategorie nicht valide automatisieren lässt. Diese Prüfung ist jedoch transparent und ohne zusätzlichen Codieraufwand umsetzbar.

Schließlich lässt sich noch ein forschungsökonomisches Argument für die Verwendung von maschinellem Lernen in der Inhaltsanalyse anbringen: Sie ist mit wenig oder keinem zusätzlichen Aufwand bzw. Kosten verbunden. Jede manuelle Codierung digitalisierter Inhalte kann quasi nebenbei daraufhin untersucht werden, ob die Klassifikation ggf. auch automatisch durchführbar ist. Durch den parallelen Einsatz automatischer Verfahren wird die manuelle Inhaltsanalyse weder gestört noch entstehen zusätzliche Kosten. Ist das maschinelle Lernen erfolgreich, lassen sich umfangreiche automatische Analysen mit minimalem Aufwand durchführen.

Ob und wie sich dieses Ziel umsetzen lässt, ist die Frage, die in den folgenden Kapiteln beantwortet werden soll. Dabei werden zunächst zentrale methodische Probleme im Kontext automatischer Analyseverfahren und vor allem des maschinellen Lernens diskutiert. Was mit überwachten Klassifikationsverfahren möglich und nicht möglich ist, soll in anschließend in den Kapiteln 5 bis 7 empirisch untersucht werden.

4 Problemfelder und Rahmenbedingungen überwachter Textklassifikation in der Online-Inhaltsanalyse

Im folgenden Kapitel werde ich die konzeptionellen und methodischen Problemfelder skizzieren, die für eine möglichst umfassend automatisierte Analyse von digitalen Medieninhalten relevant sind. Ausgehend von den im vorangegangenen Abschnitt dargestellten Überlegungen stellt dabei die überwachte Klassifikation von Textinhalten den zentralen Bezugsrahmen der Argumentation dar. Dies bedeutet jedoch nicht, dass die diskutierten Fragen nur auf Problemfelder des maschinellen Lernens abzielen. Im Gegenteil sind insbesondere Fragen der Datenerhebung und -bereinigung (Abschnitte 4.1 und 4.2) gleichermaßen zentral für manuelle und automatische Inhaltsanalysen digitaler oder digitalisierter Inhalte. Spezifisch auf den Gegenstand dieser Arbeit ausgerichtet ist hingegen Abschnitt 4.3, in dem es um die möglichst effektive und effiziente Verknüpfung manueller und automatischer Codierung geht. Diese Gedanken führen einerseits hin zu konzeptionellen Vorschlägen für ein Erhebungs- und Analyseinstrument, das im Anhang A dokumentiert ist, andererseits zu empirischen Forschungsfragen, die am Ende dieses Kapitels zusammengefasst werden und zur eigentlichen Evaluationsstudie in den Kapiteln 5 bis 7 überleiten. Da die Reliabilität und Validität der manuellen und automatischen Codierung für diese Evaluation von entscheidender Bedeutung sind, wird dem Thema der Qualitätsbestimmung ein umfangreicher Abschnitt (4.4) gewidmet.

4.1 Erhebung maschinell codierbarer Medieninhalte

4.1.1 Off- und Online-Archive

Wie im vorangegangenen Kapitel dargestellt wurde, ist die Geschichte automatischer Inhaltsanalysen aufs engste mit der Digitalisierung der Medienlandschaft verbunden. Ohne maschinenlesbare Daten funktioniert kein computergestütztes Verfahren, und die sozialwissenschaftliche Relevanz solcher Verfahren beginnt demnach erst mit der Verfügbarkeit entsprechender Dokumente. In der Anfangsphase automatischer Inhaltsanalysen war nicht nur die Kapazität von EDV-Anlagen ein limitierender Faktor, sondern auch die äußerst aufwändige Digitalisierung des Codiermaterials. Dementsprechend wurde mit vergleichsweise kleinen Textmengen, etwa Transkripten von Interviews oder einzelnen literarischen oder politischen Werken gearbeitet (Iker & Harway, 1969; Mosteller & Wallace, 1964).

Erst mit der Umstellung der Verlage auf Desktop Publishing Verfahren seit Mitte der 70er Jahre waren erstmals maschinenlesbare Inhalte ohne aufwändige Transkriptionen für die Forschung verfügbar. DeWeese (1977) konnte in einer Pionierstudie anhand der Detroit News erstmals zeigen, wie sich eine kontinuierliche Presseanalyse durch die Verwendung von (ohnehin anfallenden) digitalen Druckvorlagen automatisieren lässt. Dazu wurde ein Computer an das Drucksystem der Zeitung angeschlossen, der die Signale, die an die Druckmaschinen gingen, aufzeichnete, bereinigte und die Dokumente zur Weiterverarbeitung auf Magnetbänder sicherte.

Nachdem in den folgenden Jahrzehnten der computergestützte Textsatz in allen Verlagen Einzug hielt, wurden von den Medien selbst in großem Umfang digitale Archive ihrer Produkte angelegt. In der Folge wurden der Wissenschaft ganze Jahrgänge von Setzbändern zugänglich gemacht, etwa vom Mannheimer Morgen für das Institut für deutsche Sprache (Galliker, 1998). Seit Beginn der 90er Jahre veröffentlichen die meisten Printmedien Archiv-CD-ROMs bzw. DVDs entweder einzelner Jahrgänge (z.B. die SZ, FAZ und NZZ) oder gar des gesamten Bestan-

des (TAZ).[1] Diese sind nicht nur in den meisten Bibliotheken erhältlich, sondern oft auch für einzelne Wissenschaftler erschwinglich, so dass aus finanzieller Hinsicht die Beschaffung digitaler Medieninhalte heute unproblematisch ist. Will man allerdings die Beiträge aus verschiedenen Printmedien vergleichen, ist die Arbeit mit Archiv-CDROMs recht aufwändig, weil die Beiträge jeweils in unterschiedlichen Formaten und mit unterschiedlichen Rechercheprogrammen gebündelt werden. Vor einer übergreifenden Analyse müssen die Dokumente daher in ein einheitliches Format gebracht werden, wobei die Metadaten (Ausgabe, Seite, Platzierung) erhalten bleiben sollten.

Eine für die sozialwissenschaftliche Forschung prinzipiell höchst attraktive Alternative zur Sammlung einzelner Jahrgangsmedien liegt in der Nutzung zentraler Dokumentationsdienste, die Inhalte verschiedener Medien einheitlich archivieren und über eine Benutzerschnittstelle zur Verfügung stellen. Der bekannteste Anbieter solcher Dienste ist LexisNexis, das ursprünglich als Dienstleister für Anwaltskanzleien und Wirtschaftsunternehmen Medieninhalte sammelte, verschlagwortete und bereits Ende der 70er Jahre per Datenfernübertragung recherchierbar machte. Aktuell indiziert LexisNexis u.a. über 170 deutsche Zeitungen und Zeitschriften – von der FR bis zur Lebensmittelzeitung. Abonnenten können online auf die Volltexte aller Ausgaben zugreifen, wobei das Archiv bis in die 90er Jahre, bei englischsprachigen Zeitungen auch weiter zurückreicht (Deacon, 2007). Über ein Webinterface kann dabei sowohl mit den Mitteln der Freitextrecherche (vgl. Abschnitt 3.3.2) als auch nach einzelnen Ausgaben gesucht werden. Obwohl es theoretisch möglich ist, ganze Jahrgänge schrittweise aus dem Archiv zu extrahieren, ist dies aufgrund von Restriktionen der Benutzerschnittstelle extrem aufwändig und würde auch gegen die Nutzungsbedingungen verstoßen.[2]

1 Ausgerechnet die auflagenstärkste deutsche Tageszeitung Bild ist bislang nicht als Volltext-Archiv verfügbar. Dies mag u.a. auch der Grund dafür sein, dass diese vergleichsweise selten Gegenstand – auch konventioneller – Inhaltsanalysen ist.

2 Streng genommen verhindern die Nutzungsbedingungen jeglichen wissenschaftlichen Einsatz, da u.a. auch festgelegt wird, dass die Rechercheergebnisse nur 90 Tage gespeichert werden dürfen (§5 Abs. 3, http://www.lexisnexis.com:80/de/business/auth/displayterms.do?content=GENERAL). Da dies eine Replikation von veröffentlichten Analysen effektiv verhindern würde, ist der Nutzen von LexisNexis unter diesen Bedingungen fraglich.

Jenseits dieser rechtlichen Fragen stehen wissenschaftliche Kriterien beim Umgang mit Archiv-Diensten im Vordergrund. Deacon (2007) wirft eine große Anzahl potentieller Validitäts- und Reliabilitätsprobleme auf, mit denen LexisNexis-Nutzer konfrontiert sind. Die Validität sieht der Autor vor allem dadurch eingeschränkt, dass die Artikel ohne Kontext und nicht im Original-Layout vorliegen. Visuelle Aspekte einer Inhaltsanalyse von Printmedien gehen daher bei Volltext-Archiven verloren. Da es in dieser Arbeit jedoch ohnehin um textbasierte Verfahren geht, ist dieses Argument hier eher nachrangig. Das Validitätsproblem, dass im Zugriff durch Freitextrecherche liegt, wurde bereits in Abschnitt 3.3.2 behandelt. Mindestens genauso wichtig sind jedoch die Fragen der Reliabilität, die sich bei der Nutzung von LexisNexis und anderen Archiven stellen: Deacon (2007) kann zeigen, dass es nicht unerhebliche Lücken im Archiv gibt, sowohl auf Beitrags- als auch Ausgabeebene, dass Artikel ggf. mehrfach indiziert werden und die Zuordnung von Untersuchungseinheiten zur Printausgabe nicht immer zuverlässig funktioniert. Weaver & Bimber (2008) können zudem empirisch nachweisen, dass es weitere Lücken im Angebot von LexisNexis gibt, die sich aus der Nicht-Archivierung von Agenturmeldungen ergeben. Auf diese Probleme weisen bereits Snider & Janda (1998) hin, die einen umfassenden Überblick über die verschiedenen Anbieter von Volltext-Archiven amerikanischer Printmedien bieten.

In jüngster Zeit haben viele Zeitungs- und Zeitschriftenverlage damit begonnen, ihre Archive gänzlich der Öffentlichkeit zugänglich zu machen. Je nach Anbieter sind dabei sowohl reine Text- als auch gescannte Originalfassungen der Artikel verfügbar (vgl. Tabelle 4.1). Durch Fortschritte in der Scan- und *Optical Character Recognition*-Technologie (OCR) sind auch sehr alte Ausgaben oft vollständig durchsuchbar. Zusätzlich bieten viele Verlage auch die Möglichkeit, sich gezielt einzelne Ausgaben herausgeben zu lassen. Je nach Geschäftsmodell ist die Archivsuche vielfach kostenlos, zum Teil werden jedoch auch recht hohen Preise pro Artikel verlangt (etwa von der NY Times). In Deutschland sind vor allem Spiegel und Zeit Vorreiter auf diesem Gebiet, hier sind die vollständigen Archive seit Gründung kostenlos online verfügbar. Prinzipiell gelten jedoch auch für die selbstangelegten Archive der Verlage dieselben Vorbehalte, die Deacon (2007) gegen LexisNexis anführt. Die verlagseigenen Online-Archive

Tabelle 4.1: Online-Archive von deutschen Printmedien

Medium	Archiv seit	Text/Bild	Einzelausgaben	kostenfrei
SZ	1992	Text/Bild	ja	nein
FAZ	1993	Text/Bild	ja	nein
Welt	1995	Text/Bild	ja	nein
FR	1994	Text/Bild	nein	nein
Taz	1986	Text	ja	nein
Spiegel	1947*	Text	ja	ja
Focus	1993*	Text	ja	ja
Zeit	1946*	Text	ja	ja

* Gründungsjahr

sind jedoch bislang nicht systematisch untersucht worden, weshalb die Qualität dieser Quellen nur schwer einzuschätzen ist.

4.1.2 Erhebung von Online-Nachrichten

Eine Alternative zum Rückgriff auf bereits archivierte Datenbestände bietet die automatische Erhebung von aktuellen Online-Medieninhalten. Hierbei werden öffentlich verfügbare Quellen im Internet nach relevanten Inhalten abgesucht und diese anschließend systematisch weiterverarbeitet, d.h. bereinigt und gespeichert. Da es in diesem Kapitel nicht um grundsätzliche Fragen der Online-Inhaltsanalyse geht (vgl. dazu ausführlich McMillan, 2000; Weare & Lin, 2000; Rössler & Wirth, 2001; Rössler, 2010), seien an dieser Stelle nur die für automatische Analysen wichtigsten Probleme dieser Art der Datenerhebung genannt:[3]

1. Aktuelle Online-Inhalte sind vielfach flüchtig und dynamisch, so dass im Gegensatz zu archivierten oder Offline-Inhalten keine einheitliche kanonische Form einer Nachricht existiert (Seibold, 2002). Um replizierbare Analysen durchzuführen, müssen daher Dokumente zusammen mit

3 Insbesondere wird an dieser Stelle die Frage der passenden Untersuchungseinheit (Rössler, 2002) nicht diskutiert. Damit verbunden ist auch der Verzicht auf eine Problematisierung der Stichprobenziehung von Online-Inhalten (Meier et al., 2010).

relevanten Meta-Daten wie einer Identifikationsnummer[4] und dem Zeitpunkt des Abrufs dauerhaft archiviert werden. Speichert man die Daten im Original-Format (HTML, XML), ist gewährleistet, dass ein Dokument sowohl für die Klassifikationssoftware als auch für jeden Codierer mit gleicher Hard- und Software identisch vorliegt. Dies ist für die überwachte Klassifikation von Medieninhalten insofern von Bedeutung, als dass die Trainingsdaten manuell codiert werden müssen und dies in einem möglichst reliablen Prozess (vgl. Abschnitt 4.4) geschehen sollte.

2. Im Gegensatz zu klassischen (Text-)Emails oder anderen älteren Online-Medien ist das Angebot im World Wide Web multimedial, und dies gilt auch für Online-Nachrichten. Quandt (2008a, 140) misst bei verschiedenen Online-Nachrichten-Angeboten einen Anteil multimedialer Inhalte von rund zehn Prozent, wobei einfache Abbildungen noch nicht in diese Definition fallen. Grundsätzlich sind zwei alternative Strategien im Umgang mit multimedialen Inhalten denkbar: (A) In Anlehnung an die klassische Inhaltsanalyse werden lediglich verbale Inhalte erhoben und analysiert, dementsprechend müssen audiovisuelle Beitragsteile herausgefiltert oder schlicht ignoriert werden. (B) Folgt man der Feststellung von Rössler (2010), dass gerade die Analyse von multimedialen Inhalten ein akutes Forschungsdefizit in der Kommunikationswissenschaft darstellt, müssen diese möglichst originalgetreu erhoben und archiviert werden. Technisch stellt dies eine nicht unerhebliche Hürde dar, da multimediale Inhalte zwar im HTML/XML-Dokument referenziert, selbst aber erst durch den Browser dargestellt werden. Durch Video- und Audiostreaming sowie andere Echtzeit-Darstellungen ist es zunehmend schwierig, die eigentlichen Inhalte auf eigene Datenträger zu archivieren. Dies führt dazu, dass ggf. auch textuelle Inhalte, die in Bilddateien oder Adobe Flash-Grafiken versteckt sind, nicht als *codierbare* Untersuchungseinheiten vorliegen und damit der (automatischen) Analyse entzogen sind (Schweiger & Weber, 2010; Bock et al., 2010).

3. Durch die Hypertextualität des Mediums Internet liegen Dokumente selten linear geordnet vor. Während man eine Zeitung oder Fernsehnachrichten von Anfang bis Ende codieren kann, fällt ein solcher Zugriff

4 Wählt man als Untersuchungseinheit eine einzelne Webseite, ist dies die URL (Unique Resource Locator), bei E-Mails und Usenet-Postings die Message-ID.

auf die Daten bei Online-Angeboten schwer. Gängige Strategien für das Codieren von tageweisen Ausgaben einzelner Online-Nachrichten bedienen sich daher sog. *Web-Crawler*. Diese Programme folgen systematisch allen Links auf einer gegebenen Webseite und laden die so gefundenen Seiten ebenfalls herunter, ggf. wird dieses Verfahren rekursiv angewendet, wobei die Anzahl gefundener Dokumente exponentiell ansteigt (Quandt, 2008a). Da die Crawler-Software zumeist nur mit wenigen Regeln steuerbar ist, z.B. dass nur Dokumente unter der Ausgangsdomain zu indizieren sind, werden häufig sehr viele Dokumente archiviert, die nicht dem Zugriffskriterium entsprechen. Neben Nachrichten enthält eine Website wie Spiegel Online auch viele nichtredaktionelle Seiten, Forenbeiträge und Service-Angebote, die ebenfalls auf der Startseite verlinkt werden. Dementsprechend ist eine manuelle Auswahl relevanter Dokumente bei einer solchen Erhebungsstrategie unerlässlich.[5] Eine Alternative zum Website-Crawling stellen XML-Feeds – etwa nach dem RSS- oder ATOM-Protokoll – dar, die hochgradig strukturiert sind. Sie enthalten kaum störende Inhaltselemente, sind bereits mit Metadaten versehen und können mit standardisierten Parsing-Programmen verarbeitet werden (Kantel, 2007). Mittlerweile stehen nicht nur für Weblogs, sondern auch für viele klassische Nachrichten-Sites, Web-Foren und andere interaktive Angebote ständig aktualisierte Feeds zur Verfügung, was sie für umfangreiche automatisch Datenerhebungen sehr attraktiv macht (Erlhofer, 2010).

Zusammenfassend sind folgende Probleme bei der Entwicklung eines hochgradig automatisierten Forschungsinstruments für Inhaltsanalysen zu berücksichtigen:

1. Die Untersuchungseinheiten müssen systematisch geordnet und mit Meta-Daten versehen vorliegen. Ist kein bereits vorstrukturiertes Offline- oder Online-Archiv verfügbar, müssen die Dokumente zunächst aus potentiell wenig strukturierten Medienangeboten extrahiert werden.

5 Ein weiterer Nachteil bei der Nutzung von Crawlern liegt in der Tatsache, dass die heruntergeladenen Dateien meist nur in einer losen Ordner-Struktur vorliegen, die sich nicht für den automatischen Zugriff bei der Codierung und Analyse eignet.

2. Da viele Online-Medien nicht linear erhebbar sind, müssen inhaltliche Zugriffskriterien technisch operationalisiert werden, um zu vermeiden, dass viele irrelevante Inhalte erhoben werden, die später ggf. manuell entfernt werden müssen.
3. Die Erhebung von multimedialen Inhalten ist zur Zeit deutlich schwieriger als bei Textdokumenten, da Bilder, Audio und Video entweder auf Angebotsseite (etwa bei Volltext-Archiven) fehlen und/oder nur mit großem Aufwand archivier- und codierbar sind.

4.2 Datenbereinigung und Vorbehandlung

4.2.1 Identifikation der Untersuchungseinheiten

Natürlichsprachliche Dokumente, die in erster Linie für den menschlichen Leser gedacht sind, enthalten häufig verbale und visuelle Bestandteile, die bei der automatischen Weiterverarbeitung und Analyse nutzlos oder gar schädlich sind. Dazu gehören etwa Werbung oder andere Anzeigen, Inhaltsverzeichnisse und Logos, aber auch Navigationselemente oder Formulare auf Websites, man spricht auch von *Boilerplate-Content* (Kohlschütter et al., 2010). Mitunter ist der Anteil an irrelevanten Inhaltselementen so hoch, dass der eigentliche Artikeltext nur schwer zu erkennen ist (vgl. Abbildung 4.1). Auf der dargestellten Seite auf bild.de nimmt der eigentliche Artikeltext inklusive Überschrift und Bildunterschrift nur einen Bruchteil des sicht- und lesbaren Inhalts ein.

Während Menschen im Zuge der Lesesozialisation auch komplexe Textdokumente schnell und sicher zu dekodieren lernen (LaBerge & Samuels, 1974), ist die Identifikation relevanter Textinhalte für Computer alles andere als trivial. Folgerichtig wird das Thema der Textidentifikation bei konventionellen Inhaltsanalysen gar nicht oder nur am Rande behandelt, etwa bei der Definition der Untersuchungseinheit oder des Aufgriffkriteriums. Die eigentliche Selektion der relevanten verbalen und/oder nonverbalen Inhalte wird dagegen der Lesekompetenz und dem Urteilsvermögen der Codierer überlassen (vgl. Wirth, 2001).

Für die vollautomatische Analyse von Dokumenten, die nicht bereits dafür aufbereitet sind, stellt sich das Problem deutlicher als bei manuellen

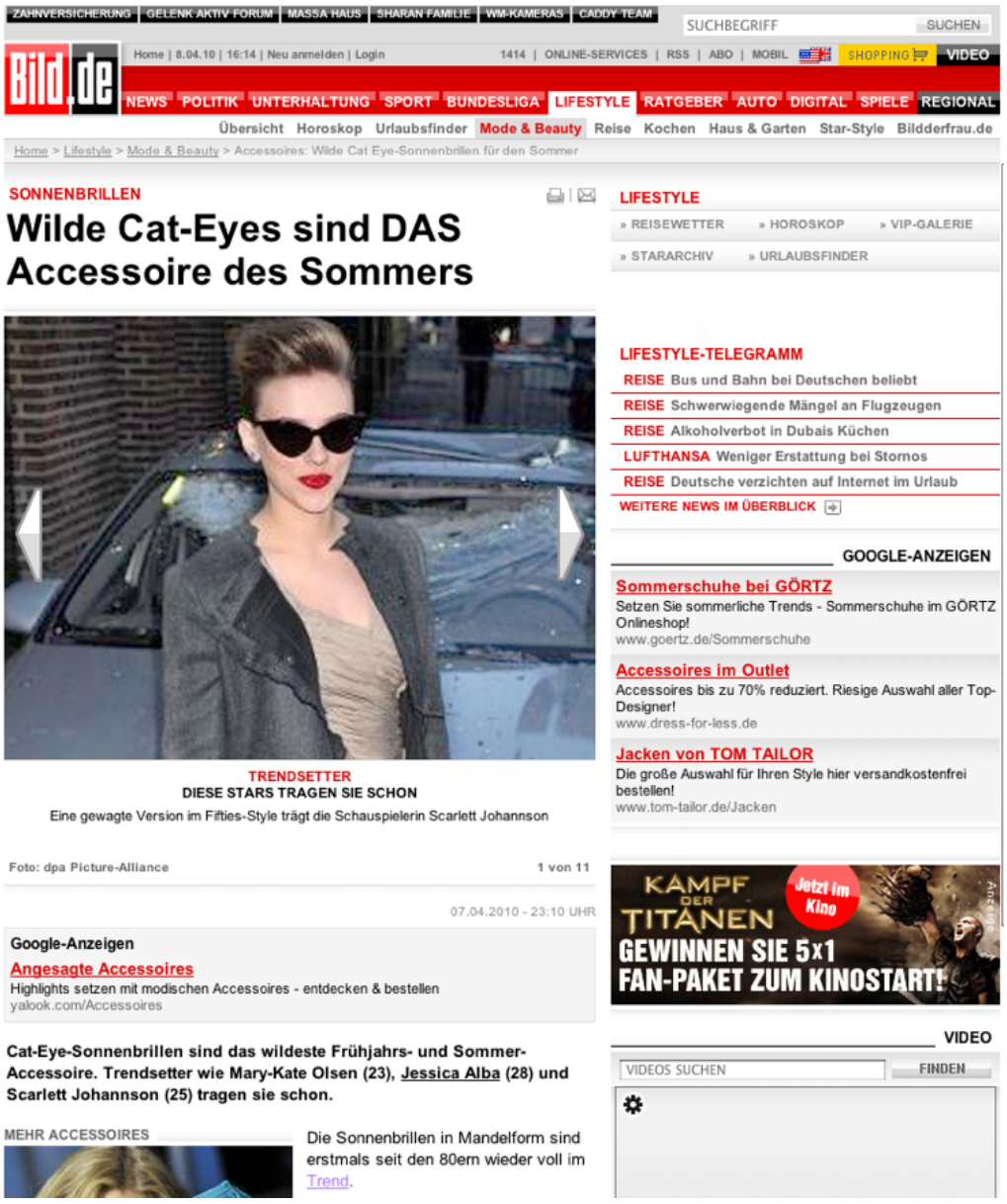

Abbildung 4.1: Screenshot eines Beitrags auf bild.de

Verfahren: Da der Algorithmus auf der Ebene einzelner Bytes und Zeichen kaum Anhaltspunkte dafür hat, was überhaupt relevanter *Text* ist, gestaltet sich die Weiterverarbeitung komplexer Dokumente schwierig. Auf Webseiten besteht dabei zusätzlich das Problem, dass rein quantitativ sowohl die nonverbalen Formatierungsanweisungen (HTML-Tags, Stylesheets) als auch für den Artikel irrelevante Textelemente (Links, Werbebanner, Kommentarfelder) den Anteil relevanten Fließtextes bei weitem übersteigen. Im dargestellten Beispiel beträgt die eigentliche Beitragslänge 778 Zeichen, während der HTML-Quelltext über 104.000 Zeichen lang ist.

Für die automatische Analyse birgt diese Komplexität der Dokumentstruktur zwei Herausforderungen: Einerseits steigt der Verarbeitungsaufwand, d.h. die benötigte Speicher- und Rechenkapazität erheblich (bei Abbildung 4.1 um den Faktor 100), andererseits wird die Analyse durch tausende irrelevante Zeichen und Wörter erschwert, was die Validität und Reliabilität der Ergebnisse u.U. stark verringern kann. Li & Ezeife (2006) untersuchen in ihrer Studie die Effektivität verschiedener Bereinigungsalgorithmen für Webseiten und deren Konsequenzen für die thematische Klassifikation der Beiträge. Dabei zeigt sich, dass die Klassifikation von unbereinigten Dokumenten zwischen 7 und 33 Prozent schlechter ausfällt als bei entsprechend aufbereiteten Inhalten.

Neben der manuellen Textextraktion sind auch automatische Verfahren in jüngster Zeit entwickelt worden. Vergleichsweise einfach umzusetzen ist eine regelbasierte Selektion anhand wiederkehrender struktureller Merkmale (vgl. Abschnitt 3.3.3). Dieses Vorgehen ist vor allem bei Daten aus einer bekannten Quelle geeignet: Mittels HTML-Parsern oder regulären Ausdrücken kann leicht definiert werden, dass beispielsweise der Artikelinhalt auf faz.net immer im Container-Element `<div class="Article">` enthalten ist oder bei LexisNexis immer in der dritten Zeile der Datei beginnt. Bei unbekannten oder häufig wechselnden Textstrukturen ist dieses Verfahren hingegen wenig hilfreich, da die Regeldefinition oft aufwändiger ist als die manuelle Extraktion bei wenigen Dokumenten. Hier sind heuristische, offene Extraktionsalgorithmen von Vorteil, die sich bekannte Strukturen von Haupttext und irrelevanten Inhalten zunutze machen.

Da die Bereinigung von elektronischen Dokumenten sowohl in der akademischen als auch kommerziellen Anwendung große Bedeutung hat, etwa für Betreiber von Suchmaschinen, wird dieses Feld in der Informatik zur Zeit intensiv erforscht. Mit dem CleanEval-Wettbewerb (Baroni et al., 2008) stehen dafür ein zentraler Anlaufpunkt, ein gemeinsamer Testkorpus sowie zahlreiche Referenzwerte zur Verfügung. Bei den erfolgreichsten heuristischen Verfahren werden sowohl linguistische, d.h. sprachabhängige, als auch visuelle oder strukturelle Informationen verarbeitet. Da die Extraktion relevanter Inhalte auch als Klassifikationsproblem verstanden werden kann, werden auch bei der Bereinigung der Texte die in Kapitel 3.4.1 beschriebenen Verfahren eingesetzt. Einen Überblick zu verschiedenen Ansätzen bieten aktuelle Arbeiten von Baisa (2009) und Kohlschütter et al. (2010).

Insgesamt hat die Extraktion relevanter Inhalte sowohl für die manuelle als auch automatische Inhaltsanalyse erhebliche Relevanz, da sie nicht nur die Anlage umfangreicher Textkorpora und Archive ermöglicht, sondern vor allem die Codierung erleichtert. Gelingt eine derartige Bereinigung ohne menschliche Unterstützung, können auch anschließende Analyseschritte automatisch durchgeführt werden, etwa textstatistische oder Co-Occurrence-Analysen. In bislang vorliegenden Studien wurden die Texte größtenteils manuell vorbereitet, was zwar in Veröffentlichungen nur am Rande erwähnt wird, jedoch einen nicht unerheblichen Teil der Arbeit ausmacht (vgl. Quandt, 2008b). Automatische Verfahren sind daher nicht nur für die Validitätssicherung von Nutzen, sondern auch forschungsökonomisch höchst wünschenswert.

Ist die Untersuchungseinheit nicht der einzelne Beitrag, sondern der Textabschnitt, Absatz oder gar die einzelne Aussage, können die genannten Verfahren ebenfalls zum Einsatz kommen. Dabei muss allerdings klar sein, dass formale oder syntaktische Aufgriffkriterien deutlich leichter automatisierbar sind als semantische. Anhand weniger Satzzeichen kann ein Text recht einfach in einzelne Sätze zerlegt werden, ebenso ist eine Extraktion von Absätzen durch Absatzmarken leicht implementierbar. Die Analyse von Aussagen setzt dagegen ein linguistisches Textverständnis voraus, dass in den seltensten Fällen reliabel automatisierbar ist (Krippendorff, 2004a, 109-110). Die gilt auch von konventionelle Inhaltsanalysen,

bei denen die Identifikation von Aussagen den Codierern in der Regel deutlich schwerer fällt als deren eigentliche Codierung (Gerhards et al. 2007, 117; Raupp & Vogelgesang 2009, 133). Für Analysen unterhalb der Beitragsebene scheint daher die Verwendung einfacher syntaktischer Regeln zur Identifikation der Analyseeinheit daher empfehlenswert, auch wenn vergleichende Methodenexperimente zu dieser Problematik bislang nicht vorliegen.

4.2.2 Preprocessing

Bevor digitale Texte automatisch analysiert werden können, bedarf es in aller Regel einer umfassenden Datenbereinigung. Dieses Preprocessing dient vor allem der Reduktion der Anzahl von Wortformen und damit Variablen im statistischen Modell. Schon ein relativ kurzer Text kann aus hunderten individuellen Wortformen bestehen, bei der Verwendung von Bigrammen und längeren Wortgruppen steigt die Zahl an Features exponentiell an (vgl. dazu den nächsten Abschnitt). Um umfangreiche Textkorpora überhaupt mit vertretbarem Zeit- und Ressourcenaufwand verarbeiten zu können, ist in vielen Fällen eine Vorbehandlung der Textdaten unerlässlich. Der Nachteil bei den meisten feature-reduzierenden Verfahren liegt in der Tatsache, dass die ggf. für die Analyse notwendige syntaktische und semantische Vielfalt so stark reduziert wird, dass die automatische Klassifikation darunter leidet (Leopold & Kindermann, 2002).

Grundsätzlich lassen sich dabei rein algorithmische und linguistische Verfahren unterscheiden. Letztere sind prinzipiell aufwändiger umzusetzen, da sie einerseits vorbereitete Wortlisten benötigen und damit sprachabhängig sind, anderseits in der Regel auch mehr Rechenaufwand bedeuten, da ggf. für jedes Wort ein oder mehrere Wortlistenvergleiche notwendig sind. In Anlehnung an die Auflistung von Hotho et al. (2005) sollen an dieser Stelle die wichtigsten Preprocessing-Verfahren vorgestellt und deren Relevanz für die eigentliche Analyse diskutiert werden. Für alle Ansätze in Tabelle 4.2 gilt, dass sie auf eine vollautomatische Verarbeitung ausgelegt sind, d.h. die vorliegenden Dokumente werden ohne Eingriffe des Forschers transformiert.

Tabelle 4.2: Verfahren für automatisches Preprocessing von Texten

Verfahren	Kurzbeschreibung	wortlistenbasiert
Textfilterung	Entfernung von Zeichen und Wörtern aus dem Dokument	teilweise
Stemming	Kürzen von gebeugten Wörtern auf einen Wortstamm	nein
Lemmatisierung	Ersetzung von gebeugten Wörtern durch deren Grundform	ja
Disambiguierung	Auflösung von Homonymen und Polysemen	ja
Anaphoren-Auflösung	Ersetzung von Pronomen durch ihr Bezugswort	teilweise
Part-of-Speech-Tagging	Zuordnung der Wortart zu jedem Wort	teilweise

Textfilterung
Ein Standardverfahren, das in fast allen Analysekontexten eingesetzt wird, ist die Filterung von Texten nach spezifischen Regeln. Beispielsweise werden fast immer Satz- und Sonderzeichen wie Bindestriche, Klammern und Absatzmarken vor der Analyse aus dem Text entfernt, so dass tatsächlich nur noch ein *Bag of Words* zur Klassifikation vorliegt. Häufig werden zudem alle Wörter in Groß- oder Kleinbuchstaben verwandelt, was jedoch u.U. in deutschsprachigen Dokumenten deutlich schwerwiegendere Konsequenzen hat als in englischen oder französischen Texten, in denen ohnehin die meisten Wörter kleingeschrieben sind.

Eine weitere oft verwendete Transformation stellt die Entfernung bestimmter Wörter aus dem Text dar. Dabei werden besonders häufige oder besonders seltene Wortformen aus dem Text entfernt. Dies geschieht aus der Überlegung heraus, dass sowohl extrem seltene als auch extrem häufige Wörter bei der Klassifizierung nicht von Nutzen sind: Wörter, die im gesamten Korpus nur in einem Dokument vorkommen, sind bei der Klassifikation genauso wertlos wie Wörter, die über alle Kategorien hinweg häufig sind. Zu den extrem häufig vorkommenden sog. Stopwörtern

zählen vor allem Personalpronomen, Präpositionen und Konjunktionen. Bei der Entfernung von Wörtern kann einerseits rein induktiv, d.h. durch Zählung von Worthäufigkeiten, vorgegangen werden, indem beispielsweise Wörter entfernt werden, die in weniger als einem oder mehr als 95 Prozent der Dokumente vorkommen. Andererseits kann für die Stopwort-Entfernung auch auf fertige Listen zurückgegriffen werden, etwa die 100 oder 1000 häufigsten Wörter im Deutschen, die vom Wortschatz-Projekt der Universität Leipzig zur Verfügung gestellt werden.[6]

Obwohl gerade die Entfernung von Stopwörtern eine weit verbreitete Praxis ist, sind deren Vorteile sowohl konzeptionell als auch empirisch zweifelhaft: Einerseits sind die meisten Klassifikationsalgorithmen sehr robust gegenüber statistischem Rauschen und ordnen Stopwörtern ohnehin niedrige Gewichte zu. Wenn die Analyse nicht Einwort- sondern N-Gramm-basiert ist, können auch häufig vorkommende Wörter entscheidende Bedeutung für die Semantik und damit die Trennschärfe eines Begriffspaars haben (Leopold & Kindermann, 2002, 438). Ein einfaches Beispiel dafür wäre eine Negation, die man nach Entfernung des häufig vorkommenden Wortes „nicht" kaum sinnvoll messen kann. Mit der Entfernung von Stopwörtern kann daher auch eine Verschlechterung der Klassifikation einhergehen (Riloff, 1995).

Stemming und Lemmatisierung

Geht man davon aus, dass die für die Klassifikation von Dokumenten notwendigen Informationen allein auf der lexikalischen, d.h. Wortebene eines Textes liegen, sollte sich die Zahl an Features pro Dokument ohne Validitätseinbußen reduzieren lassen, wenn man statt einzelner Wortformen nur eine einzige kanonische Form für die Analyse verwendet. Diese Annahme stellt die Grundlage der Verfahren *Stemming* und *Lemmatisierung* dar, die beide darauf abzielen, gebeugte Wortformen so weit wie möglich durch eine Einzelform zu ersetzen. Bei einer morphologisch vielfältigen Sprache wie Deutsch lässt sich so die Anzahl an Wortformen um bis zu 60 Prozent reduzieren (Leopold & Kindermann, 2002, 431).[7]

6 http://wortschatz.uni-leipzig.de/html/wliste.html

7 Für Volltextsuchen ergibt sich so außerdem der Vorteil, dass auch Dokumente, die eine gebeugte Form enthalten, bei der Suche mit dem ungebeugten Begriff gefunden werden können.

Tabelle 4.3: Originaltext, Stemming und Lemmatisierung im Vergleich

Original	Ich	habe	meinen	Kollegen	auf	der	Straße	gesehen
Stemming	Ich	hab	mein	Kolleg	auf	der	Strass	geseh
Lemmata	Ich	haben	mein	Kollege	auf	der	Strasse	sehen

Bei Stemming handelt es sich um ein algorithmisches, aber sprachabhängiges Verfahren, bei dem regelbasiert eine gebeugte Wortform auf einen – ggf. auch artifiziellen – Wortstamm reduziert wird. Der iterative Stemming-Algorithmus, bei dem wiederholt bekannte Suffixe aus einem Wort entfernt werden, wurde ursprünglich von Porter (1980) für die englische Sprache entworfen, ist aber mittlerweile für viele andere Sprachen implementiert (Porter, 2001). Ziel des Verfahrens ist es, eine möglichst einfache und effiziente Reduktion der Wortformen zu erreichen, die trotzdem bei Volltextabfragen gleich valide Ergebnisse garantiert. Dafür wird in Kauf genommen, dass nicht alle Suffixe korrekt erkannt und gelöscht werden und dass die entstandenen Wortstämme nicht immer sprachlich richtig sind (vgl. Tabelle 4.3).

Das linguistisch anspruchsvollere, gründlichere, aber auch deutlich aufwändigere Verfahren der Lemmatisierung ist ebenfalls sprachabhängig, benötigt aber zusätzlich ein Wörterbuch, in dem jeder gebeugten Wortform die entsprechende Grundform zugeordnet ist. Für jedes Wort im Text sind dementsprechend eine oder mehrere Abfragen aus dem Wörterbuch nötig, zudem ist die morphologische Analyse des Wortes nicht immer korrekt. Durch die Lemmatisierung wird die Anzahl an Wortformen pro Dokument je nach Sprache nochmals stärker verringert als bei algorithmischem Stemming. Ob dies allerdings den deutlich höheren Aufwand bei der Erstellung und Anwendung des Wörterbuchs rechtfertigt, bleibt eine offene empirische Frage. Tomlinson (2003) vergleicht algorithmische und lexikalische Ansätze in neun Sprachen und kann nur wenige signifikante Differenzen feststellen: Lemmatisierung führt bei der Volltextsuche in Finnisch und Deutsch zu leicht besseren Ergebnissen, in Schwedisch dagegen zu schlechteren, in Englisch, Französisch und Russisch zeigen sich keine Unterschiede. Angesichts des ungleich höheren

Aufwands scheint sich daher die lexikalische Feature-Reduktion nicht zu lohnen.

Die Frage, ob Stemming und Lemmatisierung überhaupt positive Effekte auf die Reliabilität und Validität der Analyse haben, ist ebenfalls empirisch nicht geklärt: Für sehr kurze Texte mit entsprechend wenigen Wortformen scheint sich die Reduktion auf Wortstämme zu lohnen (Tomlinson, 2003), bei längeren Dokumenten wie etwa Nachrichten ist kein positiver Effekt nachweisbar (Leopold & Kindermann, 2002; Scott & Matwin, 1999). Da zumindest algorithmisches Stemming relativ leicht umsetzbar ist, wird in dieser Arbeit der Effekt von Stemming auf die Qualität der Klassifikation zu untersuchen sein.

Weitere linguistische Verarbeitung

Neben den genannten Methoden zur Filterung und Wortformreduktion werden bei vielen automatischen Inhaltsanalysen weitere, zumeist linguistische Datenbereinigungsschritte durchgeführt. Schon der General Inquirer verfügte über sprachabhängige Preprocessing-Routinen, die die eigentliche Analyse erleichtern sollten. Rückblickend kann man sogar zu der Feststellung gelangen, dass neben der Wörterbuchentwicklung die Implementation der umfangreichen Disambiguierungsregeln die entscheidendste und aufwändigste Arbeit der General Inquirer-Entwickler darstellt (Stone, 1969b). Unter Disambiguierung versteht man in der Linguistik die Auflösung von mehrdeutigen Wortformen, also Homonymen und Polysemen, durch die Berücksichtigung der syntaktischen und semantischen Kontextes. Dazu bedarf es in der Regel umfangreicher Wörterbücher und Regelsätze, in der Form: Wenn „Hahn“ im Zusammenhang mit „krähen“ oder „Huhn“ auftaucht, geht es um den Vogel. Im Zusammenhang mit „Waschbecken“ oder „Wasser“ um die Armatur. Nach der Filterung steht dann etwa anstelle des Wortes „Hahn“ das Wort „Hahn_VOGEL“, um diese spezifische Bedeutung in der Analyse identifizierbar zu machen.

Dieselbe Logik führt das sog. *Part-of-Speech-Tagging* fort, in dem die syntaktische Funktion des Wortes identifiziert und als zusätzliches Feature in die Analyse einbringt. Dafür sind jedoch noch weitergehende sprachspezifische Wörterbücher und Regelsätze notwendig, so dass PoS-

Tagging als äußerst aufwändig und fehleranfällig gilt. Leopold & Kindermann (2002) verwenden ein relativ einfaches Ersetzungsverfahren, bei dem die PoS-Information einfach an das einzelne, ggf. lemmatisierte, Wort angehängt wird. So entsteht aus dem Wort „Kollegen“ im Beispiel aus Tabelle 4.3 ein Feature wie „Kollege_SUB_SING_MASK_AKK_OBJ“.

Insbesondere für aussagebasierte syntaktisch-semantische Analysen wird in jüngster Zeit häufig versucht, Anaphoren im Text automatisch aufzulösen. Darunter versteht man die Ersetzung von Pronomen durch die Worte, auf die sie verweisen. Dies ist vor allem dort relevant, wo möglichst exakte Subjekt-Verb-Objekt-Relationen untersucht werden sollen. Nur so kann aus einem Satz wie „Der Fernseher aus dem Katalog war defekt, deshalb habe ich ihn zurückgegeben.“ sinnvoll das Bezugsobjekt, nämlich der Fernseher, aus beiden Satzteilen extrahiert werden. Gerade im Bereich des *Opinion Mining* ist eine solche Vorbehandlungsstrategie äußerst wünschenswert. Leider sind die algorithmischen Verfahren zur Auflösung von Anaphoren bislang nicht weit genug entwickelt, um sie bedenkenlos anwenden zu können (Gürtler & Kronewald, 2010). Zudem ist der tatsächliche Nutzen bei der Klassifikation und Informationsextraktion nicht empirisch abgesichert. Bei reinen Bag-of-Words-Ansätzen, die ohne syntaktische Zusatzinformationen auskommen, sollte sich der Aufwand ohnehin nicht lohnen.

4.3 Klassifikatortraining

4.3.1 Blockweises und inkrementelles Training

Bei der Anwendung induktiver Textklassifikation kommt dem Prozess des Trainings – bzw. des Lernens aus Sicht des Klassifikationsalgorithmus – eine entscheidende Bedeutung zu. Aus den vorhandenen Daten werden die Parameter eines statistischen Modells abgeleitet, das dann für die Klassifikation neuer, d.h. uncodierter Dokumente verwendet wird. Jenseits verfahrensspezifischer Annahmen, etwa die der lokalen Unabhängigkeit der Features bei Naive Bayes, liegt der induktiven Klassifikation wie allen modellbasierten statistischen Methoden eine zentrale Annahme zugrunde: Die Daten des Trainings-Sets entstammen der selben Vertei-

lung wie die später zu klassifizierenden Dokumente. Mit anderen Worten, es wird erwartet, dass die Parameter des anhand von Beispieldaten erstellten Modells auch für alle später zu codierenden Dokumente gelten. Praktisch bedeutet dies, dass die Trainingsdaten repräsentativ für die Gesamtmenge an Dokumenten sein müssen, mithin einer Zufallsstichprobe daraus entstammen (vgl. Hopkins & King, 2010).

Ist diese Annahme plausibel, bietet es sich an, das Training des Klassifikators blockweise vorzunehmen. Bei diesem *Batch Learning* werden dem Algorithmus sämtliche vorcodierten Dokumente gleichzeitig vorgelegt. Dieser schätzt daraus in nur einem Schritt einen optimalen Parametervektor. Ein Beispiel für einen Batch-Learner ist die bereits genannte logistische Regression mit Maximum Likelihood Schätzung. Blockweise Trainingsstrategien haben zwei wichtige Vorteile für die angewandte Forschung: Erstens sind sie vollständig deterministisch, d.h. bei gegebenen Trainingsdaten ergeben sich identische Parameterschätzer, mit denen alle weiteren Codierungen komplett reproduzierbar sind. Zweitens ist Batch-Training höchst effizient, da mit einem einzigen Trainingsschritt der Klassifikator einsatzbereit ist.

In der Forschungspraxis ist die Repräsentativitätsannahme für die Trainingsdaten oft zweifelhaft oder gar offensichtlich verletzt. Einerseits stehen bei Inhaltsanalysen oft keine Zufallsstichproben zur Verfügung, entweder weil diese nicht praktikabel sind oder eine bewusste Auswahl getroffen wird (Kops, 1977). Andererseits kann der Klassifikator bei Echtzeit-Erhebungen nicht mit vollständig repräsentativen Daten trainiert werden, weil diese zum Zeitpunkt des Trainings noch nicht zur Verfügung standen (Hopkins & King, 2010). Als Beispiel sei hier die Anwendung von Klassifikatoren zum Filtern unerwünschter E-Mail-Werbung (Spam) genannt: Da sich Form und Inhalt von Spam-Mails mit der Zeit verändern, müssen auch die Klassifikatoren nachtrainiert werden. Man spricht hier von inkrementellen bzw. *On-line Learning* (Cormack & Bratko, 2006). In Situationen, in denen eine Veränderung der Dokumenteigenschaften bei gleicher Klassifikation, der sog. *Concept Drift*, bieten sich Klassifikatoren an, die schrittweise trainiert werden können. Eine mögliche, aber äußert ineffiziente Lösung des Problems kann auch mit blockweisem Training erreicht werden, indem der Klassifikator bei jedem zusätzlichen

Trainingsdokument zurückgesetzt und neu trainiert wird. Inkrementelle Algorithmen aktualisieren hingegen nur diejenigen Parameter des statistischen Modells, die vom neuen Trainingsdokument betroffen sind. Gleichzeitig ist zumeist ein Algorithmus des *Vergessens* implementiert, da sonst die Modellkomplexität unbegrenzt wachsen würde (Siefkes et al., 2004). Dies alles führt dazu, dass sich die Klassifikation eines bestimmten Dokuments im Zeitverlauf ändern kann. Auch die Reihenfolge der Trainingsdokumente kann zu unterschiedlichen Klassifikationsentscheidungen führen, so dass dieser Prozess nur bei gleichbleibender Anordnung der Trainingsdaten reproduzierbar ist.

Im Forschungsalltag ist ein inkrementeller Trainingsprozess vor allem dort von Vorteil, wo innerhalb der Feldphase noch auf Veränderungen im Codiermaterial eingegangen werden muss. Hier ist man ebenso flexibel wie bei der manuellen Codierarbeit, wo ebenfalls kurzfristig die Codieranweisungen geändert werden können. Am Beispiel des Spam-Filterns kann dies einfach illustriert werden: Ein bereits trainierter Klassifikator wird jeden Tag mit eintreffenden Emails konfrontiert und ordnet diese den Kategorien Spam/kein Spam zu. Stellt der Nutzer fest, dass eine erwünschte Mail irrtümlich als Spam klassifiziert wurde (oder umgekehrt), legt er diese Mail dem Algorithmus zum Training mit der wahren Kategorie vor. Der Klassifikator aktualisiert seine Modellparameter und wird ähnliche Mails in Zukunft richtig einordnen. Diese Strategie wird auch als *Train-on-Error* bezeichnet (Assis, 2006). Sie sorgt dafür, dass die aufwändige manuelle Codierung von Beispieldokumenten erheblich seltener benötigt wird als bei blockweisem Lernen.

Ein weiterer Vorteil inkrementellen Arbeitens liegt in der Tatsache, dass man den Umfang des Trainings-Sets relativ genau auf die zuvor definierten Reliabilitäts- und Validitätskriterien anpassen kann. Hierbei wird nach jeder Trainingseinheit ein Reliabilitätstest unternommen (vgl. den nächsten Abschnitt), um die Leistung des Klassifikators und den Bedarf an Nachschulungen abzuschätzen. Da jeder inkrementelle Algorithmus auch blockweise eingesetzt werden kann (*Single-Pass-Training*), sind diese besonders für den sozialwissenschaftlichen Forschungsalltag geeignet und werden daher im Folgenden vorrangig diskutiert.

4.3.2 Passives und aktives Lernen

Die im vorherigen Abschnitt dargestellten inkrementellen Lernstrategien waren – bezogen auf den Klassifikator – *passiv*, d.h. der Algorithmus bekommt seine Lernmaterialien zur Verfügung gestellt, ohne selbst darauf Einfluss zu nehmen. Dies ist vergleichsweise ineffektiv und ineffizient, weil nicht in jedem Dokument gleich viele Informationen zur Verbesserung der Klassifikationsleistung enthalten sind. Trainiert man einen Klassifikator mit zufällig oder bewusst ausgewählten Dokumenten, ist die Wahrscheinlichkeit groß, dass die meisten von ihnen bereits sicher eingeordnet werden können und daher keinen großen Einfluss auf die Parameter des Klassifikationsmodells haben. Auch bei der manuellen Codiererschulung werden deshalb selten Dokumente ausführlich besprochen, deren Kategorien offensichtlich sind. Vielmehr wenden sich die Mitarbeiter vor allem bei problematischen Variablen an den Untersuchungsleiter, um die korrekte Kategorisierung zu erfahren.

Dieses Prinzip des *aktiven Lernens* ist auch bei induktiven Klassifikationsverfahren von großen Nutzen, weil es den Bedarf an vorcodierten Beispielen erheblich senken kann (Lewis & Gale, 1994). Wie in Abbildung 4.2 dargestellt, übernimmt bei aktivem Lernen der Klassifikator selbst die Aufgabe, optimales Trainingsmaterial zu suchen, das zuerst manuell codiert werden soll. Ausgewählt werden solche Dokumente, bei denen der erwartete Zugewinn an klassifikationsrelevanten Informationen besonders hoch ist. Während bei blockweisem passiven Training eine zuvor definierte Menge an Beispieldokumenten verwendet wird, ist es bei inkrementellen aktiven Lernstrategien möglich und empfehlenswert, nur mit einem Minimum an Trainingsdaten (oft *Bootstrap-Sample* genannt) und einem Pool uncodierter Dokumente zu beginnen. Aus diesem wählt der Klassifikator nach Bedarf aus, bittet den menschlichen Codierer um die richtigen Kategorien und lernt mit wenigen guten Beispielen.

In den letzten Jahren hat eine rege Forschungstätigkeit auf dem Gebiet des aktiven Maschinenlernens eingesetzt, die sich vor allem mit der Frage der optimalen Auswahlentscheidung für Trainingsdokumente und deren Auswirkungen auf die Lerngeschwindigkeit und Leistung der Klassifikatoren beschäftigt. Einen umfangreichen Überblick über die

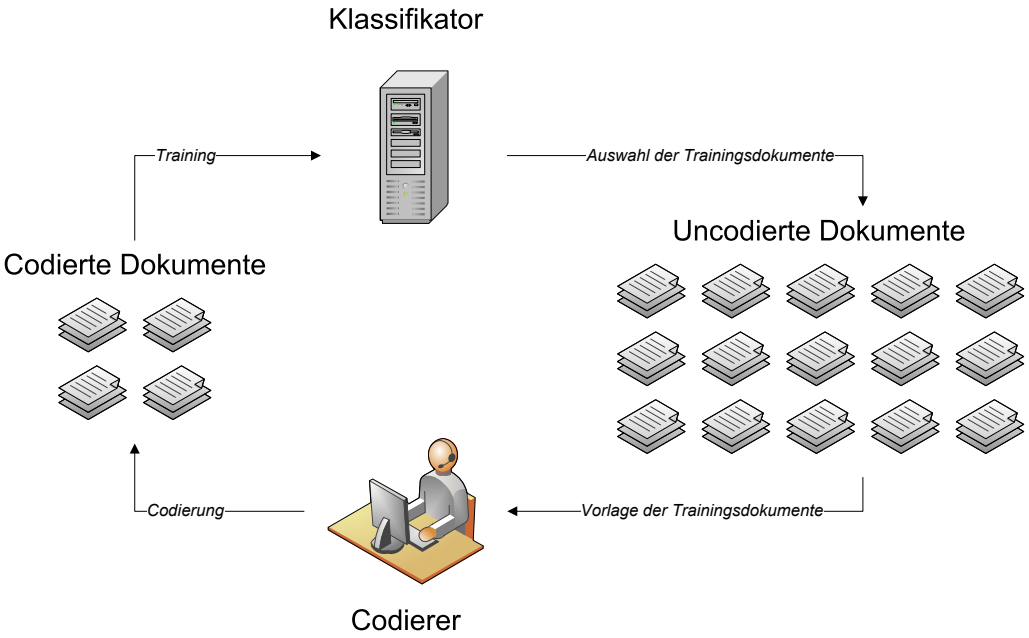

Abbildung 4.2: Ablaufschema für aktives Lernen, Darstellung nach Settles (2010)

konzeptionellen Überlegungen und die zahlreichen empirische Studien bieten Olsson (2009) und Settles (2010). Unabhängig von den verwendeten Algorithmen zeigt sich in den meisten Experimenten, dass bei aktivem Lernen eine festgelegte Reliabilität der Codierung deutlich schneller erreicht wird als bei passivem Training, d.h. zumeist zufälliger Auswahl von Beispieldokumenten (Lewis & Gale, 1994; Settles & Craven, 2008).

Praktikabel sind vor allem zwei Implementierungen aktiven Lernens: Auswahl nach Unsicherheit (*Uncertainty Sampling*) und Auswahl nach Uneinigkeit (*Query-by-Committee*). Die Funktionsweisen beider Strategien sind vergleichsweise einfach und kommen in sehr ähnlicher Weise auch bei klassischen Codiererschulungen zum Einsatz: Für ein *Query-by-Committee* sind mehrere Klassifikatoren notwendig, die alle mit den selben Beispielen trainiert werden. Diesen Klassifikatoren werden Sets von neuen Dokumenten vorgelegt, die daraufhin klassifiziert werden. Als neue Trainingsinstanzen, die dann dem menschlichen Codierer vorge-

legt werden, dienen dann diejenigen Dokumente, bei denen unter den Klassifikatoren die größte Uneinigkeit herrscht.

Aber auch mit nur einem Klassifikator ist aktives Lernen leicht realisierbar: Beim *Uncertainty Sampling* macht man sich die Tatsache zunutze, dass für alle Klassifikationsentscheidungen ein quantifizierbares Maß an Unsicherheit besteht, zum Beispiel bedingte Wahrscheinlichkeiten bei Naive Bayes-Klassifikatoren oder die Nähe zum Entscheidungsvektor bei Support-Vector-Machines (vgl. Abschnitt 3.4.1). Aus einer Menge Dokumente mit unbekannten Kategorien wählt der Algorithmus daher schlicht diejenigen aus, bei denen die Unsicherheit bezüglich der Klassenzugehörigkeit am größten ist. Gibt der menschliche Codierer dann die wahre Kategorie bekannt, wird entweder eine falsche Klassifikation korrigiert oder auch eine richtige, aber unsichere bestärkt (*Reinforcement Training*). Insgesamt hat sich diese Strategie des *Train-on-near-Error* (Assis, 2006) als äußerst leistungsfähig erwiesen und ist theoretisch und mathematisch elaborierteren Verfahren oft ebenbürtig oder sogar überlegen (Settles & Craven, 2008).

Bei allen Vorteilen, die aktives Lernen vor allem im Forschungsalltag bietet, soll abschließend jedoch darauf hingewiesen werden, dass die zentrale Prämisse der Repräsentativität von Trainingsdaten dadurch verletzt wird, dass letztlich nur Dokumente aus dem Rand der Feature-Verteilung zum Training verwendet werden. Das statistische Modell gilt daher streng genommen nur für diese schwierigen Dokumente, und die implizite Annahme, dies schade nicht bei der Klassifikation einfacher Dokumente, ist empirisch zu prüfen. Zumindest theoretisch ist die Gefahr von Overfitting der extremen Trainingsdaten und als Folge Bias in der Klassifikation gegeben. Im empirischen Teil dieser Arbeit wird daher zu prüfen sein, ob (a) aktives Lernen auch bei sozialwissenschaftlichen Kategorienschemata effektiver ist als passives und (b) ob die Trainingsstrategie Auswirkungen auf zufällige und systematische Fehlklassifikationen bei Dokumenten außerhalb des Trainings-Set hat.

4.4 Codierer- und Klassifikatorevaluation

Wie bei jeder Methode wissenschaftlicher Datenerhebung stellen Reliabilität und Validität der Messung die zentralen Gütekriterien der Inhaltsanalyse dar (Früh, 2007; Krippendorff, 2004a). Reliabilität als notwendige Voraussetzung für Validität ist bei einem relativ intransparenten (Wirth, 2001) und potentiell subjektiven Verfahren wie der Codierung von Medieninhalten von großem Interesse. Dementsprechend sind in den verbreiteten Lehrbüchern oft ganze Kapitel der Grundlogik und Anwendung von Reliabilitätstests gewidmet.[8] Da die Reliabilität streng genommen das einzige Qualitätskriterium der Inhaltsanalyse ist, das anhand der Codierungen, d.h. ohne externe Maßstäbe wie Experten oder alternative Messinstrumente, sinnvoll zu interpretieren ist (Krippendorff, 2004b), konzentriert sich die folgende Darstellung auf die wesentlichen Aspekte der Reliabilität.

Automatisierte Verfahren der Inhaltsanalyse sind im einfachsten Sinne vollständig reliabel, was von verschiedenen Autoren immer wieder als wichtiges Argument für deren Einsatz genannt wird (Rössler, 2005; Krippendorff, 2004a): Computer machen keine Flüchtigkeitsfehler, ermüden nicht und generieren unter gleichen Bedingungen gleiche Ergebnisse. Computergestützte Verfahren erreichen daher maximale Stabilität und Reproduzierbarkeit der Analysen, selbst wenn diese inhaltlich fehlerhaft sein sollten. Bezogen auf das eigentliche Ziel der Inhaltsanalyse, der Codierung von verbalen und nonverbalen Symbolen, müssen selbst deterministische Verfahren wie diktionärbasierte Analysen nicht reliabel bzw. genau sein: Allein aufgrund von Rechtschreibfehlern und Homonymen kann die Bezugnahme auf die Person „Gerhard Schröder“ durch eine Wortliste nicht vollständig erfasst werden. Bei komplexen Kategorien ist dieses Problem noch stärker ausgeprägt. Hinzu kommt, dass die in dieser Arbeit im Vordergrund stehenden Klassifikationsverfahren nicht

8 Auffällig ist dabei, dass dem Thema Reliabilität in deutschen Lehrbüchern deutlich weniger Platz eingeräumt wird als in anglo-amerikanischen (vgl. Früh, 2007; Rössler, 2005; Maurer & Reinemann, 2006; Riffe et al., 2005; Krippendorff, 2004a). Dies schlägt sich auch in der Forschungsliteratur wieder, in der auch heute oft höchst sparsam mit Informationen zur Reliabilität der Analyse umgegangen wird (Lauf, 2001; Riffe & Freitag, 1997; Dupagne et al., 2005).

deterministisch sind, d.h. durch Reihenfolge oder Komposition des Trainingsmaterials beeinflussbar sind, *und* auf der manuellen Codierung von Trainingsdaten beruhen, die wiederum selbst von der Reliabilitätsproblematik konventioneller Analysen betroffen ist. Bei der Evaluation überwachter Klassifikationsverfahren gilt es daher, die Reliabilität (und Validität) sowohl der manuellen als auch der maschinellen Codierung zu überprüfen. Dies ist für die Evaluation der auf manueller Vorcodierung aufbauenden Klassifikationsalgorithmen relevant, da davon auszugehen ist, dass unzuverlässige Trainingsdaten auch zu unzuverlässigen Klassifikationsentscheidungen führen. Mit anderen Worten: Die Qualität solcher automatischen Verfahren kann nur so gut sein wie die Codierung des Trainingsmaterials.

4.4.1 Reliabilität manueller Codierung

Die Ermittlung und Interpretation von Reliabilität inhaltsanalytischer Daten ist nicht nur in der Kommunikationswissenschaft (Lauf, 2001; Kolb, 2002; Potter & Levine-Donnerstein, 1999; Lombard et al., 2002; Krippendorff, 2004b; Hayes & Krippendorff, 2007), sondern auch in den verwandten Sozialwissenschaften (Volkens, 2007; Mikhaylov et al., 2008; Benoit et al., 2009b; Hopkins & King, 2010) und der Computerlinguistik (Eugenio & Glass, 2004; Craggs & Wood, 2005; Reidsma & Carletta, 2008; Artstein & Poesio, 2008) ein noch immer stark diskutiertes Thema. Dabei lassen sich einige Punkte erkennen, die mittlerweile als Konsens gelten, etwa dass die Untersuchungseinheiten des Reliabilitätstests einer ausreichend großen Zufallsstichprobe des Codiermaterials entstammen sollen (Lauf, 2001), während andere Fragen, vor allem nach dem angemessenen Reliabilitätsmaß, weiterhin umstritten sind oder in der angewandten Literatur wenig berücksichtigt werden. Im folgenden Abschnitt geht es nur um die für diese Arbeit relevanten Problemfelder inhaltsanalytischer Reliabilitätsmessung: (a) die Wahl des geeigneten Reliabilitätskoeffizienten, (b) die Identifikation von Ursachen mangelnder Reliabilität und (c) der Umgang mit zufälligen und systematischen Messfehlern in den Codierdaten. Nicht berücksichtigt werden hingegen Fragen der Reliabilität bei der Abgrenzung der Untersuchungseinheiten und der Stichprobenziehung

(vgl. dazu Krippendorff, 2004a), da dies nicht spezifisch für automatische und manuelle Verfahren diskutiert werden muss.

Bestimmung der Reliabilität
Definiert man in Anlehnung an die klassische Testtheorie der Psychologie Reliabilität als Anteil des wahren Testwerts τ am gemessenen Wert x, oder alternativ als Messwert x abzüglich des Messfehlers ϵ, stellt sich die Frage, wie man inhaltsanalytische Daten in einen quantitativen Kennwert überführen kann, der die Reliabilität der Messung sinnvoll wiedergibt.

$$Rel(x) = \frac{\tau}{\tau + \epsilon} \tag{4.1}$$

Da der wahre Wert τ eine latente, d.h. nicht messbare, Variable ist, kann die Reliabilität der Messung dieser Variable nur per Inferenzschluss bestimmt werden. Dafür haben sich sowohl in der psychometrischen als auch inhaltsanalytischen Literatur zwei Verfahren durchgesetzt: Die Bestimmung der Test-Retest-(Intracoder)-Reliabilität, bei dem dasselbe Phänomen zu unterschiedlichen Zeitpunkten gemessen wird, sowie der Parallel-(Intercoder)-Reliabilitätstest, bei dem mit unterschiedlichen Instrumenten bzw. Codierern dasselbe Phänomen gemessen wird. Aus der Übereinstimmung der Messungen wird dann auf deren Reliabilität geschlossen.[9] Aus der klassischen Formel für die Reliabilität lässt sich zudem ablesen, dass ein Koeffizient zwei natürliche Endpunkte hat. Er sollte 0 sein, wenn die Varianz in den Daten vollständig auf den Messfehler zurückzuführen ist, und den Wert 1 annehmen sollten, wenn die beobachtete Varianz vollständig fehlerlos gemessen wurde, d.h. der Messfehler $\epsilon = 0$. Das zentrale methodologische Problem des inhaltsanalytischen Reliabilitätstests stellt nun die Transformation der beobachteten Übereinstimmungen von Codierungen in ein solchermaßen interpretierbares Maß dar (Krippendorff, 2004b, 414-415).

9 Krippendorff (2004a) weist zu Recht darauf hin, dass Übereinstimmung nur eine notwendige, aber keine hinreichende Bedingung für Reliabilität ist: Wenn zwei Codierer (oder zwei Messgeräte) fälschlich bei verschiedenen Untersuchungsobjekten immer den gleichen Wert messen, ist damit zwar Übereinstimmung, aber keineswegs die Reliabilität der Messung gewährleistet.

Das noch immer in Lehrbüchern (z.B. Rössler, 2005) empfohlene und am häufigsten verwendete Reliabilitätsmaß ist die prozentuale Übereinstimmung von Codierern (CR), oft auch Holsti (1969) zugeschrieben. Angesichts seiner deutlichen Nachteile ist es allerdings überraschend, dass dieser Koeffizient noch Verwendung findet, da er (a) keinen interpretierbaren Nullpunkt für vollständige Nichtreliabilität hat bzw. dieser von der Anzahl Ausprägungen der Variable abhängt und (b) auch bei offensichtlich fehlender Reliabilität der Messung äußerst hohe Werte annehmen kann. Für den einfachsten Fall einer dichotomen Variablen und zwei Codierern (bzw. zwei Messzeitpunkten bei einem Codierer) gilt die folgende Formel, wobei C_a die Anzahl übereinstimmender Codierungen bezeichnet:

$$CR = \frac{2C_a}{C_1 + C_2} \tag{4.2}$$

Wenn beide Codierer die Inhalte gar nicht lesen, sondern eine Münze werfen, werden sie trotzdem in 50 Prozent der Fälle übereinstimmen, so dass $CR = .5$ den eigentlichen Nullpunkt der Reliabilität darstellt und niedrigere Werte nicht sinnvoll interpretierbar sind. Dieselbe Logik gilt auch bei multinomialen Kategorien. Prozentuale Übereinstimmung lässt sich daher nur interpretieren, wenn die Anzahl der Variablenausprägungen bekannt ist, was aus Fachartikeln häufig nicht klar hervorgeht. Viele Autoren wie Krippendorff (2004b, 413) oder Artstein & Poesio (2008, 559) kommen daher zu der Feststellung, dass Reliabilitätskoeffizienten von verschiedenen Variablen und Studien nur dann vergleichbar sind, wenn zuvor eine Zufallskorrektur stattgefunden hat.

Die prozentuale Übereinstimmung hat noch einen weiteren Nachteil, nämlich die Abhängigkeit von der Verteilung der Testdaten. Dazu ein einfaches Beispiel: Wenn zwei Codierer bei 100 Dokumenten insgesamt 98 Mal den Wert *keine Prominenz* vergeben, und jeder Codierer jeweils einmal *Prominenz* codiert, während sein Mitstreiter dies nicht tut, ergibt sich eine prozentuale Übereinstimmung von .98, die eine exzellente Reliabilität kennzeichnen würde. Faktisch wurde nicht ein einziges Mal übereinstimmend *Prominenz* codiert, so dass nicht klar ist, ob die Codierer die Definition laut Codebuch überhaupt verstanden haben. Da sich die Fragestellung der Analyse zumeist auf den Anteil von Artikeln mit

Prominenz bezieht, ist bezogen auf diese Differenz keine zuverlässige Inferenz möglich.

Entlang dieser Überlegungen wurden in den letzten Jahrzehnten dutzende Reliabilitätskoeffizienten entwickelt, die sich alle mehr oder minder der Zufallskorrektur annehmen (Krippendorff, 2004b). Dabei sind die Koeffizienten π von Scott (1955) und Krippendorffs (1980) α in der kommunikationswissenschaftlichen Literatur am weitesten verbreitet (Lauf, 2001). Krippendorffs α entspricht asymptotisch Scotts π für dichotome Variablen und zwei Codierer, kann darüber hinaus aber auch mit mehreren Codierern, fehlenden Werten und verschiedenen Skalenniveaus umgehen, so dass sich α als globales Reliabilitätsmaß anbietet. Da der Koeffizient durch die Zufallskorrektur oft – und teilweise dramatisch – niedriger ausfällt als das jahrzehntelang verwendete Maß von Holsti, wird Krippendorff häufig vorgeworfen, der Koeffizient α sei zu streng (Lombard et al., 2002).[10]

Auch wenn man sich aus guten Gründen dieser Sichtweise nicht anschließt, bleibt das Problem, dass für viele Studien nur Prozentübereinstimmungen als Reliabilitätsmaß vorliegen und man dieses zum Vergleich der Reliabilität benötigt. Da zwar hohe α-Werte immer mit hohen CR-Werten einhergehen, dies aber andersherum nicht gilt, kann man keine zufallskorrigierten Reliabilitätswerte aus diesen Angaben ableiten. Mit anderen Worten, man weiß nicht, ob die dokumentierten Reliabilitäten verlässlich sind oder auf Zufall und/oder schiefen Testdaten basieren. Aus Gründen der Vergleichbarkeit wird in dieser Arbeit neben korrigierten auch mit einfachen Prozentübereinstimmungen gearbeitet.

Quellen mangelnder Reliabilität

Obwohl die Reliabilität der Codierung zumeist nur mit einem einzelnen Koeffizienten zusammengefasst wird, sind zumeist mehrere Quellen für Messfehler verantwortlich: Neben der globalen Schwierigkeit der Codie-

10 Gerade die relativ starke Abhängigkeit von α bzw. π von der Verteilung der Variablenausprägungen ist für viele Autoren problematisch (Stegmann & Lücking, 2005). Gwet (2001) hat mit dem Koeffizienten AC1 eine Alternative entwickelt, die eher der Intuition vieler Forscher entspricht, dass häufige Übereinstimmungen in einer einzelnen Klasse auch ein Indikator guter Reliabilität seien. Obwohl AC1 eine Zufallskorrektur enthält, ist er meist höher als Krippendorffs α und oft nahe dem „liberalen“ Holsti-Wert.

rung, die mit dem Anspruch und dem Gegenstand der Kategorie variiert (Potter & Levine-Donnerstein, 1999), können auch einzelne Codierer oder einzelne Variablenausprägungen für mangelnde Reliabilität verantwortlich sein (Funkhouser & Parker, 1968). Das Ziel eines Reliabilitätstest muss neben der Quantifizierung des Messfehlers auch die Identifikation möglicher Ursachen einer unzuverlässigen Codierung sein. Dies lässt sich im einfachsten Fall von dichotomen Variablen und zwei Codierern nicht realisieren, wohl aber beim Einsatz mehrerer Codierer und komplexer Variablen.

Die mehrfache Codierung desselben Testmaterials ermöglicht die analytische Trennung von Codebuch- und Codierereinflüssen auf die Reliabilität der Inhaltsanalyse. Dabei sind zwei Möglichkeiten der Berechnung denkbar: Zum einen die paarweise Bestimmung der Coder-Coder-Reliabilität für alle beteiligten Codiererpaare, die relativ aufwändig ist und eine vollständige Mehrfachcodierung desselben Test-Sets erfordert (Kolb, 2002).[11] Der Einsatz von Krippendorffs α bietet jedoch auch für diesen Fall deutlich mehr Möglichkeiten, da durch die Verwendung einer globalen Koinzidenz-Matrix auch Testdaten mit fehlenden Werten oder unvollständiger Überschneidung der Untersuchungseinheiten für die Berechnung der Reliabilität verwendet werden können (Hayes & Krippendorff, 2007). Hierdurch kann der Test auf eine breitere empirische Basis zurückgreifen (Potter & Levine-Donnerstein, 1999). Früh (2007) schlägt vor, durch Inspektion der paarweisen Übereinstimmungsmatrix ggf. abweichende Codierer zu identifizieren, die dann nachgeschult oder aus dem Team entfernt werden können.

Einfacher und flexibler ist hingegen die Berechnung eines globalen Reliabilitätsmaßes für alle Codierer, wobei jeweils ein Codierer aus der Koinzidenzmatrix entfernt wird. Dieses Maß, das ich hier als *Reliability If Coder Omitted* (RICO) bezeichne, hilft nicht nur bei der Identifikati-

11 Von der Verwendung von Maßen, die auf Codierer-Mehrheiten oder gar vollständiger Übereinstimmung aller Codierer basieren, rät Popping (2009) aufgrund fehlender Validität ab. Zudem lässt sich nicht begründen, wie der Inferenzschluss von einer Mehrheitsentscheidung der Codierer auf die spätere Einzelcodierung zu übertragen ist. Einfache oder gewichtete Mehrheitsentscheidungen sind deshalb nur dort angebracht, wo tatsächlich auch unabhängig vom Reliabilitätstest mit Mehrfachcodierung gearbeitet wird (Carpenter, 2008).

on schlechter Codierer, sondern gibt gleichzeitig einen Reliabilitätswert an, der sich unter Beibehaltung aller anderen Codierer zeigen würde.[12] Vergleicht man diese Werte, lässt sich leicht einschätzen, ob ein einzelner Mitarbeiter den Test substanziell verschlechtert (und dies ggf. über mehrere Variablen hinweg). Aus den einzelnen RICO-Koeffizienten lässt sich nicht nur der globale Mittelwert der Reliabilität berechnen, sondern auch dessen codiererbezogene Varianz. Unabhängig von der Höhe des Reliabilitätskoeffizienten zeigt eine hohe Streuung an, dass die Codieranweisungen nicht von allen Codierern gleich gut befolgt wurden bzw. das Training nicht gleich erfolgreich war. Damit lässt sich abschätzen, ob Codebuch oder Codiererschulung verbessert werden sollten.

Auf dieselbe Art und Weise kann bei multinomialen oder ordinalen Variablen ebenfalls analysiert werden, ob und ggf. welche einzelnen Variablenausprägungen bei der Codierung Probleme verursachen. Zur Illustration sei nochmals auf Krippendorff (2004b, 426) verwiesen, der die Reliabilität einzelner Unterscheidungen durch Zusammenfassung mehrerer Kategorien und deren Auswirkung auf die Zuverlässigkeit der Gesamtcodierung misst. Eine hohe Reliabilität bei Zusammenfassung zweier Ausprägungen ist dabei als Indikator für mangelnde Trennschärfe zu interpretieren. So lässt sich feststellen, ob beispielsweise eine niedrige Reliabilität der ordinalen Variable *Prominenz* auf die Entscheidung, ob überhaupt Prominenz codiert wird, oder der Unterscheidung zwischen geringer und hoher Prominenz zurückzuführen ist. Letzteres ließe sich durch eine spätere Dichotomisierung der Daten lösen, ersteres würde ein grundsätzliches Problem im Codebuch signalisieren. Durch die Anwendung der oben vorgestellten Reliabilitätstests kann neben der globalen Zuverlässigkeit der Codierung auch die Frage nach der Fehlerquelle – Codebuch oder Codierer – beantwortet werden.

Zufällige und systematische Messfehler und deren Konsequenzen
Geht man davon aus, dass jegliche manuelle (und automatische) Inhaltsanalyse messfehlerbehaftet, d.h. nicht vollkommen reliabel ist, stellt sich die Frage: Welche Konsequenzen hat fehlende Reliabilität? Angesichts

12 Entsprechend gibt es in der psychometrischen Testung das Maß *Alpha If Item Deleted* für die Konsistenzprüfung von Skalen (Gliem & Gliem, 2003).

der recht ausführlichen Darstellung von Reliabilitätstests in den meisten Lehrbüchern fällt die Diskussion der Frage *Was tun mit den Reliabilitätswerten?* umso knapper aus. Liest man die entsprechende Literatur und auch den größten Teil der Forschungsberichte, lautet die ernüchternde Erkenntnis: Die meisten Autoren berichten, wenn überhaupt, die Ergebnisse des Reliabilitätstests und fahren dann in der Analyse fort, als ob es keinerlei Messfehler gegeben hätte (Hopkins & King, 2010). Im besten Fall wird in der Diskussion der Ergebnisse nochmals auf Reliabilitätsprobleme rekurriert, oft scheint aber der Aufwand für Reliabilitätstests ganz umsonst gewesen zu sein. Um die Frage nach den empirischen Konsequenzen mangelnder Reliabilität beantworten zu können, ist eine Unterscheidung in zufällige (*Noise*) und systematische (*Bias*) Messfehler hilfreich (Funkhouser & Parker, 1968; Krippendorff, 1970, 2009). Nach den Annahmen der klassischen Testtheorie ist der zufällige Messfehler ϵ unabhängig vom wahren Wert τ, d.h. $r_{\tau\epsilon} = 0$. Bezogen auf die Codierung bei der Inhaltsanalyse bedeutet dies, dass ein Codierer rein zufällig manchmal den falschen Code vergibt, und dies unabhängig von der Untersuchungseinheit und der Kategorie. Für die Analyse von Daten mit zufälligem Messfehler hat dies folgende Konsequenzen: Die Punktschätzer der Prozent- und Mittelwerte sind erwartungstreu, d.h. nicht von der mangelnden Reliabilität betroffen. Allerdings sind alle bi- und multivariaten Zusammenhänge von messfehlerbehafteten Variablen verzerrt, so dass dort mit entsprechend korrigierenden Analyseverfahren gearbeitet werden sollte, wie es bei Befragungsdaten seit langem der Fall ist (Fuller, 1987; Weber, 1983; Hopkins & King, 2010).

Systematische Messfehler entstehen, wenn die Wahrscheinlichkeit einer Fehlklassifikation von der wahren Ausprägung einer Variable abhängt. Dies ist immer dann der Fall, wenn Codieranweisungen unklar sind oder Codierer diese falsch verstehen. Oft werden beispielsweise die häufigsten Ausprägungen einer Kategorie relativ zuverlässig codiert, während abweichende Ausprägungen nicht nur seltener, sondern auch weniger reliabel gemessen werden. Das Hauptproblem bei systematisch messfehlerbehafteten Daten liegt in der Tatsache, dass auch die deskriptiven Mittel- und Anteilswerte sowie Intervallschätzungen verzerrt sind (Schwartz, 1985). Bi- und multivariate Zusammenhänge werden ebenfalls durch

nichtzufällige Fehlklassifikationen verzerrt und müssen entsprechend korrigiert werden (Küchenhoff et al., 2005; Hopkins & King, 2010).

Im Falle dichotomer Variablen, die einen Großteil inhaltsanalytischer Untersuchungen ausmachen, liegt so gut wie immer ein systematischer Messfehler vor, da nur unter unrealistischen Verteilungsannahmen keine verzerrte Schätzung des Prozentanteils in Stichprobe und Grundgesamtheit auftritt. Wie Schwartz (1985) eindrucksvoll zeigt, sind schon bei einer hohen Reliabilität von .9 und – in der Inhaltsanalyse recht häufig – bei schiefen Daten (z.B. einer Auftretenshäufigkeit $p = .1$) die Punktschätzungen um 75 Prozent nach oben verzerrt. Generell gilt, dass alle messfehlerbehafteten Schätzungen in Richtung $p = .5$ verzerrt sind und dieser Bias mit abnehmender Reliabilität stärker wird (vgl. Abbildung 4.3). Dadurch wird nicht nur die Varianz in der Variable verringert, sondern auch Zusammenhangs- und Unterschiedshypothesen sind nicht mehr ohne weiteres überprüfbar.

Das Problem der Messfehlerkorrektur kann bei inhaltsanalytischen Studien vergleichsweise leicht gelöst werden, da aus den Daten des Reliabilitätstests nicht nur numerische Kennwerte für systematische und zufällige Messfehler berechnet werden können, wie dies Krippendorff (1970, 2009) für intervallskalierte und kategorielle Variablen vorschlägt. Zusätzlich liegen auch empirische Daten zur konkreten Fehlklassifikation in Form von Klassifikationsmatrizen vor. Diese können dafür eingesetzt werden, die in der Normalcodierung erhobenen Daten zu korrigieren. Dabei kann nicht nur eine globale Fehlklassifikationsmatrix bzw. Krippendorffs Koinzidenzmatrix eingesetzt werden, sondern ggf. auch codiererspezifische Klassifikationstabellen, um wahrscheinliche Fehlklassifikationen einzelner Personen zu korrigieren. Wenn man beispielsweise bei einem Forscher-Codierer-Reliabilitätstest beobachtet, dass ein Codierer zwei Variablenausprägungen verwechselt oder bestimmte Codes zu selten vergibt, kann man diese Fehler anhand der Übereinstimmmungsmatrix bei der Analyse berücksichtigen. Diese Korrektur verbessert zwar nicht die Klassifikation einzelner Fälle, wohl aber die uni- und multivariaten Analysen.

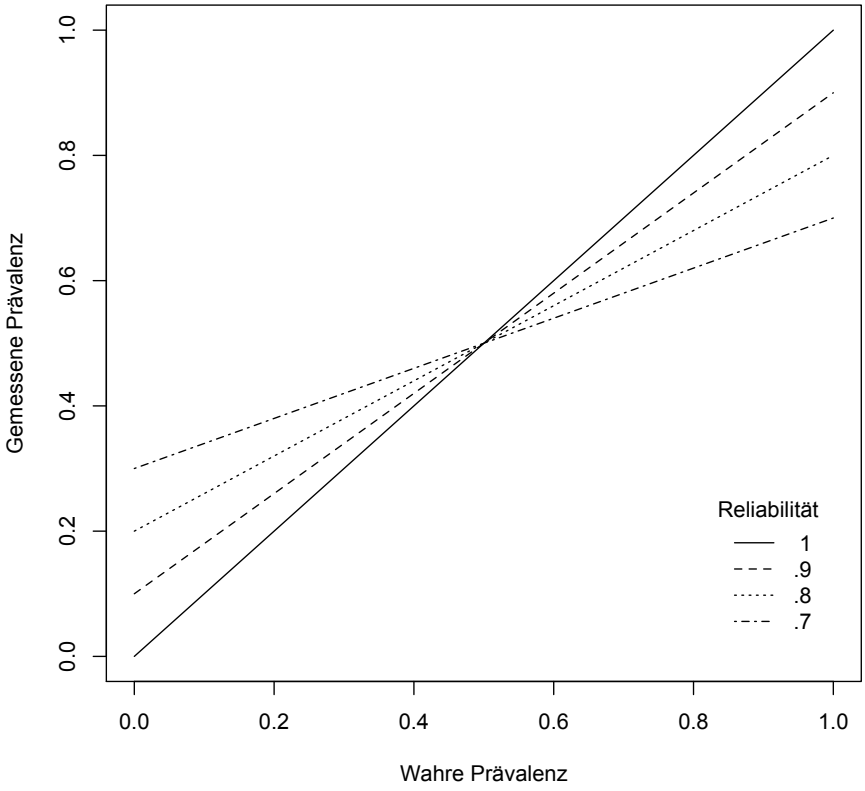

Abbildung 4.3: Zusammenhang von Reliabilität, gemessener und wahrer Verteilung bei dichotomen Variablen

4.4.2 Reliabilität und Validität automatischer Klassifikation

Die Evaluation von verschiedenen Klassifikationsalgorithmen ist für das Forschungsfeld des maschinellen Lernens von so zentraler Bedeutung, dass dem Thema nicht nur in den Standardwerken (Manning & Schütze, 1999; Witten & Frank, 2005; Manning et al., 2008; Alpaydin, 2008) viel Raum gewidmet wird, sondern die statistischen Eigenschaften verschiedener Evaluationsverfahren vielfach sehr kritisch analysiert werden

Tabelle 4.4: Konfusionsmatrix einer Klassifikationsevaluation

		Klassifikator	
		1	0
Goldstandard	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

(Wallach, 2004; Sokolova et al., 2006; Powers, 2007). Dies ist nicht zuletzt darauf zurückzuführen, dass in der Informatik häufig wissenschaftliche Wettbewerbe für bestimmte Klassifikationsaufgaben ausgeschrieben werden, deren Kriterien dann zum de facto Standard in der Forschung werden (Cormack & Lynam, 2007). Obwohl die Berechnung verschiedener Qualitätsindikatoren bei der Klassifikatorevaluation eng mit der Intercoder-Reliabilitätsbestimmung verwandt ist, sind doch einige zentrale Unterschiede hinsichtlich des Erkenntnisinteresses der Tests zu erkennen: Erstens wird eher die Validität der Codierung gemessen, zweitens werden häufig nicht alle Fehlklassifikationen als gleich relevant eingestuft.

Besonderheiten der Klassifikatorevaluation
Grundsätzlich zielen die meisten der im Folgenden vorgestellten Maße nicht streng auf die Reliabilität der Codierung ab, die bei automatischen Verfahren ohnehin perfekt ist, jedenfalls unter identischen Rahmenbedingungen. Stattdessen wird das gemessen, was in der inhaltsanalytischen Forschung, etwa bei Früh (2007), Expertenvalidität genannt wird, d.h. die Codierung des Computers wird mit einem vorgegebenen Goldstandard verglichen (Wiebe et al., 1999). Daher können die Zellen einer Konfusionsmatrix, wie in Tabelle 4.4 dargestellt, als falsch oder richtig codierte Dokumente bezeichnet und interpretiert werden. Diese Nomenklatur wird auch in anderen Forschungsfeldern, etwa der medizinischen Diagnostik, verwendet, wobei auch mehr als zwei Variablenausprägungen bzw. Klassen möglich sind. Wie bei Intercoder-Reliabilitätstests stellt die empirische Konfusionsmatrix die Grundlage für die Berechnung aller Einzelkoeffizienten dar. Das einfachste Maß für die Evaluation der Klas-

sifikationsqualität ist der Anteil an korrekt klassifizierten Dokumenten (*Accuracy*), also $TP + TN$, an der Gesamtzahl der codierten Dokumente. Dies entspricht der einfachen prozentualen Übereinstimmung, also dem Holsti-Koeffizient bei einem Reliabiltitätstest, und kann daher auch als Forscher-Klassifikator-Reliabilität bezeichnet werden. Da bei der Berechnung nicht zwischen falsch positiv und falsch negativ codierten Dokumenten unterschieden wird, kann anhand der Accuracy nur die globale Güte der Klassifikation abgeschätzt werden.

In der Forschungs- und Anwendungspraxis werden jedoch häufig die beiden möglichen Fehlklassifikationen nicht gleich wichtig eingeschätzt: Bei Suchmaschinen-Anfragen möchte etwa der Nutzer vor allem wenige falsch positive Treffer haben, da ohnehin nur ein Teil der positiv klassifizierten Treffer überhaupt angesehen wird. Da verschiedene Fehlklassifikationen oft mit unterschiedlichen Kosten verbunden sind, wurden für diesen Zweck besondere Qualitätsmaße entwickelt: *Precision* und *Recall* (vgl. Tabelle 4.5). Hat ein Klassifikator eine hohe Präzision, kann man sich darauf verlassen, dass die positiv codierten Dokumente, z.B. ob es sich um eine Sportmeldung handelt, tatsächlich auch richtig codiert sind. Dabei bleibt aber unklar, wie viele Sportmeldungen überhaupt als solche codiert wurden. Letzteres misst der Kennwert *Recall* (auch Sensitivität oder Trefferquote genannt), der den Anteil als Sportmeldung codierter Dokumente an allen Sportmeldungen wiedergibt.[13] Nutzt man z.B. eine automatische Klassifikation lediglich für eine grobe Ziehung von Untersuchungseinheiten, die dann manuell codiert werden, ist man tendenziell eher an höherem *Recall* interessiert, da irrtümlich in die Stichprobe genommene Artikel schnell bei der manuellen Analyse entfernt werden können, während nicht automatisch gefundene Beiträge gar nicht erst in die Analyse eingehen.

Will man Precision und Recall gleichrangig behandeln, bietet es sich an, deren Mittelwert als globales Gütemaß der Klassifikation zu verwenden. Mit dem *F-Measure* wird das harmonische Mittel aus beiden Werten

13 An dieser Stelle sei noch einmal auf das zentrale Problem von Freitextrecherchen erinnert (vgl. Abschnitt 3.3.2): Während sich die Präzision anhand einer Analyse der gefundenen Dokumente bestimmen lässt, ist dies für den *Recall* nicht möglich, da die Zahl nicht berücksichtigter Dokumente (FN) nicht bekannt ist.

Tabelle 4.5: Maße für die Klassifikatorevaluation

Kennwert	Formel	Bedeutung
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Wahrscheinlichkeit einer korrekten Codierung
Precision	$\frac{TP}{TP+FP}$	Wahrscheinlichkeit, mit der ein codiertes Dokument relevant ist
Recall	$\frac{TP}{TP+FN}$	Wahrscheinlichkeit, mit der ein relevantes Dokument codiert wird
F-Measure	$\frac{2 \cdot Prec \cdot Rec}{Prec+Rec}$	harmonisches Mittel aus Precision und Recall

Hinweis: Alle dargestellten Koeffizienten können Werte von $0-1$ annehmen, wobei höhere Werte höhere Validität bedeuten.

berechnet. Da es in der Praxis häufig einen *Trade-off* zwischen Precision und Recall gibt, und man dies in den meisten Fällen mit veränderten Parametern des Algorithmus manipulieren kann, wird als Gütemaß für einen Klassifikationsalgorithmus oft eine ROC-Kurve (*Receiver-Operator-Characteristic*) verwendet, in die für jeden Precision-Wert der damit einhergehende Recall-Wert abgetragen wird (Sokolova et al., 2006). Ein Klassifikator ist dann besonders gut, wenn gleichzeitig hohe Werte für beide Koeffizienten erreicht werden.

Zahlreiche Autoren weisen darauf hin, das hohe Werte einzelner Koeffizienten in Tabelle 4.5 nicht automatisch eine hohe Qualität oder Validität der Klassifikation bedeuten (Wallach, 2004): Ein Klassifikator, der stets alle Dokumente mit 1 codiert, hat automatisch einen Recall-Wert von 1. Ein Klassifikator, der bei einer Variable mit seltenem Auftreten stets 0 codiert, wird trotzdem einen hohen Wert für Accuracy haben, selbst ein Zufallsgenerator kann leicht einen Wert von 0.5 erreichen. Dieses Problem der Zufallskorrektur ist allen vorgestellten Maßen gemeinsam und wird in der Literatur nur selten thematisiert. Es führt dazu, dass die in der Forschung häufig genutzten Maße *Accuracy* und *F-Measure* oft ebenso (über-)optimistisch sind wie Prozentübereinstimmungen in inhaltsanalytischen Reliabilitätstests.

Evaluationsprozess

Um die Qualitität automatischer Klassifikationen zu beurteilen, hat sich in der Forschung ein standardisierter Evaluationsprozess etabliert, der es ermöglicht, sowohl einzelne Klassifikationsmodelle als auch verschiedene Algorithmen miteinander zu vergleichen (Sebastiani, 2002; Manning et al., 2008). Als Stichprobe für die Evaluation dient dabei stets ein Set an bereits manuell codierten Dokumenten T. Um die Klassifikation zu evaluieren, wird der Gesamtkorpus T in ein Trainings-Set T_{Tr} und ein Test-Set T_{Te} aufgeteilt. Mit den Dokumenten aus dem Trainings-Set werden die Feature-Gewichte des Klassifikators geschätzt, d.h. der Klassifikator lernt. Nach dieser Trainingsphase werden die Dokumente aus dem Test-Set vom Klassifikator codiert und diese Codierung mit der manuellen Annotation, d.h. dem Goldstandard, verglichen.[14] Aus der Übereinstimmung werden die oben beschriebenen Reliabilitäts- bzw. Validitätskoeffizienten berechnet. Dieser Training-Test-Evaluations-Prozess wird für alle zu testenden Klassifikationen wiederholt.

Da die Evaluationsergebnisse erheblich von der Komposition des Trainings- und Testmaterials abhängen, ist es notwendig, die Variabilität der Testergebnisse hinsichtlich dieser Einflussgrößen zu quantifizieren (Rodriguez et al., 2010). Dies gelingt durch wiederholtes zufälliges Ziehen von Trainings- und Test-Set, ggf. geschichtet nach der Verteilung der verschiedenen Kategorien. Ein leicht umzusetzendes, aber unsystematisches Verfahren wäre eine wiederholte einfache Ziehung des Sets, die jedoch mit geringer Präzision und ggf. Bias einhergeht (Blum et al., 1999). Eine Alternative, die Hayes & Krippendorff (2007) auch für die Bestimmung von Konfidenzintervallen bei Intercoder-Reliabilitätskoeffizienten vorschlagen, sind *Resampling*- bzw. konkret *Bootstrapping*-Verfahren, bei denen wiederholt mit Zurücklegen Daten aus dem Trainings-Set gezogen und am gleichen Test-Set evaluiert werden (Molinaro et al., 2005; Borra & Di Ciaccio, 2010). Dort bleibt allerdings das Problem bestehen, dass

14 Sebastiani (2002) weist ausdrücklich darauf hin, dass ein entscheidendes Qualitätskriterium jeder Klassifikatorevaluation darin besteht, dass keine Dokumente aus dem Trainings-Set beim Test verwendet werden und umgekehrt. Nur so ist gewährleistet, dass sich aus dem Ergebnis des Tests auch Inferenzschlüsse auf die Klassifikationsleistung bei unbekannten Dokumenten ziehen lassen.

zusätzlich das initiale Ziehen der Trainings- und Teststichprobe variiert werden muss, um auch dessen Auswirkungen beurteilen zu können.

Um höchstmögliche Präzision bei der Schätzung und vertretbare Replikationskosten zu ermöglichen, hat sich in der Literatur das Verfahren der k-fachen-Kreuzvalidierung (*k-fold cross validation*) durchgesetzt. Hierbei wird das Gesamtset an Dokumenten T mit n Elementen zufällig in k Partitionen (*folds*) der Größe n/k aufgeteilt. Eine einzelne Partition wird dann als Test-Set zurückgelegt, mit den Dokumenten aus den anderen $k-1$ Partitionen wird der Klassifikator trainiert.[15] Dieser Vorgang wird für jede Partition wiederholt und die Ergebnisse gemittelt. Das Verfahren ist deshalb besonders effizient, weil jedes Dokument genau einmal als Testfall verwendet wird, während die Trainingsdaten relativ heterogen sind. Simulationsstudien haben gezeigt, dass die dadurch gewonnenen Schätzer präzise und wenig verzerrt sind (Bengio & Grandvalet, 2004). Um den Einfluss der Partitionierung abschätzen zu können, sollten k-fache-Kreuzvalidierungen wiederholt werden (Borra & Di Ciaccio, 2010).

Zusammenfassend lässt sich festhalten, dass es für eine gleichermaßen aussagekräftige und anschlussfähige Evaluation von sozialwissenschaftlich relevanten Textklassifikationen notwendig ist, eine Vielzahl von Koeffizienten zu berichten, da von vergleichbaren Studien oft nur einzelne ausgewählte Maße vorliegen (Durant & Smith, 2007; Atteveldt et al., 2008; Hillard et al., 2007). Zudem scheint es geboten, nicht nur einzelne Punktschätzer der Qualitätsmaße zu berechnen, sondern auch die damit verbundene Variabilität sowie deren Ursachen, die in der Testprozedur sowie der Auswahl von Trainings- und Testmaterial liegen können. Aufgrund der internen Logik des Verfahrens kann die Reliabilität und Validität der automatischen Klassifikation nie besser sein als die der Trainingsdaten. In der empirischen Evaluationsstudie ist aber zu klären, ob zwischen der Qualität manueller und automatischer Codierung ein Zusammenhang besteht (Sheng et al., 2008).

15 Häufig verwendet werden 5 oder 10 Partitionen (*10-fold cross validation*, 10cv). Ein Spezialfall ist das sog. *Leave-One-Out*-Verfahren, bei dem jeweils nur ein einziges Dokument im Test-Set ist, d.h. $k = n - 1$ (Molinaro et al., 2005).

5 Überwachte Textklassifikation – eine Evaluationsstudie

In den folgenden Kapiteln wird eine empirische Evaluationsstudie zur Anwendbarkeit von Verfahren des maschinellen Lernens für die Inhaltsanalyse entworfen und deren Durchführung sowie Ergebnisse dokumentiert. Dabei werde ich aus den in Kapitel 4 skizzierten methodischen Herausforderungen zunächst allgemeine Forschungsfragen formulieren, die anschließend in Form von Hypothesen zur Klassifikationsqualität und zur Effektivität des maschinellen Lernens konkretisiert werden. In Kapitel 6 werde ich die Methode der Evaluationsstudie, die in Form eines faktoriellen Experiments mit realen Inhaltsanalysedaten durchgeführt wird, ausführlicher erläutern. Dabei stehen vor allem die Stichprobe des Untersuchungsmaterials, die Zusammensetzung des Codebuchs sowie die Untersuchungsanlage im Vordergrund. Da die Evaluation aus zwei Teilstudien besteht, werden deren Ergebnisse in Kapitel 7 jeweils in eigenen Abschnitten dargestellt.

5.1 Allgemeine Forschungsfragen

Im folgenden empirischen Teil dieser Arbeit geht es um die Evaluation des Verfahrens überwachter Textklassifikation für die sozialwissenschaftliche Inhaltsanalyse. Konkret möchte ich folgende drei allgemeine Forschungsfragen bearbeiten:

1. Wie zuverlässig und valide ist die automatische Codierung von Nachrichten durch einen manuell trainierten Klassifikationsalgorithmus?
2. Welchen Einfluss haben (a) das Codebuch bzw. dessen Variablen, (b) Umfang und Beschaffenheit des Textmaterials und (c) verschiedene Verfahren des Preprocessing auf die Qualität der Klassifikation?

3. Welchen Effekt hat die Trainingsstrategie auf die Effektivität des Lernprozesses und damit die Effizienz des Verfahrens?

Die erste Frage bezieht sich ganz allgemein auf die Eignung des Verfahrens überwachter Textklassifikation für die Analyse von deutschsprachigen Nachrichtentexten. Die Evaluation ergänzt dadurch bisherige empirische Forschungsergebnisse, die zumeist nur (a) für englischsprachige Texte vorliegen (Joachims, 2002), (b) auf ungewöhnlichem Stimulusmaterial basieren, z.B. Produktreviews (Pang & Lee, 2008) oder Gesetzestexte (Hillard et al., 2008), und vor allem (c) keine klassischen kommunikationswissenschaftlichen Kategoriensysteme verwenden.

Es stellt sich daher die Frage, ob die in den genannten Studien erzielten hohen Reliabilitäts- und Validitätswerte auch in diesem Kontext zu erzielen wären oder nur durch die Wahl optimalen Materials und einfacher Kategoriensysteme erklärbar sind. Durch eine systematische Variation solcher Einflüsse wird in dieser Arbeit eine realistische Einschätzung der Eignung überwachter Klassifikation für die sozialwissenschaftliche Inhaltsanalyse möglich gemacht.

Die Evaluationsstudie versucht außerdem zu erklären, welchen Einfluss die Komposition des Textmaterials und die in der Forschung zur automatischen Inhaltsanalysen üblicherweise empfohlene Vorbehandlung der Texte auf die Klassifikationsqualität hat. Zu diesem Zweck wird im Folgenden die Evaluation mit Online-Nachrichten durchgeführt, die vollautomatisch erhoben und bereinigt und anschließend manuell codiert werden. Sowohl der Datenerhebungs- als auch der Trainingsprozess enthalten verschiedene Fehlerquellen, die sich im Forschungsalltag nicht vollständig kontrollieren lassen. Es ist daher zu klären, welche Schritte im Forschungsprozess möglicherweise nachteilige Konsequenzen für die Klassifikation haben, und ob sich aufwändige Bereinigungsverfahren überhaupt lohnen.

Die dritte Forschungsfrage bezieht sich auf die Effektivität des Trainingsprozesses, die vor allem von großem forschungsökonomischen Interesse ist. Aufbauend auf den Studien von Hillard et al. (2007) wird überprüft, ob die Selektion von Trainingsdokumenten durch den Computer selbst, d.h. aktives Lernen, zu schnelleren Lernerfolgen führt als eine

passive Trainingsstrategie. Wäre dies der Fall, ließe sich durch gezielte manuelle Codierung relevanter Dokumente viel Aufwand, Zeit und damit Kosten bei der Analyse sparen.

Da es sich bei der ersten Forschungsfrage eher um eine übergeordnete Fragestellung handelt, werden in den folgenden Abschnitten vor allem die Forschungsfragen 2 und 3 konkretisiert, die auch in Abbildung 5.1 als Kausalmodelle dargestellt sind. Da für beide Fragestellungen ein unterschiedliches Forschungsdesign angebracht ist, werden diese in zwei separaten Teilstudien überprüft, die ich unter den Überschriften der Klassifikationsqualität und Lerneffektivität vorstellen werde. Für beide Teilstudien werden jedoch identische inhaltsanalytischen Daten, d.h. Untersuchungseinheiten und Kategorien, verwendet.

Die inhaltsanalytischen Kategorien nehmen bei der Evaluation eine Sonderrolle ein, da sie sowohl einen direkten Einfluss auf die Qualität und Effektivität der Klassifikation haben – manche Variablen lassen sich besser automatisieren als andere – als auch die Effekte der anderen Einflussgrößen moderieren können. Es ist plausibel anzunehmen, dass etwa aktives Lernen nicht bei allen Kategorien gleich effektiv ist, sondern sich vor allem bei Kategorien mit vergleichsweise schiefer Verteilung lohnen wird, da hier die seltenen Positiv-Beispiele vorrangig zum Training genutzt werden können.

5.2 Hypothesen zur Klassifikationsqualität (Teilstudie 1)

Folgt man den bisherigen empirischen Ergebnissen zur automatischen Textanalyse, hängt der Erfolg eines Verfahrens und damit die Qualität der Codierung vor allem von drei Einflussgrößen ab: den verwendeten inhaltlichen Kategorien, der Aufbereitung der Textdaten sowie deren Inhalte. Bezüglich der Kategorien lassen sich aus der Logik des Klassifikationsverfahrens mehrere Hypothesen herleiten. Zum einen kann jedes automatische Verfahren nur auf der lexikalischen und syntaktischen Textebene funktionieren, da Semantik und Pragmatik ein Textverständnis erfordern, dass sich bislang nicht erfolgreich automatisieren lässt. Deshalb ist zu vermuten, dass die Reliabilität der automatischen Klassifikation von der Lexikalität der Kategorie abhängt:

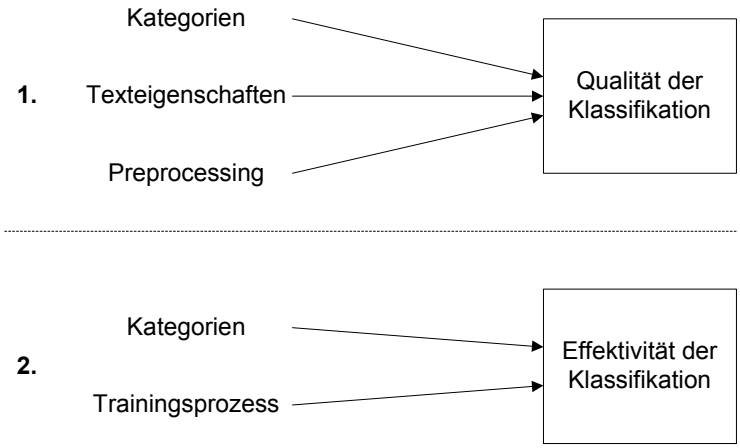

Abbildung 5.1: Kausalmodelle für die Evaluationsstudie

> **H1** Je eher die für die Klassifikation relevanten Informationen auf lexikalischer Ebene vorliegen, desto erfolgreicher wird die automatisierte Codierung sein. Umgekehrt formuliert: Je mehr implizites oder explizites Kontextwissen für die Codierung erforderlich ist, desto weniger reliabel wird die Klassifikation sein.

Aus dieser Hypothese ergibt sich die Erwartung, dass thematische Codierungen, etwa ob es sich um eine Politikmeldung handelt, deutlich zuverlässiger automatisierbar sein sollten als ein beispielsweise ein Nachrichtenfaktor wie Prominenz oder Überraschung. Da die wenigsten Meldungen explizit „Überraschung für alle: Bundespräsident tritt zurück“ oder „Prominenter Schlagersänger begeht Fahrerflucht“ lauten, sondern die Codierer meist selbst folgern, ob etwas überraschend oder jemand prominent ist, sollte dies für einen Algorithmus ohne Kontextwissen schwieriger nachzuvollziehen sein, als die Tatsache, dass Meldungen aus dem Bundestag meist politischer Natur sind.

Die zweite Hypothese zum Kategorienschema ist der Tatsache geschuldet, dass für eine erfolgreiche automatische Klassifikation entsprechend gutes Trainingsmaterial vorliegen muss:

> **H2** Je zuverlässiger das Trainingsmaterial manuell codiert wurde, desto höher ist die Wahrscheinlichkeit, dass auch die automatische Codierung reliabel ist.

Da auch die manuelle Codierung bei Variablen, die viel Kontextwissen erfordern, häufig weniger zuverlässig ist als bei einfachen formalen oder thematischen Kategorien, sind die beiden Hypothesen nur schwer getrennt zu betrachten. Während eine reliable automatische Codierung trotz unzuverlässiger Trainingsdaten prinzipbedingt äußerst unwahrscheinlich ist, kann der umgekehrte Fall jedoch als Beleg für die Lexikalitätshypothese gewertet werden.

Bezüglich des Einflusses verschiedener Preprocessing-Verfahren ist man sich in der Computerlinguistik bislang uneins, ob und welche Preprocessing-Schritte tatsächlich zu einer Verbesserung der Klassifikation führen. Während in vielen Einführungstexten meist pauschal zu einer umfangreichen Vorbehandlung geraten wird (Hotho et al., 2005; Alexa & Zuell, 2000; Popping, 2000), können die wenigen empirischen Untersuchungen in diesem Fall nur selten substanzielle Verbesserungen belegen (Felden et al., 2005; Leopold & Kindermann, 2002; Braschler & Ripplinger, 2004). Trotzdem halte ich an dieser Stelle einen positiven Einfluss für plausibler als einen negativen.

> **H3** Preprocessing-Maßnahmen verbessern die Klassifikationsqualität, da sie statistisches Rauschen aus den Textdaten entfernen.

Konkret werden in dieser Evaluationsstudie drei häufig empfohlene Preprocessing-Verfahren getestet, die in Abschnitt 4.2 beschrieben sind: (a) Die Extraktion von Klartext aus dem Original-HTML der gespeicherten Inhalte (*Text-Extraction*), (b) die Entfernung von extrem häufigen Stopwörtern, die gleichermaßen in allen deutschsprachigen Texten vorkommen und daher keine Trennschärfe besitzen (*Stopword-Removal*), (c) die Verwendung von Wortstämmen an Stelle von vollen Wortformen (*Stemming*), sodass möglichst alle Varianten eines Wortes als ein Feature für die Klassifikation eingesetzt werden. Da sowohl die Stopwortentfernung als auch das Stemming sprachabhängig sind, wäre es

forschungsökonomisch wünschenswert, ohne diese Verfahren erfolgreich Dokumente klassifizieren zu können, weil dann keine individuelle Anpassung der Klassifikationssoftware an das Stimulusmaterial erforderlich wäre.

Durch die experimentelle Manipulation des Stimulusmaterials und eine anschließende Evaluation der automatischen Klassifikation lässt sich also zusätzlich zur Frage, wie stark eine Kategorie auf lexikalischer Ebene codiert wird, auch klären, in welchen Textabschnitten die relevanten Informationen für die Inhaltsanalyse enthalten sind. Gerade bei umfangreichen Untersuchungsanlagen wäre eine Aufteilung des Materials auch mit logistischen und forschungsökonomischen Vorteilen verbunden. Da eine weitergehende Untersuchung dieser Fragestellung jedoch die Komplexität der Studie deutlich erhöhen würde, werde ich im Folgenden nur zwei Faktoren aus dem Bereich der Texteigenschaften untersuchen: die Verwendung von extrahierten Texten statt Rohdaten (HTML, s.o.) und die Berücksichtigung der Überschrift.

Folgt man der journalistischen Maxime, dass die wichtigsten Informationen jeder Meldung am Anfang stehen sollten, müssten die Überschrift und der Teaser eines Beitrags ein höheres Gewicht bei der Klassifikation haben als der restliche Text. Dies hätte zur Folge, dass die Klassifikation deutlich schlechter ausfallen müsste, wenn die Überschrift fehlt. Zumindest bezogen auf die Themenvariable lässt sich also folgende Hypothese formulieren:

> **H4** Die Klassifikationsqualität sollte bei Nachrichten mit Titel bzw. Überschrift höher sein als bei Nachrichten ohne Überschrift.

Da bei einigen Web-Angeboten die Überschriften von Artikeln grafisch dargestellt werden und ggf. aufwändig extrahiert werden müssen, stellt sich die Frage, ob dieser Schritt oder eine andere Sonderbehandlung von Titeln und Überschriften die Klassifikation überhaupt signifikant beeinflusst. Da das hier verwendete Material fast immer Titel bzw. Überschrift enthält, müssen diese für die Evaluation entsprechend entfernt werden. Anschließend wäre ein negativer Effekt auf die Klassifikation zu erwarten.

Insgesamt werde ich im Folgenden die vier genannten Hypothesen überprüfen, die sich auf den Einfluss von Textkomposition und Preprocessing-Verfahren auf die Klassifikationsqualität beziehen. Hierfür bietet sich ein faktorielles Experimentaldesign an, das im nächsten Kapitel ausführlicher darstellt wird.

5.3 Hypothesen zur Lerneffektivität (Teilstudie 2)

Die forschungsleitende Frage der zweiten Teilstudie bezieht sich auf die Effektivität des Lernprozesses durch inkrementelles Lernen. Um die Klassifikationsverfahren optimal zu nutzen und ggf. auch frühzeitig zu erkennen, wann sich ihr Einsatz (nicht) lohnt, bietet es sich an, den Klassifikator schrittweise zu trainieren und zu testen. Da die manuelle Codierung von Trainingsdokumenten den größten Aufwand in der Untersuchung verursacht, ist es aus Effizienzgründen geboten, die menschlichen Codierer möglichst sinnvoll einzusetzen. Dies gilt besonders unter Berücksichtigung der Annahme, dass sich ohnehin nur ein kleiner Teil aller Kategorien automatisch codieren lässt. Aus dieser Überlegung resultiert die Strategie, pro inhaltlicher Variable nur so viele Trainingsdokumente vorzubereiten, wie für eine ausreichend zuverlässige und gültige automatische Codierung notwendig sind. Ist die Klassifikationsqualität für die Studienziele ausreichend, können die wertvollen Codiererresourcen für schwerer oder gar nicht automatisierbare Kategorien eingesetzt werden.

Insgesamt steht angesichts der bisherigen empirischen Ergebnisse zum Trainingsprozess in der Textklassifikation zu erwarten, dass die Algorithmen sehr schnell lernen (Forman & Cohen, 2004). So konnten Dumais et al. (1998) zeigen, dass bei manchen Kategorien schon mit 20 Trainingsdokumenten hohe Reliabilitätswerte erreicht werden, die sich auch mit deutlich mehr Trainingsmaterial nicht mehr signifikant verbessern lassen. Allerdings setzt dies voraus, dass für alle relevanten Variablenausprägungen ausreichend Trainingsdaten vorliegen, d.h. die Variable möglichst ideal verteilt ist. Zweitens ist anzunehmen, dass „leichte“ Kategorien mit wenigen trennscharfen Features schneller gelernt werden können als komplexere und vielfältigere Kategorien. Analog zur oben formulierten Hypothese lässt sich also Folgendes erwarten:

> **H5** Je mehr der für die Klassifikation relevanten Informationen auf lexikalischer Ebene vorliegen, desto schneller wird der Klassifikationsalgorithmus lernen und seine maximale Leistung entfalten.

Neben der Beschaffenheit des Codebuchs spielt für die Frage nach der Trainingseffektivität die angewandte Lernstrategie eine entscheidende Rolle. Insbesondere die Verwendung von aktivem Lernen (vgl. Abschnitt 4.3) sollte die Effektivität des Trainingsprozesses erheblich steigern, wie etwa Hillard et al. (2007) zeigen können.

> **H6** Wenn der Klassifikationsalgorithmus über die Auswahl des Trainingsmaterials bestimmt (aktives Lernen), erreicht die Klassifikation eher ihre maximale Leistung.

Zusätzlich ist zu vermuten, dass aktives Lernen vor allem bei Variablen sinnvoll ist, die vergleichsweise schief verteilt sind, da sonst häufig zu wenig Trainingsmaterial für seltene Ausprägungen anfällt. In diesem Fall würde die jeweilige Kategorie einen moderierenden Einfluss auf den Effekt der Trainingsstrategie haben.

Da bislang keine empirischen Ergebnisse zur Lerneffektivität bei typischen Themenkategorien vorliegen, sind neben den o.g. Wirkungshypothesen auch die deskriptiven Ergebnisse dieser Teilstudie von Interesse: Wie viel Trainingsmaterial wird für eine zuverlässige überwachte Klassifikation benötigt? Benötigt man dutzende, hunderte oder noch mehr manuell codierte Dokumente für eine thematische Kategorie wie Sport oder Politik? Ebenso ist die Form des Lernprozesses für die generelle Verwendung überwachter Klassifikation von Bedeutung: Lernt der Algorithmus linear oder gibt es Deckeneffekte, so dass das Training sich immer weniger lohnt? Diese Fragen gilt es in der nachfolgenden Evaluation in Abschnitt 7.2 zu beantworten.

6 Methode

6.1 Beschreibung der Stichprobe

Da es in dieser Arbeit sowohl um die Umsetzbarkeit automatischer Analysen digitaler Medieninhalte als auch um die Evaluation der eigentlichen Klassifikationsverfahren geht, wurde ein Szenario angestrebt, dass den inhaltsanalytischen Forschungsalltag möglichst realistisch abbildet. Dies gilt sowohl für die Auswahl der analysierten Medien als auch den zeitlichen Rahmen der Erhebung. Die Datenerhebung und -analyse für die Evaluationsstudie wurde mit Hilfe des im vorangegangenen Kapitel vorgestellten Forschungsinstruments weitestgehend automatisiert durchgeführt.

Da es in dieser Arbeit höchstens sekundär um die eigentlichen Ergebnisse der Inhaltsanalysen geht, wurden Medien und Untersuchungszeitraum ausgewählt, ohne jedoch Anspruch auf Repräsentativität für die deutsche (Online-)Medienlandschaft zu erheben. Der Zugriff auf die Online-Medien erfolgte, wie von Zeller & Wolling (2010) vorgeschlagen, über die Offline-Pendants der Medienangebote, deren Grundgesamtheit leichter zu bestimmen ist.[1]

Auf der Seite des Medienangebots wurden für die Analyse zwölf deutsche Nachrichten-Websites (bzw. deren RSS-Feeds) bewusst ausgewählt, die ein breites Spektrum deutscher Online-Publizistik repräsentieren: Enthalten sind erstens die Webangebote der vier großen deutschen Qualitätszeitungen Frankfurter Allgemeine Zeigung, Welt, Süddeutsche Zeitung und Frankfurter Rundschau, deren Papierfassungen auch den meisten groß angelegten Presseanalysen (vgl. Pfetsch, 2004; Wilke

1 Für zukünftige Studien scheint es ratsam, einen breiteren Zugang zu wählen, um die deutsche Online-Publizistik repräsentativ analysieren zu können. Die äußerst vielversprechende Inventarisierung aller deutschen journalistischen Online-Angebote von Neuberger et al. (2009), aus der eine Stichprobenziehung möglich wäre, wurde leider erst nach Ende der Datenerhebung verfügbar.

& Reinemann, 2000) zugrundeliegen. Zweitens wurde mit Bild Online das Angebot der größten deutschsprachigen Tageszeitung erhoben, die auch zu den zehn meistbesuchten Websites in Deutschland gehört. Als Ergänzung der überregionalen Tagespresse wurden zudem die Online-Ausgabe des Berliner Tagesspiegel und das Nachrichten Portal der WAZ-Gruppe Der Westen berücksichtigt, die beide zum Zeitpunkt der Erhebung neu gestartet und in der Berichterstattung als innovative Wege des Online-Journalismus beschrieben wurden (Spiegel Online, 2007). Neben diesen sieben Tageszeitungen wurden die Online-Versionen der Magazine Spiegel und Focus sowie der Zeit in die Untersuchung aufgenommen. Angesichts der hohen Besucherzahlen (vgl. Tabelle 6.1) können nicht nur die gedruckten Fassungen, sondern auch deren Websites zu den meistgelesenen in Deutschland gezählt werden. Schließlich wurden auch die begleitenden Webangebote der beiden Hauptnachrichtensendungen Tagesschau und Heute als Untersuchungsobjekte ausgewählt. Für diese liegen leider keine IVW-Nutzerzahlen vor, es kann jedoch davon ausgegangen werden, dass die Websites als zentrale Nachrichtenangebote von ARD und ZDF im Internet eine große Reichweite haben.

Als zeitlicher Rahmen für die Untersuchung wurde ein Jahr gewählt. Die Grundgesamtheit besteht daher aus allen Artikeln, die in den RSS-Feeds der oben genannten Online-Angebote zwischen dem 1.6.2008 und dem 31.5.2009 verlinkt worden sind. Diese rund 208.000 Beiträge wurden vollständig heruntergeladen, bereinigt und in einer Datenbank archiviert. Da es sich hierbei um eine Vollerhebung der genannten Medienstichprobe handelt, kann die Grundgesamtheit der Artikel recht genau beschrieben werden. Wie aus Tabelle 6.1 hervorgeht, unterscheiden sich die verschiedenen Online-Nachrichten-Sites hinsichtlich der Menge ihrer – in den RSS-Feeds publizierten – Beiträge. Da jedoch keine Informationen darüber vorliegen, welcher Anteil an Artikeln der gesamten Website per Feed publiziert wurde, kann man die verschiedenen Angebote nur bedingt hinsichtlich des Gesamtangebotes an Inhalten vergleichen. Auffällig ist, dass Tagesspiegel, Spiegel und Bild besonders viele Beiträge publizieren, wobei die ersten beiden tatsächlich viele Artikel der gedruckten Ausgabe veröffentlichen – im Tagesspiegel ist diese Information in jedem Beitrag vermerkt –, während Bild Online vor allem auch kürzere Agen-

Tabelle 6.1: Quellen und Artikelzahl im Untersuchungszeitraum

Quelle	IVW Rang*	Anzahl Artikel Tag	Woche	Monat	Gesamt
Bild.de (BILD)	10	85	598	2592	31106
DerWesten.de (WEST)	78	20	143	622	7458
FAZ.net (FAZ)	34	46	322	1395	16735
Focus.de (FOC)	23	60	424	1835	22025
FR-online.de (FR)	–	36	251	1090	13074
Heute.de (HEU)	–	26	182	790	9485
Spiegel.de (SPON)	7	72	509	2205	26459
Sueddeutsche.de (SZ)	27	29	205	887	10647
Tagesschau.de (TAG)	–	32	227	983	11798
Tagesspiegel.de (TSP)	88	103	723	3131	37575
Welt.de (WELT)	24	23	159	690	8285
Zeit.de (ZEIT)	44	36	256	1110	13318
Gesamt		570	3999	17333	207965

*Liste der meistbesuchten Angebote laut IVW Online, Stand 06/2008

turmeldungen im RSS-Feed verlinkt. Die wenigsten Artikel wurden in den Feeds von Heute, Welt Online und dem Portal der WAZ publiziert, wobei auf den Startseiten von Heute und Tagesschau auch nur wenige Meldungen pro Tag erscheinen. Bei den Angeboten der Tageszeitungen sind hingegen viele Nachrichten nur in den ressort-spezifischen Feeds verlinkt. Berücksichtigt man diese Ressort-Feeds, die es auf Der Westen zusätzlich auch auf regionaler und lokaler Ebene gibt, erscheinen in den Online-Ausgaben der Welt und WAZ viel mehr Artikel pro Tag, als in dieser Studie erhoben wurden.

Aus diesem Korpus an Nachrichtenbeiträgen wurden für die Evaluationsstudie in einem ersten Schritt 1000 Dokumente mittels einfacher Zufallsauswahl gezogen, die die eigentliche Materialstichprobe bilden. Die gezogenen Beiträge wurden anschließend sowohl automatisch als auch manuell vor der Codierung bereinigt, um Artikel aus der Stichpro-

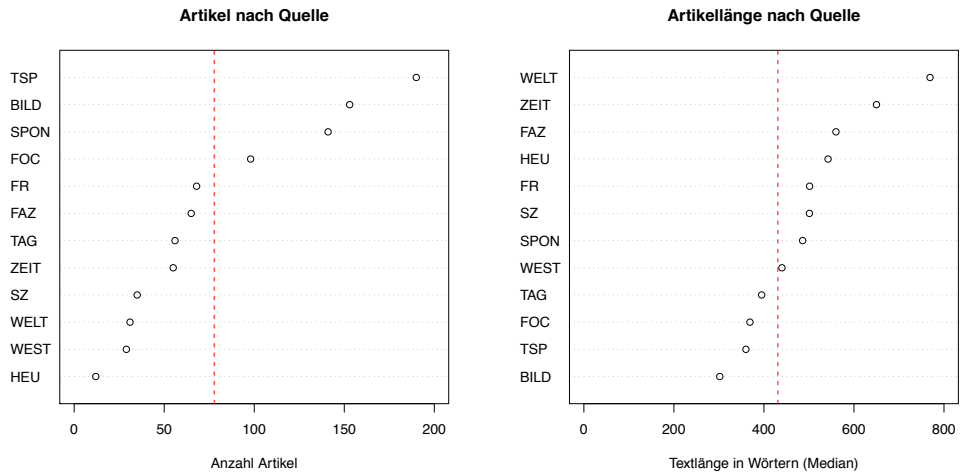

Abbildung 6.1: Artikel der Stichprobe nach Quelle, $n = 933$

be zu entfernen, die zu wenig verwertbaren Text enthielten. Dies waren nach einer ersten Sichtung des Stichprobenmaterials vor allem Beiträge mit ausschließlich audiovisuellen Inhalte. Die meisten davon ließen sich mit einer einfachen schlagwortbasierten Codierung entfernen, indem nach den Begriffen „Flash-Player installieren“, „Javascript aktivieren“ und „Mediathek“ gesucht wurde. Besonders betroffen von dieser Datenbereinigung waren die Inhalte von Heute und Spiegel Online, die zahlreiche reine Video-Beiträge im RSS-Feed verlinken.[2] Diese Filterung audiovisueller Beiträge führt dazu, dass in der Stichprobe Beiträge von Heute.de und Spiegel.de unterrepräsentiert sind, was vor allem bei der ohnehin geringen Fallzahl von heute.de ins Gewicht fällt (vgl. Abbildung 6.1). Weiterhin wurden bei der darauf folgenden manuellen Codierung alle Artikel markiert und anschließend entfernt, die weniger als einen Satz Fließtext enthielten bzw. nur aus einer Ansammlung von Links bestanden. Dies betraf vor allem Beiträge, die auf Bild Online erschienen. Durch diese Bereinigung wurden 67 der 1000 gezogenen Dokumente

2 Beim ZDF wird häufig auf die eigene Mediathek verwiesen, die vor allem Videobeiträge enthält, bei Spiegel Online gibt es u.a. die Reihe kicker.tv, die ebenfalls nur aus Videos besteht.

entfernt, so dass der finale Stichprobenumfang für die Evaluationsstudie $n = 933$ Artikel aus 12 Online-Angeboten beträgt.

Bei der automatischen textstatistischen Analyse der Stichprobe zeigen sich neben erwartbaren Ergebnissen auch einige Besonderheiten der Datenherhebung und -bereinigung. So weisen die vier Qualitätszeitungen und die Online-Ausgabe der ZEIT überdurchschnittlich lange Artikel auf (vgl. Abbildung 6.1).[3] Die Artikel der BILD sind hingegen auch in der Online-Ausgabe mit rund 300 Wörtern im Mittel sehr kurz. Die weit überdurchschnittliche Beitragslänge der WELT ONLINE ($Md = 769$) ist hingegen auf die Textextraktion zurückzuführen: Da es auf der Website keine separate Print-Version der Artikel gibt, wurden zum Teil auch die ersten ein bis zwei Kommentare von Benutzern als Beitragstext verarbeitet. Ohne manuelle Bereinigung oder spezielle Extraktionsregeln ließ sich dies nicht vermeiden, so dass die Texte der WELT ONLINE in der Folge auch nicht-redaktionelle Bestandteile enthalten, die die durchschnittliche Beitragslänge erhöhen. Für die Klassifikation hat dies allerdings keine Konsequenzen, da die Leserkommentare sich auf den konkreten Beitrag beziehen.

Bei der längsschnittlichen Betrachtung der Stichprobe in Abbildung 6.2 fallen einige größere Abweichungen von der durchschnittlichen Artikelanzahl pro Woche ($\bar{x} = 19$) auf. Diese sind einerseits technisch bzw. durch die Stichprobenziehung bedingt, anderseits auch auf die Nachrichtenlage zurückzuführen. Der niedrige Ausgangswert resultiert schlicht aus der Tatsache, dass der 1.6.2008 nicht an einem Wochenanfang lag. Die unterdurchschnittliche Artikelanzahl in den Kalenderwochen 28 und 47/2008 ist auf technische Probleme zurückzuführen, da in diesen Wochen der Server an einem bzw. zwei Tagen ausfiel und so einige Messungen nicht realisiert werden konnten. In den beiden letzten Wochen des Jahres 2008 war hingegen die Anzahl veröffentlichter Meldungen tatsächlich niedriger als im Jahresschnitt, was sicher auf die Feiertage zurückgeführt werden kann. Die positive Abweichung in der Kalenderwoche 11/2009 ist nicht ohne weiteres zu erklären: Eine plausible Möglichkeit war die

3 Da die Artikellänge bei fast allen Angeboten rechtsschief verteilt ist, und zudem einige Ausreißer enthalten sind, wird der Median statt des Mittelwertes ausgewiesen. Der Median der Artikellänge beträgt in der Stichprobe 431 Wörter, der Mittelwert 510 Wörter.

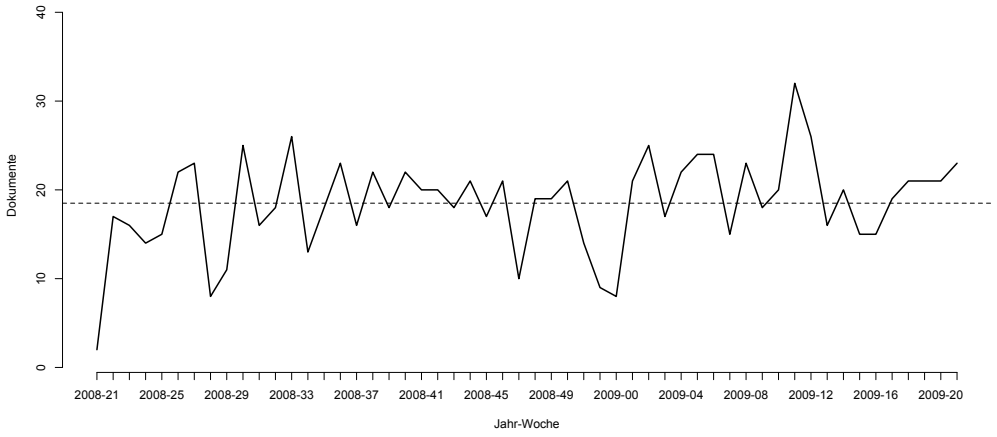

Abbildung 6.2: Artikel der Stichprobe im Längsschnitt

umfangreiche Berichterstattung über den Amoklauf von Winnenden, die insgesamt zu mehr veröffentlichten Beiträgen in dieser Woche führt. Allerdings zeigte sich dies nicht in den Themen der Stichprobenbeiträge aus dieser Woche, so dass es sich auch schlicht um Ausreißer bei der Stichprobenziehung handeln könnte.

Über die einzelnen Kalendermonate sind kaum saisonale Auffälligkeiten zu entdecken (vgl. Abbildung 6.3). Im Juni sind etwas mehr Beiträge erschienen, im Juli und September ist die Nachrichtenlage etwas dürftiger. Demgegenüber zeigen sich die erwartet großen Unterschiede in der Häufigkeit der Beiträge über die verschiedenen Wochentage hinweg (vgl. Abbildung 6.4). Die meisten Artikel erscheinen Donnerstags, die wenigsten am Wochenende. Dieses Muster zeigt sich bei fast allen Online-Angeboten, lediglich Welt, Tagesschau und Heute veröffentlichen auch am Wochenende gleich viele oder sogar mehr Artikel als an Wochentagen. Die durchschnittliche Artikellänge variiert hingegen weder über die Kalendermonate noch über einzelnen Wochentage.

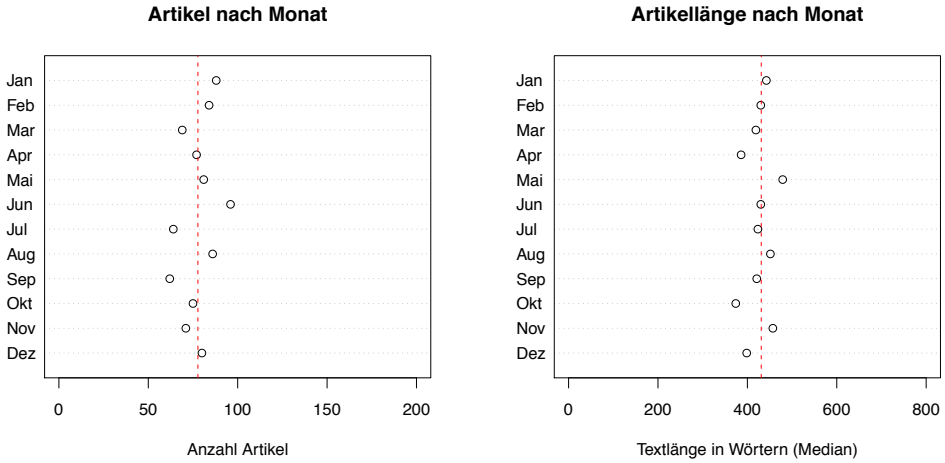

Abbildung 6.3: Artikel der Stichprobe nach Monat, $n = 933$

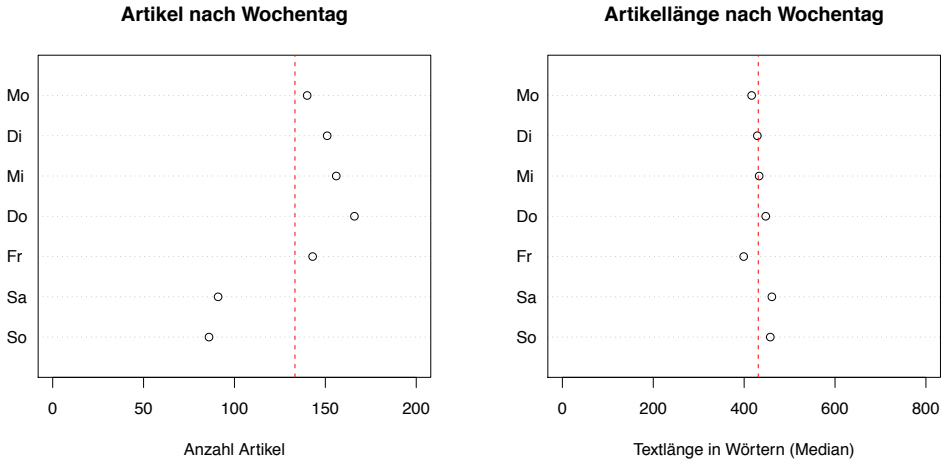

Abbildung 6.4: Artikel der Stichprobe nach Wochentag, $n = 933$

6.2 Auswahl der Kategorien

Um die Codierqualität überwachter automatischer Verfahren evaluieren zu können, müssen zuerst einmal relevante Kategorien der Inhaltsanalyse ausgewählt werden. Angesichts der Tatsache, dass sich die inhaltsanalytische Kategorienbildung stets aus dem substanziellen Forschungsinteresse und theoretischen Überlegungen ergibt, stellen sich für diese methodische Arbeit die Fragen: Welche Kategorien sind von größtmöglicher Relevanz? Nach welchen Kriterien sollen passende Kategorien gewählt werden?

Auch wenn für die Prüfung der grundsätzlichen Machbarkeit und die Bestimmung möglicher Einflussgrößen auf die Klassifikationsqualität die Wahl der inhaltlichen Kategorien streng genommen unerheblich ist, müssen letztlich einige wenige nachvollziehbar ausgewählt werden. Die Kriterien, nach denen ich die Kategorien für die Evaluation ausgesucht habe, sind folgende:

Dokumentation Um die Vergleichbarkeit der Ergebnisse zu garantieren und die Validität von Codebuch, manueller und automatischer Codierung nicht miteinander zu vermischen, werden ausschließlich gut dokumentierte und getestete Kategorien verwendet. Dies minimiert auch meinen eigenen Einfluss auf die Entwicklung des Codebuchs und die Codierung. Da es in dieser Arbeit um die Machbarkeit automatischer thematischer Inhaltsanalysen (Früh, 2007) geht, sind Kategorien aus umfangreichen und/oder Langzeitprojekten zur Themenanalyse besonders relevant.

Einschlägigkeit Auch wenn es nicht möglich ist, für die – deutsche oder internationale – Kommunikationswissenschaft repräsentative Kategorien zu finden, gibt es typische Variablentypen: Themenvariablen in unterschiedlichen Abstraktionsgraden, Nachrichtenfaktoren und Akteurscodierungen (Rössler, 2005; Fretwurst, 2008). Da sich letztere jedoch auch für einfache diktionärbasierte Analysen eignen, sind sie für die überwachte Klassifikation weniger interessant.

Codierbarkeit auf Beitragsebene Für die überwachte Klassifikation ist es zwar unerheblich, ob auf Artikel-, Absatz- oder Aussageebene

codiert wird, nur müssen die entsprechenden Codiereinheiten fertig vorliegen. Da für das Problem der zuverlässigen Identifikation von Aussagen bislang kaum manuelle und keine automatischen Lösungen in Sicht sind, wird ausschließlich auf Beitragsebene codiert. Dies wiederum hat zur Konsequenz, dass pragmatische Kategorien wie die häufig codierte Tendenz oder Bewertung nicht ausgewählt werden, weil diese m.E. nach eher auf Aussageebene zu codieren sind (vgl. Merten, 1995; Rössler, 2005).

Erwartetes Auftreten Da für die Evaluation nur begrenzte Ressourcen zur Verfügung stehen, sollen die Ergebnisse auch bei einer mittleren Fallzahl inferenzstatistisch gut abgesichert sein. Nach bisherigen Erkenntnissen ist für das Training und Testen der Klassifikation eine Häufigkeit von mindestens 10 Prozent für die seltenste Ausprägung einer Variablen anzustreben. Sind zu wenige Dokumente pro Klasse vorhanden, funktioniert weder das Training noch die Evaluation zuverlässig. Dies hat zur Folge, dass nur relativ abstrakte thematische Variablen, deren Kategorien möglichst gleich verteilt sein sollten, verwendet werden können.

Anhand der genannten Kriterien habe ich für die Evaluationsstudie zwölf Kategorien ausgewählt, die einerseits prototypisch für die kontinuierliche Nachrichtenanalyse sind, andererseits genügend Variabilität aufweisen, um sinnvoll Rückschlüsse auf ihre Eignung für eine Automatisierung durch maschinelles Lernen zu ziehen. Es liegt auf der Hand, dass jede Auswahl für diese Arbeit sich der berechtigten Kritik stellen muss, dass sie unvollständig und/oder noch immer zu homogen sei, um daraus Schlüsse auf die generelle Eignung des Analyseverfahrens zu ziehen. Aufgrund forschungsökonomischer Restriktionen ist dies nicht zu vermeiden, so dass ich an dieser Stelle und auch in der Interpretation und Diskussion der Ergebnisse nur darauf hinweisen kann, dass die Eignung weiterer Kategorien für die überwachte Klassifikation eine offene konzeptionelle und empirische Frage ist.

Rund die Hälfte der ausgewählten Kategorien besteht aus klassischen Themenvariablen, wie sie in den meisten Nachrichtenanalysen verwendet werden (vgl. z.B. Früh, 2007; Weiß, 1998). Dabei lassen sich grundsätzlich

zwei Operationalisierungstrategien unterscheiden: die häufig verwendete multinomiale Codierung und die dichotome Mehrfachcodierung von Beitragsthemen. Bei der multinomialen Codierung wird für die Variable *Thema* eine Ausprägung aus einer potentiell sehr umfangreichen Liste ausgewählt. Dies führt dazu, dass sich die Codierer bei Beiträgen mit mehreren Themenbezügen für ein Thema entscheiden müssen, was ggf. die Reliabilität der Codierung verringert (Rössler, 2005, 126). Um dieses Problem zu entschärfen, wird häufig mit zwei oder mehr Themenvariablen (z.B. Haupt- und Nebenthema) gearbeitet. Solange jedoch in einem Beitrag mehr Themenfelder behandelt werden als Themen-Variablen vorgegeben sind, müssen die Codierer eine potentiell konsequenzenreiche Selektionsentscheidung treffen, die zudem mit einem Verlust an Informationen einhergeht.

Eine alternative Kategorienbildung begegnet diesem Problem durch die Zerlegung der Themenliste mit k Ausprägungen in k dichotome Themenvariablen, die jeweils mit ja oder nein codiert werden (Bruns & Marcinkowski, 1997). Dieses Vorgehen löst das Problem der thematischen Mehrfachcodierung und vereinfacht die Analyse von Einzelthemen und Themenkombinationen. Allerdings sind mit dichotomen Indikatoren auch Probleme verbunden: Erstens muss durch entsprechende Codieranweisungen und Schulungen gewährleistet werden, dass nicht zu liberal codiert wird, in dem schon beim Auftauchen eines Schlüsselwortes oder -satzes der Code *kommt vor* vergeben wird. Zweitens ist das vollständige Abarbeiten von langen Themenlisten, in denen die Mehrzahl der Variablen mit *kommt nicht vor* codiert wird, kognitiv belastender und fehleranfälliger als die einmalige Vergabe eines einzelnen Themencodes. Hier ist eine gezielte Filterführung von abstrakteren zu konkreteren Themenvariablen unabdingbar. Da dies bei der Online-Codierung möglich ist, und zudem für die Evaluation nur wenige Themen notwendig waren, wurde für die vorliegende Studie diese Operationalisierungstrategie gewählt. Sie hat auch den Vorteil, auf die gut dokumentierten und im Rahmen umfangreicher Programmanalysen eingesetzten Kategorien von Bruns & Marcinkowski (1997) und GÖFAK Medienforschung (2010) zurückgreifen zu können.

Als Themenvariablen wurden folgende Kategorien ausgewählt: *Politik allgemein*, *bundesdeutsche Politik*, *Politik international*, *Wirtschafts- und Finanzpolitik*, *Sport* und *Kultur*. Diese relativ abstrakten Themen wurden durch die konkrete Kategorie *Bundestagswahl 2009* ergänzt, die sich jedoch durch eine recht liberale Codieranweisung auszeichnet. Alle Variablen außer *Kultur* werden dichotom codiert, die Kulturvariable hat drei Ausprägungen (keine, Hochkultur, Populärkultur). Die Variablendefinitionen und Codieranweisungen sind den Original-Codebüchern von Bruns & Marcinkowski (1997), Marcinkowski et al. (2001) und GÖFAK Medienforschung (2010) entnommen und finden sich in Anhang B.2.

Ergänzt werden die rein thematischen Kategorien durch eine Auswahl an häufig verwendeten Nachrichtenfaktoren. Die Nachrichtenwertforschung hat in der deutschen Kommunikationswissenschaft eine lange Tradition, entsprechend umfangreich sind deren inhaltsanalytische Instrumente dokumentiert (vgl. zusammenfassend Fretwurst, 2008). Nachrichtenfaktoren sind sind nicht nur für die Erklärung journalistischer Selektionsprozesse, sondern auch für die Rezeptionsforschung von Bedeutung, so dass sie häufig nicht nur im Rahmen von Programmanalysen, sondern auch in Medienwirkungsstudien erhoben werden (Eilders, 1997; Ruhrmann et al., 2003). Dass die Erhebung von Nachrichtenfaktoren auch für die Analyse von genuinen Online-Inhalten hilfreich und relevant ist, zeigt ein neuere Studie von Eilders et al. (2010), in der politische Blog-Postings analysiert werden.

Für die Evaluationsstudie wurden die Nachrichtenfaktoren *Kriminalität*, *Unglücke/Katastrophen/Unfälle*, *Kontroverse* und *Prominenz* aus dem Codebuch von Fretwurst (2008) übernommen. Gerade die letzten beiden Nachrichtenfaktoren sollten deutlich schwerer zu automatisieren sein, da die Erkennung von Prominenz oder kontroversen Standpunkten sehr viel Kontextwissen erfordert und zudem eher auf der semantischen oder pragmatischen Textebene erfolgt. Da ein Klassifikator jedoch keinerlei Vorwissen mitbringt und nur auf lexikalischer Ebene codieren kann, würde eine geringere Klassifikationsgüte nicht überraschen. Während Kriminalität und Unglücke dichotom codiert werden, sind Kontroverse und Prominenz ordinal mit drei Ausprägungen (keine, geringe, große) skaliert.

Ergänzend zu den Themenvariablen und Nachrichtenfaktoren wurde eine Kategorie in das Codebuch aufgenommen, die den journalistischen Stil einer Meldung beschreibt. Diese Variable wurde einer Studie von Trebbe (1996) entnommen und reicht in fünf Ausprägungen von sachlich-informierend bis standpunkthaft-kritisch.

Für alle Kategorien wurde auf die ursprünglichen Codebücher und, soweit vorhanden, Codieranweisungen zurückgegriffen. Lediglich bei der Variable *Prominenz* wurden zwei Ausprägungen zusammengefasst, die Einzelkategorien *Wirtschaftspolitik* und *Finanzpolitik* von Marcinkowski et al. (2001) wurden zu einer Variable fusioniert, die Variable *Sport* um eine Codieranweisung ergänzt. Insgesamt enthält das Codebuch der Evaluationsstudie damit zwölf Variablen, die für den Klassifikationsalgorithmus unterschiedliche Schwierigkeitsgrade aufweisen sollten.

6.3 Reliabilität der manuellen Codierung

Die manuelle Codierung aller Dokumente der Stichprobe wurde von insgesamt 7 Personen inklusive Forschungsleiter direkt über das NewsClassifier-Framework durchgeführt. Anhand der vorliegenden Codepläne und -anweisungen wurden alle Teilnehmer an zwei Tagen geschult, einzelne Codierer bekamen während der Feldzeit gezielte Nachschulungen, falls bedeutsame Abweichungen in der Codierung gegenüber den anderen zu verzeichnen waren. Die gesamte Codierung wurde permanent überwacht, was durch die Verwendung einer zentralen server-basierten Infrastruktur erheblich erleichtert wurde. Jeder der Codierer bekam randomisiert Dokumente aus der Stichprobe zur Codierung vorgelegt, von denen ungefähr jedes dritte von mindestens einer weiteren Person codiert wurde. Der Reliabilitätstest wurde also für die Codierer unsichtbar parallel zur Normalcodierung durchgeführt. Daher steht für die Schätzung der Reliabilität eine vergleichsweise breite Basis an Dokumenten zur Verfügung, die zudem repräsentativ für die Gesamterhebung sind.

Um das Risiko zu minimieren, durch Fehler in der Software die Codierungsarbeit zu gefährden, wurde in zwei Phasen codiert. Die Variablen *Sport* und *Politik allgemein* wurden vom Forschungsleiter und zwei weiteren Personen codiert. Nachdem diese erste Feldphase erfolgreich verlief,

wurden die weiteren Variablen in das Codebuch aufgenommen. Da eine Person während des ersten Tests unbefriedigende Leistungen zeigte, wurde diese in der zweiten Phase durch einen neuen Codierer ersetzt.

Tabelle 6.2: Intercoder-Reliabilität der manuellen Inhaltsanalyse

Variable	CR	CI_{CR}	α	CI_{α}	n_{Art}
Politik allgemein	.91	.86–.94	.79	.68–.87	178
Bundesdeutsche Politik	.90	.87–.92	.69	.60–.76	373
Politik International	.93	.90–.95	.76	.65–.82	373
Wirtschafts-/Finanzpolitik	.93	.91–.95	.74	.65–.83	373
Bundestagswahl 2009	.97	.95–.98	.48	.21–.70	373
Sport	.99	.98–1.0	.98	.93–.99	395
Kultur	.95	.92–.97	.68	.54–.79	373
Unglücke/Katastrophen/Unfälle	.95	.93–.97	.67	.54–.80	373
Kriminalität	.92	.89–.95	.67	.56–.77	373
Kontroverse	.69	.65–.74	.49	.40–.56	373
Prominenz	.71	.66–.75	.72	.66–.77	373
Journal. Stil	.53	.48–.58	.36	.25–.44	373

Ausgewiesen sind die Prozentübereinstimmung CR sowie Krippendorffs α. Die Konfidenzintervalle entsprechen Bias-Corrected Percentile Intervals (Efron & Tibshirani, 1993; Hayes & Krippendorff, 2007).

Die Ergebnisse des Reliabilitätstest sind in Tabelle 6.2 dargestellt. Sie basieren auf der Gesamtzahl an Paarvergleichen über alle Codierer hinweg, da zugunsten einer breiteren Dokumentenbasis auf vollständig überlappende Mehrfachcodierungen verzichtet wurde (Potter & Levine-Donnerstein, 1999; Krippendorff, 2004b). Da die Codebücher nicht vom Forschungsleiter selbst entwickelt wurden, wurde auf eine Überprüfung der Forscher-Codierer-Übereinstimmung verzichtet, wie sie beispielsweise Fretwurst (2008) durchführt. Stattdessen ist der Forschungsleiter gleichberechtigt am Reliabilitätstest und der Normalcodierung beteiligt. Es zeigt sich, dass die Reliabilität der Themenvariablen äußerst hoch ist, solange man die einfache prozentuale Übereinstimmung (CR) betrachtet. Da oftmals nur dieser Wert angegeben wird, lassen sich nur so

Vergleiche mit den Original-Studien ziehen. Hier zeigt sich, dass trotz relativ kurzer Schulung eine vergleichbare Zuverlässigkeit der Codierung erreicht wurde. Betrachtet man hingegen den zufalls- und prävalenzkorrigierten Koeffizienten α und dessen Konfidenzintervalle, kann man in den meisten Fällen von gerade akzeptablen Werten um .70 ausgehen.[4] Einige Variablen wie *Kontroverse* und *Journalistischer Stil* weisen dagegen eine unbefriedigende Reliabilität auf. Nimmt man die unteren Schranken des Konfidenzintervalls als konservative Schätzung, ist die Codierung nur zu 25 bzw. 40 Prozent reliabel. Besonders unbefriedigend ist die Zuverlässigkeit der Variable *Journalistischer Stil*, bei der nur in der Hälfte aller Vergleiche eine übereinstimmende Codierung gelang.[5]

Die Diskrepanz zwischen Prozentübereinstimmung nach Holsti und dem korrigierten Wert nach Krippendorff wird insbesondere bei der Kategorie *Bundestagswahl 2009* deutlich. Da die meisten Dokumente der Stichprobe übereinstimmend mit *kommt nicht vor* codiert wurden, aber die wenigen Fälle, in denen die Bundestagswahl thematisiert wurde, nicht von allen Codierern erkannt wurden, unterscheiden sich beide Koeffizienten erheblich.[6] Dies wird auch im riesigen Konfidenzintervall für dieser Variable deutlich: Da dieses auf Bootstrapping beruht, d.h. dem wiederholten Ziehen von Stichproben aus den Daten, schwankt der prävalenzkorrigierte Alpha-Wert je nach Komposition des Bootstrap-Samples erheblich (vgl. Hayes & Krippendorff, 2007). Das Konfidenzintervall der unkorrigierten Prozentübereinstimmung ist dagegen sehr klein, was als Beleg dafür gelten kann, dass auch inferenzstatistisch korrekt geschätzte Reliabilitätswerte nicht ohne weiteres als korrekt angesehen werden können.

4 Die Unterschiede in der Zuverlässigkeit der Politikvariablen lassen sich vor allem darauf zurückführen, dass in einigen Fällen Kontextwissen der Codierer gefragt war, ob ein bestimmter Politiker oder ein Politikfeld auf Ebene des Bundes oder der Länder angesiedelt war. Dies führte zu einigen Nichtübereinstimmungen.

5 Die vollständigen Koinzidenz-Matrizen aller Variablen sind in Abbildung B.1 im Anhang dargestellt.

6 Dieses Verteilungsproblem und die Diskrepanz zwischen den Koeffizienten von Holsti und Krippendorff wird auch bei der Codierung von Themen und Nachrichtenfaktoren bei Raupp & Vogelgesang (2009) deutlich.

Tabelle 6.3: Intercoder-Reliabilität nach Codierern

Codierer	A	B	C	D	E	F	G	SD
Sport	.90	.99	1.00*					.06
Politik allgemein	.86	.70	1.00*					.15
Bundesdeutsche Politik	.68	.73	.68	.65	.70	.70	.67	.07
Internationale Politik	.71	.75	.76	.77	.75	.74	.80	.03
Wirtschafts- und Finanzpolitik	.73	.75	.73	.76	.74	.73	.71	.02
Bundestagswahl 2009	.44	.42	.55	.57	.54	.46	.51	.06
Kultur	.67	.77	.57	.69	.73	.72	.60	.07
Unfälle/Unglücke	.65	.72	.62	.67	.66	.63	.76	.05
Kriminalität	.64	.69	.73	.68	.70	.65	.65	.03
Kontroverse	.49	.43	.53	.47	.50	.45	.53	.04
Prominenz	.70	.76	.70	.71	.72	.71	.68	.02
Journalistischer Stil	.28	.40	.27	.27	.31	.31	.32	.05

Ausgewiesen ist Krippendorffs α, wenn der betreffende Codierer ausgeschlossen würde. Höhere Werte bedeuten schlechtere Codierer.

∗Codierer wurde ausgetauscht, ein Ersatzmann codierte die weiteren Kategorien.

Neben dem globalen Grad an Zuverlässigkeit der Codierung ist auch die Frage nach den Ursachen fehlender Reliabilität von Bedeutung. In Tabelle 6.3 sind daher die Reliabilitäten für alle Variablen und Codierer dargestellt. Der in Kap. 4.4 vorgeschlagene RICO-Koeffizient in den Zellen bezeichnet die Reliabilität der Codierung, wenn der betreffende Codierer ausgeschlossen würde. Hohe Werte zeigen dementsprechend schlechtere Codiererleistungen an. Am deutlichsten lässt sich dies bei Codierer C veranschaulichen, der die Variablen *Politik* und *Sport* codiert hat. Ohne ihn wäre die gemessene Reliabilität dieser Variablen perfekt, d.h. alle beobachteten Nicht-Übereinstimmungen gehen auf diese Person zurück. Da eine Nachschulung nicht möglich war, wurde der Codierer kurzerhand ersetzt. In der Randspalte der Tabelle ist zudem die Standardabweichung der einzelnen RICO-Koeffizienten verzeichnet, die ein

Vergleichmaß für die Codiererabhängigkeit der Reliabilität ist. Diese ist in den meisten Fällen sehr gering, lediglich *Bundesdeutsche Politik* und *Kultur* zeigen etwas stärker differenzierte Codiererleistungen.

Insgesamt kann die manuelle Codierung von Dokumenten für die Evaluation als gelungen bezeichnet werden. Zwar wurde keinesfalls eine perfekte Reliabilität erreicht, allerdings entsprechen die ermittelten Werte in der Größenordnung denjenigen der Originalstudien und vergleichbaren Inhaltsanalyse. Da das Ziel der Evaluation darin besteht, die Möglichkeiten überwachter Klassifikation unter realistischen Bedingungen durchzuführen, habe ich auf eine nachträgliche Schulung oder Nachcodierung kritischer Dokumente verzichtet. Es ist daher anzunehmen, dass das auf diese Weise erstellte Trainingsmaterial in etwa dem gleichen Umfang messfehlerbehaftet ist, wie dies bei konventionellen Inhaltsanalysen der Fall ist. Die ermittelten absoluten Gütemaße der überwachten Klassifikation sind eher konservativ, da die Klassifikationsqualität von der Qualität der Trainingsmaterials abhängt.

6.4 Auswahl des Klassifikationsalgorithmus

Die Qualität überwachter Textklassifikation steht und fällt nicht nur mit der Auswahl der Kategorien, sondern auch mit der Leistungsfähigkeit des Klassifikationsalgorithmus. Da es sich in dieser Studie um eine sozialwissenschaftliche Evaluation handelt, verzichte ich auf einen umfassenden Vergleich verschiedener Klassifikationsverfahren, zumal dazu zahlreiche Studien vorliegen (Dumais et al., 1998; McCallum & Nigam, 1998; Joachims, 2002; Hillard et al., 2007; Durant & Smith, 2007). Diese zeigen, dass zwischen den meistverwendeten Algorithmen *Naive Bayes* und *Support Vector Machine* nur minimale Unterschiede in der Klassifikationsqualität bestehen. Für die Wahl eines Klassifikators waren daher vor allem Verständlichkeit, leichte Bedienbarkeit, Flexibilität und Verfügbarkeit vorrangig. Als Klassifikationsalgorithmus für die folgenden Analysen habe ich OSBF-Lua von Assis (2006) ausgewählt, der viele der genannten Vorteile vereint:

Verständlichkeit Bei OSBF-Lua handelt es sich im Grundsatz um einen *Naive Bayes*-Klassifikator, der allerdings über innovative Verfahren der Feature-Selektion und -Gewichtung verfügt, die die Klassifikationsleistung signifikant verbessern (Siefkes et al., 2004; Assis, 2006). Das rein statistische Funktionsprinzip ist im Vergleich zu *Support Vector Machines* leicht verständlich (vgl. Abschnitt 3.4.1), der Programmcode entsprechend kurz und nachvollziehbar.

Leistungsfähigkeit Die Entwickler von OSBF-Lua haben in den vergangenen Jahren mehrere Klassifikations-Wettbewerbe im Bereich der Spam-Filterung gewonnen, auch gegen deutlich komplexere Algorithmen.

Geschwindigkeit Die Implementation des Algorithmus ist für die schnelle Verarbeitung tausender Dokumente optimiert und stellt trotzdem geringe Anforderungen an Prozessor- und Speicherkapazität. Umfangreiche Evaluationen lassen sich daher in kurzer Zeit durchführen.

Flexibilität OSBF-Lua kann sowohl blockweise als auch inkrementell trainiert werden, so dass sich die Software auch für kontinuierliche Textanalysen eignet. Da für jede Klassifikationsentscheidung ein Wahrscheinlichkeitswert angegeben wird, eignet sich der Algorithmus auch für aktives Lernen (vgl. Abschnitt 4.3).

Verfügbarkeit Sowohl die Klassifikationsbibliothek OSBF-Lua als auch ein leicht bedienbares Front-End für die Textklassifikation sind als *Open Source*-Software erhältlich, die sich individuell anpassen lässt.[7]

OSBF-Lua und Moonfilter sind für viele UNIX- und Linux-Systeme verfügbar und lassen sich daher leicht in das im Anhang A vorgestellte Framework für manuelle und automatische Inhaltsanalysen integrieren. Nach jeder manuellen Codierung werden im Hintergrund die dazugehörigen Klassifikatoren trainiert.

7 http://osbf-lua.luaforge.net/, http://www.siefkes.net/software/moonfilter/

6.5 Untersuchungsdesign und Analysestrategie

Den Forschungsfragen in Abschnitt 5.1 folgend, kann die Evaluation überwachter Textklassifikation grob in drei Teilschritte gegliedert werden: (1) Eine grundlegende nicht-experimentelle Bestimmung der Klassifikationsqualität, (2) eine experimentelle Untersuchung des Einflusses bestimmter Texteigenschaften und Vorbehandlungsverfahren auf die Klassifikationsgüte und (3) eine experimentelle Untersuchung zum Einfluss des Trainingsprozesses auf die Entwicklung der Klassifikationsqualität. Diese drei Schritte erfordern unterschiedliche Untersuchungsdesigns, die im Folgenden erläutert werden.

Grundlegende Evaluation
Die Anwendbarkeit überwachter Textklassifikation für die Analyse von Nachrichten lässt sich empirisch durch Evaluationsverfahren des maschinellen Lernens überprüfen, wie sie in Abschnitt 4.4 dargestellt wurden. Im Zentrum der Analyse steht dabei der *Train-Test*-Ansatz, bei dem der Klassifikator mit einer Auswahl an vorcodierten Dokumenten trainiert wird und damit anschließend selbst eine neue Dokumentenstichprobe klassifiziert. Die Ergebnisse der Klassifikation können dann mit den manuellen Codierungen verglichen werden. Die Ergebnisse eines solchen *Train-Test*-Laufs bilden die Grundlage für weitere statistische Analysen. Konkret besteht ein einzelner Evaluationslauf für eine Variable V aus dem Codebuch aus folgenden Schritten:

1. Für alle Dokumente der Stichprobe wird der manuell vergebene Code für V bestimmt. Existieren aufgrund einer Mehrfachcodierung mehrere Codes nebeneinander, wird daraus per einfacher Zufallsauswahl ein Wert gezogen. Dies hat bei übereinstimmender Codierung keine Konsequenzen, bei Nichtübereinstimmung beträgt die Wahrscheinlichkeit, einen bestimmten Code x zu erhalten, der relativen Häufigkeit dieses Wertes in den Mehrfachcodierungen

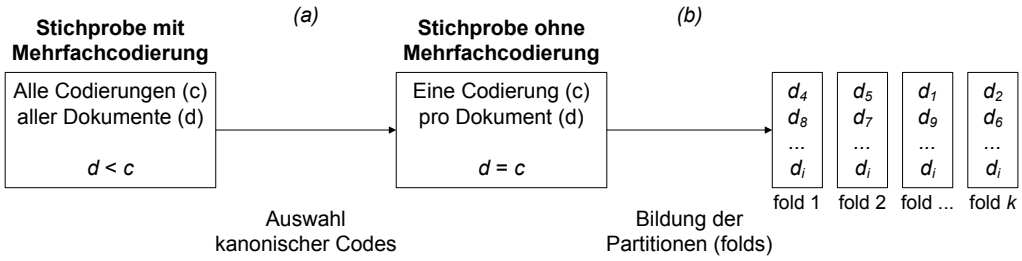

Abbildung 6.5: Zufällige Auswahlprozesse in einem Evaluationslauf

$P(x)$. Nach dieser Auswahl liegt eine Liste mit je einem als richtig definierten „kanonischen" Code pro Dokument vor.[8]

2. Die Dokumentenstichprobe wird für die *10-fold Cross Validation* zufällig in zehn Partitionen gleicher Größe aufgeteilt. Auch die Zusammensetzung dieser Partitionen variiert zwischen den einzelnen Evaluationsläufen.
3. Eine Partition wird als Test-Set zurückgelegt, mit den Dokumenten und Codes der anderen neun Partitionen wird der Klassifikator trainiert. Anschließend klassifiziert der Algorithmus die Testdokumente; die Ergebnisse werden mit den kanonischen manuellen Codes verglichen. Dieser Vorgang wird für alle Partitionen wiederholt, so dass am Ende für jedes Dokument ein Paar mit manuellen und automatisch vergebenen Codes existiert. Diese Daten werden für die Berechnung der Reliabilität nach Holsti und Krippendorff sowie der Validitätsmaße Precision und Recall verwendet. Letztere Kennwerte sind allerdings nur für dichotome Variablen sinnvoll.

Bei der Berechnung von Mittelwert und Varianz der Coder-Klassifikator-Reliabilität *Rel* sind zwei unterschiedliche Quellen von Variabilität zu berücksichtigen (vgl. Abbildung 6.5): (a) die Auswahl der kanonischen Codes bei Mehrfachcodierungen und (b) die Partionierung der Stichprobe

8 Da die Ziehung des Wertes selbst nicht-deterministisch ist, wird die Ausprägung der Test- und Trainingsdaten zwischen einzelnen Evaluationsläufen schwanken. Asymptotisch setzt sich bei mehr als zwei Codierern der Wert der Mehrheit durch, im Einzelfall kann jedoch auch der Wert eines abweichenden Codierers als kanonischer Trainings- und Testwert vorkommen.

in 10 *folds*. Um die Effekte dieser Zufallsprozesse separat abschätzen zu können, muss jeder Evaluationslauf repliziert werden, wobei man entweder die Auswahl der Codes oder die Partitionierung neu startet. Daraus ergibt sich für ein Baseline- oder Nullmodell ohne weitere Prädiktoren die Varianz für die Klassifikationsreliabilität *Rel*:

$$Var(Rel) = Var_{codes} + Var_{folds} \tag{6.1}$$

Ist diese Varianz sehr klein, kann dies als Indikator für die Robustheit der Klassifikation gegenüber der Variation der Untersuchungsbedingungen verstanden werden. Um diese Inferenzschlüsse ziehen zu können und gleichzeitig den Rechenaufwand für die Replikationen zu minimieren, wird jede 10-fach-Kreuzvalidierung $2 \times 2 = 4$ Mal wiederholt. Bei zwölf ausgewählten Variablen im Codebuch ergeben sich so 48 Evaluationsläufe, um zuverlässig Schlüsse über die Klassifikationsqualität pro Kategorie ziehen zu können.

Experimentelle Evaluation der Klassifikationsqualität (Teilstudie 1)
Um den Einfluss verschiedener Preprocessing-Verfahren und Texteigenschaften auf die Klassifikationsqualität zu prüfen, ist ein faktorielles Experiment am besten geeignet. Für jeden in Tabelle 6.4 aufgelisteten Einflussfaktor werden zu Beginn eines Evaluationslaufes die Dokumente entsprechend vorbehandelt, z.B. durch Stopwortentfernung oder Entfernung der Überschrift. Dabei ist zu beachten, dass die Treatments ausschließlich vor der automatischen Klassifikation auf die Dokumente angewendet wurden. Für die manuelle Codierung waren alle Faktoren auf ihren Standardwert 0 gesetzt, d.h. es wurden die aus dem HTML extrahierten, aber nicht weiterbehandelten Texte codiert, wobei die Überschrift für die Codierer sichtbar war. Vor allem das Stemming und die Entfernung von Stopwörtern hätten andernfalls das Verständnis der Texte erheblich erschwert.

Zur Extraktion des Fließtextes aus HTML-Dokumenten wurde der BTE-Algorithmus[9] von Finn et al. (2001) verwendet. Die Stopwortliste für die Herausfilterung der häufigsten deutschen Wörter stammt aus

9 http://github.com/aidanf/BTE

Tabelle 6.4: Faktorielles Design der Evaluation

Faktor	Stufen
Textextraktion	(0) BTE (*Body Text Extraction*) Bereinigter Text (1) Original HTML
Textfilterung	(0) ungefilterter Text (1) Entfernung der 1000 häufigsten deutschen Wörter
Stemming	(0) Text ohne Stemming (1) Text mit Porter-Stemming
Überschrift	(0) Text inklusive Überschrift (1) Text ohne Überschrift

dem Wortschatz-Projekt der Universität Leipzig[10]. Zum Stemming wurde eine Implementation des Algorithmus von Porter (1980) für die deutsche Sprache eingesetzt.[11] Alle verwendeten Softwarepakete inklusive dem eingesetzten Klassifikator sind als freie Software erhältlich, die für eigene Zwecke modifiziert werden kann.

Bei einem vollständigen 2^4-faktoriellen Design ergeben sich 16 Treatment-Kombinationen. Werden diese auf zwölf unterschiedliche Kategorien angewandt und jeder Durchlauf wiederum vier Mal repliziert ergeben sich insgesamt $16 \times 12 \times 4 = 768$ Datenreihen für die anschließende varianz- und regressionsanalytische Datenauswertung.

Experimentelle Evaluation der Lerneffektivität (Teilstudie 2)
Um den Einflusses unterschiedlicher Trainingsstrategien auf die Effektivität der Klassifikation zu prüfen, bietet sich ein anderes Forschungsdesign an, das jedoch auf der gleichen Evaluationslogik basiert. Da es relativ aufwändig wäre, die tatsächlichen Abläufe des inkrementellen Lernens durch manuelle Codierung zu untersuchen, wird dieser Prozess in Teilstudie 2 einfach simuliert. Dazu werden dem Klassifikator schrittweise Teile des Trainingsmaterials zur Verfügung gestellt und nach dem Training ein Test-Set klassifiziert. Dabei werden je nach Treatment verschiedene Stra-

10 http://wortschatz.uni-leipzig.de/Papers/top1000de.txt
11 http://github.com/aurelian/ruby-stemmer

tegien der Selektion von Trainingsdaten angewandt – passives Lernen, aktives Lernen und eine Mischform aus beiden (vgl. Abschnitt 4.3). Die konkrete Versuchsanordnung besteht auf folgenden Schritten:

1. Aus der Dokumentenstichprobe werden 233 zufällig ausgewählte Dokumente als Test-Set zurückgehalten, die anderen 700 Artikel werden wie zuvor zum Training verwendet. Der Klassifikator wird zu Beginn mit einem zufällig daraus ausgewählten Initial-Set von 50 Dokumenten trainiert.
2. Dem Klassifikator werden schrittweise 50 weitere Trainingsdokumente vorgelegt, wobei je nach Experimentalbedingung (a) alle Dokumente zufällig aus dem Trainings-Set ausgewählt werden (passives Lernen), (b) der Klassifikator aus den noch nicht verwendeten Dokumenten 50 auswählen kann, deren Codierung dann bekannt gemacht wird (aktives Lernen) oder (c) je 25 Dokumente aktiv und passiv gelernt werden.
3. Nach jedem Trainingsschritt werden alle Dokumente des Test-Sets klassifiziert und die Evaluationsergebnisse gespeichert. Die Klassifikationsentscheidungen fließen aber nicht in das Training des Klassifikators ein, so dass dieser die Testdokumente immer wieder als neu behandelt.

Die Schritte 2 und 3 werden solange wiederholt, bis alle Dokumente des Trainings-Sets verwendet wurden. Auf diese Weise entsteht ein geschachtelter Datensatz mit Messwiederholungen. Um auch hier die Effekte der Zufallsauswahlen zu berücksichtigen, wird die gesamte Simulation für alle 12 Variablen 20 Mal repliziert. Insgesamt ergeben sich für das Experiment mit drei Faktorstufen $3 \times 20 \times 12 = 720$ Durchläufe.

Statistische Analyse
Aufgrund des faktoriellen Designs der beiden Teilstudien bieten sich für die Auswertung der Daten klassische Varianz- und Regressionsanalysen mit dichotomen Treatment-Prädiktoren an, wobei die eigentlich sehr einfache Modellierung durch zwei Besonderheiten der Evaluation verkompliziert wird: Erstens werden stets Modelle für vier abhängige Variablen geschätzt, von denen zwei als Reliabilitätsmaße (CR nach Holsti, Krippendorffs α) und zwei als semantische Validitätsmaße (Precision

und Recall) angesehen werden können (vgl. Abschnitt 4.4). Zweitens sind die Experimentaldaten jeweils für alle zwölf Variablen des Codebuchs verfügbar, und es ist dabei nicht unbedingt von gleichgerichteten Effekten auszugehen. Will man die Variabilität der Effekte pro Klassifikationsvariable sinnvoll einschätzen, müssten jeweils 12 unterschiedliche Modelle gerechnet werden. Multipliziert mit vier abhängigen Variablen ergeben sich so 48 verschiedene Regressionsmodelle. Solche eine Vielzahl einzelner Modelle ist nicht nur schwer darzustellen, die Modelle haben wegen der geringen Fallzahl auch relativ wenig Power. Besser geeignet sind für diesen Zweck hierarchische Regressionsmodelle mit variierenden Koeffizienten, wobei die Gruppierungs- oder Level-2-Variable die Codebuch-Variable darstellt. Diese Mehrebenenanalysen ergeben bei ausbalancierten Designs mit wenigen Gruppen zwar ähnliche Ergebnisse wie separate Regressionen, sind aber insgesamt effektiver, da alle Informationen aus den Daten simultan in die Schätzung eingehen (Gelman & Hill, 2007).

Für jede der vier abhängigen Variablen wird dementsprechend ein hierarchisches Regressionsmodell geschätzt, deren Parameter zwischen den Variablen des Codebuchs variieren dürfen. So kann genau bestimmt werden, ob ein Treatment etwa nur für bestimmte Klassifikationsvariablen, z.B. *Sport* oder *Politik*, effektiv ist.

Da in der zweiten Teilstudie nicht die absolute Größe der Qualitätsindikatoren von Interesse ist, sondern deren Entwicklung durch Training im Zeitverlauf, bieten sich für die Datenanalyse Wachstumskurvenmodelle (*Growth Curves*) an. Mit diesen lassen sich gleichzeitig intra-individuelle Entwicklungsverläufe (*Trajectories*) pro Variable, inter-individuelle Unterschiede bei diesen Verläufen sowie deren Ursachen (z.B. differentielle Trainingsstrategien) und Auswirkungen schätzen (Urban, 2002). Die Parametrisierung von Wachstumskurven geht von zwei zentralen Größen aus: einem Ausgangswert (*Intercept*) α und einer Steigung (*Slope*) β.

$$y_{jt} = \alpha_j + \beta_j \lambda_t + \epsilon_{jt} \tag{6.2}$$

Jeder Messwert y des Klassifikators j nach t Trainingsdokumenten ist also eine Funktion des Ausgangswertes und des bisherigen Wachstums-

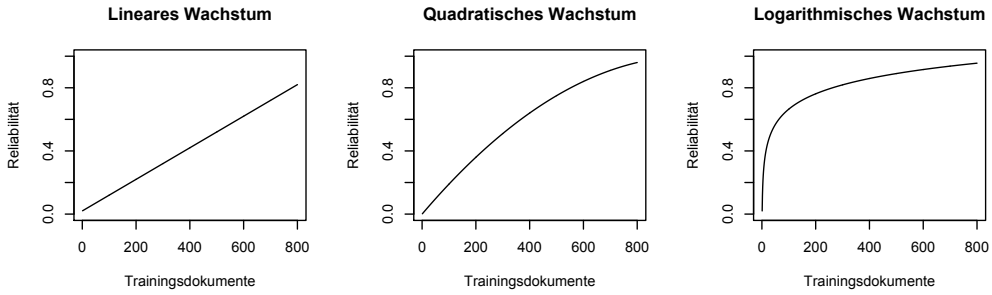

Abbildung 6.6: Typische Ausprägungen von Wachstumskurven

verlaufes, in diesem Fall einer linearen Steigung. Um inter-individuelle Unterschiede modellieren zu können, wird dieses Modell nicht mit zwei fixen Koeffizienten α und β geschätzt, sondern mit individuell variierenden Werten, von denen sich nicht nur der globale Mittelwert, sondern auch die individuelle Varianz bestimmen lassen. Statistisch lassen sich solche Modelle ebenfalls als hierarchische Regressionsanalysen mit jeweils variierenden Intercepts und Slopes schätzen (Bryk & Raudenbush, 1987; Hox & Stoel, 2005).

Grundsätzlich können inter-individuelle Unterschiede (a) im Ausgangswert, (b) in der Steigung oder (c) in beiden Parametern vorliegen. Sollte einer der Koeffizienten keine signifikante Varianz aufweisen, kann das Modell entsprechend reduziert werden. Eine für diese Studie plausible Erweiterung des linearen Wachstumsmodells wären nicht-lineare Verläufe, die sich entweder durch Transformation der Zeitmetrik, z. B. durch Logarithmieren $log(\lambda_t)$, oder durch die Hinzunahme weiterer Polynome, z.B. eines quadratischen Terms $\beta_j \lambda_t^2$, modellieren lassen. Diese alternativen Modelle sind idealtypisch in Abbildung 6.6 dargestellt. Da in der Evaluationsstudie ein natürlicher Nullpunkt im Ausgangswert vorliegt, bietet sich sogar eine Modellspezifikation ohne Intercept-Term an, bei der nur ein oder mehrere Steigungs-Parameter geschätzt werden.

Um den Einfluss der Lernstrategien auf den Lernprozess zu schätzen, ist es angesichts der drei Faktorstufen, die sich ggf. sogar auf einen dichotomen Prädiktor *aktives Lernen* verdichten lassen, am einfachsten, einen

entsprechenden Interaktionsterm zum Modell hinzuzufügen. Dieses hat für eine lineares Wachstum ohne Intercept-Term folgende Form:

$$y_{jt} = \beta_j \lambda_t + \beta_{jtr} \lambda_t treat + \epsilon_j \tag{6.3}$$

Die Klassifikationsqualität der Variable j nach t Trainingsdokumenten ist also eine Funktion des individuellen Anstiegs β_j, der je nach Lernstrategie *treat* steiler oder flacher ausfällt. Bei quadratischen Wachstumsmodellen wird auch der quadratische Effekt noch durch eine Treatment-Interaktion ergänzt, so dass dann insgesamt vier Parameter plus Fehlerterm pro Klassifikator geschätzt werden.

Neben einer grafischen Analyse der Wachstumskurven und der oben dargestellten statistischen Modellierung des Lernprozesses erscheint mir eine dritte Auswertungsstrategie vielversprechend, die stärker auf die forschungspraktische Interpretation der Ergebnisse abzielt. Eine häufige Frage beim Umgang mit überwachten Klassifikationsverfahren lautet: Wie viele Trainingsdokumente sind für eine zufriedenstellende Qualität notwendig? Dies lässt sich beantworten, indem mit Hilfe der Modellkoeffizienten ein typischer Verlauf des Lernprozesses simuliert und aus den vorhergesagten Werten eine Kennzahl abgelesen wird, z.B. nach wie vielen Dokumenten der Klassifikator eine Reliabilität von .7 oder etwa 80 Prozent der maximal möglichen Reliabilität der Variable erreicht wird. Dieser Wert ist ggf. leichter zu interpretieren als die Parameter eines Wachstumsmodells, auch wenn dabei Informationen zum Verlauf des Trainingsprozesses verloren gehen.

7 Ergebnisse

7.1 Teilstudie 1: Klassifikationsqualität und deren Determinanten

7.1.1 Klassifikationsqualität der Kategorien

Die erste Forschungsfrage dieser Evaluation gilt der grundsätzlichen Qualität der überwachten Klassifikation. Im Folgenden werden dafür vier Kennwerte ausgewertet, die jeweils die Reliabilität und Validität der Codierung ausdrücken. Bevor die einzelnen Koeffizienten diskutiert werden, will ich kurz auf die Zuverlässigkeit der Evaluationsergebnisse eingehen. Es hat sich nach Abschluss aller Train-Test-Läufe gezeigt, dass die Ergebnisse der 10-Fold-Kreuzvalidierung sehr stabil sind. Die Standardabweichung der Alpha-Koeffizienten über die 4 Replikationen lag im Mittel bei weniger als .025, die Werte nach Holsti sowie Precision und Recall streuten noch weniger. Lediglich bei den Variablen *Bundestagswahl 2009* (SD=.08) und *Unglücke/Unfälle* (SD=.05) schwankten die Ergebnisse etwas stärker. Dafür war zu etwa gleichen Teilen die Auswahl kanonischer Trainingsdaten sowie die Komposition der Partitionen für die Variabilität der Werte verantwortlich (vgl. Abschnitt 6.5).

In Tabelle 7.1 ist die Reliabilität der überwachten Codierung für alle zwölf verwendeten Variablen der Inhaltsanalyse wiedergegeben. Bei der Betrachtung der Punktschätzer und Konfidenzintervalle fällt auf, dass auch bei der automatischen Codierung die Maße von Holsti und Krippendorff zu unterschiedlichen Bewertungen der Klassifikationsgüte führen. Folgt man den Empfehlungen von Früh (2007, 193) oder Rössler (2005) kann man für fast alle Variablen ein sehr gute Reliabilität konstatieren. Der Klassifikator erreicht bei den Themenvariablen durchgängig prozentuale Übereinstimmungen von über .85 mit der manuellen Codierung. Die Nachrichtenfaktoren *Prominenz* und *Kontroverse* sind vom Algorithmus weniger zuverlässig codiert worden, liegen aber noch in

Tabelle 7.1: Reliabilität der überwachten Klassifikation

Variable	CR	CI_{CR}	α	CI_{α}	$P(c_1)$
Politik	.86	.85–.87	.65	.64–.66	28
Bundesdeutsche Politik	.86	.85–.87	.55	.53–.57	19
Politik International	.89	.88–.90	.61	.59–.63	19
Wirtschafts-/Finanzpolitik	.90	.89–.91	.61	.58–.65	15
Bundestagswahl 2009	.98	.97–.99	.18	.06–.30	2
Sport	.96	.95–.97	.84	.81–.87	15
Kultur	.91	.90–.92	.09	.04–.13	9
Unglücke/Katastrophen/Unfälle	.93	.92–.94	.17	.09–.25	7
Kriminalität	.86	.86–.87	.36	.34–.38	16
Kontroverse	.62	.60–.64	.30	.26–.34	41
Prominenz	.60	.57–.62	.45	.42–.47	50
Journal. Stil	.45	.44–.47	.31	.28–.34	47

$n = 933$ (Dokumente), 10-Fold-Cross-Validation, 4 Replikationen
$P(c_1)$: Anteil der Beiträge außerhalb der Standardkategorie c_0 in Prozent

einem Bereich, der auch bei konventionellen Analysen nicht selten ist (Eilders et al., 2010). Zudem war auch bei der manuellen Codierung die Intercoder-Reliabilität dieser Variablen vergleichsweise niedrig (vgl. Tabelle 6.2). Dies gilt ebenso für die Variable *Journalistischer Stil*, die nur unzuverlässig automatisiert werden konnte. Trotzdem scheint sich bei der Betrachtung der prozentualen Übereinstimmung die überwachte Klassifikation tatsächlich als zuverlässige Alternative für Inhaltsanalysen zu empfehlen. Angesichts dieser Werte kann man erwarten, dass zumeist über 80 Prozent aller Dokumente von der Software übereinstimmend mit den Codierern klassifiziert werden. Für viele Fragestellungen ist diese prognostische Qualität sicher ausreichend. Bezogen auf die Reliabilität nach Holsti fällt die automatische Codierung im Mittel nur wenig (-.05) schlechter aus als die manuelle (vgl. Tabelle 7.2).

Betrachtet man die Koeffizienten für Krippendorffs α in Tabelle 7.1, relativieren sich diese optimistischen Erwartungen erheblich. So sind die Werte nicht nur für alle Variablen grundsätzlich niedriger, sondern die Re-

Tabelle 7.2: Vergleich von Intercoder- und Klassifikationsreliabilität

Variable	CR_a	$CR_a - CR_m$	α_a	$\alpha_a - \alpha_m$
Politik	.86	-.05	.65	-.14
Bundesdeutsche Politik	.86	-.04	.55	-.14
Politik International	.89	-.04	.61	-.15
Wirtschafts-/Finanzpolitik	.90	-.03	.61	-.13
Bundestagswahl 2009	.98	.01	.18	-.30
Sport	.96	-.03	.84	-.14
Kultur	.91	-.04	.09	-.59
Unglücke/Katastrophen/Unfälle	.93	-.02	.17	-.50
Kriminalität	.86	-.06	.36	-.31
Kontroverse	.62	-.07	.30	-.19
Prominenz	.60	-.11	.45	-.27
Journal. Stil	.45	-.08	.31	-.05

Subskripte a und m bezeichnen die automatische bzw. manuelle Codierung.

liabilität streut auch erheblich stärker über die Variablen. Würde man die gleichen Daumenregeln für die Interpretation der Alpha-Koeffizienten wie für die Prozentübereinstimmung befolgen, kann man nur die automatische Codierung der Variable *Sport* als zuverlässig bezeichnen. Die übrigen Themenvariablen werden mittelmäßig reliabel codiert, die Nachrichtenfaktoren nochmals weniger zuverlässig. Insgesamt fällt die automatische Codierung im Durchschnitt erheblich schlechter aus als die manuelle (-.24). Angesichts dieser unterschiedlichen Ergebnisse stellen sich zwei Fragen: Wie kommt die Differenz zwischen den Maßen zustande und welchen Werten sollte man beim Vergleich mit anderen inhaltsanalytischen Studien folgen?

Letzteres lässt sich einfach beantworten: Solange in Veröffentlichungen nur die Werte nach Holsti angegeben werden (Lauf, 2001), kann man auch nur diese Koeffizienten mit den hier ermittelten vergleichen. Sinnvoller ist jedoch der Vergleich der korrigierten Koeffizienten, da diese sehr viel stärker differenzieren und auch ohne weitere Angaben, etwa zur Verteilung der Variablen, interpretierbar sind.

Die beobachteten Differenzen zwischen Prozentübereinstimmung und den Koeffizienten für Alpha weisen in zweifacher Weise auf die bereits erwähnte Problematik hin: Erstens wird durch die Zufalls- und Verteilungskorrektur bei Krippendorff die Reliabilität der Codierung anders begründet. Es zählt nicht mehr die Wahrscheinlichkeit, ein beliebiges Dokument richtig zu klassifizieren, sondern die Wahrscheinlichkeit, ein Dokument aus der seltensten Kategorie richtig zu codieren. Dies zeigt sich deutlich bei den Variablen mit einer niedrigen Auftretenshäufigkeit, zum Beispiel *Bundestagswahl 2009*: Da 98 Prozent aller Dokumente nicht die Wahl thematisieren, erreicht selbst ein Klassifikator, der immer 0 codiert, eine hohe Übereinstimmung. Damit ist allerdings nichts über die Qualität des Verfahrens gesagt, die sich erst bei der richtigen Codierung der wenigen Dokumente, in denen die Wahl thematisiert wird, zeigt. Und hier versagt der Klassifikationsalgorithmus, wie später noch deutlicher wird.

Zweitens gibt es auch einen substanziellen Grund, warum die Reliabilität bei den Variablen mit schiefer Verteilung schlechter sein könnte als bei gleichmäßig verteilten Variablen: Es fehlt schlicht an ausreichend Trainingsmaterial, um ein statistisches Modell der Kategorienzuordnung zu entwickeln. Für die Variable *Bundestagswahl* lagen nur 18 positive Dokumente vor, von denen pro Evaluationslauf ca. 15 zum Training zur Verfügung standen und drei als Testdokumente verwendet wurden. Die Fehlklassifikation dieser drei Dokumente wiegt bei der Berechnung von Krippendorffs α deutlich schwerer als die richtige Einordnung aller anderen Beiträge. Das seltene Vorkommen der Kategorie hat somit zweifach negative Konsequenzen.

Der Mangel an Trainingsdaten ist jedoch keinesfalls die einzige Ursache für unterschiedliche Reliabilitäten, wie ein Vergleich der Themenvariablen und der Nachrichtenfaktoren zeigt. Hypothese 1 ging davon aus, dass letztere weniger zuverlässig automatisch zu codieren sind, da Nachrichtenfaktoren stärker auf dem Kontextwissen der Codierer basieren, das vom Klassifikator nur langsam im Training erworben werden kann. Diese Erwartung wird durch die Ergebnisse bei den Variablen *Kontroverse* und *Unglücke/Unfälle* bestätigt, die für den Klassifikationsalgorithmus deutlich schwerer zu codieren waren. Die Tatsache, dass ein Ereignis

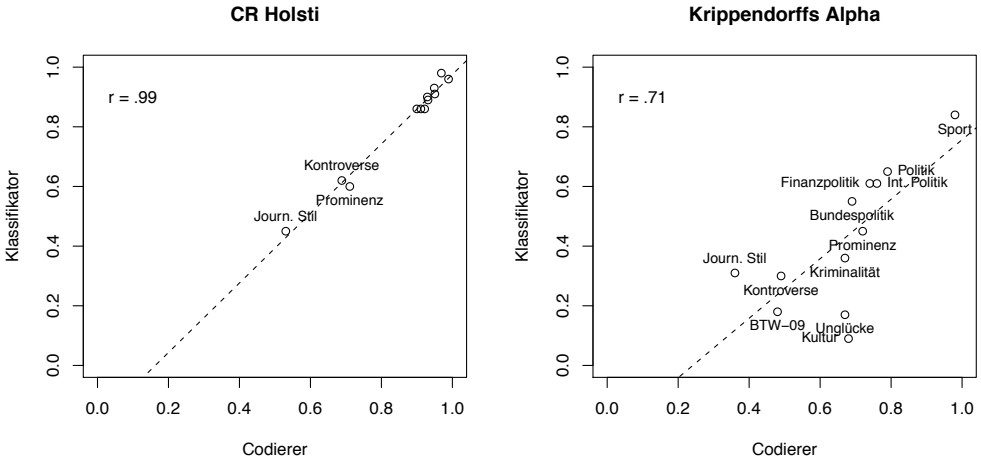

Abbildung 7.1: Zusammenhang zwischen Intercoder- und Klassifikationsreliabilität

kontrovers oder eine Handlung kriminell ist, lässt sich offenbar nur unzureichend auf der Wortebene ermitteln. Diese semantischen Feinheiten der Texte gehen bei der automatischen Klassifikation verloren.

Interessant ist die Variable *Journalistischer Stil*, weil diese zwar in absoluter Höhe unzuverlässig vom Klassifikator codiert wurde, die automatische Reliabilität jedoch nicht viel schlechter ist als die der manuellen Codierung. Hypothese 2 besagte, dass die Reliabilität manueller und automatischer Codierung zusammenhängt. Um dies für alle Variablen des Codebuchs zu testen, habe ich den Zusammenhang zwischen Intercoder- und Klassifikationsreliabilität untersucht. Wie auf der linken Seite von Abbildung 7.1 zu erkennen ist, zeigt sich bei der Verwendung der Prozentübereinstimmung nach Holsti ein fast perfekter linearer Zusammenhang. Die Ergebnisse bezogen auf Krippendorffs α sind wiederum deutlich differenzierter: Zwar gibt es auch hier einen linearen Zusammenhang, dieser ist jedoch nicht ganz so stark ($r = .71$). Indes sind auch zahlreiche Abweichungen von dieser Regel zu beobachten. Die Themenvariablen aus *Politik* und *Sport* sind zumeist zuverlässiger automatisierbar als vorausgesagt, während die Reliabilität der Variablen *Unglücke/Unfälle* und *Kultur*

deutlich schlechter ist als man angesichts der Intercoder-Reliabilität erwarten konnte. Auch hier spielt mit hoher Wahrscheinlichkeit die geringe Menge an Trainingsmaterial eine Rolle. Grundsätzlich kann jedoch Hypothese 2 als bestätigt angesehen werden.

Um die Stärken und Schwächen der überwachten Klassifikation besser einschätzen zu können, bietet sich ein Blick auf die Maße Precision und Recall an, die nach Krippendorff (2004a) als Indikatoren für die semantische Validität der Analyse gelten können. Beide Koeffizienten basieren auf der Annahme, dass die manuelle Codierung einen Goldstandard darstellt, den der Klassifikationsalgorithmus bestmöglich replizieren muss.[1] Auch wenn diese Annahme angesichts der messfehlerbehafteten manuellen Codierung nicht haltbar ist, bieten die Validitätsmaße wertvolle zusätzliche Informationen zur Klassifikationsqualität. Da die Berechnung beider Koeffizienten nur für dichotome Variablen sinnvoll ist, wurden bei den ordinalen Variablen alle Codes größer Null zusammengefasst, die Variable *Journalistischer Stil* wurde aus diesem Grunde aus der Analyse ausgeschlossen.

Aus den ersten beiden Spalten von Tabelle 7.3 geht hervor, dass die Klassifikation bei den meisten Kategorien recht präzise ist. Präzision bezeichnet hier die Wahrscheinlichkeit, dass ein automatisch positiv codierter Fall auch tatsächlich positiv ist. Die vom Klassifikator positiv codierten Dokumente der Kategorie *Unglücke/Unfälle* waren zu 78 Prozent auch von den Codierern so eingeschätzt worden. Umgekehrt ist zu erkennen, dass bei der Klassifikation von *Kultur*, *Kontroverse* und *Bundestagwahl* bis zu 40 Prozent falsch positive Dokumente anfallen.[2] Insgesamt kann auch bei den unzuverlässigeren Nachrichtenfaktor-Variablen davon ausgegangen werden, dass die vom Klassifikator positiv codierten Dokumente auch tatsächlich zur Positiv-Kategorie gehören.

Anders fallen die Ergebnisse bei der Betrachtung des Recall, d.h. der Trefferquote aus, die die Wahrscheinlichkeit bezeichnet, dass ein tatsäch-

1 Diese Annahme gilt auch für den Koeffizienten *F*, der den harmonischen Mittelwert von Precision und Recall darstellt.

2 An den großen Konfidenzintervallen der Variablen *Kultur* und *Bundestagswahl* sieht man, dass die Werte erheblich schwanken, was wiederum auf die Komposition von Trainings- und Test-Set und letztlich die Verteilungsproblematik zurückzuführen ist.

Tabelle 7.3: Precision und Recall der überwachten Klassifikation

Variable	Prec.	CI_{Pr}	Rec.	CI_{Re}	F	CI_F
Politik	.73	.72–.74	.78	.77–.79	.75	.75–.76
Bundesdeutsche Pol.	.65	.63–.67	.63	.61–.64	.64	.63–.65
Politik International	.77	.74–.81	.60	.58–.62	.68	.66–.70
Wirtschaftspolitik	.65	.63–.67	.69	.66–.73	.67	.65–.70
Bundestagswahl 2009	.59	.31–.87	.11	.04–.19	.19	.07–.30
Sport	.94	.91–.98	.80	.78–.81	.85	.84–.88
Kultur*	.61	.27–.94	.08	.05–.10	.14	.09–.18
Unglücke/Unfälle	.78	.64–.91	.12	.06–.18	.21	.12–.29
Kriminalität	.66	.62–.70	.32	.29–.35	.43	.41–.46
Kontroverse*	.62	.60–.64	.52	.49–.54	.56	.54–.59
Prominenz*	.73	.72–.74	.63	.61–.65	.68	.66–.69

Ausgewiesen sind Werte für dichotome bzw. dichotomisierte (*) Variablen.
$n = 933$

lich positives Dokument auch positiv codiert wird. Die teils dramatisch niedrigen Recall-Werte zeigen deutlich, dass in vielen Fällen relevante (positive) Dokumente nicht vom Klassifikator erkannt werden. Besonders augenfällig ist dies wiederum bei der Variable *Unglücke/Unfälle*, die eine hohe Präzision, aber einen niedrigen Recall aufweist. Offenbar gelingt es dem Algorithmus nicht, alle Unglücksfälle zuverlässig zu identifizieren, die Falsch-Negativ-Rate ist also sehr hoch. Dies ist vor allem dann problematisch, wenn die automatische Klassifikation zur Selektion von Beiträgen für eine manuelle Inhaltsanalyse genutzt werden soll. Während eine falsch positive Klassifikation in der Regel einfach von den Codierern oder dem Forschungsleiter korrigiert werden kann, die das Dokument dann aus der Stichprobe entfernen, führt ein geringer Recall zu einer systematischen Unterrepräsentation relevanter Dokumente. Wenn diese gar nicht weiterverarbeitet werden, ist die Validität der Analyse entsprechend gering. Im Gegensatz zur Schlagwortsuche in Datenbanken (Hagen, 2001) lässt sich bei der Nutzung von NewsClassifier die Trefferquote zumindest bestimmen.

Da die Reliabilitäts- und Validitätsmaße auf denselben Übereinstimmungsdaten basieren, ist es wenig verwunderlich, dass gerade die zuverlässigen Themenvariablen auch ausgeglichene Precision-Recall-Werte aufweisen. Der Anteil an falsch positiven und falsch negativen Dokumenten ist über alle Politikvariablen hinweg ausgeglichen, lediglich im Bereich *internationaler Politik* werden deutlich mehr Beiträge falsch negativ klassifiziert. Bei allen anderen Variablen ist der Recall deutlich niedriger als die Präzision. Wäre der Klassifikator ein normaler Codierer, könnte man sagen, er codiere sehr vorsichtig. Im Zweifelsfall werden nur eindeutig positive Dokumente erkannt und die übrigen in die Nullkategorie eingeordnet.

7.1.2 Einfluss von Preprocessing und Texteigenschaften

In den Hypothesen 3 und 4 geht es um die Frage, von welchen Faktoren die Qualität der überwachten Textklassifikation abhängt. Hierzu habe ich in einem faktoriellen Experiment den Einfluss von vier unterschiedlichen Treatments auf die Klassifikation der zwölf Variablen aus dem Codebuch untersucht. Da es sich hierbei um ein Datenmodell mit zwei Ebenen handelt (Treatments und Codebuch-Variablen), stellt eine hierarchische Regression das adäquate statistische Verfahren zur Analyse des Experiments dar (Gelman & Hill, 2007). Mit diesem Modell lässt sich in einem ersten Schritt prüfen, ob die Experimentalbedingungen oder das Codebuch für die Variation in der abhängigen Variable, z.B. dem Reliabilitätskoeffizienten Alpha, verantwortlich sind. Separiert man die drei Varianzanteile des Modells (Treatments, Variablen, Residuen), zeigt sich hier ein eindeutiges Ergebnis, das die alte Weisheit der Inhaltsanalyse bestätigt: Sie steht und fällt mit den Variablen des Codebuchs. Konkret lassen sich 91 Prozent der Varianz in der Reliabilität auf die unterschiedlichen Variablen zurückführen, während die vier Treatments Text-Extraktion, Stemming, Stopwort- und Überschriftentfernung nur rund 4 Prozent der Variabilität erklären können.[3] Dies bedeutet, dass sich

3 Dies bedeutet auch, dass das hierarchische Regressionsmodell mit Krippendorffs α als abhängiger Variable rund 95 Prozent der Varianz aufklären kann und somit die Daten hervorragend abbildet. Die R^2 Werte für die anderen drei abhängigen Variablen CR (Holsti), Precision und Recall liegen sogar noch höher.

die Reliabilität der überwachten Codierung nur wenig durch Preprocessing beeinflussen lässt, oder positiv formuliert, dass die Klassifikation mit OSBF-Lua robust gegenüber einer Vorbehandlung des Stimulusmaterials ist.

Die im Folgenden vorgestellten Detailanalysen basieren alle auf dem gleichen hierarchischen Modell (mit jeweils unterschiedlicher abhängiger Variable), das in der Terminologie von Gelman & Hill (2007) als *Varying-Intercept-Varying-Slope*-Modell mit den Codebuchkategien als Gruppenvariable j bezeichnet werden kann. Bei dieser Modellklasse dürfen sowohl die Mittelwerte (Intercept-Terme) als auch die Regressionsgewichte in Abhängigkeit der jeweiligen Kategorie variieren. Die Ergebnisse dieses *Random-Effects*-Modells gliedern sich grob in drei Gruppen von Koeffizienten:

1. Die sog. *Fixed Effects*, d.h. der über alle Gruppen geschätzte Intercept-Wert α, der die mittlere Reliabilität des Modells ohne Treatments wiedergibt, sowie die vier Treatment-Haupteffekte β_i. Zusätzlich zu den Haupteffekten habe ich auch alle Zwei-Wege-Interaktionseffekte geschätzt, wie dies bei faktoriellen Designs üblich ist (Box et al., 1978). Aus Gründen der Modellsparsamkeit habe ich jedoch auf die Schätzung von Interaktionseffekten höherer Ordnung verzichtet.
2. Die sog. *Random Effects*, d.h. der für jede Codebuch-Variable spezifische Intercept-Term α_j sowie die gruppenspezifischen Treatment-Effekte β_{ij}, wobei auch hier sowohl Haupt- als auch Interaktionsterme geschätzt werden. Diese variierenden Koeffizienten zeigen, ob bestimmte Treatments verschiedene Auswirkungen auf die Klassifikationsqualität der unterschiedliche Variablen haben. Dies wäre z.B. der Fall, wenn die Verwendung von Stemming bei *Sport* einen positiven, bei *Kultur* jedoch einen negativen Einfluss auf die Klassifikation hätte.
3. Die *Korrelationen* zwischen den variierenden Koeffizienten, insbesondere zwischen Intercept und Regressionsgewichten. Anhand dieser lässt sich beispielsweise erkennen, ob bestimmte Treatments eher bei gut oder bei schlecht automatisierbaren Kategorien einen Effekt haben. Da diese Korrelationen bei nur zwölf Kategorien nicht

sehr aussagekräftig sind, werden sie im Folgenden nicht weiter behandelt. Die Ergebnisse sind jedoch im Anhang B.1 dokumentiert.

Obwohl die Analyse dieser Evaluationsstudie auf einem relativ einfachen Regressionsmodell mit lediglich 10 Prädiktoren (4 Haupt- und 6 Interaktionseffekte) basiert, gestaltet sich die Darstellung der Ergebnisse aufgrund der Vielzahl von Koeffizienten und deren Standardfehlern recht schwierig. Ich werde daher zuerst die Fixed Effects erläutern, die den Ergebnissen eines globalen Regressionsmodells über alle Kategorien entsprechen. Anschließend werden die vier Faktoren des Experiments detaillierter hinsichtlich ihres Einflusses auf die Klassifikationsqualität bei den einzelnen Kategorien dargestellt. Da die tabellarische Darstellung aller Koeffizienten und ihrer Konfidenzintervalle die Interpretation der Ergebnisse eher erschwert, werde ich zudem weitgehend auf eine grafische Präsentation in Form von Koeffizientenplots zurückgreifen und damit den Empfehlungen von Gelman et al. (2002) und Kastellec & Leoni (2007) folgen. In Abbildung 7.2 sind die unstandardisierten Regressionsgewichte der Treatment-Variablen und deren 95-Prozent-Konfidenzintervalle abgebildet. Diese Darstellung hat den Vorteil, dass man sowohl die Effekte untereinander vergleichen als auch auf den ersten Blick erkennen kann, ob die Null-Achse außerhalb des Konfidenzintervalls liegt und der Koeffizient damit statistisch signifikant ist.

Allgemeiner Einfluss des Preprocessing

Die in Hypothese 3 formulierte Erwartung, dass das Preprocessing einen positiven Effekt auf die Klassifikationsqualität hat, zeigt sich nicht in den Ergebnissen: Auf den ersten Blick fällt auf, dass fast alle Koeffizienten negativ sind. Dies bedeutet nichts anderes, als dass das Baseline-Modell mit bereinigtem Fließtext, aber ohne weiteres Preprocessing, über alle Variablen hinweg die höchste Reliabilität verspricht. Der einzige Schritt im Prozess der Dokumentenverarbeitung, der einen eindeutig positiven Effekt auf die Zuverlässigkeit der überwachten Klassifikation hat, ist die Extraktion des Fließtextes aus dem rohen HTML-Code[4]. Verzichtet man auf diesen Bereinigungsschritt, fällt die Reliabilität der Klassifikation im

4 Da dies der Standard bei der Codierung ist, muss man die stark negativen Koeffizienten der Rohtextcodierung entsprechend umgekehrt interpretieren.

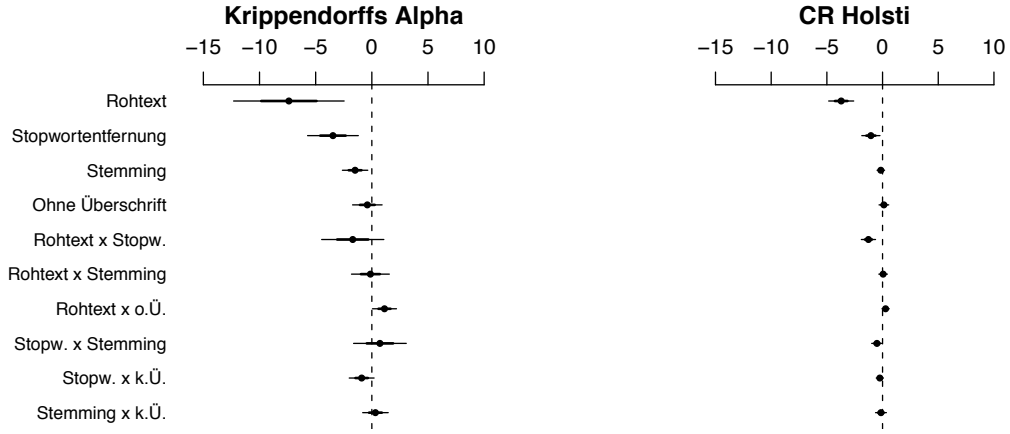

Abbildung 7.2: Fixe Effekte der Treatments auf die Klassifikationsreliabilität, unstandardisierte Regressionskoeffizienten und 95%-Konfidenzintervalle

Mittel rund 7 Prozent (bzw. 4 Prozent bezogen auf die Prozentübereinstimmung nach Holsti) niedriger aus.[5] Vergleicht man linke und rechte Seite von Abbildung 7.2, sind die Konfidenzintervalle bei der Prozentübereinstimmung nach Holsti deutlich kleiner. Das liegt daran, dass die Varianz dieses Wertes über alle Treatments und Kategorien hinweg deutlich geringer ist als bei der Verwendung von Krippendorffs α. Diese fehlende Sensitivität ist neben der Zufalls- und Verteilungsproblematik ein weiterer Grund, der gegen die Verwendung des Holsti-Koeffizienten für Reliabilitätsanalysen spricht.

Das Herausfiltern der häufigsten deutschen Wörter hat insgesamt einen signifikant negativen Einfluss auf die Reliabilität der Klassifikation. Dies kann man damit erklären, dass die Wörter für sich genommen nicht zwischen den unterschiedlichen Kategorien unterscheiden helfen, aber in Kombination mit anderen Wörtern durchaus Bedeutung haben

5 Um die Interpretation der Ergebnisse zu erleichtern, habe ich die abhängige Variable in allen Modellen mit 100 multipliziert, so dass die Kennziffern als Prozentwerte zu interpretieren sind.

können, die spezifisch für bestimmte Kategorien sind. Da der verwendete Algorithmus auf Basis dieser N-Gramme ein Modell entwickelt, schadet die Entfernung dieser Stopwörter eher bei der Klassifikation.

Das vielfach zur Reduktion der Komplexität empfohlene Stemming-Verfahren, bei dem alle Wortformen auf einen (künstlichen) Stamm gekürzt werden, hat im Mittel einen leicht negativen bzw. gar keinen Effekt auf die Klassifikation. Dies hat vor allem für die Forschungspraxis Bedeutung, da das Stemming nicht nur relativ aufwändig, sondern auch sprachspezifisch ist. Angesichts der Ergebnisse dieser Evaluation könnte man ohne weiteres dazu raten, diesen Preprocessing-Schritt auszulassen, was die Verwendung von NewsClassifier für mehrsprachige Inhaltsanalysen noch leichter macht.

Insgesamt lässt sich zumindest über alle Variablen des Codebuchs hinweg zusammenfassen, dass sich kein positiver Effekt von Stemming und Stopwortentfernung nachweisen lässt, sondern eher ein negativer. Dagegen kann man davon ausgehen, dass sich die Textextraktion positiv auswirkt, auch weil der Speicherbedarf und die Klassifikationsgeschwindigkeit durch diese Maßnahme deutlich gesenkt werden. Die Preprocessing-Maßnahmen sind daher nur zum Teil sinnvoll, jedenfalls wenn man die *Fixed Effects* des Modells betrachtet.

Die letzte experimentelle Variation bezieht sich auf die Komposition der codierten Mitteilungen. In Hypothese 4 bin ich davon ausgegangen, dass die Überschrift besonders wichtig für die Klassifikation der Artikel ist, da in dieser die Substanz der Meldung zusammengefasst wird. Dies zeigt sich allerdings nicht in den Ergebnissen. Zwar ist die Klassifikation minimal schlechter, wenn die Überschrift nicht bei der Codierung einbezogen wird, der Effekt ist jedoch nicht von zufälligen Abweichungen zu unterscheiden. Kurz gesagt macht es keinen Unterschied, ob die Überschriften der Beiträge mitcodiert werden oder nicht. Offenbar steckt die für die Klassifikation relevante Information vor allem im Fließtext der Meldung.

Die Interpretation der fixen Interaktionseffekte kann an dieser Stelle sehr knapp gehalten: Keine Treatment-Kombination hat einen substanziell positiven Effekt, die meisten haben keinen signifikanten Einfluss auf die Reliabilität der Klassifikation. Auch für die Validitätskoeffizi-

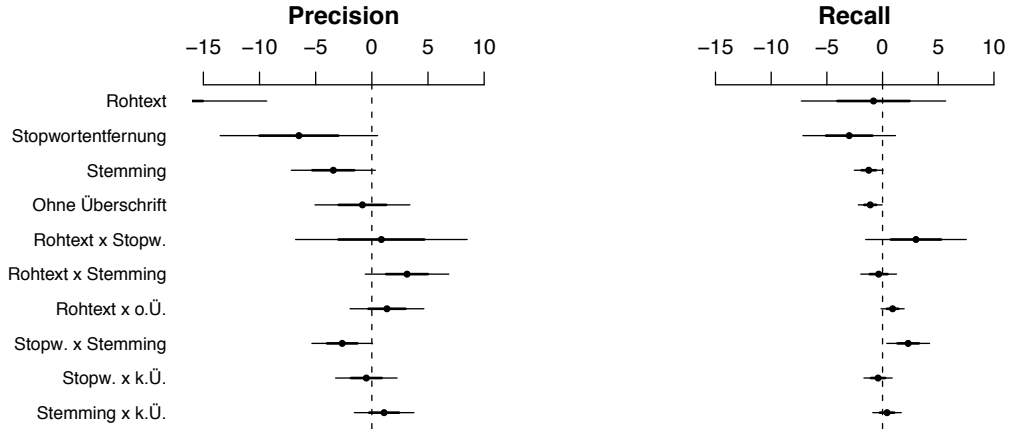

Abbildung 7.3: Fixe Effekte der Treatments auf die Klassifikationsvalidität, unstandardisierte Regressionskoeffizienten und Konfidenzintervalle

enten Precision und Recall lassen sich keine bedeutsamen Treatment-Kombinationen identifizieren, wie in Abbildung 7.3 ersichtlich. Lediglich die gemeinsame Verwendung von Stopwortentfernung und Stemming hat im Mittel einen leicht positiven Einfluss auf den Recall, der die negativen Haupteffekte der beiden Preprocessing-Verfahren jedoch nicht kompensieren kann. Gerade für die Trefferquote ist zu konstatieren, dass diese sich durch kein Treatment bedeutend verbessern lässt. Das Risiko einer falsch negativen Klassifikation lässt sich über alle Variablen hinweg mit den hier untersuchten Mitteln nicht senken. Hingegen profitiert die durchschnittliche Präzision von der Textextraktion: Der Anteil falsch positiv codierter Dokumente ist bei der Verwendung unbereinigter HTML-Seiten rund 21 Prozent höher als bei bereinigten Texten.

Effekte bei individuellen Variablen des Codeplans

Ein Vorteil von Mehrebenenmodellen liegt in der Schätzung von Regressionskoeffizienten für jede Ausprägung der Gruppenvariable. In dieser Studie werden daher neben den mittleren Treatment-Effekten auch deren Einflüsse auf die Klassifikation für jede Codebuchvariable

Tabelle 7.4: Standardabweichungen der Random Effects

	α	CR Holsti	Precision	Recall
Rohtext (HTML)	8,38	1,89	18,64	10,67
Stopwortentfernung	3,59	1,34	11,27	6,77
Stemming	1,19	0,35	5,51	1,71
Keine Überschrift (k.Ü.)	1,66	0,59	6,40	1,21
Rohtext × Stopw,	4,53	1,00	12,34	7,38
Rohtext × Stemming	2,47	0,47	5,47	2,32
Stopw. × Stemming	1,00	0,32	4,65	1,13
Rohtext × k.Ü.	3,75	0,67	3,44	2,91
Stopw. × k.Ü.	1,12	0,30	3,52	1,66
Stemming × k.Ü.	1,20	0,70	3,35	1,68

geschätzt. Dies ermöglicht es, zu untersuchen, ob die oben dargestellten gruppenübergreifenden Effekte homogen sind, d.h. auf alle Variablen der Inhaltsanalyse zutreffen, oder es unterschiedliche Auswirkungen gibt, die sich im Mittel ausgleichen. Da die Darstellung von zehn variierenden Regressionskoeffizienten bei zwölf Kategorien und vier Kriteriumsvariablen, also insgesamt 480 Punktschätzer und ebenso vielen Standardfehlern, den Blick auf die relevanten Ergebnisse eher erschwert, beginne ich mit einer zusammenfassenden Darstellung der Variabilität der Effekte.

In Tabelle 7.4 sind die Treatment-Variablen sowie die Standardabweichung ihrer Koeffizienten dargestellt. Anders als die in Abbildung 7.2 dargestellten Standardfehler haben diese keine inferenzstatistische Bedeutung, sondern stellen die empirische Variabilität der Koeffizienten über die zwölf inhaltlichen Variablen dar. Eine große Standardabweichung besagt dabei, dass sich die Effekte der Treatments je nach inhaltlicher Kategorie erheblich unterscheiden, etwa indem gleichermaßen positive und negative Koeffizienten vorkommen. Ein kleiner Wert für die Standardabweichung der Effekte ist als Beleg für eine homogene Wirkung zu verstehen, d.h. die Wirkung der Treatments hängt nicht von den Klassifikationskategorien ab.

Die größte Heterogenität in der Wirkung lässt sich bei der Verwendung von Rohtext beobachten, sowohl der Haupteffekt als auch die Interaktionseffekte mit anderen Treatments schwanken erheblich. Trotz des im Mittel negativen Effekts gibt es offenbar auch Kategorien, die sich besser anhand der unbereinigten HTML-Seiten klassifizieren lassen. Während der Effekt von Stemming relativ homogen ist, hat die Stopwortentfernung keine uniforme Wirkung auf die Klassifikationsqualität. In manchen Fällen hat sie einen stark negativen Einfluss auf die Präzision der Klassifikation, in anderen einen stark positiven. Das Weglassen der Überschrift entfaltet dagegen kaum eine differenzierte Wirkung.

Für die detaillierte Darstellung der Random Effects habe ich wiederum auf Koeffizientenplots zurückgegriffen, da diese alle relevanten Information des Regressionsmodells anschaulich wiedergeben. In den folgenden vier Abbildungen sind die Effekte der Treatments bezogen auf die Reliabibilitäts- (links) und die Validitätsmaße (rechts) dargestellt.[6] Da es zudem jeweils zwei abhängige Variablen gibt, sind die Koeffizienten für die Reliabilität nach Holsti rot (im Druck grau) dargestellt, Krippendorffs α schwarz. In der rechten Hälfte ist entsprechend Recall rot bzw. grau und Precision schwarz.

Effekte der Rohtext-Klassifikation

Wie Abbildung 7.4 zeigt, lässt sich für alle Variablen außer *Sport* und *Kultur* ein negativer bzw. zumindest kein positiver Einfluss der Berücksichtigung des HTML-Quelltextes bei der Klassifikation der Nachrichten nachweisen. In fast allen Fällen zeigt sich dasselbe Ergebnis wie bei den Fixed Effects, dass nämlich die Verwendung von unbereinigtem HTML der Klassifikationsreliabilität schadet. Der abweichende positive Effekt bei der Variable *Kultur* lässt sich einerseits statistisch durch die schiefe Verteilung und damit unsichere Reliabilitätsschätzung erklären, andererseits ist es plausibel, dass sich bei der Nutzung von HTML-Daten die Ressort-Angabe Kultur im Header oder der Artikel-URL befindet, die dann natürlich ein hilfreicher Indikator für die Klassifikation ist.[7] Da

6 Angesichts der schwachen fixen Interaktionseffekte und deren geringer Variabilität verzichte ich auf ein ausführliche Darstellung.

7 Diese Vermutung gilt auch für *Sport*, hier ist jedoch die Baseline-Reliabilität höher und der Zuwachs durch Preprocessing ingesamt geringer.

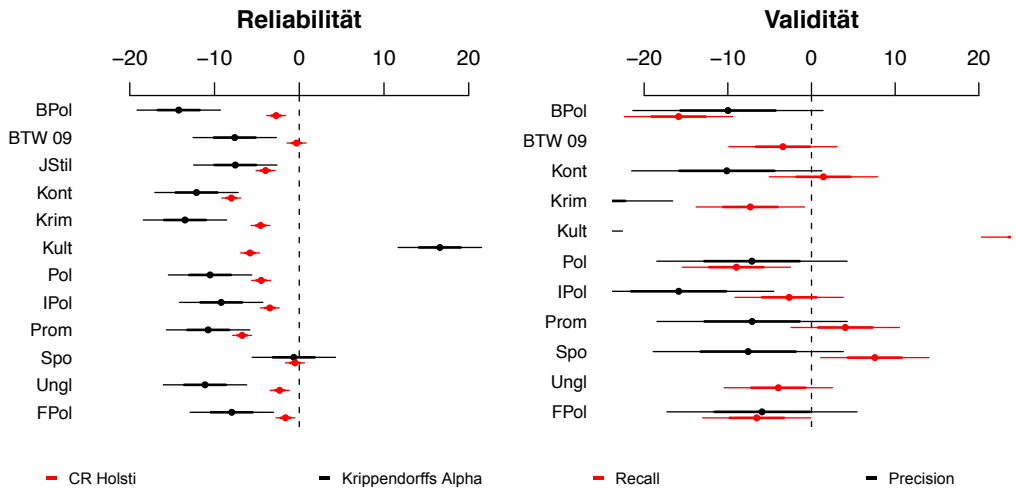

Abbildung 7.4: Effekte der Rohtextcodierung auf die Klassifikationsqualität, unstandardisierte Regressionskoeffizienten und Konfidenzintervalle

die Ausgangsreliabilität nach Krippendorff zudem bei *Kultur* fast Null ist, führt auch eine Steigerung um 15 Prozentpunkte noch nicht zu einer zuverlässigen Klassifikation.

Auf der rechten Seite der Abbildung wird ersichtlich, dass die Verwendung von unbereinigtem HTML-Text fast durchgängig negative Auswirkungen auf Precision und Recall hat: Die Präzision ist bei allen Variablen niedriger, die Trefferquote verbessert sich lediglich bei *Sport* und *Kultur* signifikant. Für die Variable *Kultur* lässt sich anhand der rechten Abbildung der sog. Precision-Recall-Tradeoff erkennen: Während die Präzision um rund 24 Prozent abnimmt, wird die Trefferquote (die im Baseline-Modell bei nur 8 Prozent liegt) um 27 Prozent gesteigert. Beide Werte sind als Ausreißer nur noch andeutungsweise in der Abbildung zu sehen. Für die Variable *Kultur* lässt sich so ein insgesamt positiver Effekt bei der Verwendung von Rohtexten konstatieren, für alle anderen Variablen empfiehlt sich die Verwendung von extrahierten Texten, die zu reliableren und valideren Klassifikationen führt.

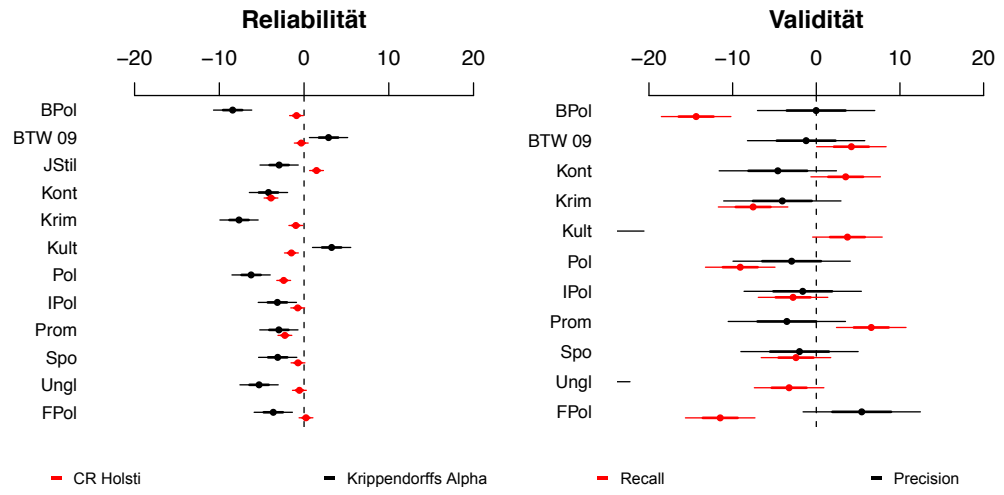

Abbildung 7.5: Effekte der Stopwortentfernung auf die Klassifikationsqualität, unstandardisierte Regressionskoeffizienten und Konfidenzintervalle

Effekte der Stopwortentfernung

Die Entfernung der 1000 häufigsten deutschen Wörter hat insgesamt weniger drastische Auswirkungen auf die Klassifikationsreliabilität als die Textextraktion, ist aber für viele Variablen eher schädlich (vgl. Abbildung 7.5) . Die einzig positiven Ausnahmen bilden die ohnehin unzuverlässigen Variablen *Kultur* und *Bundestagswahl 2009*, die aber auch nach dem Preprocessing deutlich unbefriedigende Reliabilitätswerte aufweisen. Im Bereich der Validitätsmaße sind etwas stärkere Effekte zu beobachten, die allerdings nur für wenige Variablen gelten. Die Entfernung häufiger Wörter hat insgesamt negative Auswirkungen auf die Trefferquote, vor allem bei den Politikvariablen. Nur bei *Prominenz* sinkt durch die Maßnahme der Anteil falsch negativer Klassifikationen. Die Präzision der Klassifikation ist ebenfalls durchgängig niedriger, wenn häufige Wörter ausgefiltert werden. Dies gilt insbesondere für die Variablen *Kultur* und *Unglücke/Unfälle*, deren Klassifikation mehr als 20 Prozent weniger präzise ausfällt.

Angesichts dieser Ergebnisse spricht kaum etwas für eine generelle Stopwortentfernung, auch wenn diese in Einzelfällen leicht positive Auswirkungen auf die Qualität der Klassifikation haben kann. Dies gilt umso mehr für Situationen, in denen eine Filterliste nicht ohne weiteres verfügbar ist, z.B. für fremd- oder gemischtsprachliche Dokumente. Solange die Daten automatisch weiterverarbeitet werden und die Feature-Zahl keine große Rolle spielt, sollte auf eine Entfernung häufiger Wörter verzichtet werden.

Effekte des Stemming
Die Random Effects des Stemming in Abbildung 7.6 zeigen nur geringe Variation um den Wert Null, so dass selbst bei statistisch signifikanten Koeffizienten kaum von substanziellen Wirkungen ausgegangen werden kann. Stemming verändert die Reliabilität der automatischen Klassifikation nur minimal. Auch bezogen auf die Validität zeigen sich keine bedeutsamen Koeffizienten, lediglich auf die Präzision bei den Variablen *Bundestagswahl 2009* und *Unglücke/Unfälle* hat das Stemming einen stark negativen Einfluss. Im Gegensatz zu den Ergebnissen von Braschler & Ripplinger (2004) ist für das Gebiet der Textklassifikation kein positiver Effekt des Stemmings festzustellen. Angesichts des Rechenaufwands für das algorithmische Stemming kann die Schlussfolgerung für die behandelten Variablen nur lauten, diesen Schritt schlicht wegzulassen. Dies deckt sich mit den Ergebnissen von Leopold & Kindermann (2002, 438), die statt des einfachen Stemming sogar eine echte sprachspezifische Lemmatisierung verwenden.

Auswirkung der Klassifikation ohne Überschrift
Die Ergebnisse bezüglich der Einbeziehung der Überschrift bei der Textklassifikation ähneln stark denjenigen beim Stemming: Die Effekte sind durchgängig sehr klein, so dass Hypothese 4 als widerlegt gelten kann. Auch ohne Verwendung der Überschrift ist die Klassifikation so zuverlässig und valide wie mit dieser. Relevante Textinformationen sind augenscheinlich in allen Teilen des Dokuments enthalten. Wiederum stellt die Variable *Kultur* eine Ausnahme dar, hier lässt sich tatsächlich eine niedrigere Präzision beobachten, wenn die Überschrift dem Klassifikationsalgorithmus nicht vorgegeben wird. Bei *Unglücken/Unfällen* ist

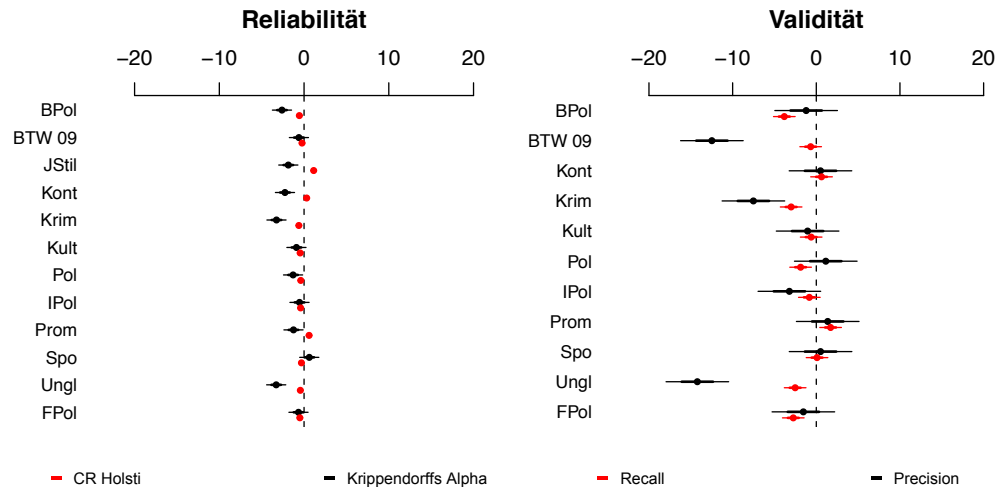

Abbildung 7.6: Effekte des Stemming auf die Klassifikationsqualität, unstandardisierte Regressionskoeffizienten und Konfidenzintervalle

hingegen eine höhere Präzision zu beobachten, wenn die Überschrift nicht berücksichtigt wird.

Zusammenfassung der Ergebnisse zum Preprocessing
Angesichts der Vielzahl an Einzelergebnissen möchte ich das faktorielle Experiment nachfolgend kurz zusammenfassen: Von den vier Treatments hat sich nur die Textextraktion als besonders wirkungsvoll erwiesen. Allerdings variieren die Wirkungen der Preprocessingmaßnahmen teilweise erheblich. Die Wirkung des Preprocessing hängt nicht nur von der jeweiligen Variable des Codebuchs ab, sondern auch von der Frage, ob man ggf. gezielt die Trefferquote oder die Präzision optimieren möchte. Die Ergebnisse des Experiments zeigen zudem, dass sowohl Stemming als auch die besondere Berücksichtigung der Beitragsüberschriften keine substanziellen Folgen für den Klassifikationsprozess haben. Bei der Betrachtung der vier verschiedenen abhängigen Variablen muss man konstatieren, dass das Preprocessing nur selten positive Effekte hat, und dann oft gegenläufige für Precision und Recall. Damit zeigen sich in

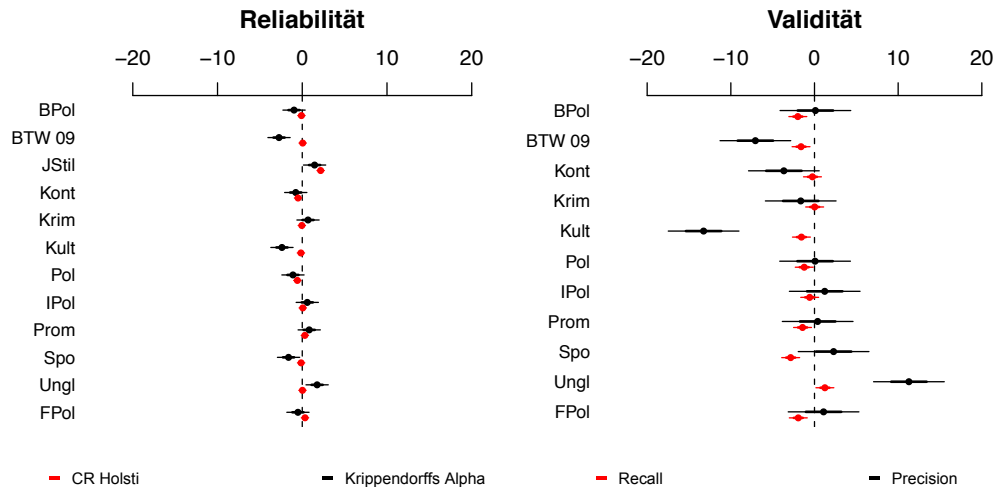

Abbildung 7.7: Effekte fehlender Überschriften auf die Klassifikationsqualität, unstandardisierte Regressionskoeffizienten und Konfidenzintervalle

dieser Studie ähnliche Ergebnisse wie bei Felden et al. (2005). Nicht zuletzt zeigen die Daten der ersten Teilstudie, dass sich die Reliabilität und Validität der Klassifikation technisch kaum steigern lassen. Die Qualität der überwachten Klassifikation hängt sehr viel stärker von der Wahl der Kategorien und der Qualität der manuellen Codierung ab als von Textkomposition und Preprocessing.

7.2 Teilstudie 2: Effektivität des Trainingsprozesses

7.2.1 Beschreibung des Lernprozesses

Im vorangegangenen Abschnitt war die forschungsleitende Frage, wie zuverlässig und valide die Ergebnisse überwachter Textklassifikation für sozialwissenschaftliche Inhaltsanalysen sind. In der zweiten Teilstudie liegt der Focus auf dem dazugehörigen Trainingsprozess, der mit der manuellen Codierung einhergeht. Zu diesem Zweck habe ich, wie in Abschnitt 6.5 beschrieben, aus den vorhandenen Codierungen einen

typischen Prozess simuliert, in dem schrittweise Dokumente für das Training des Klassifikators freigegeben werden. Nach jedem Trainingsabschnitt werden wie zuvor die Reliabilität und Validität gemessen und die entsprechenden Werte gespeichert.

Für das Baseline-Modell mit passivem Lernen, bei dem zufällig Trainingsdaten ausgewählt werden, ergeben sich bei 20 Replikationen pro Variable insgesamt 240 Verlaufskurven. Diese werden in Abbildung 7.8 zunächst grafisch dargestellt, um einen ersten Eindruck von dem maschinellen Lernprozess des Klassifikators zu vermitteln. Anschließend werde ich ein statistisches Modell der Prozesse zu entwickeln, um damit die Lernkurven unter verschiedenen Trainingsstrategien vergleichen zu können.

In Abbildung 7.8 sind die mittleren Verlaufskurven der Reliabilität und Validität pro Variable abgebildet. Die graue Linie bezieht sich hierbei auf den Reliabilitätskoeffizienten nach Holsti, die schwarze auf Krippendorffs α, die gestrichelt auf Precision bzw. die gepunktete Recall. Auf den ersten Blick wird ersichtlich, dass die Prozentübereinstimmung nur minimal mit der Anzahl an Trainingsdokumenten ansteigt und faktisch ab der ersten Messung sehr hoch ist. Dies macht nochmals deutlich, dass mit der Reliabilität nach Holsti nicht viel mehr als die Verteilung der Kategorien abgebildet wird, so dass sich das Maß nicht für weitergehende Analysen eignet. Dagegen folgt der Verlauf der Reliabilität nach Krippendorff den erwarteten Mustern, wobei sich natürlich Unterschiede zwischen den Variablen beobachten lassen. Bei allen Variablen steigt die Präzision deutlich schneller als die Trefferquote (Recall), die wiederum einen ähnlichen Verlauf wie die Reliabilität nimmt.

Am schnellsten lernt der Klassifikationsalgorithmus die Codierung der Variable *Sport* sowie der *Politik*-Variablen, wobei in jedem Fall mindestens 200 bis 300 Trainingsdokumente notwendig sind, um relativ zuverlässige Klassifikationen zu erhalten. Dasselbe Lernverhalten zeigt sich – wenn auch auf niedrigerem Niveau – bei den Variablen J*ournalistischer Stil*, *Kontroverse* und *Prominenz*. Setzt man den Anstieg der Lernkurve zum finalen Reliabilitätswert in Beziehung, ist bei diesen Variablen eine positive Korrelation erkennbar: Je höher die maximal erreichbare Zuverlässigkeit der Klassifikation, desto schneller lernt der Klassifikator auch.

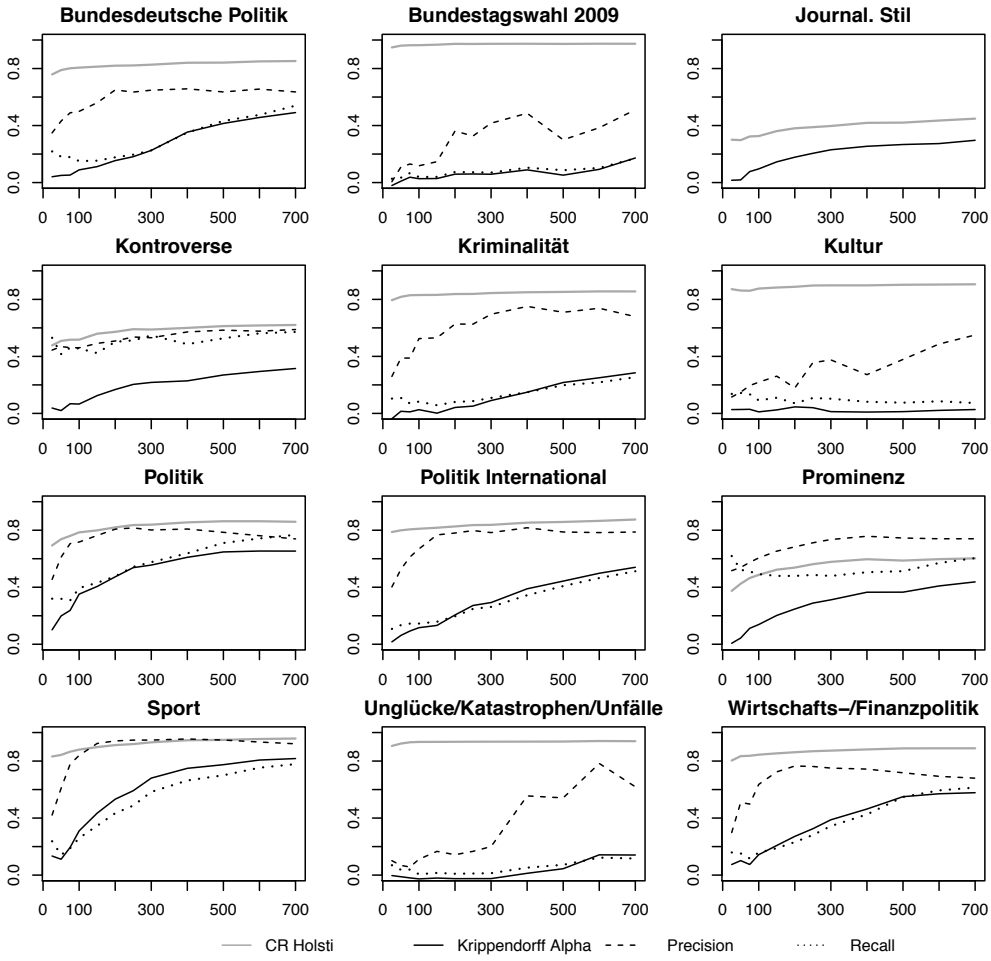

Abbildung 7.8: Entwicklung der Klassifikationsqualität mit zunehmender Anzahl an Trainingsdokumenten (passives Lernen)

Diese Erkenntnis ist für die Machbarkeitsprüfung einer überwachten Klassifikation relevant, denn anhand des Kurvenverlaufs für die ersten Trainingssätze lässt sich relativ gut abschätzen, wie hoch die Reliabilität schließlich ausfallen wird. Zudem kann so eine Entscheidung getroffen werden, ob sich zusätzliche manuelle Codierungen noch lohnen.

Während bei *Sport* und *Politik* bereits nach 500 Trainingsdokumenten kaum noch ein Zuwachs an Reliabilität zu erkennen ist, muss man für andere Variablen deutlich mehr Trainingsmaterial verwenden. So lässt sich sowohl bei den Variablen *bundesdeutsche* und *internationale Politik* als auch bei *Kriminalität* noch kein Saturierungseffekt erkennen. Ich interpretiere dies als Folge der Komplexität der Variablen: Während die Entwicklung von Klassifikationsregeln aufgrund weniger Schlüsselbegriffe bei *Sport* und *Politik allgemein* recht schnell zu guten Ergebnissen führt, sind die Unterschiede bei den anderen Kategorien subtiler. Hinzu kommt gerade bei der Variable *Kriminalität*, dass recht wenige Positiv-Fälle im Trainingsmaterial enthalten sind, so dass die Differenzierung zwischen den Kategorien schwieriger ist.

Schließlich zeigt sich auch bei der Betrachtung der Lernkurven, dass sich Variablen mit extrem schief verteilten Kategorien nicht für die überwachte Klassifikation eignen. Solange nicht eine Mindestmenge an Positiv- und Negativ-Fällen pro Kategorie vorhanden ist, kann kein statistisches Modell für die Klassifikation entwickelt werden. Dies erkennt man leicht an den Variablen *Kultur*, *Unglücke/Unfälle* und *Bundestagswahl 2009*. Hier steigt – wenn überhaupt – erst bei über 500 Trainingsdokumenten die Reliabilität substanziell an. Im Fall von *Unglücke/Unfälle* sind dann erst 35 Positiv-Dokumente in Trainings-Set, bei Bundestagswahl 2009 sogar nur zehn (vgl. die *P*-Werte in Tabelle 7.1). Um bei derart seltenen Ausprägungen überhaupt einen Lernprozess in Gang zu setzen, ist eine geschichtete oder gezielte Auswahl an Trainingsdokumenten unabdingbar. In einer begrenzten Stichprobe ist dies unter Umständen gar nicht möglich, so dass es ggf. unklar bleibt, ob eine Variable grundsätzlich schwierig zu automatisieren ist oder es nur an Trainingsmaterial mangelt.

Betrachtet man nochmals die beiden Validitätskoeffizienten Precision und Recall, ist bei vielen Variablen eine Konvergenz der beiden Werte zu beobachten, wobei die Präzision oft von Beginn an relativ hoch ist.

Das zusätzliche Training führt also eher zur Reduktion falsch negativer Klassifikationen, während der Anteil falsch positiver Codierungen oft konstant bleibt. Je nach Fragestellung kann so die notwendige Trainingsmenge erheblich variieren: Schon nach 100 Trainingsdokumenten kann man recht sicher sein, dass mit der Kategorie *Politik* codierte Dokumente auch politische Inhalte haben. Allerdings sind mehr als 700 Trainingsdokumente notwendig, um mit derselben Sicherheit auch sagen zu können, dass alle Politik-Beiträge als solche klassifiziert werden.

Angesichts der Form der meisten Lernkurven in Abbildung 7.8 scheint eine quadratische Wachstumsmodellierung am besten zu der Vielfalt von Verlaufskurven zu passen: bei *Sport* und *Politik* ist der Koeffizient für den quadratischen Effekt eher stark und negativ, was für einen substanziellen Saturierungseffekt spricht, während er bei linearem Wachstum, z.B. bei *Kriminalität*, gegen Null tendiert. Die Modellgüte des univariaten quadratischen Modells ist in den meisten Fällen sehr gut ($R^2 > .8$), lediglich bei *Kultur*, *Bundestagswahl 2009* und *Unglücke/Unfälle* fällt die Varianzaufklärung niedriger aus. Ähnliches gilt, wenn Precision und Recall als abhängige Variablen verwendet werden. Aus Gründen der Übersichtlichkeit verzichte ich hier auf eine ausführliche Darstellung der Wachstums-Koeffizienten, da das univariate Modell im nächsten Abschnitt lediglich durch zwei Treatment-Interaktionen erweitert wird.

7.2.2 Einfluss der Trainingsstrategie

Angesichts der theoretischen Argumente und der Studienergebnisse, die in Abschnitt 4.3 dargestellt wurden, lässt sich ein positiver Effekt des aktiven Lernens auf den Verlauf der Klassifikationsqualität erwarten. Bei dieser Strategie wählt der Klassifikator aus den Trainingsdokumenten solche aus, deren testweise Klassifikation am unsichersten ist und deren Kenntnis mit dem größten Informationsgewinn einhergehen würde. Stimmt also Hypothese 6, ist bei aktivem Lernen eine steilere Lernkurve und damit Klassifikationsqualität zu erwarten.

Wie in Teilstudie 1 bietet sich für die statistische Überprüfung der Frage, welchen Einfluss die Trainingsstrategie auf die Entwicklung der Klassifikationsqualität hat, ein hierarchisches Modell mit variierenden

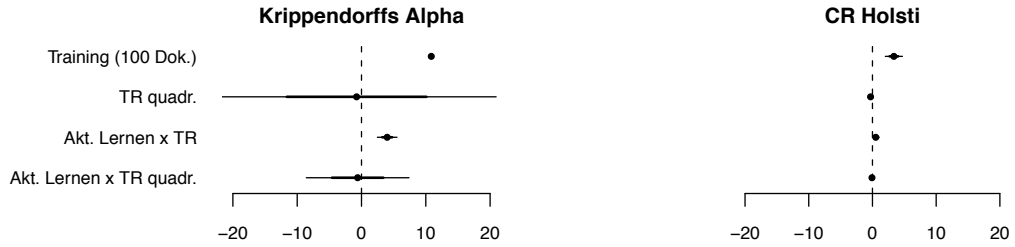

Abbildung 7.9: Einflüsse auf die Entwicklung der Klassifikationsreliabilität, unstandardisierte Regressionskoeffizienten u. Konfidenzintervalle

Effekten an. Die Auswertungslogik ist daher dieselbe wie im vorangegangen Kapitel und erfordert auch hier die Inspektion der fixen und variierenden Effekte im Modell. Dieses besteht aus vier Parametern (bzw. fünf, wenn man den Intercept-Term berücksichtigt): dem linearen und quadratischen Effekt des Trainingsumfangs sowie den Interaktionstermen mit dem Treatment Trainingsstrategie.[8]

In Abbildung 7.9 sind die Ergebnisse des hierarchischen Regressionsmodells mit den abhängigen Variablen Krippendorffs α und Prozentübereinstimmung nach Holsti dargestellt. Zur besseren Lesbarkeit habe ich die Variable Training so skaliert, dass der Wert 1 einem Satz von 100 Trainingsdokumenten entspricht. Über alle Variablen hinweg steigt die Reliabilität nach Krippendorff pro Trainingssatz linear um rund 11 Punkte, wenn eine passive Lernstrategie verfolgt wird.[9] Dieser Prozess wird durch einen negativen quadratischen Effekt gedämpft, der allerdings insgesamt nicht statistisch signifikant ist.

Wie in Hypothese 6 formuliert, wird ein deutlich positiver Einfluss der Trainingsstrategie sichtbar. Im Mittel steigt die Reliabilität nach Krippen-

8 Da die Werte der gemischten Trainingsstrategie in allen Fällen genau zwischen den beiden Varianten aktives und passives Lernen liegen, verzichte ich im Folgenden auf die Auswertung dieser Misch-Strategie und arbeite mit einer dichotomen Variable *aktives Lernen* weiter. Dies erleichtert u.a. die Interpretation der Haupt- und Interaktionseffekte des Modells.

9 Wie in den Abbildungen oben schon erkennbar wurde, ist der Anstieg der Reliabilität nach Holsti deutlich geringer, dafür ist der Intercept-Term sehr hoch.

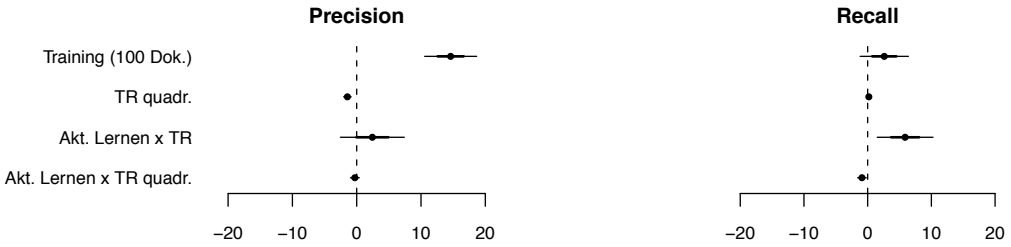

Abbildung 7.10: Einflüsse auf die Entwicklung der Klassifikationsvalidität, unstandardisierte Regressionskoeffizienten u. Konfidenzintervalle

dorff bei aktivem Lernen um fast 5 Prozentpunkte mehr je 100 Trainingsdokumente, dies bedeutet eine Steigerung der Lerneffektivität um mehr als ein Drittel. Aktives Lernen hat hier einen deutlich positiven Effekt, der bei umfangreichen Trainingsdaten nur minimal nachlässt.

Ein ähnliches Ergebnis zeigt sich auch bei den Validitätsmaßen Precision und Recall (vgl. Abbildung 7.10). Auf beide Werte hat die aktive Selektions von Trainingsdaten durch den Klassifikationsalgorithmus einen positiven Einfluss auf den Lernprozess. Dieser ist jedoch über alle Codebuch-Variablen nur bei der Trefferquote signifikant. Auch bei Precision und Recall ist ein leichter Saturierungseffekt zu erkennen, der sich bereits in den Verlaufskurven zeigte.

Effekte der Trainingsstrategie pro untersuchter Variable

Bei der Betrachtung der Random Effects der hierarchischen Regressionsmodelle zeigt sich wie in Teilstudie 1, dass der Effekt aktiven Lernens sehr stark variiert. Dies lässt sich am besten in den Verlaufskurven der Abbildungen 7.11 bis 7.13 erkennen, in denen die durchgezogene Linie den gemittelten Verlauf bei passivem, die gestrichelte Linie den Verlauf bei aktivem Lernen wiedergibt. Der Effekt aktiven Lernens auf das lineare Wachstum ist zusätzlich angegeben, allerdings in der Original-Metrik der abhängigen Variablen, die von 0 bis 1 reicht. Dieser Effekt lässt sich als Differenz im linearen Zuwachs der Klassifikationsqualität pro Trainingssatz von 100 Dokumenten interpretieren.

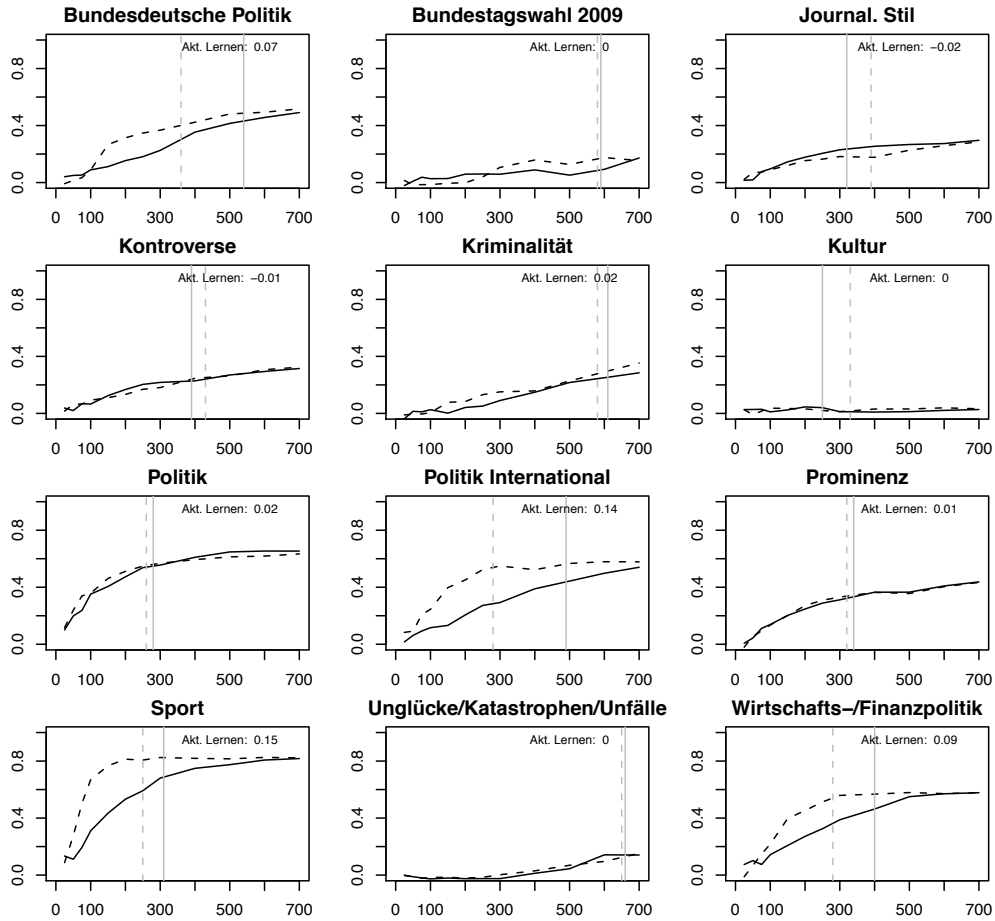

Abbildung 7.11: Entwicklung von Krippendorffs α bei aktivem und passivem Lernen

Insgesamt zeigt sich bei der Inspektion der Reliabilitätswerte nach Krippendorff, dass eine aktive Lernstrategie fast immer zu effektiverem Lernen führt und die Reliabilität dadurch schneller wächst. Dies ist besonders auffällig bei den thematischen Variablen, etwa *Sport* (+.15) oder *Wirtschaftspolitik* (+.09). Bei keiner Variable führt aktives Lernen zu einer signifikant schlechteren Entwicklung der Reliabilität – die Anwendung dieser Strategie kann augenscheinlich nicht schaden. Auffällig ist jedoch, dass aktives Lernen vor allem in den Fällen positiv wirkt, in denen die Klassifikationsqualität insgesamt hoch ist. Dies belegt die recht hohe positive Korrelation (r =.54) zwischen dem allgemeinen Anstieg der Reliabilität und dem Treatment-Effekt. Anders formuliert: Wenn der Klassifikator ohnehin kaum lernt, hilft auch die aktive Selektion von Trainingsmaterial nur wenig. Dies überrascht insofern nicht, als dass ein möglicher Grund mangelnder Klassifikationsqualität in zu homogenem Trainingsmaterial liegt. Die Zusammensetzung des Trainings-Sets kann auch durch aktive Selektion der Dokumente nicht verändert werden, weshalb die Trainingsstrategie ihre Wirkung nicht entfalten kann.

Ein Blick auf den Verlauf von Präzision und Trefferquote bei der Klassifikation beantwortet die Frage, auf welche Art und Weise die Klassifikationsqualität von der Trainingsstrategie beeinflusst wird. In Abbildung 7.12 ist zu erkennen, dass die Präzision durch aktives Lernen sowohl schneller – bei *Unglücke/Unfälle* oder *Bundestagswahl 2009* – als auch langsamer wachsen kann, etwa bei den Politikvariablen. Im Gegensatz zur Reliabilitätsentwicklung ist aber die Korrelation zwischen dem allgemeinem Wachstumsterm und der Treatmentvariable negativ (r =-.47), d.h. aktives Lernen hilft vor allem in den Fällen, bei denen der Lernzuwachs eher langsam verläuft.

Eindeutig ist dagegen die Entwicklung der Trefferquote: Diese profitiert fast durchgängig von der Anwendung aktiven Lernens. Kann der Klassifikator selbst das Trainingsmaterial auswählen, steigt der Recall bei der Variable *Sport* pro Trainingssatz fast um den Wert .20. Die Kurven ähneln insgesamt denjenigen bei der Reliabilität nach Krippendorff, und auch die Korrelation zwischen dem Wachstum und der Lernstrategie ist hier positiv (r =.74). Vom aktiven Lernen profitieren eher Kategorien, bei denen die Trefferquote relativ schnell ansteigt. Ist man an einer hohen

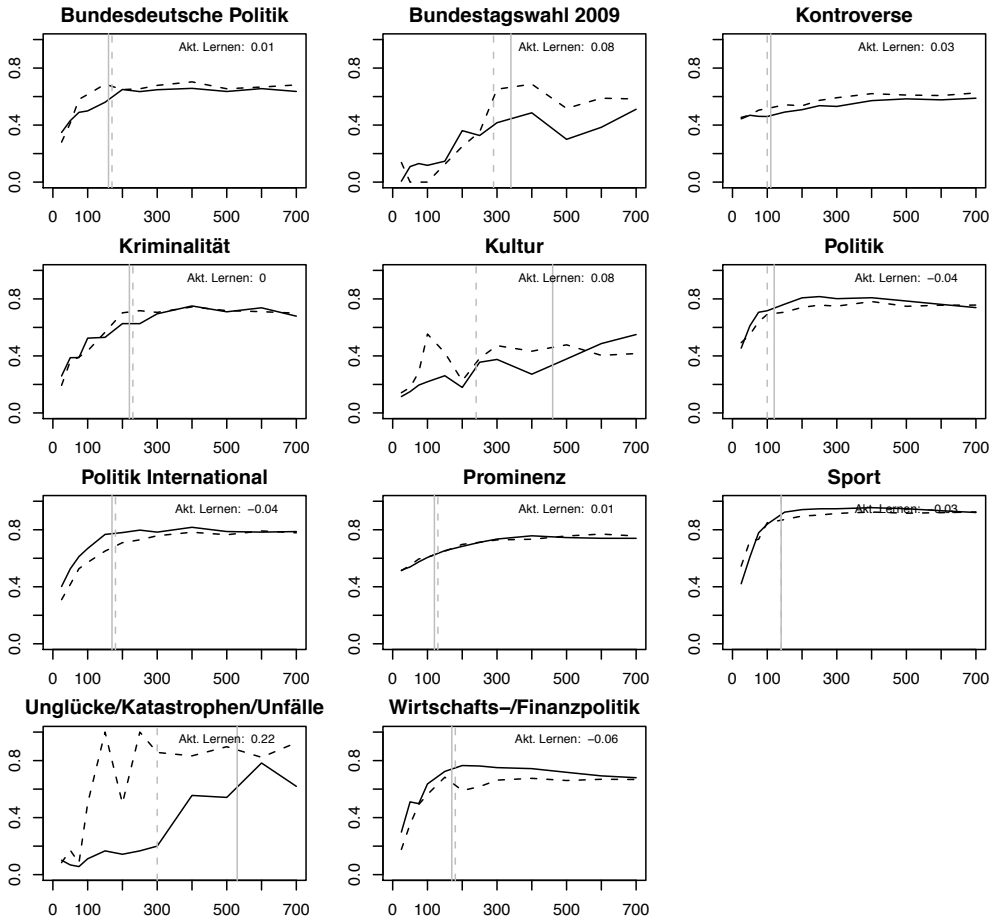

Abbildung 7.12: Entwicklung der Precision bei aktivem und passivem Lernen

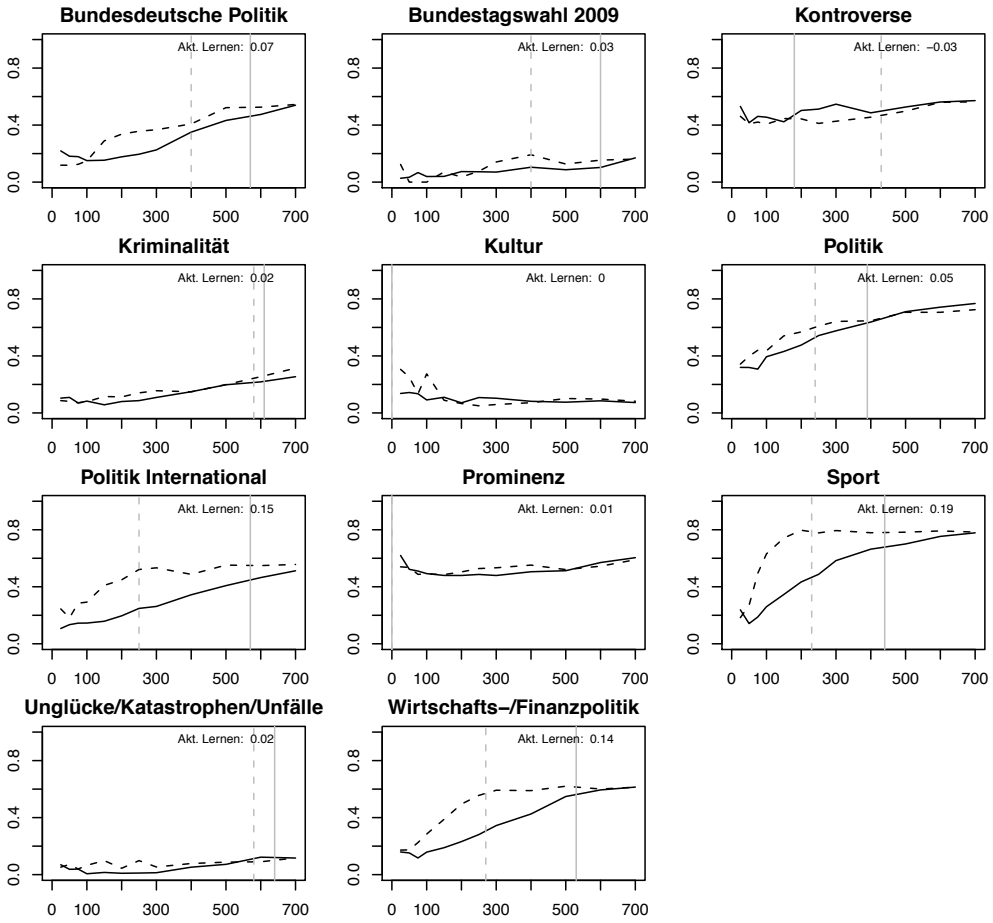

Abbildung 7.13: Entwicklung des Recalls bei aktivem und passivem Lernen

Trefferquote interessiert, etwa um die automatischen Klassifikationen als Aufgriffskriterium einer nachgelagerten Analyse zu verwenden, lohnt es sich in jedem Fall, eine aktive Lernstrategie beim Klassifikatortraining zu verfolgen.

Wie viel Trainingsmaterial wird benötigt?
Am Ende von Abschnitt 6.5 habe ich eine alternative Darstellung des Trainingsverlaufs vorgeschlagen, die bei der Beantwortung der Frage nach dem mininimal benötigten Trainingsmaterial hilft. Hierzu werden aus den Koeffizienten der hierarchischen Regressionsmodelle typische Verlaufskurven simuliert und daraus bestimmte *Cut-off*-Werte abgeleitet. In den Abbildungen 7.11 bis 7.13 sind jeweils zwei vertikale Linien eingezeichnet, die den Schwellenwert markieren, an dem 80 Prozent des maximal erreichbaren Reliabilitäts- oder Validitätswertes der Variable erreicht werden. Auch wenn diese Werte je nach Anpassungsgrad der Regressionsgleichung nicht ganz genau sind, bieten sie doch Richtlinien für die Menge benötigten Trainingsmaterials.

In Abbildung 7.11 kann man erkennen, dass für eine relativ zuverlässige Klassifikation von Sport- oder Politikmeldungen bei aktivem Lernen ca. 300 Trainingsdokumente ausreichen. Danach steigt die Reliabilität nur noch geringfügig an. Bei den Variablen *Politik International* oder *Bundesdeutsche Politik* wird die Auswirkung der Trainingsstrategie deutlich: Bei konventionellem passiven Training müssen fast 200 Dokumente mehr manuell codiert werden, um dieselbe Reliabilität zu erreichen.

Auch die Frage, wie viele Trainingsdokumente für eine Reliabilität oder einen Recall von .7 notwendig wären, lässt sich anhand der interpolierten Verläufe schätzen, wobei die Saturierungseffekte nur schwer vorauszusagen sind. Es ist daher nicht sicher, ob die Variable *Kriminalität* nach 1500 Trainingsdokumenten tatsächlich eine entsprechende Reliabilität aufweist. In vielen Fällen steht aber zu vermuten, dass auch bei aktivem Lernen die hier verwendeten 700 Trainingsdokumente noch nicht ausreichen, um die maximale Qualität der überwachten Klassifikation auszuschöpfen. Dies ist eine entscheidende Einschränkung der gesamten Evaluationsstudie, die ich im folgenden Abschnitt nochmals resümieren werde.

7.3 Zusammenfassung und Kritik der Evaluation

Im Mittelpunkt der empirischen Evaluationsstudie standen die Forschungsfragen, wie die Qualität überwachter Textklassifikation in einem kommunikationswissenschaftlichen Anwendungsszenario einzuschätzen ist, welche Faktoren die Klassifikation beeinflussen, und wie sich die Lerneffektivität des Klassifikators steigern lässt.

Die Ergebnisse der ersten Teilstudie zur Klassifikationsqualität zeigen, dass – wie in Hypothesen 1 und 2 postuliert – gerade Variablen, die sich zuverlässig manuell codieren lassen und auf lexikalischer Ebene angesiedelt sind, sehr gut automatisch klassifizierbar sind. Dies gilt insbesondere für Themenvariablen wie *Sport* und *Politik*, die man zukünftig nicht mehr manuell codieren muss. Andererseits fällt es dem Klassifikationsalgorithmus schwer, Nachrichtenfaktoren oder andere komplexe Kategorien zu lernen. Dies liegt erstens daran, dass schon das Trainingsmaterial weniger zuverlässig ist, zweitens an der teilweise zu geringen Menge an Positiv-Beispielen, drittens auch daran, dass die relevanten Informationen für die Codierung auf semantischer oder pragmatischer Ebene liegen und damit nicht mit statistischen Bag-of-Words-Ansätzen zu analysieren sind.

Bezogen auf die Hypothesen 3 und 4 zeigt sich, dass außer der Extraktion von Fließtext aus komplexen HTML- oder XML-Dokumenten kein Preprocessing-Schritt die Klassifikationsqualität substanziell verbessern kann. Anders formuliert: Die Klassifikation kann als robust gelten, technisch aufwändige Vorarbeiten sind nicht unbedingt notwendig. Betrachtet man die Ergebnisse der hierarchischen Regressionsmodelle, zeigt sich, dass über 90 Prozent der Varianz in den Daten auf Unterschiede zwischen den Variablen des Codebuchs zurückgeht. Deshalb sollten alle Ressourcen darauf konzentriert werden, die manuelle und automatische Codierung zu verbessern.

In der zweiten Teilstudie ging es um die Frage, wie effektiv der Trainingsprozess bei der Anwendung überwachter Textklassifikation ist. Auch hier zeigt sich, dass es auf die Variablen im Codeplan ankommt. Einfache Themenvariablen werden nicht nur zuverlässig, sondern auch schnell gelernt, während Regeln für komplexe Variablen nur sehr lang-

sam aus den Beispieldaten zu extrahieren sind. Zudem ist die Lernkurve bei einseitig verteilten Trainingsdaten deutlich flacher als bei Variablen, in denen alle Kategorien etwa gleich häufig sind. Die Hypothese, dass eine aktive Selektion von Trainingsmaterial durch den Computer den Lernprozess beschleunigt, lässt sich eindeutig als bestätigt ansehen. In einigen Fällen ist der Effektivitätsgewinn sehr groß, vor allem aber schadet aktives Lernen in keinem Fall der Klassifikation. Die Empfehlung kann hier nur lauten, wann immer möglich eine entsprechende Selektionsstrategie zu verfolgen.

Betrachtet man die Ergebnisse insgesamt im Kontext bisheriger Studien, zeigt sich, dass auch im hier gewählten Anwendungsszenario – deutsche Nachrichtentexte, einfache bis komplexere kommunikationswissenschaftliche Kategorien – überwachte Klassifikationsverfahren grundsätzlich ein großes Potential besitzen. Zwar ist die Qualität der Klassifikation nur bedingt geeignet, manuelle Codierungen insgesamt obsolet zu machen, doch muss man berücksichtigen, dass die Prüfung der *Möglichkeit* der Automatisierung auch nur mit minimalen Zusatzaufwand verbunden ist. Der Vergleich mit den quantitativen Ergebnissen früherer Studien ist insofern schwierig, als dass häufig die Übereinstimmung nach Holsti als Qualitätsindikator verwendet wird, dieser jedoch nicht unbedingt geeignet ist, problematische Variablen zu identifizieren. Verglichen mit der Klassifikation von Gesetzestexten bei Hillard et al. (2007) fallen die Ergebnisse in dieser Studie deutlich schlechter aus, was aber auch darauf zurückzuführen ist, dass dort ausschließlich eng definierte thematische Variablen codiert wurden. Vergleicht man die Klassifikationsqualität in dieser Arbeit aber mit der inhaltlich näherliegenden Studie von Durant & Smith (2007), die politische Blog-Postings analysieren, zeigen sich vergleichbare Ergebnisse, was die prozentuale Übereinstimmung in der manuellen und automatischen Codierung betrifft.

Einschränkungen der Studie

Sowohl in der deskriptiven Darstellung der Klassifikationsqualität als auch in den eigentlichen Kausalanalysen zeigt sich ein methodisches Problem dieser Evaluationsstudie: Die verwendeten Koeffizienten nach Holsti bzw. Krippendorff sowie Precision und Recall eignen sich nur

bedingt als abhängige Variablen. Die Prozentübereinstimmung nach Holsti gibt die Reliabilität der Klassifikation stark nach oben verzerrt wieder und reagiert dabei so wenig auf substanzielle Veränderungen, dass sie eigentlich für die Fragestellungen dieser Arbeit ungeeignet ist. Krippendorffs α reagiert andererseits extrem sensibel auf die Verteilung der Kategorien und deren Schwankungen, so dass nicht immer klar ersichtlich wird, ob eine substanzielle Veränderung in der Klassifikation vorliegt oder sich nur die Materialkomposition geändert hat. Precision und Recall gehen per Definition von einem manuellen Goldstandard aus und betrachten daher jede Nichtübereinstimmung als Fehler des Klassifikators. Dies ist angesichts des nachweislich fehlerbehafteten Trainings- und Testmaterials eine unplausible Annahme. Da keiner der vier Koeffizienten statistisch *und* diagnostisch günstige Eigenschaften hatte, war es angebracht, mehrere Indikatoren zu verwenden.

Eine weitere Einschränkung der empirischen Evaluation betrifft die Prototypik des Codeplans. Während ich mit einiger Gewissheit davon ausgehen kann, dass die Ergebnisse weder durch die Auswahl noch die Beschaffenheit des Stimulusmaterials relativiert werden müssen, ist die Auswahl der Kategorien deutlich problematischer. Einerseits sind die verwendeten Variablen aus bereits existierenden Codebüchern entnommen, so dass diese potentielle Ursache fehlender externer Validität ausscheidet. Andererseits basieren die Ergebnisse der Evaluation auf Ebene des Codebuchs eben nur auf zwölf Variablen, so dass nicht ausgeschlossen werden kann, dass Preprocessing bei anderen Variablen substanzielle Auswirkungen auf die Klassifikation hat. Zudem habe ich mich auf größtenteils dichotome Variablen konzentriert und auf klassische Bewertungs-Variablen verzichtet. Die Evaluationsergebnisse lassen sich deshalb nicht ohne weiteres auf Variablen mit vielen Ausprägungen oder Codierungen von Tendenzen etc. generalisieren. Hier sind weitere Untersuchungen notwendig, wobei sich das Vorgehen in dieser Arbeit so bewährt hat, dass man es mit entsprechendem Trainingsmaterial wiederholen kann.

In der Materialmenge liegt auch ein weiteres Problem dieser Studie: Da manche Variablen extrem schief verteilt waren, ist selbst nach 1000 Dokumenten noch nicht klar, ob sich die unzureichende Klassifikationsqualität auf mangelndes Trainingsmaterial oder auf tatsächliche Probleme bei der

Regelextraktion zurückführen lässt. Diese Frage ist nur mit geschichteten oder sehr viel umfangreicheren Stichproben zu klären.

Eine letzte Einschränkung der empirischen Ergebnisse liegt in der Wahl des Klassifikationsalgorithmus. Da es bei dieser Arbeit gerade nicht um die Frage ging, welcher Klassifikator unter welchen Bedingungen besser abschneidet, habe ich mich für einen einzigen Algorithmus und eine Implementation entschieden. Angesichts der zahlreichen vergleichenden Studien zur Klassifikatorwahl (Felden et al., 2005; Durant & Smith, 2007; Dumais et al., 1998; Joachims, 2002) liegt die Vermutung nahe, dass bei den hier verwendeten Variablen andere Algorithmen ggf. unterschiedliche Ergebnisse erbracht hätten, jedoch keinesfalls in der Größenordnung wie die Unterschiede zwischen den Variablen. Nichtsdestotrotz kann diese offene Frage relativ leicht beantwortet werden, in dem die entsprechende Klassifikationssoftware in NewsClassifier ausgetauscht wird.

8 Diskussion und Ausblick

Am Anfang dieser Arbeit stand die Forschungsfrage, ob und wie man durch Automatisierung den inhaltsanalytischen Forschungsprozess transparenter, effizienter und zuverlässiger machen kann, um den zunehmenden methodologischen und forschungspraktischen Herausforderungen der Kommunikationswissenschaft begegnen zu können. Ich habe dabei argumentiert, dass sich durch die methodische Weiterentwicklung der Inhaltsanalyse einerseits die Reichweite und Genauigkeit der Inferenzen erhöhen lässt, andererseits viele neue Forschungsfragen erst im großen Maßstab überhaupt empirisch fassbar werden, so dass an einer zumindest teilweisen Automatisierung der Analyse kein Weg vorbei führen wird.

Für die Aufarbeitung des Forschungsstandes zu automatischen Verfahren der Textanalyse hat sich eine einfache Typologie bewährt, in der fast alle in den letzten 50 Jahren entwickelten Ansätze verortet werden können. Vollautomatische, explorative bzw. unüberwachte Ansätze auf Wort- und Dokumentebene wie Co-Occurrence-Analysen oder Document-Clustering sind dabei technisch relativ weit entwickelt, eignen sich aber nur selten zur Hypothesen-Prüfung und noch seltener als Ersatz für klassische manuelle Analysen. Nichtsdestotrotz erfüllen vollautomatische Verfahren der Inhaltsanalyse, gerade in Verbindung mit einer grafischen Aufbereitung, einen wichtigen Zweck im Forschungsprozess: Auf keine andere Weise lässt sich so schnell und kostengünstig eine sonst unüberschaubare Menge an (Text-)Daten strukturieren und zusammenfassen. Dies wiederum erleichtert die induktive Kategorienbildung und Codebuchentwicklung (Früh, 2007). Aufgrund ihres hohen Standardisierungsgrades ist es nicht nur sinnvoll, sondern auch leicht umsetzbar, die häufigsten unüberwachten Analyseverfahren in den inhaltsanalytischen Forschungsprozess zu integrieren. Dies ist allerdings

nur dann möglich, wenn auch die entsprechende Datenerhebung und -aufbereitung vollautomatisch verläuft.

Bei den hypothesengeleiteten und damit nicht vollautomatischen Analyseverfahren dominieren bislang deutlich die deduktiven Ansätze, bei denen der Forscher explizit umfassende und trennscharfe Codierregeln, entweder auf lexikalischer oder syntaktischer Ebene, entwickeln muss, die dann streng deterministisch umgesetzt werden. Auf die Problematik dieses Ansatzes habe ich bereits hingewiesen: Erstens ist Sprache nicht einfach algorithmisch fassbar, jedenfalls nicht auf der Ebene von intersubjektiv geteilten Bedeutungen. Zweitens erfordert die Regelbildung, sei es bei Diktionären oder Extraktionsregeln, nicht nur eine streng formale Theorie von Sprache, sondern Programmierkenntnisse, die man als Sozialwissenschaftler nur mühsam erwerben oder teuer zukaufen muss. Je weiter sich dabei die Computerlinguistik spezialisiert, desto größer wird der Graben zu den Forschern, die weiterhin manuell codieren (lassen). Drittens sind deduktive Verfahren fest an eine Sprache oder Textsorte gebunden, was komparative Forschung ggf. deutlich erschwert.

Angesichts des Potentials induktiv-statistischer Analyseverfahren stehen diese im Zentrum der vorangegangenen Kapitel. Insbesondere die induktive Textklassifikation, die sich Algorithmen aus dem überwachten maschinellen Lernen bedient, verspricht, die manuelle und automatische Analyse stärker zu integrieren. Hinzu kommt die Tatsache, dass sie mit wenig zusätzlichem Aufwand umsetzbar, sprach- und gegenstandsunabhängig sowie algorithmisch gut erforscht ist. All dies sprach und spricht dafür, die Möglichkeiten des Einsatzes überwachter Klassifikationsverfahren für die Inhaltsanalyse fruchtbar zu machen. Der Nutzen des Verfahrens lässt sich dabei auf zwei Ebenen bestimmen: Im engeren Sinne geht es um die Frage, ob und wie reliabel und valide sich die manuelle Codierung nach konventionellen Regeln automatisieren lässt. Im weiteren Sinne kann das Verfahren auch dann von Nutzen sein, wenn die eigentliche Codierung noch nicht den Ansprüchen des Forschers genügt. Dies lässt sich am deutlichsten anhand der Nutzung überwachter Klassifikation für die Selektion problematischer Dokumente im Rahmen der Codebuchentwicklung nachweisen.

Betrachtet man die Ergebnisse der Evaluationsstudie, kommt man erstens zu dem Schluss, dass die überwachte Textklassifikation keinesfalls als Allzweckmittel zur Substitution menschlicher Codierer taugt. Zwar lassen sich vor allem Themenvariablen relativ gut automatisch codieren, an komplexeren Kategorien scheitert der Computer jedoch. Ein zweites empirisches Ergebnis dieser Arbeit liegt darin, dass sich auf technischem Wege die automatische Codierung nicht verbessern, allerdings auch nur selten substanziell verschlechtern lässt. Der teilweise erhebliche Aufwand bei der Datenvorbehandlung, der oft bei vollautomatischen Analysen notwendig ist, lässt sich bei überwachten Klassifikationsverfahren einsparen. Hier ähnelt sie in den Anforderungen eher konventionellen Inhaltsanalysen. Drittens zeigt sich, dass der Klassifikationsalgorithmus schneller lernt, wenn man das Trainingsmaterial gezielt auswählt bzw. auswählen lässt.

Dieser letzte Punkt ist für die Beurteilung, ob der Einsatz der Klassifikationssoftware nun einen messbaren Nutzen bringt, von entscheidender Bedeutung: Angesichts der empirischen Ergebnisse ist davon auszugehen, dass zumindest Teile vieler Codebücher sich für die Automatisierung durch überwachte Klassifikation eignen.[1] Unabhängig davon, ob nun eine erfolgreiche Regelextraktion für ein Kategoriensystem gelingt, lässt sich durch die gezielte Nutzung automatischer Klassifikationen, die im Hintergrund und ohne Zusatzaufwand erfolgen können, der inhaltsanalytische Forschungsprozess transparenter gestalten. Versteht man den Klassifikator als ungeschulten, naiven Codierer, aus dessen Fehlern man Konsequenzen für die Codebuchentwicklung ziehen kann oder der auf schwierige Dokumente hinweist, kann auch bei einer unbefriedigenden Klassifikationsleistung der inhaltsanalytische Forschungsprozess substanziell verbessert werden. Im Kontext der Dokumentation des Tools NewsClassifier (vgl. Anhang A) habe ich weitere Vorschläge für einen sinnvollen Einsatz von automatischen Softwarelösungen für den Forschungsalltag der Inhaltsanalyse gemacht, die alle letztlich das Ziel der

1 Allerdings ist es durchaus möglich, dass sich verschiedene Variablengruppen unterschiedlich automatisieren lassen, etwa Akteursvariablen durch Diktionäre und Themenvariablen durch überwachte Klassifikation.

Transparenz und Effizienz haben. Dies würde mehr Raum für substanzielle Fragen der empirischen Kommunikationsforschung schaffen.

Anwendungen jenseits von Texten

Während die eben geschilderten forschungspraktischen Vorteile automatischer Verfahren für die Inhaltsanalyse erst im Laufe der empirischen Arbeit hervorgetreten sind, gab es von Anfang an ein zentrales Argument dafür, die überwachten Klassifikationsalgorithmen aus der Informatik der sozialwissenschaftlichen Forschung zugänglich(er) zu machen: Die Verfahren sind in fast unveränderter Form für eine Vielzahl von Anwendungsfeldern jenseits der Textanalyse einsetzbar. Zwar stand in dieser Arbeit der häufigste Anwendungsfall, die Analyse von Nachrichtentexten, im Mittelpunkt des Interesses, Naive-Bayes- oder SVM-Klassifikatoren eignen sich jedoch auch für anderes Stimulusmaterial. Da alle Algorithmen intern lediglich mit numerischen Variablen arbeiten, ist nur ein einziger Schritt im Analyseprozess substanziell zu verändern: die Feature-Extraktion, d.h. die Zuordnung von Codes zu sprachlichen, auditiven oder visuellen Merkmalen des Stimulusmaterials. Ist dieser zentrale Transformationsschritt gelungen, lassen sich mit der gleichen Infrastruktur, dem gleichen Trainingsprozess und den gleichen Evaluationsstrategien beliebige Inhalte codieren. Mit anderen Worten: Kennt man die Grundlagen überwachter Klassifikationsalgorithmen, erschließt sich die Logik vieler automatisierter Analyseverfahren in und jenseits der Kommunikationswissenschaft. Einige Anwendungsfälle möchte ich als Ausblick hier kurz ansprechen, um die Möglichkeiten der Automatisierung jenseits des Textcodierung zu illustrieren.

Ein leicht verständliches, wenn auch technisch komplexes Beispiel für die Möglichkeiten überwachter Klassifikation ist die Genre-Erkennung bei Musikstücken. Bei einer klassischen manuellen Inhaltsanalyse von Musikstücken werden die Codierer anhand typischer Merkmale wie Rhythmus, Melodie, Instrumentierung, Geschwindigkeit oder Dynamik jedes Lied mehr oder minder zuverlässig einem Genre zuordnen können. Wie Scaringella et al. (2006) in ihrem Überblicksartikel zeigen, lassen sich dieselben überwachten (und auch unüberwachten) Algorithmen für die Analyse von auditiven Stimuli anwenden. Auch hier werden dem Klassi-

fikator vorcodierte Beispielstücke vorgelegt, anhand dieser die relevanten Features für die Genre-Bestimmung lernt. Ein anderes Anwendungsfeld für überwachte Klassifikationsalgorithmen liegt in der Erkennung von besonderen Ereignissen in audiovisuellen Stimuli. Xu et al. (2003) können auf diese Weise anhand typischer Geräusche – Pfiffe, Applaus, Ausrufe des Kommentators – zuverlässig die Höhepunkte eines Fußballspiels aus der Aufzeichnung extrahieren (vgl. auch Radhakrishan et al., 2004). Da auch in diesem Fall die eigentliche Semantik der auditiven Signale für die Klassifikation irrelevant ist, lässt sich dieser Ansatz auch problemlos auf andere Forschungsfragen, etwa in der Analyse von Spielfilmen, TV- oder Parlamentsdebatten anwenden.

Einen Schritt weiter gehen Cutler & Davis (2002), die mit einem multimodalen Ansatz die Sprecher in audiovisuellem Material identifizieren. Hierbei werden sowohl auditive als auch visuelle Features verwendet, da die Kombination beider Informationsquellen die beste Erkennungsgrundlage bietet. Da überwachte Klassifikationsverfahren ohnehin nur mit numerischen Codes arbeiten, kann man bei der Feature-Extraktion beliebig verbale, auditive und visuelle Merkmale verwenden. Dies erlaubt eine große Flexibilität bei der Operationalisierung relevanter Konzepte. Auch auf rein visueller Ebene können überwachte Klassifikatoren eingesetzt werden, wie etwa Goela et al. (2007) zeigen, die auf diese Weise automatisch Beitragssegmente in Fernsehprogrammen identifizieren. Die Beitragssegmentierung ist ein Problem, das sich in der klassischen Programmforschung (Weiß, 1998) in fast identischer Weise zeigt. Angesichts des hohen Aufbereitungs- und Codieraufwands wäre eine Teilautomatisierung dieses Prozesses durchaus wünschenswert. Die Arbeit von Snoek et al. (2006) geht noch einen Schritt weiter und klassifiziert einzelne Segmente thematisch, um etwa Wetterberichte oder Moderationen zu identifizieren. Da noch unklar ist, wie detailliert eine solche Klassifikation sein kann, gibt es hier Bedarf für eine ähnliche Evaluationsstudie, wie sie in dieser Arbeit durchgeführt wurde.

Ein letztes Beispiel betrifft die Codierung von Mimik und Gestik, die insbesondere in der medienpsychologischen Forschung zunehmend Anwendung findet. Schon zu Beginn der 90er Jahre wurden erste Versuche unternommen, das FACS-Codiersystem von Ekman et al. (1978), das im

Normalfall von gut geschulten Codierern angewendet wird, mittels überwachter Klassifikationsalgorithmen zu automatisieren (Kaiser & Wehrle, 1992). Mittlerweile gehört die Mimikerkennung zu den meisterforschten Gebieten des maschinellen Lernen jenseits der Textklassifikation, so dass nicht nur zahlreiche Studien (Bartlett et al., 1999; Fasel & Luettin, 2003; Sebe et al., 2007) zu diesem Problem durchgeführt wurden, sondern mittlerweile auch einfach zu bedienende Software für diesen Zweck erhältlich ist, etwa FaceReader (Den Uyl & Van Kuilenburg, 2005) oder Shore[2]. Hier gilt ähnlich wie bei inhaltsanalytischen Fragestellungen, dass der Computer (noch) nicht in der Lage ist, alle komplexen Kategorien zuverlässig zu codieren, dass aber umgekehrt für viele Fragestellungen die manuelle Codierung schlicht unrealisierbar ist. Gerade wenn es um die kontinuierliche Erhebung mimischer Reaktionen, etwa während der Medienrezeption, geht, ist die automatische Codierung geradezu alternativlos. Zudem lohnt es sich bei längerfristigen Beobachtungen, den Klassifikator individuell anhand der Versuchpersonen zu trainieren, um so die Codierqualität zu erhöhen. Der Aufwand für das Training ist dabei minimal, da lediglich vorcodierte Bilder vorgelegt werden müssen.

Schlussbetrachtung
In der Kommunikationswissenschaft werden Verfahren des maschinellen Lernens bisher recht selten verwendet, weil es einerseits an Know-How im Umgang mit den entsprechenden Softwarepaketen fehlt und weil andererseits bislang nur wenig über die Leistungsfähigkeit der Verfahren im Forschungsalltag bekannt ist. Genau hier soll die vorliegende Arbeit ein Anstoß für eine stärkere Auseinandersetzung und ggf. Berücksichtigung automatischer inhaltsanalytischer Verfahren in der Forschungspraxis sein. Auch wenn die hier untersuchte Textsorte und vor allem die verwendeten Kategorien nur einen sehr kleinen Ausschnitt aus dem reichen Fundus inhaltsanalytischer Forschung darstellen, hoffe ich, dass die vorgestellten konzeptionellen Überlegungen und Lösungsvorschläge sowie die Anlage der Evaluationsstudie Anlass für eine verstärkte Methodenforschung im Bereich der Inhaltsanalyse geben können.

2 http://www.iis.fraunhofer.de/EN/bf/bv/kognitiv/biom/dd.jsp.

Angesichts der von Krippendorff (2004a) formulierten Herausforderungen können automatische Verfahren, und insbesondere die überwachte Textklassifikation, die Inhaltsanalyse auf vielfältige Weise bereichern: Sie unterstützen den Forscher bei der Entwicklung des Codebuchs und machen es erstmals praktikabel, verschiedene Operationalisierungsstrategien systematisch und unter gleichen Bedingungen zu testen. Hierbei können Codierer und verschiedene Klassifikationsalgorithmen flexibel eingesetzt werden, da der Umgang mit beiden weitgehend den gleichen Regeln folgt. Letztlich wird auf diese Weise nicht nur die Entwicklung von Codeplänen, sondern auch der Schulungs- und Codierprozess transparenter und leichter reproduzierbar.

Ein zusätzlicher Nutzen überwachter Klassifikation liegt in den Möglichkeiten, an bestehende Studien anzuknüpfen und diese fortzusetzen. Durch die Wiederverwendung eines trainierten Klassifikators, dessen „Wissen“ in einer einzelnen Datei gespeichert ist, kann nicht nur das dokumentierte Codebuch, sondern die eigentliche Codierung für neue Studien nutzbar gemacht werden. Zudem können auch die Daten vollständig manuell durchgeführter Studien als Trainingsmaterial für die überwachte Klassifikation eingesetzt werden. Hierfür werden nur die vergebenen Codes und das digitale Untersuchungsmaterial benötigt. Die Verwendung maschinellen Lernens eröffnet daher völlig neue Wege der Replikation und Weiterentwicklung von Inhaltsanalysen.

Überwachte Klassifikationsverfahren können jedoch nicht nur die Qualität der Codierung steigern, sondern vor allem auch deren Quantität und damit die Reichweite der Analyse. Sobald ein Klassifikator so weit trainiert ist, dass eine zuverlässige automatische Codierung gelingt, können viele neue Dokumente mit minimalem Aufwand verarbeitet werden. Eine einzelne Analyse von wenigen hundert Beiträgen lässt sich so fast beliebig erweitern. Auf diese Weise können gerade bei der Untersuchung von Online-Inhalten kontinuierliche Analysen eines breiten Medienangebots durchgeführt werden. Die hier vorgestellten Erhebungs- und Klassifikationsverfahren ermöglichen umfangreiche Studien bei vertretbarem Aufwand, beispielsweise die von Zeller & Wolling (2010) konzipierten Strukturanalysen des publizistischen Online-Angebotes.

Auch wenn die Leistungsfähigkeit überwachter Textklassifikation in hohem Maße von der inhaltlichen Gestaltung des Codeplanes abhängt, lässt sich zusammenfassend festhalten, dass sich ihr Einsatz bei der Analyse digitaler Inhalte lohnt. Zwar ist es nicht immer angebracht, alle Variablen tatsächlich automatisch codieren zu lassen, doch lässt sich die Reliabilität und Validität der Klassifikation leicht bestimmen. Da der Anteil systematischer und zufälliger Fehlklassifikationen für jede Variable bekannt ist, kann man dies bei der Analyse und den darauf aufbauenden Inferenzschlüssen berücksichtigen. Die Konsequenzen der Verwendung bestimmter Klassifikationsverfahren werden auf diese Weise transparent. Ob die Automatisierung im Einzelnen die Qualität der Inferenzen verbessert hat, kann dann anhand empirischer Daten beurteilt werden. Schon aus diesem Grund lohnt es sich, zukünftig häufiger automatische Verfahren bei der Inhaltsanalyse einzusetzen.

Literatur

Aas, K., & Eikvil, L. (1999). Text categorisation: A survey. *Raport NR 941*.

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, *20*(5), 67–75.

Abbott, M., & Fisher, M. (2010). *The Art of Scalability*. Amsterdam: Addison-Wesley Longman.

Adam, S. (2008). Medieninhalte aus der Netzwerkperspektive. *Publizistik*, *53*(2), 180–199.

Adams, S. (2004). Statement Analysis: Beyond the Words. *FBI Law Enforcement Bulletin*, *73*(4).

Aldenderfer, M., & Blashfield, R. (1984). *Cluster Analysis*. Beverly Hills: Sage.

Alexa, M., & Zuell, C. (2000). Text Analysis Software: Commonalities, Differences and Limitations: The Results of a Review. *Quality and Quantity*, *34*(3), 299–321.

Alpaydin, E. (2008). *Maschinelles Lernen*. Oldenbourg Wissenschaftsverlag.

American Educational Research Association (1985). *Standards for educational and psychological testing*. American Psychological Association.

Apté, C., Damerau, F., & Weiss, S. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, *12*(3), 233–251.

Arlt, D., Hoppe, I., & Wolling, J. (2010). Klimawandel und Mediennutzung. *Medien und Kommunikationswissenschaft*, *58*(1), 3–25.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Assis, F. (2006). OSBF-Lua-A text classification module for Lua–the importance of the training method. In *Proceedings of the 15th international conference on WWW TREC*.

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

Atteveldt, W. v., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles. *Political Analysis*, *16*(4), 428–446.

Baisa, V. (2009). *Web Content Cleaning*. Master's thesis, Masaryk University, Brno.

Bamberger, R., & Vanecek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache*. Wien: Jugend und Volk.

Baroni, M., Chantree, F., Kilgarriff, A., & Sharoff, S. (2008). Cleaneval: a competition for cleaning web pages. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, *36*(2), 253–263.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *The Journal of Machine Learning Research*, *5*, 1089–1105.

Bengston, D., & Xu, Z. (2009). Changing national forest values: A content analysis. In K. Krippendorff, & M. A. Bock (Hrsg.) *The Content Analysis Reader*, (283–294). Thousand Oaks: Sage.

Benoit, K., Bräuninger, T., & Debus, M. (2009a). Challenges for Estimating Policy Preferences: Announcing an Open Access Archive of Political Documents. *German Politics*, *18*(3), 441–454.

Benoit, K., & Laver, M. (2007). Benchmarks for Text Analysis: A Response to Budge and Pennings. *Electoral Studies*, *26*(1), 130–135.

Benoit, K., Laver, M., & Mikhaylov, S. (2009b). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, *53*(2), 495–513.

Berelson, B. (1952). *Content Analysis in Communication Research*. New York: The Free Press.

Berendt, B., Schlegel, M., & Koch, R. (2008). Die deutschsprachige Blogosphäre: Reifegrad, Politisierung, Themen und Bezug zu Nachrichtenmedien. In A. Zerfaß, M. Welker, & J. Schmidt (Hrsg.) *Kommunikation, Partizipation und Wirkungen im Social Web*, (72–96). Köln: Halem.

Berg, H. v. d., & Veer, K. v. d. (2000). Computerized Decision Support Systems and Text Analysis: Evaluating CETA. *Quality and Quantity*, *34*, 65–86.

Best, K. (2006). Sind Wort-und Satzlänge brauchbare Kriterien zur Bestimmung der Lesbarkeit von Texten? In S. Wichter, & A. Busch (Hrsg.) *Wissenstransfer-Erfolgskontrolle und Rückmeldungen aus der Praxis*, (21–31). Frankfurt a.M.: Peter Lang.

Best, M. (1997). Models for Interacting Populations of Memes: Competition and Niche Behavior. *Journal of Memetics-Evolutionary Models of Information Transmission*, *1*.

Bjornsson, C. (1983). Readability of Newspapers in 11 Languages. *Reading Research Quarterly*, *18*(4), 480–97.

Blum, A., Kalai, A., & Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, (203–208). Santa Cruz: ACM.

Bock, A., Isermann, H., & Knieper, T. (2010). Herausforderungen bei der quantitativen (visuellen) Inhaltsanalyse von Online-Inhalten. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (224–239). Köln: Halem.

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis*, *54*(12), 2976–2989.

Box, G., Hunter, W., & Hunter, J. (1978). *Statistics for experimenters. An introduction to design, data analysis, and model building*. New York: Wiley.

Braschler, M., & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval? *Information Retrieval*, *7*, 291–316.

Bräuninger, T., & Debus, M. (2008). Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern. *Politische Vierteljahresschrift*, *49*(2), 309–338.

Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., & Yergeau, F. (2000). Extensible markup language (XML) 1.0. *W3C recommendation*, *6*.

Broder, A., Glassman, S., Manasse, M., & Zweig, G. (1997). Syntactic clustering of the web. *Computer Networks and ISDN Systems*, *29*(8-13), 1157–1166.

Brodley, C., & Friedl, M. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, *11*(1), 131–167.

Bröker, E. (1984). Computerunterstützte Inhaltsanalyse der internationalen Berichterstattung der Massenmedien: Erfahrungen bei der Analyse großer Datenmengen mit einem komplexen Wörterbuch. In H.-D. Klingemann (Hrsg.) *Computerunterstützte Inhaltsanalyse in der empirischen Sozialforschung*, (155–171). Frankfurt a.M.: Campus.

Brooks, C., & Montanez, N. (2006). Data mining classification: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th international conference on World Wide Web WWW*, (625–632). Edinburgh.

Brosius, F. (2006). *SPSS 14*. Heidelberg: MITP.

Brosius, H., Haas, A., & Koschel, F. (2009). *Methoden der empirischen Kommunikationsforschung*. Wiesbaden: VS Verlag.

Bruns, T., & Marcinkowski, F. (1997). *Politische Information im Fernsehen*. Opladen: Leske + Budrich.

Bryk, A., & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158.

Budge, I., & Pennings, P. (2007a). Do They Work? Validating Computerised Word Frequency Estimates against Policy Series. *Electoral Studies*, *26*, 121–129.

Budge, I., & Pennings, P. (2007b). Missing the Message and Shooting the Messenger: Benoit and Laver's Response. *Electoral Studies*, *26*, 136–141.

Busemann, K., & Gscheidle, C. (2010). Web 2.0: Nutzung steigt – Interesse an aktiver Teilhabe sinkt. *Media Perspektiven*, *7-8*, 359–368.

Carley, K. M. (1997). Network text analysis: The network position of concepts. In C. Roberts (Hrsg.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, (79–100). Mahwah: Lawrence Erlbaum Associates.

Carpenter, B. (2008). Multilevel Bayesian Models of Categorical Data Annotation. Tech. rep., Alias-i.

Casella, G., & Berger, R. (2002). *Statistical inference*. Pacific Grove: Duxbury.

Coleman, M., & Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284.

Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism and Mass Communication Quarterly*, *83*(1), 186.

Cormack, G., & Bratko, A. (2006). Batch and on-line spam filter comparison. In *Proceedings of the Third Conference on Email and Anti-Spam (CEAS)*. Mountain View.

Cormack, G. V., & Lynam, T. R. (2007). Online Supervised Spam Filter Evaluation. *ACM Transactions on Information Systems*, *25*(3), 11.

Craggs, R., & Wood, M. M. G. (2005). Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, *31*, 289–296.

Cuilenburg, J. v., Kleinnijenhuis, J., & De Ridder, J. (1986). A theory of evaluative discourse: Towards a graph theory of journalistic texts. *European Journal of Communication*, *1*(1), 65–96.

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

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

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *Proceedings of the 2004 Usenix Conference*. Boston.

Debole, F., & Sebastiani, F. (2005). An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and Technology*, *56*(6), 584–596.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Deichsel, A. (1975). *Elektronische Inhaltsanalyse: zur quantitativen Beobachtung sprachlichen Handelns*. Berlin: Volker Spiess.

Dekel, O., & Shamir, O. (2009). Good learners for evil teachers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, (233–240). ACM.

Den Uyl, M., & Van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. In *Proceedings of the Measuring Behaviour conference*, (589–590).

Deutsche Forschungsgemeinschaft (1999). *Qualitätskriterien der Umfrageforschung*. Berlin: Akademie Verlag.

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

Di Giacomo, E., Didimo, W., Grilli, L., & Liotta, G. (2007). Graph visualization techniques for web clustering engines. *IEEE Transactions on Visualization and Computer Graphics*, *13*(2), 294–304.

Diefenbach, D. (2001). Historical Foundations of Computer-Assisted Content. In M. West (Hrsg.) *Theory, method, and practice in computer content analysis*, (13–42). Westport: Ablex.

Diesner, J., Frantz, T., & Carley, K. (2005). Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory*, *11*(3), 201–228.

Doerfel, M., & Barnett, G. (1996). The use of Catpac for text analysis. *Cultural Anthropology Methods Journal*, *8*(2), 4–7.

DuBay, W. (2004). *The principles of readability*. Costa Mesa: Impact Information.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, (148–155). New York: ACM.

Dupagne, M., Carroll, T., & Campbell, K. (2005). Trends in Content-Analytic Research Practices in the Journal of Broadcasting & Electronic Media, 1956-2001. *Feedback*, *46*(5), 4–11.

Durant, K., & Smith, M. (2007). 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, Webkdd 2006*, (187–206). Philadelphia: Springer-Verlag New York Inc.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC.

Eilders, C. (1997). *Nachrichtenfaktoren und Rezeption: eine empirische Analyse zur Auswahl und Verarbeitung politischer Information*. Opladen: Westdeutscher Verlag.

Eilders, C., Geißler, S., Hallermayer, M., Noghero, M., & Schnurr, J.-M. (2010). Zivilgesellschaftliche Konstruktionen politischer Realität. Eine vergleichende Analyse zu Themen und Nachrichtenfaktoren in politischen Weblogs und professionellem Journalismus. *Medien und Kommunikationswissenschaft*, *58*(1), 46–62.

Ekman, P., Friesen, W., & Hager, J. (1978). *Facial action coding system*. Palo Alto: Consulting Psychologists Press.

Ellison, J. (1965). Computers and the Testaments. In E. Bowles (Hrsg.) *Computers in Humanistic Research: Readings and Perspectives*, (64–74). Englewood Cliffs: Prentice-Hall.

Erbring, L., Goldenberg, E. N., & Miller, A. H. (1980). Front-Page News and Real-World Cues: A New Look at Agenda-Setting by the Media. *American Journal of Political Science*, *24*(1), 16–49.

Erlhofer, S. (2010). Datenerhebung in der Blogosphäre: Herausforderungen und Lösungswege. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (144–166). Köln: Halem.

Eugenio, B. D., & Glass, M. (2004). The Kappa Statistic: A Second Look. *Computational Linguistics*, *30*(1), 95–101.

Evans, M., McIntosh, W., Lin, J., & Cates, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039.

Fan, D. (1988). *Predictions of public opinion from the mass media: Computer content analysis and mathematical modeling*. New York: Greenwood Pub Group.

Fan, D. (1997). Computer content analysis of press coverage and prediction of public opinion for the 1995 sovereignty referendum in Quebec. *Social Science Computer Review*, *15*(4), 351–366.

Fan, D., & McAvoy, G. (1989). Predictions of public opinion on the spread of AIDS: Introduction of new computer methodologies. *Journal of Sex Research*, *26*(2), 159–187.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, *36*(1), 259–275.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Felden, C., Bock, H., Gränding, A., & Molotowa, L. e. a. (2005). Evaluation von Algorithmen zur Textklassifikation. Tech. rep.

Finn, A., Kushmerick, N., & Smyth, B. (2001). Fact or fiction: Content classification for digital libraries. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*. Dublin.

Flesch, R. (1948). A new readability yardstick. *Journal of applied Psychology*, *32*(3), 221–233.

Forman, G., & Cohen, I. (2004). Learning from little: Comparison of classifiers given little training. In *Proceedings of Knowledge Discovery in Databases: PKDD 2004*, (161–172). Springer.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2005). Visualization of Text Document Corpus. *Informatica*, *29*(4), 497–502.

Franzosi, R. (1995). Computer-Assisted Content Analysis of Newspapers. *Quality and Quantity*, *29*(2), 157–172.

Fretwurst, B. (2008). *Nachrichten im Interesse der Zuschauer. Eine konzeptionelle und empirische Neubestimmung der Nachrichtenwerttheorie*. Konstanz: UVK Verlag.

Friedl, J. (2006). *Mastering regular expressions*. Sebastopol: O'Reilly Media, Inc.

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

Fuller, W. (1987). *Measurement error models*. New York: Wiley.

Funkhouser, G., & Parker, E. (1968). Analyzing coding reliability: The random-systematic-error coefficient. *Public Opinion Quarterly*, *32*(1), 122–128.

Galliker, M. (1998). Von der manuellen zur elektronischen Datenerhebung: Informationsquellen und Textanalysen. *ZUMA-Nachrichten*, *43*, 45–72.

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

Gehrau, V., Fretwurst, B., Krause, B., & Daschmann, G. (2005). *Auswahlverfahren in der Kommunikationswissenschaft*. Köln: Halem.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's Practice What We Preach: Turning Tables into Graphs in Statistic Research. *The American Statistician*, *56*(2), 121–130.

Gerbner, G., Holsti, O., Krippendorff, K., Paisley, W., & Stone, P. (1969). *The analysis of communication content*. New York: Wiley.

Gerhards, J., Offerhaus, A., & Roose, J. (2007). Die öffentliche Zuschreibung von Verantwortung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *59*(1), 105–124.

Ghosh, J., & Strehl, A. (2006). Similarity-Based Text Clustering: A Comparative Study. In J. Kogan, C. Nicholas, & M. Teboulle (Hrsg.) *Grouping Multidimensional Data*, (73–97). Berlin: Springer.

Gliem, J., & Gliem, R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In *Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, (82–88). East Lansing.

Goela, N., Wilson, K., Niu, F., Divakaran, A., & Otsuka, I. (2007). An SVM framework for genre-independent scene change detection. In *Proceedings of the 2007 IEEE International Conference on Multimedia*, (532–535). IEEE.

Goertz, L. (2009). Wie die Medien die Fragmentierung des Publikums verhindern. In C. Holtz-Bacha, G. Reus, & L. B. Becker (Hrsg.) *Wissenschaft mit Wirkung*, (65–72). Wiesbaden: VS Verlag.

GÖFAK Medienforschung (2010). Fernsehanalyse zum Bundestagswahlkampf 2009. Methodenbericht GLES1401 der German Longitudinal Election Study. http://www.gesis.org/fileadmin/upload/dienstleistung/forschungsdatenzentren/gles/SecureDownload/frageboegen/GLES1401_Pre1.0%20-%20Methodenbericht.pdf.

Gottschalk, L. (2000). The application of computerized content analysis of natural language in psychotherapy research now and in the future. *American Journal of Psychotherapy*, *54*(3), 305–311.

Gottschalk, L., & Gleser, G. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press.

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

Grimmer, J., & King, G. (2009). Quantitative Discovery from Qualitative Information: A General-Purpose Document Clustering Methodology. http://gking.harvard.edu/files/discov.pdf.

Gürtler, K., & Kronewald, E. (2010). The Automated Analysis of Media: prime web. Analysis: A Case Study. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (365–386). Köln: Halem.

Gwet, K. (2001). *Handbook of Inter-Rater Reliability*. Gaithersburg: StatAxis Publishing.

Haas, A., Keyling, T., & Brosius, H. (2010). Online-Diskussionsforen als Indikator für interpersonale (Offline-)Kommunikation? Methodische Ansätze und Probleme. In N. Jackob, T. Zerback, O. Jandura, & M. Maurer (Hrsg.) *Methoden der Online-Forschung: Das Internet als Forschungsinstrument und -gegenstand der Kommunikationswissenschaft.*, (63–85). Köln: Halem.

Hagen, L. M. (2001). Freitextrecherche in Mediendatenbanken als Verfahren zur computerunterstützten Inhaltsanalyse. Beschreibung, theoretische und praktische Überlegungen zur Validität und ein Anwendungsbeispiel. In W. Wirth, & L. M. Hagen (Hrsg.) *Inhaltsanalyse. Perspektiven, Probleme, Potentiale*, (337–352). Köln: Halem.

Hammer, M., & Salzinger, K. (1964). Some Formal Characteristics of Schizophrenic Speech As a Measure of Social Deviance. *Annals of the New York Academy of Sciences*, *105*(15), 861–889.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89.

Henzinger, M. (2001). Hyperlink analysis for the web. *IEEE Internet Computing*, *5*(1), 45–50.

Hillard, D., Purpura, S., & Wilkerson, J. (2007). An Active Learning Framework for Classifiying Political Text. In *Annual Meeting of the Midwest Political Science Association*. Chicago.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Holicki, S., & Brosius, H. (1988). Der Einfluß von Filmmusik und nonverbalem Verhalten der Akteure auf die Wahrnehmung und Interpretation einer Filmhandlung. *Rundfunk und Fernsehen*, *36*(2), 189–206.

Hollanders, D., & Vliegenthart, R. (2008). Telling What Yesterday's News Might be Tomorrow: Modeling Media Dynamics. *Communications*, *33*(1), 47–68.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, *13*(3), 111–117.

Holsti, O. (1966). External Conflict and Internal Consensus: The Sino-Soviet Case. In P. Stone, D. Dunphy, M. Smith, & D. Ogilvie (Hrsg.) *The General Inquirer: A Computer Approach to Content Analysis*, (343–358). Cambridge: MIT Press.

Holsti, O. (1969). *Content analysis for the social sciences and humanities*. Reading: Addison-Wesley.

Hoover, D. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, *37*(2), 151–178.

Hopkins, D., & King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, *54*(1), 229–247.

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

Hox, J., & Stoel, R. (2005). Multilevel and SEM approaches to growth curve modeling. *Encyclopedia of statistics in behavioral science*, *3*, 1296–1305.

ICH (1996). Drug Administration. Guidance for industry, E6 good clinical practice: consolidated guidance. http://www.cc.nih.gov/ccc/clinicalresearch/guidance.pdf.

Iker, H., & Harway, N. (1969). A computer systems approach toward the recognition and analysis of content. In G. Gerbner (Hrsg.) *The Analysis of Communication Content. Developments in Scientific Theories and Computer Techniques.*, (381–405). New York: Wiley.

Jackman, S. (2006). Data from the Web into R. *The Political Methodologist*, *14*(2), 11–15.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM computing surveys*, *31*(3), 265–323.

Joachims, T. (1999). Transductive inference for text classifcation using support vector machines. In *Proceeding of the International Conference on Machine Learning (ICML)*, vol. 1999.

Joachims, T. (2002). *Learning to classify text using support vector machines*. Boston: Kluwer Academic Publishers.

Kaczmirek, L. (2009). *Human-survey interaction: Usability and nonresponse in online surveys*. Halem.

Kaiser, S., & Wehrle, T. (1992). Automated coding of facial behavior in human-computer interactions with FACS. *Journal of Nonverbal Behavior*, *16*(2), 67–84.

Kantel, J. (2007). *RSS und Atom–kurz & gut*. Köln: O'Reilly.

Kastellec, J., & Leoni, E. (2007). Using graphs instead of tables in political science. *Perspectives on Politics*, *5*(4), 755–771.

Kelly, E., & Stone, P. (1975). *Computer recognition of English word senses*. Amsterdam: North-Holland.

Kepplinger, H. (2009). *Politikvermittlung*. Wiesbaden: VS Verlag.

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

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*(3), 444–452.

King, G. (2003). The future of replication. *International Studies Perspectives*, *4*(1), 72–107.

King, G., & Lowe, W. (2003). An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization*, *57*(3), 617–642.

King, G., Schlozman, K., & Nie, N. (2009). *The future of political science: 100 perspectives*. New York: Routledge.

Klebanov, B. B., Diermeier, D., & Beigman, E. (2008). Lexical Cohesion Analysis of Political Speech. *Political Analysis*, *16*(4), 447–463.

Kleinnijenhuis, J., De Ridder, J., & Rietberg, E. (1997). Reasoning in economic discourse: An application of the network approach to the Dutch press. In C. Roberts, B. Jennings, & D. Zillmann (Hrsg.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, (191–209). Mahwah: Lawrence Erlbaum Associates.

Klemmensen, R., Hobolt, S., & Hansen, M. (2007). Estimating policy positions using political texts: An evaluation of the Wordscores approach. *Electoral Studies*, *26*(4), 746–755.

Klimt, B., & Yang, Y. (2004). Introducing the Enron corpus. In *First conference on email and anti-spam (CEAS)*. Mountain View.

Klingemann, H.-D. (1984). Computerunterstützte Inhaltsanalyse und sozialwissenschaftliche Forschung. In H.-D. Klingemann (Hrsg.) *Com-*

puterunterstützte Inhaltsanalyse in der empirischen Sozialforschung, (7–14). Frankfurt a.M.: Campus.

Klingemann, H.-D., Höhe, J., Mohler, Philip, P., Radermacher, K., & Züll, C. (1984). TEXTPACK: Ein Programmsystem für sozialwissenschaftliche Inhaltsanalyse. In H.-D. Klingemann (Hrsg.) *Computerunterstützte Inhaltsanalyse in der empirischen Sozialforschung*, (15–34). Frankfurt a.M.: Campus.

Kohlschütter, C., Fankhauser, P., & Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, (441–450). New York: ACM.

Kolb, S. (2002). Verlässlichkeit von Inhaltsanalysedaten. *Medien und Kommunikationswissenschaft*, *3*, 335–354.

Kops, M. (1977). *Auswahlverfahren in der Inhaltsanalyse*. Meisenheim am Glan: Hain.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. London: Sage.

Krippendorff, K. (2004a). *Content analysis: An introduction to its methodology*. London: Sage, 2. ed.

Krippendorff, K. (2004b). Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433.

Krippendorff, K. (2009). Inferring the Readibility of Text. In K. Krippendorff, & M. A. Bock (Hrsg.) *The Content Analysis Reader*, (202–208). Thousand Oaks: Sage.

Küchenhoff, H., Mwalili, S., & Lesaffre, E. (2005). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*(1), 85–96.

Kuckartz, U. (2007). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS Verlag.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*(2), 293–323.

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

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

Lasswell, H., Lerner, D., & de Sola Pool, I. (1952). *The comparative study of symbols: An introduction*. Stanford: Stanford University Press.

Lasswell, H., & Namenwirth, J. (1968). The Lasswell Value Dictionary.

Lauf, E. (2001). ". 96 nach Holsti" Zur Reliabiliät von Inhaltsanalysen und deren Darstellung in kommunikationswissenschaftlichen Fachzeitschriften. *Publizistik*, *46*(1), 57–68.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, *97*(2), 311–331.

Lebert, M. (2005). Project Gutenberg, from 1971 to 2005. http://www.etudes-francaises.net/dossiers/gutenberg_eng.htm.

Lemaire, B., & Denhière, G. (2006). Effects of High-Order Co-occurrences on Word Semantic Similarity. *Current psychology letters*, *1*(18).

Lemnitzer, L., & Zinsmeister, H. (2006). *Korpuslinguistik: eine Einführung*. Tübingen: Gunter Narr Verlag.

Leopold, E., & Kindermann, J. (2002). Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning*, *46*(1-3), 423–444.

Leopold, E., Kindermann, J., & Paaß, G. (2007). Analysis of E-Discussions Using Classifier Induced Semantic Spaces. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *22*(1), 21–27.

Lewis, D., & Gale, W. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, (3–12). Dublin: Springer-Verlag New York, Inc.

Lewis-Beck, M., Bryman, A., & Liao, T. (2004). *The Sage encyclopedia of social science research methods*, vol. 1. Sage Publications, Inc.

Li, J., & Ezeife, C. (2006). Cleaning web pages for effective web content mining. In *Database and Expert Systems Applications*, (560–571). Berlin, Heidelberg: Springer.

Lisch, R., & Kriz, J. (1978). *Grundlagen und Modelle der Inhaltsanalyse: Bestandsaufnahme und Kritik*. Reinbek: Rowohlt.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587–604.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis*, *16*(4), 356–371.

Lücke, S. (2007). *Ernährung im Fernsehen: Eine Kultivierungsstudie zur Darstellung und Wirkung*. Wiesbaden: VS Verlag.

Luzar, K. (2004). *Inhaltsanalyse von webbasierten Informationsangeboten: Framework für die inhaltliche und strukturelle Analyse*. Norderstedt: Books on Demand.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

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

Marcinkowski, F., Greger, V., & Hüning, W. (2001). Stabilität und Wandel der Semantik des Politischen: Theoretische Zugänge und empirische Befunde. In F. Marcinkowski (Hrsg.) *Die Politik der Massenmedien. Heribert Schatz zum 65. Geburtstag*, vol. 65, (12–114). Köln: Halem.

Martin, L., & Vanberg, G. (2008). A Robust Transformation Procedure for Interpreting Political Text. *Political Analysis*, *16*(1), 93–100.

Martindale, C. (1975). *Romantic progression: The psychology of literary history*. New York: Halsted Press.

Matthes, J. (2007). *Framing-Effekte: Zum Einfluss der Politikberichterstattung auf die Einstellungen der Rezipienten*. R. Fischer.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, *58*(2), 258–279.

Maurer, M., & Reinemann, C. (2006). *Medieninhalte: eine Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Mayntz, R., Holm, K., & Hübner, P. (1974). *Einführung in die Methoden der empirischen Soziologie*. Opladen: Westdeutscher Verlag.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 Workshop on Learning for Text Categorization*, (41–48). Madison.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, *30*, 249–272.

McComas, K., & Shanahan, J. (1999). Telling stories about global climate change: Measuring the impact of narratives on issue cycles. *Communication Research*, *26*(1), 30–57.

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., Sable, C., Schiffman, B., & Sigelman, S. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster.

In *Proceedings of the second international conference on Human Language Technology Research*, (280–285).

McMillan, S. (2000). The Microscope and the Moving Target: The Challenge of Applying Content Analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*, *77*(1), 80–98.

McTavish, D. (1997). Scale validity: A computer content analysis approach. *Social Science Computer Review*, *15*(4), 379–393.

McTavish, D., Litkowski, K., & Schrader, S. (1997). A computer content analysis approach to measuring social distance in residential organizations for older people. *Social Science Computer Review*, *15*(2), 170–180.

Meier, S., Wünsch, C., Pentzold, C., & Welker, M. (2010). Auswahlverfahren für Online-Inhalte. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (102–123). Köln: Halem.

Merten, K. (1995). *Inhaltsanalyse: Einführung in die Theorie, Methode und Praxis*. Opladen: Westdeutscher Verlag.

Mikhaylov, S., Laver, M., & Benoit, K. (2008). Coder Reliability and Misclassification in Comparative Manifesto Project Codings. In *66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, April*. Chicago.

Miller, M. M. (1997). Frame Mapping and Analysis of News Coverage of Contentious Issues. *Social Science Computer Review*, *15*(4), 367–378.

Mirza, D., & Scharkow, M. (2009). Through the Eyes of the Spectator: A Content Analysis of User-Comments on the Internet Movie Database. In *Paper presented at the Society for Cognitive Studies of the Moving Image 2009 Conference*. Copenhagen.

Molinaro, A., Simon, R., & Pfeiffer, R. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307.

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

Moosbrugger, H. (2007). Klassische Testtheorie (KTT). In H. Moosbrugger, & A. Kelava (Hrsg.) *Testtheorie und Fragebogenkonstruktion*, (99–112). Springer.

Morris, M., & Ogan, C. (1996). The Internet as mass medium. *Journal of Communication*, *46*(1), 39–50.

Morris, R. (1994). Computerized content analysis in management research: A demonstration of advantages & limitations. *Journal of Management*, *20*(4), 903–931.

Morton, A. (1963). A computer challenges the church. *The Observer*, *3*, 21.

Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Reading: Addison-Wesley.

Muhr, T. (1991). ATLAS/ti—A prototype for the support of text interpretation. *Qualitative Sociology*, *14*(4), 349–371.

Muslea, I. (1999). Extraction patterns for information extraction tasks: A survey. In *The AAAI-99 Workshop on Machine Learning for Information Extraction*. Orlando.

Neuberger, C., Nuernbergk, C., & Rischke, M. (2009). Journalismus–neu vermessen: Die Grundgesamtheit journalistischer Internetangebote–Methode und Ergebnisse. In C. Neuberger, C. Nuernbergk, & M. Rischke (Hrsg.) *Journalismus im Internet*, (197–230). Wiesbaden: VS Verlag.

Neuberger, C., vom Hofe, H. J., & Nuernbergk, C. (2010). *Twitter und Journalismus. Der Einfluss des Social Web auf die Nachrichten*. Düsseldorf: LfM.

Neuendorf, K. (2002). *The content analysis guidebook*. Thousand Oaks: Sage.

North, R., Holsti, O., Zaninovich, M., & Zinnes, D. (1963). *Content analysis: A handbook with applications for the study of international crisis*. Evanston: Northwestern University Press.

Nunally, J., & Bernstein, I. (1978). *Psychometric theory*. New York: McGraw-Hill.

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. Tech. rep., Technical report, Swedish Institute of Computer Science.

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

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Park, H. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, *25*(1), 49–61.

Park, H., & Thelwall, M. (2003). Hyperlink analyses of the World Wide Web: A review. *Journal of Computer-Mediated Communication*, *8*(4).

Pennebaker, J., & Chung, C. (2009). Computerized text analysis of al-qaeda transcripts. In K. Krippendorff, & M. A. Bock (Hrsg.) *The Content Analysis Reader*, (452–466). Thousand Oaks: Sage.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The development and psychometric properties of LIWC2007. *LIWC.Net*.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Pennings, P., & Keman, H. (2002). Towards a New Methodology of Estimating Party Policy Positions. *Quality and Quantity*, *36*(1), 55–79.

Pfetsch, B. (2004). The Voice of the Media in European Public Sphere: Comparative Analysis of Newspaper Editorials. http://europub.wz-berlin.de.

Pool, I. d. S. (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.

Popping, R. (2000). *Computer-assisted text analysis*. Thousand Oaks: Sage.

Popping, R. (2009). Some views on agreement to be used in content analysis studies. *Quality and Quantity*, *44*(6), 1067–1078.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Porter, M. (2001). Snowball: A language for stemming algorithms. snowball.tartarus.org.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284.

Powers, D. M. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. http://david.wardpowers.info/BM/ECAIrej-Significance_Confidence.pdf.

Purpura, S., & Hillard, D. (2006). Automated classification of congressional legislation. *Proceedings of the 2006 international conference on Digital government research*, (219–225).

Quandt, T. (2008a). Neues Medium, alter Journalismus? Eine vergleichende Inhaltsanalyse tagesaktueller Print-und Online-Nachrichtenangebote. In T. Quandt, & W. Schweiger (Hrsg.) *Journalismus online-Partizipation oder Profession?*, (131–155). Wiesbaden: VS Verlag.

Quandt, T. (2008b). (No) News On The World Wide Web? *Journalism Studies*, *9*(5), 717–738.

Quasthoff, U. (1998). Projekt Der deutsche Wortschatz . In G. Heyer, & C. Wolff (Hrsg.) *Linguistik und neue Medien*. Wiesbaden: DUV.

Quinn, K., Monroe, B., Colaresi, M., Crespin, M., & Radev, D. (2006). An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th US Senate. In *In Midwest Political Science Association Meeting*.

Radhakrishan, R., Xiong, Z., Divakaran, A., & Ishikawa, Y. (2004). Generation of sports highlights using a combination of supervised & unsupervised learning in audio domain. In *Proceedings of the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 2, (935–939). IEEE.

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

Raykar, V., Yu, S., Zhao, L., Jerebko, A., Florin, C., Valadez, G., Bogoni, L., & Moy, L. (2009). Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, (889–896). ACM.

Reidsma, D., & Carletta, J. (2008). Reliability measurement without limits. *Computational Linguistics*, *34*(3), 319–326.

Rice, R. (1994). Network analysis and computer-mediated communication systems. In S. Wassermann, & J. Galaskiewicz (Hrsg.) *Advances in social network analysis: Research in the social and behavioral sciences*, (167–203). Thousand Oaks: Sage.

Riffe, D., & Freitag, A. (1997). A content analysis of content analyses: Twenty-five years of Journalism Quarterly. *Journalism and Mass Communication Quarterly*, *74*(3), 515–524.

Riffe, D., Lacy, S., & Fico, F. (2005). *Analyzing media messages: Using quantitative content analysis in research*. Mahwah: Lawrence Erlbaum.

Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, (130–136). New York: ACM.

Ritsert, J. (1972). *Inhaltsanalyse und Ideologiekritik: ein Versuch über kritische Sozialforschung*. Frankfurt a.M.: Athenäum Fischer.

Roberts, C. (1997a). Introduction. In C. Roberts (Hrsg.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, (1–8). Mahwah: Lawrence Erlbaum Associates.

Roberts, C. (1997b). Semantic Text Analysis. In C. Roberts (Hrsg.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, (55–78). Mahwah: Lawrence Erlbaum Associates.

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

Rodriguez, J., Perez, A., & Lozano, J. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575.

Rosen, D., & Corbit, M. (2009). Social network analysis in virtual environments. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, (317–322). Torino: ACM.

Rosenberg, S., Schnurr, P., & Oxman, T. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment*, *54*(1), 298–310.

Rössler, P. (2002). Content analysis in online communication: A challenge for traditional methodology. In B. Batinic, U.-D. Reips, & M. Bosnjak (Hrsg.) *Online Social Sciences*, (291–307). Toronto: Hofgrefe & Huber.

Rössler, P. (2005). *Inhaltsanalyse*. Konstanz: UVK.

Rössler, P. (2010). Das Medium ist nicht die Botschaft. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (31–43). Köln: Halem.

Rössler, P., & Wirth, W. (2001). Inhaltsanalysen im World Wide Web. In W. Wirth, & E. Lauf (Hrsg.) *Inhaltsanalyse. Perspektiven, Probleme, Potentiale.*, (280–302). Köln: Halem.

Rüdiger, K., & Welker, M. (2010). Redaktionsblogs deutscher Zeitungen. Über die Schwierigkeiten diese inhaltsanalytisch zu untersuchen – ein Werkstattbericht. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (448–468). Köln: Halem.

Ruhrmann, G., Woelke, J., Maier, M., & Diehlmann, N. (2003). *Der Wert von Nachrichten im deutschen Fernsehen*. Opladen: Leske+ Budrich.

Salisbury, J. (2001). Using neural networks to assess corporate image. In M. D. West (Hrsg.) *Applications of Computer Content Analysis*, (65–86). Westport: Ablex Pub.

Saris-Gallhofer, I. N., Saris, W. E., & Morton, E. L. (1978). A validation study of Holsti's content analysis procedure. *Quality and Quantity*, *12*(2), 131–145.

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE*, *23*(2), 133–141.

Scharkow, M. (2007). Scraping Youtube with Beautiful Soup. http://underused.org/2007/12/11/scraping-youtube-with-beautiful-soup/.

Scharkow, M. (2010a). Crowdsourcing von Inhaltsanalysen im Word Wide Web? In N. Jackob, T. Zerback, O. Jandura, & M. Maurer (Hrsg.) *Methoden der Online-Forschung: Das Internet als Forschungsinstrument und -gegenstand der Kommunikationswissenschaft.*, (301–315). Köln: Halem.

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

Scharrer, E. (2002). An Improbable Leap: a content analysis of newspaper coverage of Hillary Clinton's transition from first lady to Senate candidate. *Journalism Studies*, *3*(3), 393–406.

Scheufele, B. (2003). *Frames-Framing-Framing-Effekte: Theoretische und methodische grundlegung des Framing-Ansatzes sowie empirische Befunde zur nachrichtenproduktion*. Wiesbaden: VS Verlag.

Scheufele, B., & Engelmann, I. (2009). *Empirische Kommunikationsforschung*. Konstanz: UVK.

Schönbach, K. (1978). Nachrichtenwerte und computerunterstützte Inhaltsanalyse. *ZUMA Nachrichten*, *2*, 3–11.

Schönbach, K. (1982). The Issues of the Seventies. *Publizistik*, *27*(1-2), 129–140.

Schrodt, P., Davis, S., & Weddle, J. (1994). Political Science: KEDS – A Program for the Machine Coding of Event Data. *Social Science Computer Review*, *12*(4), 561–587.

Schrodt, P., & Donald, C. (1990). Machine Coding of Events Data. In *International Studies Association meetings*. Washington.

Schulz, W. (1976). *Die Konstruktion von Realität in den Nachrichtenmedien: Analyse der aktuellen Berichterstattung*. Freiburg, München: Alber.

Schwartz, J. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research*, *13*(4), 435–466.

Schweiger, W., & Weber, P. (2010). Strategische Kommunikation auf Unternehmens-Websites. Zur Evaluation der Kommunikationsleistung durch eine Methodenkombination von Online-Inhaltsanalyse und Logfile-Analyse. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (267–290). Köln: Halem.

Schweitzer, E. J. (2010). Politische Websites als Gegenstand der Online-Inhaltsanalyse. In M. Welker, & C. Wünsch (Hrsg.) *Die Online-Inhaltsanalyse*, (44–102). Köln: Halem.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In *Proceedings of ICML-99, 16th International Conference on Machine Learning*, (379–388). San Francisco.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321–325.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Sebe, N., Lew, M., Sun, Y., Cohen, I., Gevers, T., & Huang, T. (2007). Authentic facial expression analysis. *Image and Vision Computing*, *25*(12), 1856–1863.

Seibold, B. (2002). Die flüchtigen Web-Informationen einfangen. *Publizistik*, *47*(1), 45–56.

Settles, B. (2010). Active Learning Literature Survey. Tech. rep., University of Wisconsin–Madison.

Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (1070–1079). Waikiki: Association for Computational Linguistics.

Shapiro, G. (1997). The Future of Coders. In C. Roberts (Hrsg.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*, (225–239). Mahwah: Lawrence Erlbaum Associates.

Shellman, S. M. (2008). Coding Disaggregated Intrastate Conflict: Machine Processing the Behavior of Substate Actors Over Time and Space. *Political Analysis*, *16*(4), 464–477.

Sheng, V., Provost, F., & Ipeirotis, P. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, (614–622). Las Vegas: ACM.

Siefkes, C. (2007). *An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models*. Ph.D. thesis, Freie Universität Berlin.

Siefkes, C., Assis, F., Chhabra, S., & Yerazunis, W. S. (2004). Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, (410–421). Pisa.

Simon, A. F., & Xenos, M. (2004). Dimensional Reduction of Word-Frequency Data as a Substitute for Intersubjective Content Analysis. *Political Analysis*, *12*(1), 63–75.

Smith, M. (2003). Measures and maps of Usenet. In C. Lueg, & D. Fisher (Hrsg.) *From Usenet to CoWebs: Interacting with Social Information Spaces*, (47–78). London: Springer.

Snider, J., & Janda, K. (1998). Newspapers in Bytes and Bits: Limitations of Electronic Databases for Content Analysis. In *Paper presented at the annual meeting of the American Political Science Association*. Boston.

Snoek, C., Worring, M., & Hauptmann, A. (2006). Learning rich semantics from news video archives by style analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, *2*(2), 91–108.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (254–263). Association for Computational Linguistics.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, *34*(1), 233–272.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *Lecture Notes in Computer Science*, *4304*, 1015.

Spiegel Online (2007). Ganz, ganz tief im Westen. SPIEGEL ONLINE, 26.10.2007 http://www.spiegel.de/netzwelt/web/0,1518,513770,00.html.

SPIEGEL Verlag (2007). SPIEGELnet und Wissen Media starten SPIEGEL Wissen. http://www.spiegelgruppe.de/spiegelgruppe/home.nsf/pmwebaktuell/1619A963C27E7741C12573B4003468B9.

Stegbauer, C., & Rausch, A. (2006). *Strukturalistische Internetforschung Netzwerkanalysen internetbasierter Kommunikationsräume*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Stegmann, J., & Lücking, A. (2005). Assessing reliability on annotations (1): Theoretical considerations. Tech. rep.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, vol. 34, (35).

Stephen, T. (1999). Computer-assisted concept analysis of HCR's first 25 years. *Human Communication Research*, *25*(4), 498–513.

Stewart, B., & Zhukov, Y. (2009). Use of force and civil–military relations in Russia: an automated content analysis. *Small Wars & Insurgencies*, *20*(2), 319–343.

Stone, P. (1969a). Confrontation of Issues: Excerpts from the Discussion Session at the Conference. In G. Gerbner, O. Holsti, K. Krippendorff, W. Paisley, & P. Stone (Hrsg.) *The Analysis of Communication Content*, (523–537). New York: Wiley.

Stone, P. (1969b). Improved quality of content analysis categories: Computerized-disambiguation rules for high-frequency English words. In G. Gerbner, O. Holsti, K. Krippendorff, W. Paisley, & P. Stone (Hrsg.) *The Analysis of Communication Content*, (199–221). New York: Wiley.

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

Stone, P., Dunphy, D., Smith, M., & Ogilvie, D. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: The MIT Press.

Suckfüll, M. (1997). *Film erleben: Narrative Strukturen und physiologische Prozesse-"Das Piano"von Jane Campion*. Berlin: Edition Sigma.

Taddicken, M. (2008). *Methodeneffekte bei Web-Befragungen*. Köln: Halem.

Tankard, J., Hendrickson, L., & Lee, D. (1994). Using Lexis/Nexis and other databases for content analysis: Opportunities and risks. In *Annual meeting of the Association for Education in Journalism and Mass Communication*, (2006). Atlanta.

Tankard, J. W. (2001). Using the Computer to Identify Unknown Authors. In M. D. West (Hrsg.) *Applications of Computer Content Analysis*, (51–64). Westport: Ablex Pub.

Teichert, T., & Schöntag, K. (2009). Exploring consumer knowledge structures using associative network analysis. *Psychology and Marketing*, *27*(4), 369–398.

Tomlinson, S. (2003). Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer TM at CLEF 2003. In *4th Workshop of the Cross-Language Evaluation Forum*. Trondheim.

Trebbe, J. (1996). *Der Beitrag privater Lokalradio-und Lokalfernsehprogramme zur publizistischen Vielfalt*. München: Fischer.

Urban, D. (2002). Prozessanalyse im Strukturgleichungsmodell: Zur Anwendung latenter Wachstumskurvenmodelle in der Sozialisationsforschung. *ZA-Information*, *51*, 6–37.

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

Van Der Linden, W., & Glas, C. (2000). *Computerized adaptive testing: Theory and practice*. Springer Netherlands.

Volkens, A. (2007). Strengths and Weaknesses of Approaches to Measuring Policy Positions of Parties. *Electoral Studies*, *26*(1), 108–120.

Vu, H. N. N., & Gehrau, V. (2010). Agenda Diffusion: An integrated model of agenda setting and interpersonal communication. *Journalism Mass Communication Quarterly*, *87*(1), 100–116.

Wallach, H. (2004). Evaluation metrics for hard classifiers. *Unpublished note (http://www. inference. phy. cam. ac. uk/hmw26/papers/evaluation. ps)*.

Weare, C., & Lin, W. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review*, *18*(3), 272–292.

Weaver, D., & Bimber, B. (2008). Finding News Stories: A Comparison of Searches Using LexisNexis and Google News. *Journalism & Mass Communication Quarterly*, *85*(3), 515–530.

Weber, R. P. (1983). Measurement Models for Content Analysis. *Quality and Quantity*, *17*(2), 127–149.

Weber, R. P. (1984). Computer-aided content analysis: A short primer. *Qualitative sociology*, *7*(1), 126–147.

Weber, R. P. (1990). *Basic content analysis*. Newbury Park, London, New Delhi: Sage.

Weiß, H. (1998). *Auf dem Weg zu einer kontinuierlichen Fernsehprogrammforschung der Landesmedienanstalten*. Berlin: Vistas.

Welker, M., Werner, A., & Scholz, J. (2005). *Online-Research: Markt-und Sozialforschung mit dem Internet*. Heidelberg: dpunkt.

Wenger, L., Malone, R., & Bero, L. (2001). The cigar revival and the popular press: a content analysis, 1987-1997. *American Journal of Public Health*, *91*(2), 288–291.

Wessler, H. (2008). *Transnationalization of public spheres*. Basingstoke: Palgrave Macmillan.

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

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 2009 Neural Information Processing Systems (NIPS) Conference*.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, *20*(2), 233–287.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, (246–253). College Park, Maryland: Association for Computational Linguistics.

Wilke, J., & Reinemann, C. (2000). *Kanzlerkandidaten in der Wahlkampfberichterstattung: Eine vergleichende Studie zu den Bundestagswahlen 1949-1998*. Köln: Böhlau.

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

Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Pub.

Wolling, J. (2002). Methodenkombination in der Medienwirkungsforschung. Der Entscheidungsprozess bei der Verkn
üpfung von Umfrage-und Inhaltsanalysedaten. *ZUMA-Nachrichten*, *50*, 54–85.

Xu, M., Maddage, N., Xu, C., Kankanhalli, M., & Tian, Q. (2003). Creating audio keywords for event detection in soccer video. In *Proceedings of the 2003 International Conference on Multimedia*, vol. 2. IEEE.

Zeller, F., & Wolling, J. (2010). Struktur- und Qualitätsanalyse publizistischer Onlineangebote. *Media Perspektiven*, *3*, 143–153.

Zerback, T., Schoen, H., Jackob, N., & Schlereth, S. (2008). Zehn Jahre Sozialforschung mit dem Internet—eine Analyse zur Nutzung von Online-Befragungen in den Sozialwissenschaften. In N. Jackob, H. Schoen, & T. Zerback (Hrsg.) *Sozialforschung im Internet*, (15–31). Springer.

Zipf, G. (1965). *The psycho-biology of language*. Cambridge: MIT Press.

Züll, C., & Alexa, M. (2001). Automatisches Codieren von Textdaten. Ein Überblick über neue Entwicklungen. In W. Wirth, & E. Lauf (Hrsg.) *Inhaltsanalyse–Perspektiven, Probleme, Potenziale*, (303–317). Halem.

Züll, C., & Landmann, J. (2002). Computerunterstützte Inhaltsanalyse: Literaturbericht zu neueren Anwendungen. ZUMA Methodenbericht 2002/02.

Züll, C., Weber, R., & Mohler, P. (1989). Computer-Assisted Text Classification: The General Inquirer III.

A NewsClassifier – eine Software zur manuellen und automatischen Inhaltsanalyse

Grundlegende Überlegungen

Im Folgenden wird die Konzeption und Implementation eines computergestützten Forschungsinstruments für Inhaltsanalysen dargestellt, mit dem sich manuelle und automatische Verfahren sinnvoll kombinieren lassen. Zu diesem Zweck habe ich ein Forschungsinstrument (oder eher *Framework*) entwickelt, mit dessen Hilfe sich konventionelle und automatische Inhaltsanalysen von digitalen Texten durchführen lassen. Dieses Framework, im Folgenden nach seinem primären Zweck NewsClassifier genannt, wurde parallel zu dieser Arbeit entwickelt und auch als Grundlage für die empirische Evaluation verwendet. Wie in vergleichbaren Monographien zu automatischen Analyseverfahren (Luzar, 2004; Atteveldt, 2008) dient dieses Kapitel zwei Zwecken: Einerseits als Methodendokumentation, die das Verständnis der eigentlichen empirischen Studie erleichtern soll, andererseits als *Proof-of-Concept* für die Möglichkeiten der Automatisierung in einem typischen Forschungsprojekt.

Da sich viele konzeptionelle Ideen nicht nur leichter anhand konkreter Umsetzungsvorschläge nachvollziehen lassen, sondern häufig erst im Zuge der tatsächlichen technischen Umsetzung entwickelt werden können, erscheint es mir sinnvoll, beides gemeinsam darzustellen (vgl. Muhr, 1991).

Bei Entwurf und Umsetzung des inhaltsanalytischen Frameworks standen drei Gedanken im Vordergrund: Erstens ist es sowohl methodologisch als auch forschungsökonomisch wünschenswert, alle Schritte im Forschungsprozess zu automatisieren, die nicht direkt mit der konkreten Untersuchungsanlage bzw. Forschungsfrage verknüpft sind und daher ein Eingreifen von Forschungsleiter oder Codierer erfordern. Dies minimiert potentielle Fehlerquellen, gewährleistet eine hohe Replikati-

onsfähigkeit der Inhaltsanalyse und ermöglicht es dem Forscher, Umfang und Art der Codierung gezielt zu steuern.

Zweitens soll das Instrument für unterschiedlichste Untersuchungsdesigns einsetzbar sein und mit den wissenschaftlichen und infrastrukturellen Anforderungen der Analyse skalieren. Ziel ist es, den Forscher mit Hilfe der Software zu unterstützen und soweit zu entlasten, dass dieser sich weitestgehend der theoretischen und empirischen Forschungsarbeit widmen kann. Die Analyse kann dabei auf mindestens zwei Wegen verbessert werden, die ich in Abschnitt 2.2.1 diskutiert habe: Durch eine breitere Messung, d.h. die Verwendung mehrerer Operationalisierungen, mehrerer Codierer oder Klassifikationsverfahren, kann die Reliabilität und Validität der Analyse erhöht werden. Eine umfangreichere Codierung reduziert zudem die Unsicherheit der statistischen Inferenzen.

Drittens kann und soll es bei einer konzeptionellen Weiterentwicklung nicht darum gehen, den inhaltsanalytischen Forschungsprozess gänzlich neu zu gestalten. Vielmehr sollen für die in Kapitel 4 formulierten methodischen Herausforderungen automatisierte Lösungen gefunden werden. Die Anschlussfähigkeit an die über Jahrzehnte weiterentwickelte Forschungspraxis, etwa in der kontinuierlichen Programm- (Weiß, 1998; Bruns & Marcinkowski, 1997) oder Presseanalyse (Wilke & Reinemann, 2000; Pfetsch, 2004) erhält daher oberste Priorität. Da für die Online-Inhaltsanalyse bislang weniger Erfahrungen vorliegen (vgl. Rössler, 2010; Rüdiger & Welker, 2010), werden im Folgenden gerade für online-spezifische Forschungsprobleme einige neue Lösungen vorgestellt, die sich im Rahmen dieser Arbeit bewährt haben.

NewsClassifier ist als integriertes Framework konzipiert, das von der automatischen Datenerhebung und -bereinigung über die Stichprobenziehung, die Organisation der Feldarbeit und die Durchführung von Reliabilitätstests bis hin zur eigentlichen manuellen und/oder automatischen Codierung reicht. Ein Schwerpunkt liegt dabei, dem Kernthema dieser Arbeit folgend, in der Anwendung von Verfahren aus dem maschinellen Lernen für die Analyse von Texten sowie der Verknüpfung manueller und automatischer Codierung. Der gesamte Forschungsprozess, der sich so abbilden lässt, ist in Abbildung A.1 auf der nächsten Seite in Form eines Flussdiagramms dargestellt.

Konkret umgesetzt wurde dieses Framework als server-basierte Online-Applikation, die sowohl für Ad-hoc-Analysen mit kleinen Stichproben und wenigen Variablen als auch für umfangreiche und kontinuierliche Studien eingesetzt werden kann. Dabei kann es sich beim Untersuchungsmaterial einerseits um genuine Webinhalte wie HTML-Seiten, Emails oder Twitter-Mitteilungen handeln, andererseits sind auch digitalisierte Inhalte von klassischen Offline-Medien analysierbar. Da in absehbarer Zukunft online-basierte Kommunikation noch zunehmen wird, sollte sich das Instrument für eine Vielzahl von Forschungsfragen eignen.

Wie in Abbildung A.1 dargestellt, besteht das Framework aus drei großen Komponenten: Datenerhebung, klassische inhaltsanalytische Codierung und überwachte Textklassifikation, die im Zentrum dieser Arbeit steht. In Anlehnung an die in Kapitel 4 formulierten Problemfelder werden diese drei Bestandteile von NewsClassifier in den nächsten Abschnitten behandelt. Die zentralen Fragen hierfür sind dementsprechend: Wie lassen sich Datenerhebung und -management automatisieren? Wie kann der Computer die manuelle Codierung unterstützen? Wie lässt sich die überwachte Textklassifikation optimal in den inhaltsanalytischen Forschungsprozess integrieren?

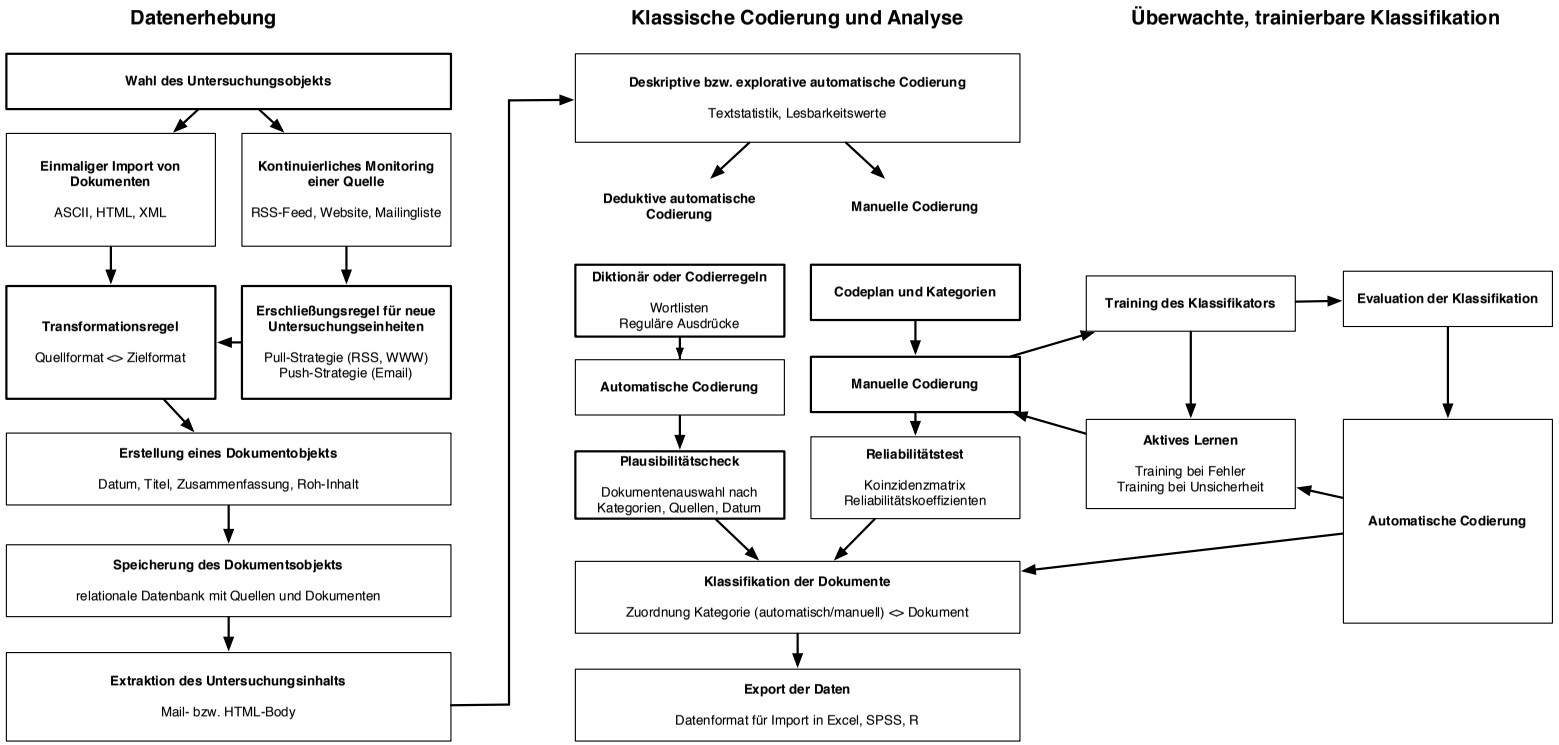

Abbildung A.1: Inhaltsanalytischer Forschungsprozess mit NewsClassifier

Online-Datenerhebung und -archivierung

Datenmodell und Erhebungsstrategien

Jede Inhaltsanalyse beginnt mit der Definition des Untersuchungsgegenstands, der Eingrenzung des Untersuchungsmaterials sowie der Auswahl- und Analyseeinheiten (Rössler, 2005). Auch bei der Entwicklung des hier vorgestellten Forschungsinstruments standen diese Überlegungen an erster Stelle. Grundlage des Datenmodells von NewsClassifier ist die Analyse- oder Codiereinheit, die in den meisten Fällen sowohl bei der Codierung als auch bei der Auswertung der Ergebnisse den einzelnen Fall definiert. Dieses Objekt wird im Folgenden allgemein als *Dokument* bezeichnet, unabhängig davon, ob es sich um eine Email, einen Blog-Eintrag, ein Video oder einen Nachrichtenartikel handelt.

Alle Dokumente werden in einer relationalen Datenbank abgelegt, in der neben dem eigentlichen Inhalt, d.h. Überschrift, Teaser bzw. Zusammenfassung und Hauptinhalt, auch verschiedene Meta-Daten wie der Erscheinungszeitpunkt, Umfang in Zeichen, Position oder Seitenzahl sowie eine eindeutige URL oder ID erfasst wird. Jedem Dokument wird zudem eine sog. *Quelle* zugeordnet, d.h. zumeist das Medienangebot, zu dem es gehört. Die Objekte der Klasse *Quelle* enthalten neben einem Namen auch Angaben dazu, wie die Dokumente aus ihnen erhoben werden können. Grundsätzlich sind dadurch beliebig viele Dokumente aus beliebig vielen Quellen effizient archivier- und analysierbar. Die Eigenschaften von Dokument und Quelle sind in Abbildung A.2 dargestellt. Durch die Speicherung von Meta-Daten im Dokument-Objekt lassen sich vielfältige Abfragen und Möglichkeiten der Stichprobenziehung realisieren, etwa nach Zeiträumen, Wochentagen, Medienangeboten oder Schlagwörtern, die in Titel oder Teaser vorkommen.

Wichtiger als die Frage nach der reinen Software-Umsetzung ist die nach der tatsächlichen Datenerhebung: Welche Arten von inhaltsanalytisch relevanten Daten lassen sich wie mit einem Online-Tool erheben? Dabei sind zwei forschungspraktische Problemfelder der Datenerhebung von Bedeutung: die Strukturiertheit des Untersuchungsmaterials und die Frequenz der Erhebung.

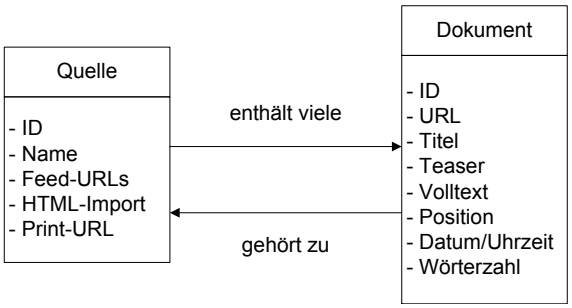

Abbildung A.2: Aufbau von Quell- und Dokumentobjekten

Für den ersten Aspekt der Strukturierheit spielt die Frage nach der digitalen Dokumentenform eine wichtige Rolle. Da Texte in der Regel nur durch die Syntax der Sprache und typografische Hilfsmittel strukturiert sind, müssen diese entweder aufwändig in strukturierte Dokumente überführt oder lediglich als geordnete Reihe von Wörtern behandelt werden.[1] Für die automatische Verarbeitung der Dokumente eignen sich daher stärker strukturierte Datenformate wie XML besser, da hierfür eigene regelbasierte Parser existieren, die zuverlässig auf Titel, Autor, Teaser und Haupttext eines Dokuments zugreifen können. Vor dem eigentlichen Import der Daten ist es bei unstrukturierten Dokumenten, etwa dem ASCII-Text-Output von LexisNexis, Email-Archiven im mbox-Format oder digitalisierten Magazinartikeln in PDF-Form, notwendig, eigene Transformationsroutinen zu entwerfen, mit denen diese in das Zielformat der Datenbank konvertiert werden. Für Datenbanken wie LexisNexis oder Factiva existieren bereits solche Importfilter, bei denen als Zwischenformat die Auszeichnungssprache XML (Bray et al., 2000) dient, die sich flexibel einsetzen lässt. Das zweite relevante Problemfeld der Datenerhebung mit NewsClassifier ist die Frequenz der Erhebung. Während die meisten Inhaltsanalysen mit archiviertem Material arbeiten, das einmalig importiert und dann codiert wird, ist es bei Online-Inhaltsanalysen auch möglich – und ggf. notwendig –, kontinuierlich auf Inhalte aus dem Inter-

1 In diesem Abschnitt geht es nur um Texte, nicht aber um Bilder, Ton- und Videobeiträge, für die ohnehin andere – und in jedem Fall aufwändigere – Transformationsschritte notwendig sind, um sie automatisch analysieren zu können.

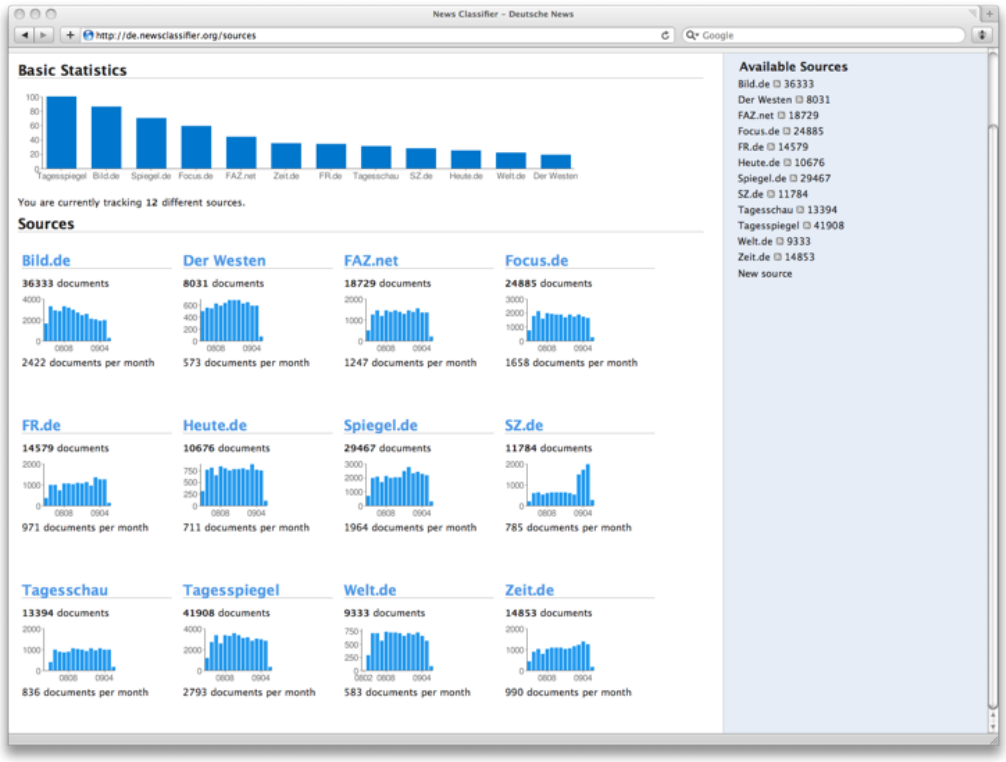

Abbildung A.3: NewsClassifier: Übersichtsseite der Quellenverwaltung

net zuzugreifen. Im Unterschied zu passiven Importen kann die Software dann dazu genutzt werden, aktiv die Inhalte anderer Websites herunterzuladen und zu archivieren. Die gesamte Erhebung ist dabei über ein zentrales Web-Interface für einen oder mehrere Projektmitarbeiter kontrollierbar (vgl. Abbildung A.3)

Der einmalige Import von reinen Textdaten ist vergleichsweise unkompliziert: Die Dokumente werden gesammelt, transformiert und anschließend in der Datenbank gespeichert. Dies ist beispielsweise der Fall, wenn bereits vorliegende Artikel aus Datenbanken wie LexisNexis exportiert und anschließend in NewsClassifier importiert werden. Für ein noch laufendes Pilotprojekt zur Wirtschaftsberichterstattung im Wall Street Journal wurde diese Strategie bereits erfolgreich angewandt.

Dabei wurden rund 50.000 Artikel zuerst aus der Factiva-Datenbank heruntergeladen, in das XML-Format konvertiert und anschließend in ein Projekt importiert. Die Konvertierung und der Import erfolgten dabei ohne manuelle Eingriffe und waren in wenigen Minuten abgeschlossen.

Da die Erhebung von genuinen Online-Inhalten wie Webseiten, Blog-Postings, Twitter-Meldungen oder Foren-Beiträgen deutlich größere Anforderungen stellt, will ich im folgenden Abschnitt einen innovativen Ansatz skizzieren, mit dem sich Online-Nachrichten und andere aktuelle Inhalte fast in Echtzeit im World Wide Web sammeln, archivieren und für die sofortige Codierung vorbereiten lassen.

Erhebung von Online-Nachrichten

Online-Inhaltsanalysen zeichnen sich vor allem durch die Dynamik, Hypertextualität und Flüchtigkeit des Untersuchungsmaterials aus (Seibold, 2002). Während die Flüchtigkeit der Inhalte auf Websites durch häufige Aktualisierungen für die Verwendung automatisierter Erhebungsformen spricht, stellt die große Gestaltungsfreiheit von Websites ein Hindernis für ein solches Vorgehen dar. Dies zeigt sich bereits bei der Auswahl der Untersuchungseinheiten, die bei Online-Inhaltsanalysen zumeist mehrstufig erfolgt.

Auf der ersten Stufe müssen Webangebote ausgewählt werden, die entweder als Stichprobe oder Vollerhebung die gewünschte Grundgesamtheit abdecken. Da eine echte zufällige Stichprobenziehung für Online-Angebote nicht realisierbar ist, ist die gängige Forschungspraxis durch bewußte Auswahlentscheidungen geprägt (Rössler & Wirth, 2001; Meier et al., 2010). Dies ist allerdings bei den meisten Inhaltsanalysen von Print- oder Rundfunkangeboten ebenso der Fall (Rössler, 2005). Da die Stichprobenziehung von Medienangeboten ein prinzipielles und kein technisches Problem ist, kann dieser Schritt nicht sinnvoll durch eine Softwarelösung unterstützt werden. Relevant für die Anwendung automatischer Erhebungsverfahren ist daher vor allem die zweite Ebene des Sampling, in der die eigentlichen Untersuchungseinheiten ausgewählt werden.

Gürtler & Kronewald (2010, 372) unterscheiden grundsätzlich zwei Ansätze, mit denen Online-Inhalte automatisiert erhoben und archiviert

werden können: Zum einen können Webcrawler eingesetzt werden, die den HTML-Code und zugehörige Dateien einer Website speichern, zum anderen können RSS-Feeds heruntergeladen werden, die bereits die relevanten Inhalte eines Webangebots enthalten.

HTML-Crawling
Crawling-Software wird bislang bevorzugt für die Analyse von Online-Nachrichten, kompletten Webauftritten oder einzelnen Websites eingesetzt. Ausgehend von einer oder mehreren Start-Adressen werden nicht deren HTML-Inhalte gespeichert, sondern auch den dort gesetzten Links gefolgt, deren Zielseiten wiederum heruntergeladen werden. Diese Rekursion kann bis zu einer beliebigen Tiefe fortgesetzt werden, allerdings wird in der Praxis meist nicht weiter als zwei Ebenen (oder Klicks) von der Startseite „gecrawlt" (Quandt, 2008b). Die heruntergeladenen Seiten können dann für manuelle Analysen auch offline betrachtet und ggf. durchsucht werden. Das Verfahren des Webcrawlings hat eine Reihe von Vorteilen: Die Seiten werden sicher archiviert, das Verfahren eignet sich für alle HTML-basierten Angebote, es ist eine Vielzahl von Programmen verfügbar, die sich leicht bedienen lassen (Rüdiger & Welker, 2010). Aus diesen Gründen hat sich der Einsatz von Crawling-Software mittlerweile zum de facto Standard der Online-Inhaltsanalyse entwickelt (vgl. auch Schweiger & Weber, 2010; Schweitzer, 2010). Allerdings ist das Verfahren nicht ohne Nachteile: Die Software sammelt beim Crawling oft zahlreiche irrelevante Inhalte, die später aufwändig aus den Daten entfernt werden müssen. Auch können nicht alle Inhalte, die der Nutzer im Browser sieht, auch gespeichert werden. Dies gilt vor allem für Adobe Flash Inhalte wie Videos oder interaktive Grafiken. Drittens sind die gespeicherten Websites für automatische Textanalysen nicht gut geeignet, da sie vergleichsweise frei gestaltet und unstrukturiert sind (vgl. Abschnitt 4.2.1).

Für die Erhebung von Online-Nachrichten hat das rekursive Crawling von Nachrichten-Sites wie Spiegel Online zudem den Nachteil, das potentiell relevante Beiträge, die nicht auf der Startseite verlinkt sind, nicht gefunden werden. Die Anwendung dieses Schneeball-Verfahrens zur Nachrichtensammlung führt daher bei liberalen Einstellungen, d.h. großer Rekursionstiefe, zu vielen falsch positiven Untersuchungseinhei-

ten. Bei relativ strengen Einstellungen, z.B. der ausschließlichen Erhebung von Beiträgen, die auf der Startseite verlinkt sind, wird nur ein Bruchteil aller aktuellen Nachrichten gefunden. Zudem bestimmt der Zeitpunkt der Erhebung maßgeblich die Menge gefundener Beiträge (vgl. Quandt, 2008b; Rüdiger & Welker, 2010).

RSS-Feeds

Für die Publikation neuer Online-Inhalte hat sich seit einigen Jahren das Format der RSS- oder ATOM-Feeds durchgesetzt, in denen in strukturierter Form aktuelle Inhalte von Webangeboten zur Archivierung und Weiterverarbeitung angeboten werden (Kantel, 2007). Strukturell sind RSS-Feeds wie ein Nachrichten-Ticker aufgebaut, so dass aktuelle Meldungen immer zuerst im Dokument stehen. Die Feeds werden kontinuierlich aktualisiert und können ebenso kontinuierlich abgefragt werden, was die Datenerhebung sehr vereinfacht.

Durch die zunehmende Verwendung von professionellen Content-Management-Systemen bieten heute so gut wie alle Nachrichtenmedien im World Wide Web eigene Feeds für ihre Inhalte an. Dies gilt für Medienangebote wie Spiegel Online und die Tagesschau ebenso wie für Weblogs, Foren, Social Networking Sites wie Facebook oder Twitter und die Webauftritte von öffentlichen Institutionen und Konzernen.[2] Viele Nachrichtenangebote veröffentlichen sogar ressort-spezifische Feeds, so dass sich entsprechende Aufgriffkriterien, etwa dass nur Beiträge im Politikteil der Seite erhoben werden, unproblematisch umsetzen können.

Standardisierte RSS-Feeds haben für die automatisierte Datenerhebung den großen Vorteil, dass die Inhalte beliebiger Feeds mit denselben Regeln archivierbar sind (Erlhofer, 2010). Der Datenerhebungsprozess mit NewsClassifier beschränkt sich in diesem Fall darauf, pro Medienangebot die Feed-URL zu speichern und diese dann regelmäßig automatisch herunterladen zu lassen.[3] So gut dieses Verfahren bei der Erhebung von

2 Zum Zeitpunkt der Datenerhebung im Jahr 2008 hatte von allen publikumsstarken Print- und Rundfunkmedien lediglich die TAZ keinen RSS-Feed. Dieser wurden jedoch kurze Zeit später eingeführt, so dass m.W. aktuell kein Angebot von umfangreichen Online-Nachrichten ohne RSS- oder ATOM-Feeds existiert.

3 Selbst die Eingabe der Feed URL lässt sich noch vereinfachen, da der Hauptfeed stets auf der Startseite des Webangebots verlinkt wird und damit automatisch zu finden ist.

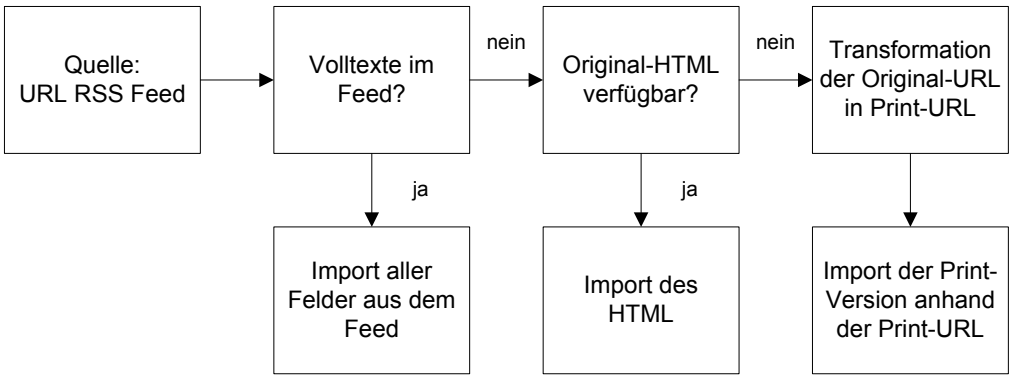

Abbildung A.4: Entscheidungen bei der Datenerhebung mit NewsClassifier

Weblogs und anderen Inhalten, z.B. Pressemitteilungen, funktioniert, scheitert es zur Zeit an der Veröffentlichungspolitik der Medienunternehmen, die in ihren Feeds nicht die Volltexte der Beiträge veröffentlichen, sondern lediglich Teaser sowie den Link auf die eigentliche Webseite. Um diesem Problem flexibel zu begegnen, ist zusätzlich ein gestuftes Erhebungsverfahren implementiert, das ich nachfolgend kurz beschreibe.

Synthese aus Feed-Parsing und Crawling
Nachdem die Auswahl der Medienangebote erfolgt ist, werden die Regeln des Erhebungsverfahrens in NewsClassifier mit wenigen Entscheidungen festgelegt, die in Abbildung A.4 dargestellt sind. Als erstes müssen eine oder mehrere URLs für die zu erhebenden Feeds angegeben werden. Anschließend gilt es zu bestimmen, ob die Dokumente aus dieser Quelle nur einmalig heruntergeladen oder kontinuierlich überwacht werden sollen. Ist letzteres der Fall, wird der Import-Prozess in regelmäßigen Abständen, z.B. alle 8 Stunden, wiederholt, wobei bereits importierte Dokumente nicht neu geladen werden.

Anschließend muss der Software mitgeteilt werden, ob bei der Quelle in den Feeds die Volltexte der Beiträge enthalten sind und nur noch gespeichert werden müssen. Dies ist vor allem bei Inhaltsanalysen von Weblogs, Twitter-Meldungen, Foren-Beiträgen oder ausgewählten Medienangeboten sinnvoll, die diese Volltexte im Feed publizieren, etwa

der britischen Tageszeitung GUARDIAN. Sind die Volltexte dagegen nicht im Feed enthalten, wird der Feed-Inhalt als Teaser gespeichert, während gleichzeitig per Crawling-Befehl der vollständige HTML-Inhalt der im Feed verlinkten Artikelseite heruntergeladen und als Hauptinhalt importiert wird.[4] Dieser Schritt ist zur Zeit bei fast allen deutschen Nachrichtenangeboten notwendig.

Eine weitere Verfeinerung dieser Erhebungsmethode liegt in der Verwendung der Druckfassungen der Online-Artikel. Diese werden von den meisten Content-Management-Systemen automatisch erstellt und haben gegenüber den normalen HTML-Seiten der Webangebote einige Vorteile: Sie enthalten deutlich weniger störende Elemente wie Navigation, Werbung und Linklisten. Zudem werden lange Beiträge auf einer Seite dargestellt, während sie auf den normalen Webseiten meist paginiert, d.h. in mehrere durchzuklickende Segmente aufgeteilt werden. Da die URLs dieser Printfassungen in den meisten Fällen leicht aus den Original-URLs ableitbar sind, ist pro Quelle nur eine Ersetzungsregel in Form eines Regulären Ausdrucks nötig, um anschließend das HTML der Printfassung herunterzuladen und zu archivieren. Wurden die Dokumente nicht direkt aus dem Feed extrahiert, sondern das HMTL der Seite heruntergeladen, steht als nächstes die automatische Bereinigung des gespeicherten HTML-Codes an, um daraus verwertbare Textdateien zu gewinnen. Hierfür wird das in Abschnitt 4.2.1 vorgestellte BTE-Toolkit verwendet, das bei den getesteten Nachrichtenangeboten zuverlässig arbeitete. Am Ende dieses Prozesses steht dann ein bereinigtes und korrekt gespeichertes Textdokumente, unabhängig davon, ob nun RSS-Feeds direkt verwendet oder die Webseiten der Artikel heruntergeladen wurden.

Um zu prüfen, welche Unterschiede sich bei der Datenerhebung im Vergleich zu manuellen Online-Inhaltsanalysen von Nachrichten-Sites ergeben, sind in Tabelle A.1 die Stichprobendaten von Quandt (2008b) und Rüdiger & Welker (2010), die von der Startseite ausgehen, und die Daten meiner eigenen Erhebung, die auf RSS-Feeds basiert, dargestellt (vgl. ausführlich in Abschnitt 6.1). Grundlage dabei sind tagesaktuelle

4 Da es in dieser Arbeit nur um die Analyse von Textdokumenten geht, werden zu diesem Zeitpunkt weder Bilder noch audiovisuelle Inhalte archiviert.

Tabelle A.1: Erhebung von Online-Nachrichten durch Crawling und Feeds

	Artikel pro Tag		Mittlere Wörterzahl	
Medienangebot	RSS-Feed	Startseite*	RSS-Feed	Startseite*
FR Online	36	60	520	k.A.
Tagesspiegel	103	104	396	k.A.
SZ Online	29	18	569	569
FAZ Online	46	21	743	609
Spiegel	72	14	612	611

* Quellen: Quandt (2008b, 138), Rüdiger & Welker (2010, 458)

Erhebungen der Online-Auftritte von FAZ, SZ, FR, Tagesspiegel und Spiegel Online.

Die Anzahl der aus den RSS-Feeds extrahierten Artikel liegt zwischen der von Rüdiger & Welker (2010), die jeden auf der Startseite verlinkten Artikel als Untersuchungseinheit gezogen haben, und der von Quandt (2008b), bei dem zusätzlich mindestens ein Teasertext auf der Startseite vorhanden sein musste. Berücksichtigt man auch die ressort-spezifischen Feeds der Angebote, steigt die Zahl erhobener Beiträge nochmals erheblich an, da nicht alle Artikel in den Hauptfeed aufgenommen werden. Ein Indikator für die Validität der Erhebung ist auch die durchschnittliche Textlänge der Beiträge. Hier zeigen sich starke Übereinstimmungen mit den manuell erhobenen Daten von Quandt (2008b), zumal in meiner Stichprobe zwei überlange Ausreißer enthalten sind. Insgesamt lässt sich aus diesem Vergleich der Schluss ziehen, dass mit der hier vorgeschlagenen Erhebungsstrategie vergleichbar valide Stichproben von Online-Nachrichten gezogen werden können. Die Reliabilität der vollautomatischen Erhebung dürfte zudem erheblich höher sein als bei den manuell gestarteten Crawling-Verfahren.

Manuelle Inhaltsanalyse

Im Rahmen einer manuellen Analyse digitaler Inhalte stellt NewsClassifier im einfachsten Fall ein Interface für die Dateneingabe dar, die andernfalls häufig auf Papier oder mit spezialisierter Software wie SPSS Data Entry durchgeführt wird. Durch die Integration des Dokumentmanagements und der Möglichkeit, mehrere Codierer in einem Projekt zu beschäftigen, können darüber hinaus jedoch viele Aufgaben des inhaltsanalytischen Forschungsalltags durch die Software automatisiert werden.

Gestaltung des Codeplans
Um die Erstellung eines Codeplans und dessen Anwendung in NewsClassifier zu erklären, ist ein Blick auf die dahinter liegende Datenstruktur hilfreich. Diese enthält vier Bestandteile: die Variable V, deren k Ausprägungen oder Kategorien, der Codierer C und die eigentlichen Codierung oder Klassifikation Cl. Bei der manuellen Codierung weist ein Codierer einem Dokument eine Kategorie zu, die jeweils zu einer Variable des Codeplans gehört (vgl. Abbildung A.5).

Die Entwicklung eines Codeplans folgt in NewsClassifier weitestgehend den Abläufen, die auch bei der Nutzung von Papier und Stift zu befolgen wären: Zunächst wird eine Variable erstellt, die einen Namen und eine Beschreibung erhält. Diese sind sowohl bei der Codierung als auch im automatisch erstellten Codebuch für die Codierer sichtbar. Anschließend werden die verschiedenen Ausprägungen oder Kategorien der Variable erstellt, wobei stets eine kurze Bezeichnung, eine verbale Erläuterung und ein numerischer Code vergeben werden kann (vgl. Abbildung A.6). Im Screenshot sieht man, dass die Variable zwei Ausprägungen hat und mit 0 und 1 codiert wird. Grundsätzlich können beliebig viele Variablenausprägungen erstellt werden, um nominale oder ordinale Variablen zu erstellen. Für metrische Variablen eignet sich dieses Verfahren hingegen nicht, da hier meist eine offene Eingabe effektiver ist.

Für jede Variable kann zudem festgelegt werden, auf welche Bestandteile des Beitrags (Überschrift, Teaser, Fließtext) sie sich beziehen soll. Bei der Codierung werden entsprechend auch nur diese Textteile angezeigt

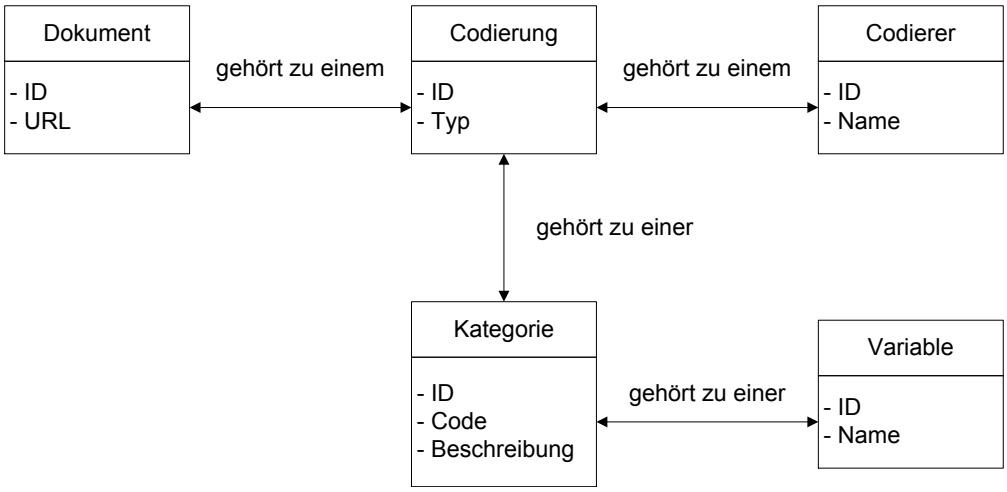

Abbildung A.5: Objektstruktur für die Codierung in NewsClassifier

bzw. die kleinste für die Codierung der Variablen benötigte Schnittmenge. Die Angabe, auf was sich die Codierung beziehen soll, ist später auch für den Einsatz der überwachten Klassifikation wichtig, da auch das Trainings- und Codiermaterial entsprechend aufbereitet wird.

Organisation der Feldarbeit

Da NewsClassifier eine zentralisierte Web-Anwendung ist, ist die Verteilung von Codieraufgaben und -material auf viele Mitarbeiter deutlich einfacher umzusetzen als bei dezentralen Desktop-Programmen wie Excel oder Data Entry. Ein erster Punkt betrifft die Arbeitsteilung unter den Codierern. Rössler (2005) schlägt vor, bei umfangreichen Untersuchungen einzelne Codierteams zu bilden, die sich auf bestimmte Abschnitte des Codebuchs oder Medienangebote konzentrieren. Dies entlastet die Codierer bei der Aufgabe, ein ggf. sehr umfangreiches Kategoriensystem zu memorieren und bietet die Möglichkeit, sich stattdessen auf bestimmte Bereiche, etwa Nachrichtenfaktoren, Akteure oder Themen zu spezialisieren. Da eine kognitive Überlastung zu einer heuristischen statt systematischen Verarbeitung der Inhalte und damit schlechteren Codierungen führt, ist ein solches Vorgehen häufig (vgl. Wirth, 2001). NewsClassifier ermöglicht es, jede Variable spezifisch einem oder mehreren Codierern

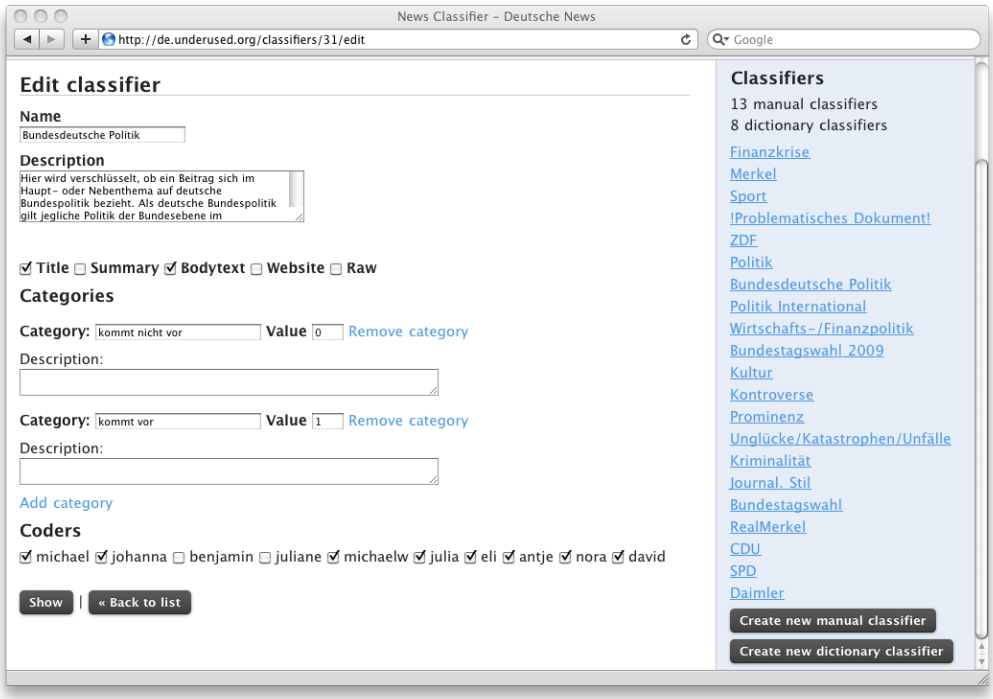

Abbildung A.6: NewsClassifier: Erstellung einer Variablen

zuzuordnen (vgl. Abbildung A.6), und nur diese bekommen das entsprechende Eingabefeld überhaupt angezeigt. Mit einer kategorienbasierten Arbeitsteilung sind jedoch Einschränkungen in der Effizienz verbunden, da mehrere Codierer dasselbe Dokument lesen müssen, was bei umfangreichen Dokumenten wertvolle Zeit bei der Codierung in Anspruch nimmt.

Eine zweite zentrale Aufgabe bei der inhaltsanalytischen Feldarbeit ist die Verteilung des Untersuchungsmaterials auf die Codierer. Während es bei der Analyse von Printmedien oder Fernsehsendungen fast immer sinnvoll ist, das Material als Klumpenstichprobe einzelner Ausgaben oder Sendungen auf die Codierer zu verteilen (Rössler, 2005, 172f.), ist dies bei digitalen Dokumenten nicht notwendig. Eine einfache bzw. geschichtete Zufallsauswahl hat gegenüber fixen Verteilungsplänen einige Vorteile: Erstens kann durch die Randomisierung sichergestellt

werden, dass alle systematischen Verzerrungen, die sich aus bestimmten Codierer-Material-Kombinationen ergeben, vermieden werden. Zweitens ist nur durch eine Zufallsauswahl die Unabhängigkeit der einzelnen Codierungen gewährleistet. Dies ist ein m.E. wichtiger Punkt, der klar der Empfehlung von Rössler (2005, 173) zuwiederläuft, das Material in chronologischer Reihenfolge vorzulegen, da es „zum richtigen Verständnis der Medienberichterstattung hilfreich [sei], früheres Geschehen zu kennen."[5] Dies ist aus zwei Gründen problematisch: Zunächst ist es ein klares Defizit des Codebuchs, wenn Artikel B nur nach Lektüre von Artikel A codierbar ist, da das für die Codierung nötige Kontextwissen aus den Codeerläuterungen und nicht der Erfahrung des Codierers hervorgehen sollte. Außerdem wird die Annahme der Unabhängigkeit der Analyseeinheiten verletzt, die in vielen statistischen Verfahren, etwa der linearen Regression, vorausgesetzt wird. Aus diesem Grund wird in der Standardeinstellung von NewsClassifier jedem Codierer per Zufall ein Dokument aus der Stichprobe vorgelegt, das dieser dann zu codieren hat (vgl. Abbildung A.7).

Reliabilitätsbestimmung

Eng mit der Verteilung des Codiermaterials ist auch die Überprüfung der Reliabilität verbunden, da hierfür Dokumente mehrfach codiert werden müssen. Grundsätzlich sind dabei zwei Strategien möglich: explizite Codiersitzungen zur Überprüfung der Reliabilität oder kontinuierliche Überprüfung während der eigentlichen Feldphase. Vor allem aus forschungspraktischen Überlegungen ist sind separate Reliabilitätstests die Regel (Früh, 2007; Rössler, 2005). Dabei wird vor der eigentlichen Codierung eine Stichprobe der Daten gezogen, die dann von mehreren Personen codiert werden. Anschließend wird die Reliabilität berechnet und erst dann mit der eigentlichen Feldarbeit begonnen oder weitere Schulungen durchgeführt. Dieses Verfahren ist zwar unkompliziert durchzuführen, hat aber den Nachteil, dass die Bedingungen, unter denen der Test abläuft, nicht denen in der Feldphase entsprechen. Unter anderem ist es wahrscheinlich, dass sich die Codierer im Reliabilitätstest eher

5 Bei kontinuierlichen Analysen mit ständig neu erhobenen Beiträge lässt sich natürlich nicht verhindern, dass früher erhobenes Material auch zuerst codiert wird.

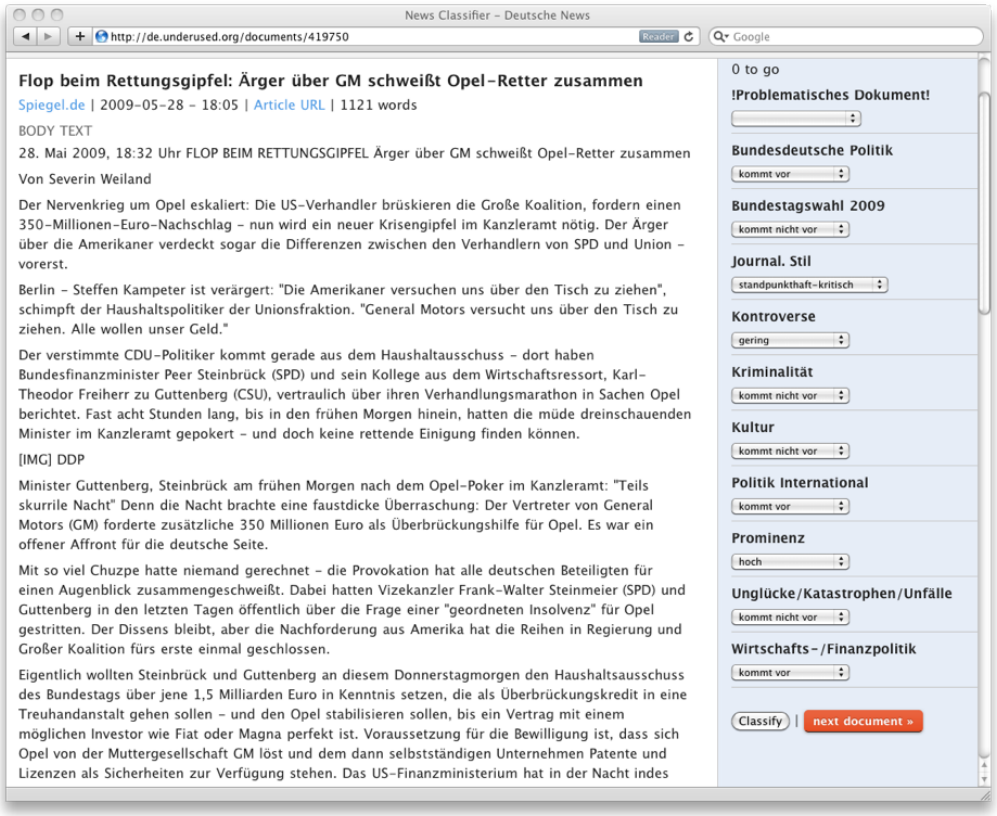

Abbildung A.7: NewsClassifier: Codierung eines Dokuments

anstrengen als bei der Normalcodierung. Die erzielten Werte würden in diesem Fall die Reliabilität der eigentlichen Codierung überschätzen.

Als Ergänzung oder Alternative bietet es sich an, während der Feldarbeit Reliabilitätsdaten zu erheben, um eine kontinuierliche Codierqualität sicherzustellen. Aus diese Weise lassen sich zudem schlechte Codierer gezielt nachschulen oder aus der Untersuchung entfernen. NewsClassifier bietet die Möglichkeit, den Codierern gezielt Dokumente zur Reliabilitätsberechnung vorzulegen, ohne dass diese in ihrem normalen Arbeitsablauf gestört werden. Damit ist nicht nur sichergestellt, dass die für die Reliabilitätsberechnung verwendeten Dokumente repräsentativ aus den Daten gezogen werden, sondern auch dass die Bedingungen, un-

ter denen die Dokumente codiert werden, repräsentativ für die gesamte Feldarbeit sind. Praktisch wird dies umgesetzt, in dem der Projektleiter einen Prozentwert festlegt, der den Anteil mehrfach zu codierender Einheiten bestimmt und daher Werte von 0 bis 100 annehmen kann. Bei einem Wert von 0 wird jedes Dokument genau einem Codierer zugespielt, so dass keine Übereinstimmung gemessen werden kann. Bei einem Coverage-Wert von 100 werden hingegen zuerst alle bereits von anderen codierten Beiträge vorgelegt und erst dann neue Dokumente ausgewählt, wenn keine vorcodierten Artikel mehr vorhanden sind.

In Kombination mit der Möglichkeit, beliebig große Stichproben aus dem Datenmaterial zu ziehen, bietet sich in NewsClassifier folgender Ablauf für die Codierung an:

1. Eine kleine Stichprobe an Dokumenten wird für die Schulung gezogen, nach deren Abschluss entweder die Dokumente oder zumindest deren Codierungen aus dem Projekt entfernt werden.
2. Ist ein separater Pretest angebracht, wird eine Stichprobe an Dokumenten für den Reliabilitätstest gezogen und der Parameter für Mehrfachcodierungen auf 100 Prozent gesetzt. Anschließend werden sämtliche Dokumente von allen Personen codiert. Nach Abschluss des Pretests wird für die Reliabilität der Codierung berechnet.
3. Für die Normalcodierung wird der Coverage-Parameter auf einen plausiblen Wert, z.B. 20 Prozent, gesetzt, so dass jedes fünfte Dokument mehrfach codiert wird und so in die kontinuierliche Reliabilitätsbestimmung eingeht.

Liegt der Coverage-Parameter bei unter 100 Prozent, werden bei mehr als zwei Codierern nicht alle möglichen paarweisen Übereinstimmungen realisiert. Stattdessen ist das Verfahren darauf optimiert, eine möglichst breite Basis für die Reliabilitätsbestimmung zu schaffen, so dass in der Regel eher viele Dokumente von nur zwei Personen codiert werden statt wenige von allen Codierern. Da auch die Zuordnung von Dokumenten zur Mehrfachcodierung zufällig erfolgt, können aber Dokumente manchmal auch drei oder mehr Codierern vorgelegt werden. Statt einfacher paarweiser Codierer-Vergleiche oder den von Kolb (2002) vorgestellten

Verfahren ist für die Auswertung eher der Ansatz von Krippendorff (2004a) geeignet, bei dem alle Mehrfachcodierungen in einer Koinzidenzmatrix zusammengefasst werden, die ggf. fehlende Werte für einzelne Dokument-Codierer-Paare aufweist.

Da die Codierungen in NewsClassifier zentral an einer Stelle gespeichert werden, ist es zu jedem Zeitpunkt der Untersuchung möglich, einen Reliabilitätstest mit den dann vorhandenen Mehrfachcodierungen durchzuführen. Die Berechnung der Koeffizienten nach Holsti und Krippendorff erledigt dabei die Software. Zusätzlich können für jede Variable auch die Koinzidenzmatrizen für eigene Berechnungen exportiert werden.

Alle in diesem Abschnitt vorgestellten Maßnahmen dienen dem Zweck, sowohl Forschungsleiter als auch Codierer von allen handwerklichen Aufgaben zu entlasten, so dass Ressourcen für die eigentliche Codierung der Inhalte frei werden. Die Automatisierung von Stichprobenziehung, Materialverteilung und Reliabilitätsbestimmung ist jedoch nicht nur aus forschungsökonomischen Erwägungen sinnvoll, sondern dient vor allem der Transparenz und Replikationsfähigkeit der Analyse. Damit geht die Annahme einher, dass qualitätssichernde Maßnahmen eher durchgeführt und deren Ergebnisse berichtet werden, wenn diese praktisch ohne Zusatzaufwand in den Forschungsprozess zu integrieren sind.

Verknüpfung manueller und automatischer Codierung

Training, Test und kontinuierliche Verbesserung des Klassifikators

Neben der Verknüpfung von Datenerhebung, -bereinigung und manueller Codierung liegt die zentrale Aufgabe der Software in der Integration von überwachten Textklassifikationsverfahren. Die Überlegungen zu dieser Verknüpfung manueller und automatischer Analyse basieren auf folgenden Grundannahmen:

1. Für die meisten inhaltsanalytischen Forschungsprojekte stellt die manuelle Codierung den Normalfall dar, so dass diese in NewsClassifier möglichst reibungslos durchgeführt werden sollte.

2. Für die Integration der in Abschnitt 3.4.1 vorgestellten Klassifikationsverfahren aus dem Bereich des maschinellen Lernens sind lediglich zwei Ressourcen nötig: eine Klassifikationssoftware und möglichst viele korrekt codierte Texte.
3. Da jede manuelle Codierung ohnehin die benötigten Trainingsdaten liefert, ist es ohne weiteres möglich, mit diesen einen überwachten Klassifikator im Hintergrund zu trainieren. Hierfür sind keine manuellen Eingriffe nötig, so dass der Trainingsprozess vollständig automatisch im Hintergrund ablaufen kann.
4. Zu jedem Zeitpunkt in der Feldphase lässt sich in NewsClassifier sowohl die Intercoder-Reliabilität – sofern mehrere Codierer vorhanden sind – als auch die bisher erreichte Klassifikationsgüte bestimmen. Für letztere werden, wie in Abschnitt 4.4 dargestellt, die manuell zugewiesenen Codierungen als Vergleichsmaßstab benötigt.
5. Fällt der Qualitätstest für die automatische Klassifikation zufriedenstellend aus, kann man ggf. die Variable aus der manuellen Codierung entfernen, d.h. keinem Codierer mehr zuweisen, und die noch ausstehende Dokumente automatisch codieren lassen.

Konkret wurden diese Anforderungen folgendermaßen umgesetzt: Wird eine neue Variable zum Codeplan eines Projektes hinzugefügt oder eine bestehende modifiziert, bereitet die Software im Hintergrund einen neuen Klassifikator vor, der jeweils die Ausprägungen der Variable als Klassen enthält.[6] Da die meisten Klassifikationsalgorithmen nur für nominale Variablen entworfen werden, können Variablen mit ordinalen Kategorien nur multinomial umgesetzt werden. Dies hat zur Folge, dass die Ordinalität der Ausprägungen bei Training und Test der automatischen Klassifikation verloren geht.

Wird im Laufe der Feldphase ein bestimmtes Dokument einer Kategorie zugeordnet, führt NewsClassifier im Hintergrund den betreffenden Beitrag dem Klassifikator unverzüglich als Trainingsdokument zu. Bei

6 Für die Evaluation habe ich bislang eine einzige Klassifikationssoftware (OSBF-Lua, Assis 2006) eingebunden, die Verwendung anderer Lösungen ist jedoch nur mit geringem Mehraufwand verbunden.

einem Codeplan mit 20 Variablen laufen nach der Codierung eines Dokuments folglich bis zu 20 Trainingsschritte ab. In der aktuellen Version der Software geht unkonditional jede Codierung in das Training des Klassifikators ein. Angesichts aktueller Forschungsarbeiten zur statistischen Modellierung von Codierungen (Carpenter, 2008; Dekel & Shamir, 2009; Raykar et al., 2009; Whitehill et al., 2009) sind jedoch auch alternative Optionen denkbar, etwa dass nur Dokumente verwendet werden, deren Codierung wahrscheinlich korrekt ist. Die Verwendung aller Codierungen für die überwachte Klassifikation hat auch zur Folge, dass sich widersprechende Mehrfachcodierungen alle gleichwertig in die Klassifikation eingehen. Da es sich bei NewsClassifier um einen inkrementellen Trainingsprozess handelt, haben zeitlich spätere Codierungen ein höheres Gewicht als frühere. Da dies nicht in allen Situationen sinnvoll ist, etwa wenn bei Nichtübereinstimmung bestimmte Codierer ihre Entscheidungen nachträglich korrigieren müssen, lässt sich der Klassifikator auch zurücksetzen und mit den vorhandenen Codierungen in zufälliger Reihenfolge neu trainieren. Dies entspricht einem Batch-Learning (vgl. Abschnitt 4.3).

Für jede Variable des Codeplans lässt sich entweder auf Anfrage oder in festgelegten Abständen sowohl die Intercoder- als auch die Coder-Klassifikator-Reliabilität berechnen. Dies geschieht ebenfalls automatisch, so dass in der Übersicht der Variablen (vgl. Abbildung A.8) schnell deutlich sind, wo es ggf. Defizite in der manuellen und/oder automatischen Codierung gibt. Als Standard werden hierzu wiederum die Koeffizienten von Holsti und Krippendorff sowie die Fallzahl für den Test angegeben. Detaillierte Analysen lassen sich ebenfalls auf Basis der Konfusions- bzw. Koinzidenzmatrizen durchführen, die in der Einzelansicht jeder Variable zum Export verfügbar sind.

Die Diagnose der Reliabilität führt unweigerlich zur Frage: Was tun, wenn die Qualität der Codierung unzureichend ist? Bei der manuellen Analyse liegt die Lösung meist in der Nachschulung oder notfalls im Ausschluss der nicht zufriedenstellenden Codierer, wobei dann Nachcodierungen notwendig sind (Früh, 2007, 199). Für die überwachte Klassifikation sind prinzipiell drei unterschiedliche Ursachen für eine unbefriedigende Klassifikationsqualität denkbar: (a) die Trainingsdaten sind

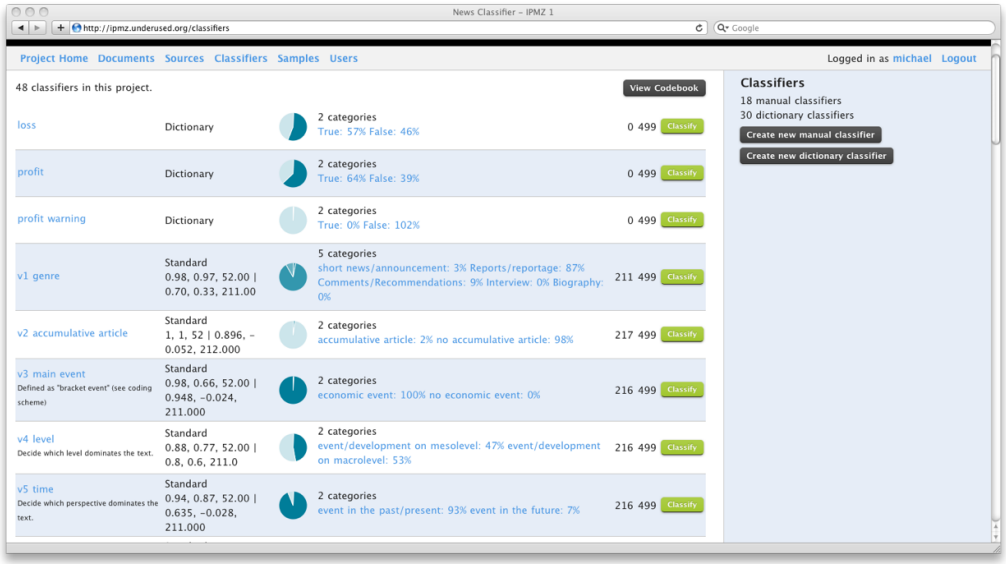

Abbildung A.8: NewsClassifier: Übersicht über die Variablen des Codebuchs

nicht reliabel und valide, (b) es sind zu wenige Trainingsdokumente vorhanden, (c) die betreffende Variable eignet sich grundsätzlich nicht für die Automatisierung (vgl. Tabelle A.2). Zumindest die erstgenannten zwei Problemfälle lassen sich ohne Umstände diagnostizieren und ggf. durch zusätzliche Arbeit lösen. Bleibt die Klassifikation trotz qualitativ und quantitativ ausreichendem Trainingsmaterial unbefriedigend, z.B. weil die Variable auf pragmatischer Ebene ansetzt oder viel Kontextwissen erfordert, ist eine manuelle Codierung aller Dokumente die einzige Lösung.

Die beiden erstgenannte Problemstellungen treten deshalb auf, weil für eine zuverlässige Klassifikation auch zuverlässige Trainingsdaten für alle Kategorien vorliegen müssen. In beiden Fällen kann die Klassifikationsgüte durch zusätzliche manuelle Codierungen erhöht werden. Die Frage ist dabei: Sollen eher neue Dokumente codiert werden, um den Umfang des Trainingsmaterials zu erhöhen, oder Mehrfachcodierungen des bestehenden Materials durchgeführt werden, um zu gewährleisten, dass die Trainingsdaten weitestgehend fehlerfrei sind? Die empirische

Tabelle A.2: Ursachen niedriger Klassifikationsgüte

Problem	Indikator	Lösung
unzuverlässige Trainingsdaten	niedrige Intercoder-Reliabilität	Nachschulung, Mehrfachcodierung
zu wenige Trainingsdaten	schiefe Verteilung der Kategorien	gezielte Codierung neuer Dokumente
fehlende Automatisierbarkeit	niedrige Klassifikationsqualität	Re-Operationalisierung, manuelle Codierung

Ergebnislage zu diesem Thema ist nicht eindeutig. So können Sheng et al. (2008) zeigen, dass in bestimmten Situationen beide Strategien die Klassifikationsqualität verbessern können. Generell scheint es bei hoher Intercoder-Reliabilität eher sinnvoll, neue Dokumente codieren zu lassen, während sich bei schlechterer Codiererleistung eher eine Mehrfachcodierung lohnt, da die Codierungen bei Mehrheitsentscheidungen deutlich zuverlässiger sind als diejenigen der einzelnen Codierer (Snow et al., 2008). Diese Einschätzungen sind jedoch nur als Entscheidungsheuristiken zu verstehen, ggf. ist eine Einzelfalldiagnose unumgänglich.

Sollen weitere manuelle Codierungen vorgenommen werden, unterstützt die Software den Forscher bei der Auswahl geeigneter Dokumente. Hierfür sind verschiedene Selektionsverfahren implementiert, die nicht nur bei der Optimierung der überwachten Klassifikation, z.B. durch aktives Lernen, sondern auch für rein manuelle Inhaltsanalysen von Nutzen sein können. Diese Funktionen erläutere ich im nächsten Abschnitt, um nochmals die Möglichkeiten der Integration automatischer und manueller Verfahren herauszustellen.

Nutzen des Klassifikators für die manuelle Codierung

Durch die integrierte Anwendung manueller und automatischer Verfahren ergeben sich zahlreiche Möglichkeiten, die Qualität der Analyse vor und während der Feldarbeit zu überwachen und zu verbessern. Dies lässt sich vor allem durch die gezielte Auswahl von Stimulusmaterial

erreichen, das (a) nicht-übereinstimmend mehrfachcodiert wurde, (b) bei dem automatische und manuelle Codierung voneinander abweichen oder (c) dessen Codierung im Vorfeld den stärksten Erkenntniszuwachs verspricht. Diese drei Selektionsstrategien für Dokumente können sowohl die Reliabilität der Analyse erhöhen und auch Möglichkeiten zu konzeptionellen und operationalen Verbesserungen in der Kategorienbildung aufzeigen.

Das Auffinden von uneinheitlich codierten Dokumenten ist seit langem ein einfaches und effektives Hilfsmittel der Qualitätssicherung bei Inhaltsanalysen. Während der Codiererschulung und des Pretests kann der Computer helfen, potentiell schwierige Dokumente zu erkennen, an deren Beispiel sich Unklarheiten bezüglich des Codeplans ausmachen und korrigieren lassen. Folgt man der Empfehlung von Früh (2007), einen Anteil des Materials auch während der Feldphase mehrfach codieren zu lassen, kann nicht nur der Messfehler besser eingeschätzt werden, sondern widersprüchliche Codierungen auch durch den Forschungsleiter aufgelöst werden. Dies erhöht gleichzeitig die Qualität des Trainingsmaterials für den Klassifikator. Da die Codierungen in NewsClassifier zentral gespeichert werden, können uneinheitlich codierte Dokumente noch während einer Schulungssitzung oder der Normalcodierung identifiziert werden.

Der eigentliche Nutzen eines überwachten Klassifikationsalgorithmus für die manuelle Codierung besteht aber in der Möglichkeit, diesen als automatischen Tester einzusetzen, der problematische Codierungen identifizieren hilft. Die betroffenen Dokumente können anschließend von Forschungsleiter bewertet und ggf. mit einer Mastercodierung versehen werden. Das Vorgehen bei dieser automatischen Fehlersuche ist dabei denkbar einfach: Alle codierten Dokumente werden dem bereits trainierten Klassifikator vorgelegt und diejenigen, bei denen sich automatische und manuelle Codierung unterscheiden, werden im Anschluss nochmals den Codierern vorgelegt. Im Prinzip wird also der Klassifikationsalgorithmus als zusätzlicher Codierer verwendet.

Das Framework unterscheidet bei der Nutzung der automatischen Klassifikation zwei Arten von problematischen Dokumenten. Im ersten Fall werden, Brodley & Friedl (1999) folgend, bei der Evaluation durch

10-Fold-Cross-Validierung mit einem Trainings-Set aus 90 Prozent aller Codierungen die Klassen der übrigen 10 Prozent abgefragt. Die Dokumente aus dem Test-Set, bei denen die Klassifikation nicht der manuellen Codierung entspricht, werden anschließend als problematisch gespeichert. Ist die Variable nicht zuverlässig automatisch codierbar, werden sehr viele Dokumente als problematisch markiert sein. Eine strengere Selektion ergibt sich aus der zweiten Identifikationsstrategie: Hier wird der Klassifikator mit allen manuell codierten Dokumenten in zufälliger Reihenfolge trainiert und anschließend anhand derselben Dokumente getestet. Wird bei einem probabilistischen Klassifikator ein Dokument trotz Vorgabe der korrekten Kategorie später falsch klassifiziert, kann dies als Indikator gelten, dass diese manuelle Codierung implizit allen anderen widerspricht, aus denen das statistische Modell abgeleitet wurde. Dieser Fall ist recht selten, aber äußerst folgenreich, da durch dieses potentiell widersprüchliche Trainingsmaterial die Entwicklung eines guten Klassifikationsmodells behindert wird. Es lohnt sich daher, diese besonders problematischen Codierungen einzeln zu prüfen und ggf. zu korrigieren. Auch wenn man gar kein substanzielles Interesse an einer automatischen Klassifikation hat, kann auf diese Weise die Software zur Verbesserung der manuellen Inhaltsanalyse beitragen.

Gerade bei der Verfeinerung des Codebuchs kann auch eine dritte Selektionsstrategie hilfreich sein. Diese zielt im Gegensatz zu den bereits vorgestellten auf die Auswahl von bislang noch uncodierten Dokumenten. Wie im Abschnitt 4.3 erläutert, kann ein bereits trainierter Klassifikator dafür eingesetzt werden, neues Trainingsmaterial zu wählen. Dazu werden alle neuen Dokumente dem Klassifikator vorgelegt, der dann nicht nur die wahrscheinlichste Kategorie bestimmt, sondern auch die Wahrscheinlichkeit der gewählten Klassenzugehörigkeit quantifizieren kann. Die Dokumente, bei denen der Klassifikator unsicher ist, sind zumeist auch schwieriger manuell zu codieren. Die manuelle Sichtung und Codierung dieser Dokumente trägt entsprechend nicht nur zum Trainingserfolg des Klassifikators bei, sondern hilft ebenfalls bei der Verfeinerung und Konkretisierung der Codieranweisungen. Insgesamt kann der Einsatz überwachter Klassifikation daher auch bei ausschließlich manuellen Inhaltsanalysen hilfreich sein.

Verbesserung des Forschungsprozesses durch NewsClassifier?

Worin liegen nun die Möglichkeiten der Automatisierung für die Forschungspraxis? Im Bereich der Datenerhebung profitiert man vor allem von der Standardisierung der Datenformate, die jeder automatischen Weiterverarbeitung zu Grunde liegt. Da viele Akteure im Internet, etwa Produzenten, Mediennutzer, Anbieter von Software oder Marktforschungsunternehmen, ein Interesse daran haben, dass digitale Inhalte möglichst reibungslos verbreitet und verarbeitet werden können, kann man sich beim Import auf wenige Formate wie HTML, RSS oder JSON konzentrieren und damit einen großen Teil aller verfügbaren Inhalte analysierbar machen. Ein direkter Vorteil der Automatisierung ist bei der kontinuierlichen Analyse von Online-Inhalten zu verzeichnen: Mit minimalem Aufwand können durch die Nutzung von RSS-Feeds sowohl klassische Medieninhalte als auch nutzergenerierte Mitteilungen wie Twitter- oder Facebook-Meldungen kontinuierlich überwacht werden. Bis hin zur eigentlichen Codierung lässt sich dies ohne manuelle Eingriffe realisieren.

Die automatische Durchführung von Stichprobenziehung, Materialverteilung, Reliabilitätstests und anderen Aufgaben ermöglicht es, sowohl für die eigene Projektabwicklung als auch in der Dokumentation ohne Zusatzaufwand weitgehende Überprüfbarkeit zu gewährleisten. Dass diese Praxis bei Inhaltsanalysen sehr wichtig ist, gleichzeitig aber viel zu selten geschieht, demonstriert Rössler (2005) eindrucksvoll: Nachdem in seinem Lehrbuch die Qualitätskriterien der Inhaltsanalyse ausführlich dargestellt wurden, muss bei der Vorstellung der meisten Beispielstudien festgestellt werden, dass keinerlei Informationen zur Qualitätssicherung gegeben wurden. Bei einer softwaregestützten Inhaltsanalyse fallen diese Daten ohnehin an, so dass sie nur noch dokumentiert und interpretiert werden müssen. Die automatische Durchführung von Reliabilitätstests während der Normalcodierung sollte daher ein Anreiz für mehr Qualitätskontrolle und bessere Dokumentation sein.

Durch die Einbindung der überwachten Klassifikation bei der Selektion von Codiermaterial kann die Codebuchentwicklung auf eine breitere

empirische Basis gestellt werden. Als Ergänzung zur klassischen Codierersitzung ist die aktive Selektion von problematischen Dokumenten zu verstehen: Da das Klassifikatortraining bei jeder manuellen Analyse im Hintergrund läuft, ist zu jedem Zeitpunkt sowohl ein Klassifikationstest als auch, darauf aufbauend, eine gezielte Auswahl von schwierigen Dokumenten möglich, an denen der Algorithmus scheitert. Diese Informationen helfen bei der Verfeinerung der Codieranweisungen ebenso wie bei der Identifikation von menschlichen Codierfehlern.

Letztlich dienen gerade die im Hintergrund vollautomatisch ablaufenden Reliabilitäts- und Klassifikationstests dazu, sowohl bei der Codebuchentwicklung und dem Pretest als auch während der Feldarbeit schnell auf Fehler und Probleme reagieren zu können. Eine umfangreiche automatische Qualitätskontrolle ermöglicht und ermutigt zu einer breitangelegten Operationalisierungsstrategie, bei der verschiedene Messinstrumente, Codierer und Klassifikationsverfahren auf ein relevantes Konstrukt angesetzt werden.

Wie jedes Forschungsinstrument eignet sich auch das hier vorgestellte Framework nicht uneingeschränkt für alle Forschungsfragen und Untersuchungsgegenstände. Die wichtigsten Einschränkungen sein daher abschließend an dieser Stelle aufgezählt:

Beschränkung auf digitale Inhalte Dokumente, die nicht online bzw. nicht einmal in digitaler Form vorliegen, lassen sich mit dem Instrument nicht analysieren. Für einige Forschungsgegenstände, etwa ältere Druckerzeugnisse, Ton- und Videoaufzeichnungen, wird sich der Aufwand der Digitalisierung und Archivierung vermutlich nicht lohnen.

Optimierung auf Text-Material Obwohl sich grundsätzlich auch Youtube-Videos oder Audio-Dateien analysieren lassen, sind doch viele Bestandteile des Frameworks für die Verarbeitung von Textdaten optimiert. Dies trifft vollständig auf die automatische Codierung und Klassifikation zu, ebenso auf das Preprocessing und Teile des Datenmanagements.

Analyseeinheit = Codiereinheit In der aktuellen Entwicklungsphase eignet sich die Software nur für Analysen, bei denen die Ana-

lyseeinheit der Codiereinheit entspricht. Die Vergabe von Codes erfolgt immer bezogen auf ein Dokument, egal ob Email, Website, Forenbeitrag oder Twitter-Meldung. Eine Zuweisung von Codes zu einzelnen Textelementen, etwa Aussagen oder Sätzen, ist bislang nicht möglich.

Während einige Einschränkungen grundsätzlicher Natur sind, können andere durch eine Weiterentwicklung des Frameworks behoben werden. In allen Fällen, in denen sich die Software sinnvoll einsetzen lässt, scheint sie jedoch nach den bisherigen Erfahrungen dazu beizutragen, Inhaltsanalysen effektiver und effizienter zu machen. Insbesondere skaliert NewsClassifier mit den Anforderungen: Sowohl kleine Studien im Rahmen studentischer Projekte als auch umfangreiche kontinuierliche Analysen lassen sich mit fast gleichbleibendem Aufwand umsetzen. Zudem wird durch die Automatisierung der Feldarbeit und der Reliabilitätsprüfung stets ein Mindestmaß an Qualitätssicherung gewährleistet. Dies führt in der Regel zu einer zuverlässigeren und valideren Messung.

B Anhang

B.1 Ergebnistabellen

Abbildung B.1: Intercoder-Reliabilität der manuellen Codierung, standardisierte Koinzidenzmatrizen, $n = 373$

	0	1
0	63	5
1	5	27

(a) Politik

	1	2
1	75	5
2	5	15

(b) Bundespolitik

	0	1
0	79	4
1	4	14

(c) Politik International

	0	1
0	81	3
1	3	12

(d) Wirtschaftspolitik

	0	1
0	95	2
1	2	2

(e) Bundestagswahl

	0	1
0	83	0
1	0	17

(f) Sport

	0	1	2
0	89	0	2
1	0	3	0
2	2	0	3

(g) Kultur

	0	1
0	90	2
1	2	5

(h) Unglücke/ Unfälle

	0	1
0	82	4
1	4	10

(i) Kriminalität

	0	1	2
0	49	10	2
1	10	19	3
2	2	3	1

(j) Kontroverse

	0	1	2
0	43	6	2
1	6	9	7
2	2	7	19

(k) Prominenz

	1	2	3	4	5	6
1	37	2	5	3	5	1
2	2	2	1	1	1	0
3	5	1	6	1	0	1
4	3	1	1	4	2	0
5	5	1	0	2	2	0
6	1	0	1	0	0	2

(l) Journ. Stil

Abbildung B.2: Standardisierte Konfusionsmatrizen zwischen Codierern (Zeilen) und Klassifikator (Spalten), $n = 933$

	0	1	
0	64	8	72
1	6	22	28
	70	30	100

(a) Politik

	0	1	
1	74	7	81
2	7	12	19
	81	19	100

(b) Bundespolitik

	0	1	
0	77	3	81
1	8	11	19
	85	15	100

(c) Politik International

	0	1	
0	79	6	85
1	5	10	15
	84	16	100

(d) Wirtschaftpolitik

	0	1	
0	97	1	98
1	2	1	2
	99	1	100

(e) Bundestagswahl

	0	1	
0	84	1	85
1	3	12	15
	87	13	100

(f) Sport

	0	1	2	
0	91	0	0	91
1	3	0	0	3
2	5	1	0	6
	99	1	0	100

(g) Kultur

	0	1	
0	93	0	93
1	6	1	7
	99	1	100

(h) Unglücke/Unfälle

	0	1	
0	82	3	84
1	11	5	16
	92	8	100

(i) Kriminalität

	0	1	2	
0	46	13	0	59
1	17	15	0	32
2	3	5	1	9
	66	33	1	100

(j) Kontroverse

	0	1	2	
0	39	4	7	50
1	11	6	7	24
2	7	3	16	26
	57	13	30	100

(k) Prominenz

	1	2	3	4	5	6	
1	29	1	4	16	2	0	53
2	1	3	3	1	0	0	8
3	2	2	5	5	1	0	14
4	3	0	2	7	1	0	12
5	1	0	1	5	1	0	9
6	0	1	1	1	0	1	5
	36	7	17	34	4	2	100

(l) Journ. Stil

Tabelle B.1: Anteilswerte der Kategorien nach Quelle

	BILD	WEST	FAZ	FR	FOC	HEU	SZ	SP	TAG	TSP	WELT	ZEIT	Ges.
n	153	29	65	68	98	12	35	141	56	190	31	55	933
Poltik allg.	7	31	23	26	29	17	29	38	52	28	52	31	30
Bundespolitik	7	38	17	24	21	58	17	21	20	19	26	24	24
Int. Politik	5	14	22	22	14	33	17	28	45	16	32	20	22
Wirtschaftspolitik	5	17	18	18	16	50	17	19	14	13	13	22	19
Bundestagswahl	2	7	0	3	6	0	9	2	2	2	3	0	3
Sport	27	3	20	10	6	8	11	12	4	22	6	5	11
Kultur,	7	14	9	18	6	8	3	8	0	12	3	11	8
davon													
Hochkultur	2	0	6	6	0	0	3	3	0	5	3	7	3
Popkultur	5	14	3	12	6	8	0	5	0	7	0	4	5
Unglücke/Unfälle	8	10	5	12	8	0	6	4	9	7	3	7	7
Kriminalität	8	17	11	24	15	8	11	19	23	15	26	16	16
Kontroverse,	27	34	35	43	37	50	43	48	39	43	61	44	42
davon													
gering	23	34	26	35	27	50	40	35	29	35	39	42	35
heftig	5	0	9	7	10	0	3	13	11	8	23	2	7
Prominenz,	56	38	52	47	53	42	63	49	52	48	61	40	50
davon													
gering	36	14	29	31	23	17	23	20	23	24	23	16	23
hoch	20	24	23	16	30	25	40	29	29	25	39	24	27

B.2 Codebuch der Evaluationsstudie

Stichprobe

Erhebungsgrundlage sind alle Beiträge, die in den Haupt-RSS-Feeds der folgenden Nachrichtenangebote verlinkt wurden: spiegel.de, focus.de, zeit.de, bild.de, welt.de, faz.net, sueddeutsche.de, fr-online.de, derwesten.de, tagesspiegel.de, tagesschau.de, heute.de.

Zeitlicher Rahmen
Ausgewählt wurden per einfacher Zufallsauswahl 1000 Beiträge, die zwischen dem 1.6.2008 und dem 31.5.2009 erschienen sind.

Aufgriffskriterium
Codiert werden alle vorliegenden Beiträge, die mehr als drei Sätzen Fließtext enthalten und nicht aus ausschließlich audiovisuellen Inhalten bestehen. Codiereinheit ist der ganze Beitrag inklusive Überschrift.

Formale Variablen

Folgende Variablen werden automatisch bei der Erhebung codiert:

1. Identifikationsnummer
2. URL des Artikels
3. Zeitpunkt der Veröffentlichung, entweder im RSS-Feed angegeben oder auf den Zeitpunkt der Speicherung gesetzt
4. Umfang des bereinigten Fließtexts in Wörtern und Sätzen

Inhaltliche Variablen

Bundesdeutsche Politik
Hier wird verschlüsselt, ob ein Beitrag sich auf deutsche Bundespolitik bezieht. Als deutsche Bundespolitik gilt jegliche Politik der Bundesebene im Zusammenhang mit der Regelung von (sozialen, wirtschaftlichen, politischen usw.) Verhältnissen innerhalb der Bundesrepublik (Innenpolitik, z.B. innere Sicherheit, Gesundheitspolitik, Arbeitsmarktpolitik) oder im Zusammenhang mit Beziehungen der Bundesrepublik zu einem oder

mehreren anderen Staaten oder internationalen Organisationen (Außenpolitik, z.B. Besuche deutscher bundespolitischer Politiker im Ausland, Gipfelkonferenzen).

0 kommt nicht vor
1 kommt vor

Quelle: GÖFAK Medienforschung (2010)

Politik international
Nationale Politik anderer Länder (außer Außenpolitik, die sich auf Deutschland bezieht), die Beziehungen zwischen anderen Ländern ohne Bezug zu Deutschland sowie die Politik internationaler Organisationen ohne Bezug zu Deutschland, wobei der Bezug zu Deutschland auch über das Thematisierung deutscher bundespolitischer Akteure hergestellt werden kann.

0 kommt nicht vor
1 kommt vor

Quelle: GÖFAK Medienforschung (2010)

Wirtschaftspolitik/Finanzpolitik
Ereignisse oder Maßnahmen mit dominanter Beteiligung staatlicher Institutionen, Organisationen oder politischer Funktionsträger bzw. auf diese gerichtete Aktivitäten, die der Festlegung des Wirtschaftsablaufes, der Ordnung der Wirtschaftsstruktur und der Gestaltung der wirtschaftlichen Rahmenbedingungen dienen (z.B. Kartellgesetze, Konjunkturpolitik, Arbeitsmarktpolitik, Außenhandelspolitik, Subventionen, Strukturpolitik, Verbraucherpolitik) *oder* die den Bereich der Geld- und Vermögensverwaltung des Staates (wie z.B. Steuern, Finanzierung öffentlicher Vorhaben, Projekte, Institutionen) und Ausgleichsmaßnahmen und wirtschaftsstabilisierende Eingriffe in den Staatshaushalt betreffen.

0 kommt nicht vor
1 kommt vor

Quelle: Bruns & Marcinkowski (1997)

Bundestagswahl 2009
Bezugnahme auf die Bundestagswahl 2009, z.B. Kandidaten, Wahlprogramm, TV-Duell, Umfragen, Wahlkampf, etc.

0 kommt nicht vor
1 kommt vor

Quelle: GÖFAK Medienforschung (2010)

Sport
Ereignisse und Maßnahmen, die sich in einem nichtpolitischen Kontext den Bereich des Leistungs- und Breitensports betreffen, z.B. Sportveranstaltungen, nationale und internationale Wettkämpfe. Doping wird nur dann als Sport codiert, wenn tatsächlich ein Bezug zu einem Sportereignis vorhanden ist. Nicht als Sport codiert werden Meldungen, die in denen Sportler lediglich als Prominente, Unfallopfer, Kriminelle etc. vorkommen.

0 kommt nicht vor
1 kommt vor

Quelle: Bruns & Marcinkowski (1997), eigene Ergänzungen

Kultur
Ereignisse und Maßnahmen, die kulturelle Einrichtungen (Theater, Oper, Schauspiel, Festivals, Musik etc.) in einem nichtpolitischen Sinn betreffen (Theateraufführungen, Musikveranstaltungen, Kunstausstellungen etc.).

0 kommt nicht vor
1 Hochkultur kommt vor (Theater, Ballett, Oper, Kunstausstellungen, Museen, Klassische Musik, Jazz, Literatur)
2 Populärkultur kommt vor (Pop- und Rockmusik, Film, Musical, Comedy, Comics, Zirkus)

Quelle: Bruns & Marcinkowski (1997), eigene Ergänzungen

Unglücke/Katastophen/Unfälle
Ereignisse und Maßnahmen, die sich in einem nichtpolitischen Kontext einzelne Unglücksfälle und größere Unglücksvorkommnisse betreffen (Erdbeben, Autounfälle, Explosionen).

0 kommt nicht vor
1 kommt vor

Quelle: Bruns & Marcinkowski (1997)

Kriminalität
Ereignisse und Maßnahmen, die sich in einem nichtpolitischen Kontext ein Verbrechen betreffen (Mord, Raub, Überfall, Steuerhinterziehung, Geiselnahmen, (pol.) Korruption); in Afghanistan ist Konversion zum Christentum kriminell; Krieg ist nicht kriminell, aber Kriegsverbrechen! Menschenrechtsverletzungen sind nicht automatisch Kriminalität.

0 kommt nicht vor
1 kommt vor

Quelle: Bruns & Marcinkowski (1997); Fretwurst (2008)

Kontroverse
Unter Kontroverse wird die erkennbare Darstellung von Meinungsunterschieden verstanden (= mind. 2 Parteien/Akteure (müssen nicht unbedingt beide genannt werden), aber: tätliche Gewalt wird nicht codiert). Die Darstellung der Meinungs- verschiedenheiten kann entweder von Journalisten thematisiert werden oder auch durch Zitate oder O-Töne erfolgen. Gegenschnitt der „Gegner". Nicht die bloße Möglichkeit von Meinungsverschiedenheiten. Meinungsumfragen sind keine Kontroverse. Auch politische Forderungen von Terroristen werden nicht als Kontroverse codiert, weil hier kein Streit zwischen Diskussionspartnern vorliegt, sondern Forderungen unter Gewaltandrohung.

0 keine Kontroverse erkennbar
1 geringe Kontroverse (sachlich, institutionalisiert). Sachliche Darstellung divergierender Ansichten bzw. Vorwürfe, ohne die Lauterkeit der Kontrahenten Personen oder die Rechtmäßigkeit des Verhaltens zu bestreiten.
2 heftige Auseinandersetzung (persönlich, beleidigend). Vorwürfe, bei denen anderen die Lauterkeit abgesprochen oder die Rechtmäßigkeit ihres Verhaltens bestritten wird. Gerichtliche Auseinanderset-

zungen (auch Ankündigungen) hier codieren. Hierzu gehört auch die Anklageerhebung und eine Festnahme!

Quelle: Fretwurst (2008)

Prominenz
Unter Prominenz wird der Grad der Bekanntheit einer namentlich erwähnten Person verstanden, unabhängig von ihrer politischen/ wirtschaftlichen Macht. Unter Prominenz können alle vorkommenden Personen (auch Pop-Gruppen o.ä.) codiert werden, unabhängig davon, ob sie als Handelnde, sich Äußernde oder Betroffene vorkommen und unabhängig davon, ob sie nur am Rande im Beitrag erwähnt werden, oder ob über sie vorwiegend berichtet wird. Es werden jedoch nur Personen codiert, die genannt werden. Es wird nur die Person mit der höchsten Prominenz codiert. Werden in einem Segment mehrere Gleichprominente erwähnt, so wird die erstgenannte Person codiert.

0 keine prominente Person genannt
1 geringe Prominenz, d.h. nur (durch Massenmedien) in nationalen bzw. Teilöffentlichkeiten bekannte Person
2 hohe Prominenz, d.h. auf internationaler Ebene bekannte Persönlichkeit aus Sport, Kultur, Unterhaltung, Wirtschaft, Politik. Bei deutschen Politikern nur (aktueller oder ehemaliger) Bundespräsident, Bundeskanzler, Außenminister. Sport-Mannschaften/Einzelsportler in populären internationalen Wettbewerben mit hoher Medienpräsenz (z.B. Michael Schumacher, Boris Becker, Kati Witt). Pop- und Film/TV-Stars (Beatles, Sinatra, Brando, Gottschalk, Karajan etc.)

Quelle: Fretwurst (2008)

Journalistischer Präsentationsstil
1 Sachlich informierend (faktenorientiert). Dieser Code wird in der Regel für Nachrichtenbeiträge gelten. Die Darstellung ist sachlich, die Informationsfunktion steht im Vordergrund.
2 Reißerisch informierend (Boulevard-Stil). Inhaltlich stehen Informationen bei der Berichterstattung im Vordergrund. Der Ton ist jedoch eher sensationsheischend. Die Neuigkeit der Information wird explizit in den Vordergrund gestellt.

3 Unterhaltsam informierend (humoristisch-plaudernd). In dieser Kategorie sollten Beiträge aufgenommen werden, die, wie der Bericht über den ins Wasser gefallenen Beginn der Badesaison, in einer lockeren Sprache präsentiert werden.

4 Analysierend-kritisch (argumentativ). Der Beitrag enthält Wertungen und Argumente des Journalisten und stellt schon eher eine Auseinandersetzung des Journalisten mit dem Thema, Problem oder Ereignis dar.

5 Standpunkthaft-kritisch (explizite Stellungnahmen). Beiträge, in denen explizit kommentiert und Stellung bezogen wird. Der Journalist äußerst seine Meinung oder benutzt zur Präsentation des Beitrags zitierte Stellungnahmen, die er mit Kontextfunktionen einander gegenüberstellt.

6 Lobpreisend (unreflektiert). Beiträge mit emotionalisierendem Charakter, die in der Regel unreflektiert positive Einschätzung vermitteln, z.B. Beiträge, die sich mit der regionalen Wirtschaft befassen und in denen die Grenze zur PR oder Schleichwerbung sehr fließend verläuft.

Quelle: Trebbe (1996)