Matthias Lemke
Gregor Wiedemann *Hrsg.*

Text Mining in den Sozialwissenschaften

Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse

PEER REVIEWED

Springer VS

Text Mining in den Sozialwissenschaften

Matthias Lemke
Gregor Wiedemann (Hrsg.)

Text Mining in den Sozialwissenschaften

Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse

Springer VS

Herausgeber
Matthias Lemke
Helmut-Schmidt-Universität/
Universität der Bundeswehr Hamburg
Deutschland

Gregor Wiedemann
Universität Leipzig
Deutschland

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

DLR Projektträger

ISBN 978-3-658-07223-0 ISBN 978-3-658-07224-7 (eBook)
DOI 10.1007/978-3-658-07224-7

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer VS

Lektorat: Jan Treibel, Stefanie Loyal

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Vorwort

In jenem Zweig sozialwissenschaftlicher Forschung, der sich die digitale Analyse großer Datenbestände zur Aufgabe gemacht hat, konkurrieren gegenwärtig zahlreiche, ebenso abgrenzende wie beschreibende Begriffe um Deutungsmacht: *Computational Social Science*, *e-Social Science* oder *Digital Social Science* versuchen, die neuen Entwicklungen bei der Erfassung und Auswertung digitaler Daten in den Sozialwissenschaften begrifflich zu fassen. Dabei betonen sie jeweils unterschiedliche Aspekte interdisziplinärer Methodenanwendung. Wichtiger als die Debatte darüber, welcher Begriff das Verhältnis von informationswissenschaftlicher Auswertungskapazität und menschlicher Interpretationsleistung am besten beschreibt, ist uns jedoch der Austausch über Erfahrungen mit der praktischen Anwendung computerunterstützter Analyseverfahren. Es ist dieser Aspekt des noch jungen Überschneidungsbereichs von *Digital Humanities*[1] und Sozialwissenschaften, für den dieser Band ein Orientierungsangebot sein will.

Zur bestmöglichen Einlösung dieses Anspruches operiert er in einem quasi-experimentellen Setting: ForscherInnen aus unterschiedlichen Disziplinen der *Humanities* haben zu ihren je eigenen Erkenntnisinteressen mit derselben Analyse-Infrastruktur, dem Leipzig Corpus Miner (LCM), auf denselben Daten, einem Textkorpus aus ca. 3,5 Millionen Zeitungsartikeln, gearbeitet. Für die Ausrichtung des Bandes hat das ebenso einfache, wie weitreichende Konsequenzen: Die Anwendung von Text Mining in den Sozialwissenschaften kann so organisiert sein und kann so funktionieren, wie wir es hier vorschlagen. Aber, und dieses ‚aber' ist uns wichtig, für eine gelingende Forschungspraxis sozialwissenschaftlicher

1 Wir verwenden den Begriff „Digital Humanities" in diesem Band vor dem Hintergrund des angelsächsischen Verständnisses von „Humanities", der Geistes- und Sozialwissenschaften als einen gemeinsamen Bereich von den Naturwissenschaften („Science") abgrenzt.

Textanalysen sind in den *Digital Humanities* auch teilweise und vielleicht sogar gänzlich andere Wege denkbar.

Insoweit wollen wir unser Orientierungsangebot keinesfalls als *die* autoritative Best Practice, wohl aber als *eine* Best Practice angewandten Text Minings verstanden wissen. Gerade in einem so jungen Forschungsfeld, so unsere tiefe Überzeugung, sollte es nicht darum gehen, eine Position gegen andere derart zu behaupten, dass alternative Wege ausgeschlossen oder abgeschnitten werden. Anstelle methodischer Grabenkämpfe braucht es einen konstruktiven Dialog, der an einer gelingenden Integration von qualitativen und quantitativen Analyseperspektiven zur Erschließung (textueller) Bedeutungsrepräsentationen aus digitalen Datenbeständen interessiert ist.

Nachdem wir – in aller hier gebotenen Kürze – angedeutet haben, welches wissenschaftsstrategische Ziel wir mit dem vorliegenden Band verfolgen, noch einige nicht minder wichtige Worte des Dankes. Wie bei jedem komplexen Forschungsprojekt – und das gilt insbesondere für Kooperationsprojekte zwischen bislang so wenig verwandten Disziplinen wie den Sozialwissenschaften und der Informatik – waren auch an diesem Band zahlreiche Personen beteiligt, deren ganz unterschiedliche Beiträge in ihrer individuellen Bedeutung kaum zu ermessen, geschweige denn adäquat zu würdigen sind. Sicher ist indes, dass sie alle maßgeblich zum Zustandekommen und zum Gelingen des Bandes beigetragen haben. Besonders hervorheben möchten wir zum einen die externen Autorinnen und Autoren, ohne die nicht nur der zweite Teil dieses Buches nicht möglich gewesen, sondern ohne deren kritische Fragen im Rahmen unseres Workshops der *Leipzig Corpus Miner* und unser Überlegungsstand zum Konzept des *blended reading* nicht dort angekommen wären, wo beide heute stehen. Viel verdanken wir auch den Gutachterinnen und Gutachtern, die in unglaublich kurzer Zeit bereit waren, viele wichtige, scharfsinnige und vor allem konstruktive Kommentare zu den hier versammelten Beiträgen zu liefern. Schließlich gilt unser Dank all jenen, die uns im Kontext des Bandes, des Projektes „ePol – Postdemokratie und Neoliberalismus“ und auch darüber hinaus überaus tatkräftig, nachhaltig und akribisch auf ganz vielfältige Art und Weise unterstützt haben und von denen wir sicher sind, dass sie wissen, dass sie hier gemeint sind.

Hamburg und Leipzig, im Mai 2015
Matthias Lemke und Gregor Wiedemann

Inhalt

Teil 2 Fallstudien

Teil 3 Fazit und Ausblick

Einleitung
Text Mining in den Sozialwissenschaften
Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse

Matthias Lemke und Gregor Wiedemann

1 Computergestützte Textanalysen zwischen Qualität und Quantität

Qualitative Methoden, die durch die Analyse von Texten Aussagen über die soziale Wirklichkeit ermöglichen sollen, gehören zweifelsohne zum zeitgenössischen Kanon sozialwissenschaftlicher Forschung (vgl. dazu Stulpe / Lemke in diesem Band). Wissenssoziologie und Hermeneutik sind ebenso einschlägige Konzepte, wie Grounded Theory, Diskursanalyse oder Qualitative Inhaltsanalyse, die als konkrete Auswertungsmethoden weite Verbreitung gefunden haben. Seit den 1980er Jahren unterstützen Computerprogramme wie MAXQDA oder ATLAS.ti SozialwissenschaftlerInnen beim *Datenmanagement* in der Anwendung dieser Methoden – zunächst noch mit größerer Skepsis ob ihres Einflusses auf den Forschungsprozess verwendet, sind sie heutzutage jedoch fest etabliert (vgl. Kelle 2008). Wichtig dabei ist die Feststellung, dass sich die Leistung von Software für die Analyse qualitativer Daten in den Sozialwissenschaften in der Regel auf die Bereitstellung von Werkzeugen zur manuellen *Organisation und Strukturierung* von Dokumenten beschränkt. Für die eigentliche *Datenanalyse* kommen Computerprogramme dagegen nur selten zum Einsatz. Trotz aller sozialwissenschaftlichen Affinität zu Texten als Datenquelle tauchen Verfahren des Text Mining in einschlägigen Handbüchern (vgl. Diaz-Bone / Weischer 2015) bisher nicht auf (vgl. Puchinger in diesem Band) – ein Befund, der erstaunt. Denn die in den letzten Jahren immer weiter voranschreitende Digitalisierung hat unter anderem auch dazu geführt, dass immer größere Mengen von Textdaten digital vorliegen. Angefangen bei retrodigitalisierten Büchern über Zeitungsartikel bis hin zu diversen Social Media Daten bieten sie ein vielschichtiges Reservoir, das für sozialwissenschaftliche Forschung Verwendung finden kann – und mittlerweile auch in zunehmendem Maße findet.

M. Lemke, G. Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften,*
DOI 10.1007/978-3-658-07224-7_1,

Um dieser Entwicklung Rechnung zu tragen hat das Bundesministerium für Bildung und Forschung (BMBF) im Jahr 2011 ein Programm zur Förderung der „e-Humanities“[1] aufgelegt, in dem explizit „qualitativ arbeitende sozialwissenschaftliche Fächer“ aufgerufen waren, „neue Forschungsmethoden insbesondere unter Nutzung von Informatikmethoden in ihren Fachgebieten [zu] entwickeln“ (BMBF 2011). In diesem Zusammenhang stellen sich zwei große Fragen: Erstens – Warum sollen ausgerechnet jetzt qualitative SozialforscherInnen computergestützte Textanalysen nutzen lernen, wo doch quantitative Inhaltsanalysen, zum Beispiel in den Medienwissenschaften, bereits seit den 1960er Jahren eine etablierte Methode darstellen? Und – zweitens – wie können SozialwissenschaftlerInnen den methodischen Umgang mit eben solchen Verfahren erlernen, ohne selbst tiefgehende Kenntnisse in Informatik und automatischer Sprachverarbeitung erwerben zu müssen?

Einen wichtigen Aspekt zur Antwort auf die *erste* Frage haben wir bereits skizziert: Es ist die rasant gewachsene Verfügbarkeit digitaler Daten, die in retro-digitaler Form zudem auch einen Blick in die Vergangenheit erlauben, der in dieser umfassenden und detaillierten Art über reine Stichproben aus analogen Archiven bislang nicht möglich war. Einen zweiten, nicht minder wichtigen Aspekt stellt die Entwicklung computergestützter Auswertungsverfahren dar. Im Gegensatz zur einfachen Zählung von Schlüsselbegriffen, wie sie in den computergestützten Medieninhaltsanalysen seit den 1960er Jahren durchgeführt werden konnten, erlauben aktuelle Verfahren des Text Mining die Einbeziehung unterschiedlicher Grade von Kontext in die Analyse. Im Gegensatz zu früheren Ansätzen ermöglicht es diese Weiterentwicklung der Verfahren, komplexere semantische Bedeutungsstrukturen in Texten aufzufinden. In einer schon fast sprichwörtlichen gewordenen Definition von Bedeutung schreibt John Firth: „You shall know a word by the company it keeps“ (1957: 11). Vor dem Hintergrund der strukturalistischen Semantik verweist Firth damit auf die Bedeutung von lokalem Kontext, also dem sprachlichen Kontext eines Wortes innerhalb einer Kontexteinheit (z. B. eines Satzes oder Dokuments), der als Sprachregelmäßigkeit bei der Beobachtung einer ganzen Kollektion von Kontexteinheiten (also vielen Sätzen oder Dokumenten) einen spezifischen globalen Kontext in Form aggregierter lokaler Kontexte hervorbringt. Text Mining-Verfahren, wie lexikometrische Ansätze zur Schlüsselwort-Extraktion und Kookkurrenzmessung oder Ansätze des maschinellen Lernens, machen sich die Beobachtung dieser

1 Im Gegensatz zu Digital Humanities rückt der Begriff e-Humanities nicht nur die Arbeit mit digitalisierten Daten in den Vordergrund, sondern will auch stärker auf die Nutzung computergestützter Analyse-Werkzeuge hinweisen. Im Kampf um die Hegemonie zur Beschreibung des Forschungsfeldes scheint sich der Begriff aber gegenüber dem der Digital Humanities zuletzt nicht entscheidend durchgesetzt zu haben.

globalen Kontexte zur Modellierung von Bedeutung zunutze. Insbesondere über Verfahren des maschinellen Lernens ist es möglich, komplexe Wissensstrukturen in Analogie zu Modellen menschlicher Kognition aus großen Textkollektionen heraus zu extrahieren (McNamara 2011).[2] Die Modelle maschinellen Lernens erlauben dabei nicht nur eine Extraktion globaler Kontexte aus der gemeinsamen Beobachtung einer Vielzahl von Dokumenten – beispielsweise die Identifikation eines thematischen Clusters als latenter Sinnzusammenhang in einer Textkollektion. Sie modellieren darüber hinaus gleichzeitig die Beziehungen zwischen den globalen Kontexten und den lokalen Kontexteinheiten – zum Beispiel der Anteil eines Themas an jedem einzelnen Dokument in der Textkollektion. In die Modellberechnungen kann zusätzlich zu den Beobachtungen lokaler Dokumentkontexte auch (qualitatives) menschliches Wissen über Feedback-Schleifen oder textexterne Datenressourcen einfließen. Durch die wechselseitige Betrachtung von Einzelfällen und globaler Kollektion fallen notwendig qualitative und quantitative Perspektiven auf die zu analysierenden Texte in den Modellen maschinellen Lernens zusammen. Es ist eben dieser Umstand, welcher in der Tat auch einer qualitativ ausgerichteten Sozialforschung neue Möglichkeiten zur Integration großer Datenmengen in ihr Methodenrepertoire ermöglicht. Heyer et al. (2006) fassen unter dem Terminus Text Mining solche computergestützten Verfahren für die „semantische Analyse von Texten" zusammen, „welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen" (3). Für die AnwenderInnenperspektive formulieren Mehler und Wolff (2004) in Bezug auf Text Mining den Bedarf an Technologien, „die ihren Benutzern nicht nur ‚intelligente' Schnittstellen zur Textrezeption anbieten, sondern zugleich auf inhaltsorientierte Textanalysen zielen, um auf diese Weise aufgabenrelevante Daten explorieren und kontextsensitiv aufbereiten zu helfen" (1f). Diese Anforderungen im Zusammenhang mit der gestiegenen Verfügbarkeit digitaler Textressourcen für eine Anwendung in den Sozialwissenschaften umzusetzen, ist Gegenstand derzeitiger Forschungsbemühungen.

Die *zweite* Frage hingegen stellt zusätzlich zur technischen Herausforderung eine Herausforderung aus sozialwissenschaftlicher Perspektive dar. Noch steckt die sozialwissenschaftliche Anwendung von Text Mining-Verfahren, die eine Auswertung beliebig großer Datenmengen erlauben, in den Anfängen. Die Gründe hierfür sind

2 Unterschieden werden im Wesentlichen zwei Arten des maschinellen Lernen: 1) datengetriebenes Clustering von Texteinheiten (unsupervised learning) und 2) Klassifikation anhand extern vorgegebener Kategorien (supervised learning). Ausführlichere Informationen zum maschinellen Lernen finden Sie im Beitrag von Dumm/Niekler in diesem Band.

vielfältig: Die semi-automatische Auswertung von Textdaten bedarf umfangreicher Arbeitsschritte. Diese reichen von der zeit- und kostenintensiven Erschließung von Textkorpora, der Entwicklung oder Anpassung von geeigneten Analyseinfrastrukturen zur automatischen Sprachverarbeitung, der Entwicklung oder Anpassung geeigneter Text Mining-Algorithmen, der Durchführung aufwändiger Evaluationsmaßnahmen bis hin zur Generierung und Interpretation mehr oder weniger komplexer Ergebnisvisualisierungen. Einerseits gilt es, SozialwissenschaftlerInnen überhaupt in die Lage zu versetzen, solche Auswertungsmethoden auf Texte anzuwenden. Gewisse Grundkenntnisse in computerlinguistischen Verfahrensweisen und Ansätze der automatischen Sprachverarbeitung sind unseres Erachtens dafür unumgänglich. Gleichzeitig erscheint es schwierig, an SozialwissenschaftlerInnen, die große Mengen qualitativer Daten analysieren wollen, die Anforderung zu stellen, ihre Analysen durch Programmierung weitgehend eigenständig umzusetzen. In den Digital Humanities wird dazu seit einigen Jahren um das Motto „more hack; less yack" eine Debatte geführt (Nowviskie 2014), bei der die Forderung im Mittelpunkt steht, Potenziale, Möglichkeiten und Ansätze zur Auswertung digitaler Ressourcen in interdisziplinärer Zusammenarbeit nicht nur diskursiv auszuloten, sondern als autonomer Digital Humanist die hierfür erforderlichen Werkzeuge selbst in die Hand zu nehmen. KritikerInnen dieser Haltung beklagen dagegen hohe Einstiegshürden, Theorievergessenheit und Mangel an methodischen Standards (Cecire 2011). Um die Anschlussfähigkeit von Methoden des Text Mining für die Analyse qualitativer Daten sicher zu stellen, muss unserer Meinung nach auf die lange Tradition von Debatten zur Qualitätssicherung empirischer Sozialforschung besonders Rücksicht genommen werden. Anstelle der „more-hack"-Haltung, die ihre BefürworterInnen in doppelqualifizierten Ein-Personen-Projekten zu finden meinen und die für die experimentelle Entwicklung neuer Auswertungstechniken durchaus wichtig und notwendig erscheinen, braucht es für eine nachhaltige Etablierung von Text Mining in den Sozialwissenschaften eine Einigung auf bestimmte, einfach zu handhabende Standards, die den Forschenden eine stärkere Konzentration auf die Forschungsinhalte und weniger auf die Methoden erlauben. Sobald die technische Umsetzung nicht mehr im Vordergrund der Debatte um den Einsatz der neuen Technologien steht, wird ein Austausch über den systematischen Methodeneinsatz und die einschlägigen Qualitätskriterien für die empirische Forschung möglich. Mit dieser doppelten Herausforderung hat sich das Projekt „ePol – Postdemokratie und Neoliberalismus", aus dessen Kontext dieser Band hervorgegangen ist, konfrontiert gesehen. Der vorliegende Band ist unser Versuch, eine Antwort auf diese Herausforderungen zu formulieren.

2 Qualitative Sozialforschung trifft Text Mining

Im ePol-Projekt wird nicht nur eine politikwissenschaftliche Fragestellung mit Hilfe von Text Mining-Verfahren beantwortet. Vielmehr ist eine Analyseinfrastruktur für die Speicherung, das Suchen und die Analyse großer Dokumentbestände entstanden, die für verschiedenste Forschungsfragen produktiv einsetzbar ist. Im folgenden Abschnitt wollen wir aus Gründen der Transparenz und Nachvollziehbarkeit daher kurz den eigentlichen Projekthintergrund zur Analyse neoliberaler Plausibilisierungen in der politischen Öffentlichkeit und, damit zusammenhängend, unsere Datengrundlage und Analyseinfrastruktur vorstellen. Jenseits unseres eigenen Erkenntnisinteresses zum empirischen Nachweis neoliberaler Hegemonie bilden die Daten und die Analyseinfrastruktur die Ausgangsbasis aller in diesem Band versammelten Fallstudien.

2.1 Das „ePol"-Projekt

Das vom BMBF geförderte Projekt „ePol – Postdemokratie und Neoliberalismus"[3] hat sich zum Ziel gesetzt, den Wandel von politischen Begründungsstrukturen in 60 Jahren bundesdeutscher Zeitungsberichterstattung mit Hilfe von Text Mining-Verfahren zu untersuchen (Wiedemann et al. 2013). Die dem Projekt zugrunde liegende Hypothese lautet, dass sich im Zuge einer Verbreitung politischer Leitideen des Neoliberalismus der Chicago School, die von der Postdemokratiedebatte als neoliberale Hegemonie beschrieben wird (Crouch 2008; 2011), in zunehmendem Maße ökonomie- und marktaffine Begründungen für kollektiv bindende Entscheidungen zur Implementierung beliebiger policies in immer mehr Politikfeldern durchgesetzt haben und nach wie vor durchsetzen. Diese müssen, so unsere aus der konjunkturellen Logik einer politischen Debatte entlehnte Annahme, zu Anfang noch intensiv begründet werden, wohingegen die Begründungsdichte mutmaßlich abnimmt, sobald sich eine Hegemonie neoliberaler Vorstellungen im betroffenen Politikfeld bzw. in der politischen Öffentlichkeit etabliert hat. Eine solche, im Medium der Sprache nachvollziehbare neoliberale Ökonomisierung rekurriert – wie bereits angedeutet – in der Politischen Theorie auf kritische Gegenwartsdiagnosen, die den zeitgenössischen repräsentativen Demokratien westlichen Typs eine zunehmende Abnahme der Demokratiequalität attestieren. Diese Postdemokratien zeichnen

3 Verbundprojekt des Lehrstuhls für Politische Theorie, Helmut-Schmidt-Universität/Universität der Bundeswehr Hamburg (FKZ 01UG1231A) und der Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig (FKZ 01UG1231B).

sich dabei auf sprachlicher Ebene unter anderem durch die behauptete Einengung politischer Möglichkeitsräume aufgrund vermeintlicher ökonomischer Sachzwänge aus. Die zunehmende Relevanz neoliberaler, in der politischen Öffentlichkeit präsenter Leitideen für diese Tendenz zur Postdemokratisierung untersuchen wir im ePol-Projekt für den Zeitraum von 1949 bis 2011 anhand einer Diskursanalyse, bei der neben qualitativen Ausprägungen vor allem auch quantitative Verteilungen und Veränderungen politischer Plausibilisierungsmuster analysiert werden.

Neben der Erkenntnis über diskursive Entwicklungen neoliberaler Plausibilisierungsmuster zielt das Projekt vor allem auf die Entwicklung und Etablierung von geeigneten Werkzeugen für die Analyse großer Textmengen in den Sozialwissenschaften. Die Identifikation möglicher Vorgehensweisen zum Einsatz von Text Mining und die Erfahrungen damit sind ein wichtiges Zwischenergebnis des Projekts, das in diesem Band seinen Ausdruck finden soll. Um die Diskussion hierüber nicht allein projektintern zu führen und damit etwaigen Verfahrensblindheiten zu erliegen, haben wir im September 2014 in Hamburg einen Workshop unter dem Titel „Text Mining in der Politikwissenschaft" veranstaltet. 25 WissenschaftlerInnen aus Deutschland, Österreich und den Niederlanden waren eingeladen, mit der Analyseinfrastruktur des Projekts, dem *Leipzig Corpus Miner* (LCM), und den Daten von ePol möglichst neue Antworten und Einsichten zu eigenen Forschungsfragen zu finden. In vier Tagen wurden die Grundlagen und Anwendungsmöglichkeiten des LCM vermittelt, so dass die TeilnehmerInnen in der Lage waren, autonom mit der Analyseinfrastruktur zu arbeiten. Damit haben wir zwei Ziele verfolgt: Zum einen wollten wir über den Kontext unseres eigenen Projekts hinaus Beispiele generieren, die unterschiedliche Anwendungen – von der Politikwissenschaft über die Soziologie bis in die Zeitgeschichte – sozialwissenschaftlichen Text Minings zeigen und so als Anschauungs- und Orientierungsmaterial dienen können. Zum anderen haben uns diese thematisch sehr unterschiedlichen Fallstudien im Sinne eines *proof of concept* gezeigt, inwiefern die Analyseinfrastruktur nicht nur maßgeschneidert für das Erkenntnisinteresse des ePol-Projekts funktioniert, sondern darüber hinaus für beliebige andere Fragestellungen verwendbar ist. In diesem Zusammenhang werden auch Anforderungen an die methodische Integration von Text Mining und qualitativer Sozialforschung sichtbar, die über unser eigenes Projekt hinausgehen. Die Fallstudien, die auf Grundlage der im Rahmen dieses Workshops erworbenen Kompetenzen entwickelt wurden, sowie die Erfahrungen aus der Anwendung von Text Mining-Verfahren durch vornehmlich qualitativ arbeitende SozialwissenschaftlerInnen sind in Teil II dieses Bandes dokumentiert.

2.2 Daten

Die Daten, die dem ePol-Projekt für die Analyse neoliberaler Ökonomisierung in Plausibilisierungsmustern der politischen Öffentlichkeit als Grundlage dienen, bestehen aus insgesamt ca. 3,5 Millionen Artikeln dreier deutscher Tages- und einer Wochenzeitung: Neben der *Frankfurter Allgemeine Zeitung* (FAZ), der *tageszeitung* (taz) und der *Süddeutsche Zeitung* (SZ) komplettiert *Die Zeit* das Datenpaket (vgl. Tabelle 1). Die Tageszeitungen taz und SZ sowie die Wochenzeitung Die Zeit stehen in den angegebenen Zeiträumen als Vollkorpora zur Verfügung. Von der FAZ liegt eine repräsentative Stichprobe aus den Ressorts „Politik", „Wirtschaft" und „Feuilleton" vor. Diese Strichprobe wurde anhand der folgenden Vorgehensweise aus dem vollständigen Archiv der FAZ der Jahrgänge 1959 bis 2011 gezogen:

1. Wähle alle Artikel der Kategorie „Meinung", die in den Ressorts „Politik", „Wirtschaft" oder „Feuilleton" erschienen sind als Stichprobenmenge aus.
2. Wähle zusätzlich alle Artikel, die in den Ressorts „Politik", „Wirtschaft" oder „Feuilleton" erschienen sind, die jedoch nicht der Kategorie „Meinung" oder „Rezension" zugeordnet sind,
   - ordne sie aufsteigend nach Datum, und
   - wähle jeden zwölften Artikel aus dieser sortieren Liste zur Erweiterung der Stichprobenmenge aus 1. aus.

Die Auswahl des Zeitraums, der Ressorts sowie die Berücksichtigung aller in der Rubrik „Meinung" erschienenen Artikel für die FAZ-Stichprobe richtete sich an den Erfordernissen des ePol-Projekts aus.[4] Über die dargestellte Strategie zur Auswahl der Stichprobe wurden ca. 15 Prozent aller in den drei genannten Ressorts erschienenen Artikel ausgewählt. Das Verfahren stellt sicher, dass die Artikelmengen in den einzelnen Jahrgängen sowie den einzelnen Ressorts direkt proportional zu den vollständigen Artikelmengen entlang dieser Kriterien verteilt sind. Die Tatsache, dass die Verlage für das Projekt ihre gepflegten Volltextarchive bzw. Stichproben davon zur Verfügung stellten, stellt ferner sicher, dass sich keine Doubletten in den Daten befinden, wie es häufig der Fall ist, wenn größere Mengen

4 Untersucht werden soll die Ökonomisierung in politischen Plausibilisierungsmustern vordringlich ab den 1960er Jahren, weil ein zunehmender Einfluss der in der Postdemokratiedebatte hierfür als federführend eingestuften Chicago School um Milton Friedman, Gary S. Becker und anderen erst ab Ende der 1970er Jahre – etwa in Form des Thatcherismus oder der Reaganomics – angenommen wird. Wir gehen weiterhin davon aus, in Kommentaren und Meinungsartikeln von Tageszeitungen eine höhere Dichte von Begründungen für Politik vorzufinden, als in anderen Artikelformaten.

von Zeitungsdaten über Drittanbieter bezogen werden. Zudem enthalten die von den Verlagen gelieferten Artikel bereits wertvolle Metadaten-Annotationen, wie Erscheinungsdatum, Ressort, Autorenname und Seitenzahl in der Druckausgabe, die für spätere Analysen wertvolle Kriterien zur Selektion von Subkorpora liefern. Innerhalb der einzelnen Texte sind Informationen zu Überschrift, Unterüberschrift und die Trennung in Absätze annotiert, so dass Analysenverfahren auf diese einzelnen Kontexteinheiten zugreifen können. Ältere Jahrgänge der Zeitungen, im Wesentlichen alle Ausgaben vor 1990, liegen in retro-digitalisierter Form vor. Das heißt, die archivierten gedruckten Ausgaben wurden eingescannt und anschließend mit einer OCR-Texterkennungssoftware in maschinenlesbaren Text umgewandelt. Die unterschiedliche Qualität dieser OCR-Vorgänge und der Archivvorlagen führen dazu, dass vor allem in älteren Jahrgängen hin und wieder Fehler durch falsch erkannten Text in den Daten auftauchen. Die Verlage, welche die Digitalisierung ihrer Archive durchgeführt geliefert haben, garantieren jedoch eine Genauigkeit der Texterkennung von über 99 Prozent.

Tab. 1 Daten des ePol-Projekts.[5]

Publikation	Zeitraum	Ausgaben	Artikel	Speicher
Die Zeit	1946–2011	3.841	397.729	4,5 GB
Frankfurter Allgemeine Zeitung (FAZ)	1949–2011	16.185	200.398	1,1 GB
Tageszeitung (TAZ)	1986–2011	7.821	1.391.981	3,7 GB
Süddeutsche Zeitung (SZ)	1992–2011	6.027	1.505.714	5,2 GB
Gesamt		33.874	3.495.822	14,5 GB

Quelle: ePol – eigene Zusammenstellung.

2.3 Analyseinfrastruktur

Der *Leipzig Corpus Miner* (LCM) ist eine Webanwendung, die im Rahmen von ePol entwickelt wurde, um SozialwissenschaftlerInnen computergestützte Auswertungsalgorithmen für sehr große Textmengen zur Verfügung zu stellen (ausführlich dazu der Beitrag von Wiedemann/Niekler in diesem Band). Dazu bündelt

5 Die Zeit, TAZ und SZ liegen als vollständige Korpora vor. Von der FAZ wurde eine repräsentative Stichprobe (ca. 15 % der redaktionellen Beiträge aus Politik, Wirtschaft und Kultur jeder Ausgabe) gezogen.

die LCM-Architektur verschiedene Text Mining-Verfahren für den Zugriff, die Verwaltung, die Analyse und die Visualisierung qualitativer Daten. Durch eine einfach zu bedienende Benutzeroberfläche ermöglicht der LCM Volltextzugriff auf die 3,5 Millionen Zeitungstexte, die nach Suchbegriffen und Metadaten zu Subkollektionen gefiltert werden können. Auf dem Gesamtdatenbestand sowie auf den Subkollektionen können verschiedene computergestützte Auswertungsverfahren angewendet und zu Analyseworkflows kombiniert werden, darunter Frequenz- und Kookkurrenzanalysen, Topic Modelle und überwachte Text-Klassifikation. Damit ermöglicht der LCM die empirische Analyse sozialwissenschaftlicher Fragestellungen auf Basis großer Dokumentkollektionen, wobei qualitative und quantitative Analyseschritte miteinander verschränkt werden können. Das Arrangement des Workshops – WissenschaftlerInnen arbeiten zu je eigenen Fragestellungen auf den gleichen Daten mit dem gleichen Werkzeugkasten – erlaubt es Erfahrungen darüber zu sammeln, wie vornehmlich qualitativ arbeitende SozialwissenschaftlerInnen die neuen Möglichkeiten des Text Mining für ihre Erkenntnisinteressen einsetzen.

Der Vorteil der Nutzung des LCM als vergleichsweise einfach zu handhabende Infrastruktur für eine systematische Anwendung komplexer Auswertungsverfahren besteht unzweifelhaft darin, dass sich die Forschenden vorrangig auf die von Ihnen untersuchten inhaltlichen Fragestellungen konzentrieren können. Die in diesem Band versammelten Fallstudien haben damit weniger den Charakter von methodischer Exemplifizierung sondern entsprechen eher einem *real world scenario* sozialwissenschaftlicher Forschung. Dieses Setting erlaubt es, einen vergleichenden Blick auf die angewandten Methoden und Forschungsabläufe zu werfen, der seinerseits Aufschluss über Anforderungen, Zugangsschwellen und Kompatibilität von verschiedenen Text Mining-Verfahren im Zusammenhang mit etablierten Abläufen der qualitativen empirischen Sozialforschung geben kann.

3 Aufbau des Bandes

Unser Band, dessen Ziel es ist, anhand praktisch angewandter Analyseverfahren Orientierungsangebote zur Integration von Text Mining-Verfahren in sozialwissenschaftliche Forschungskontexte zu eröffnen, gliedert sich in drei Teile.

Der erste Teil umfasst all das, was unsere eigene Positionsbestimmung ausmacht. Hier führen wir – im Beitrag von *Alexander Stulpe und Matthias Lemke* – aus, warum und wie Text und soziale Wirklichkeit analytisch aufeinander bezogen werden können, was sich durch die Digitalisierung an diesem Verhältnis ändert und wie genau blended reading als Anwendungsstrategie der Text Mining-Verfahren des LCM

funktionieren kann. *Gregor Wiedemann und Andreas Niekler* erläutern, aufgrund welcher Erwägungen wir unsere Analyseinfrastruktur, den LCM, hinsichtlich der enthaltenen Funktionalitäten so gestaltet haben, wie er gegenwärtig aussieht und wie seine Bedienung funktioniert. Zudem reflektieren *Sebastian Dumm und Andreas Niekler* die Anschlussmöglichkeiten zwischen etablierten sozialwissenschaftlichen Methodenstandards an die vom LCM ermöglichten Analyseverfahren sowie an ihre Visualisierungen. Ein Überblick über den gegenwärtigen Stand (2015) der Anwendungen von Text Mining-Verfahren in der sozialwissenschaftlichen Literatur im Beitrag von *Carmen Puchinger* rundet den ersten Teil ab.

Der zweite Teil des Bandes enthält Fallstudien aus dem sozialwissenschaftlichen Fächerspektrum aus dem deutschen und internationalen Kontext. Die einzelnen Beiträge machen praxisnah Anforderungen an die methodische Integration von Text Mining und qualitativer Sozialforschung sichtbar, die über unser eigenes Projekt hinausgehen. Da es uns genau um diese Orientierungsleistung geht, die darin besteht zu zeigen, wie diese Integration gelingen kann, haben wir die Texte nach dem Grad der Intensität und Komplexität der eingesetzten Verfahren sortiert. Dabei stehen jene Fallstudien am Anfang, die einen eher zurückhaltenden Gebrauch von Text Mining-Verfahren machen und diese stärker als Instrument zur Erschließung des Korpus, etwa hinsichtlich geeigneter Einstiegspunkte für ein *close reading*, verwenden. Am Ende dieses Spektrums finden sich wiederum Beiträge, die nicht nur auf die textstatistischen Verfahren des LCM zurückgegriffen haben, sondern diese auch noch durch eigene Berechnungen oder Datenquellen ergänzen oder mit solchen kombinieren. *Annette Förster* geht in ihrem Beitrag der Frage nach, wie genau der Absolutheitsanspruch hinsichtlich des Verbots von Folter nach den Anschlägen vom 11. September 2001 und der Entführung Jakob von Metzlers 2002 in Frage gestellt worden ist. Mit Hilfe eines durch Frequenz- und Kookkurrenzanalysen vorbereiteten close readings zeigt der Beitrag, wie sich die Zeitungsberichte zur Folterdebatte vor und nach den genannten Ereignissen in quantitativer und inhaltlicher Hinsicht unterscheiden und wie in diesem Zusammenhang der Einfluss wissenschaftlicher Debatten auf den öffentlichen Diskurs zu werten ist. Ein ähnliches Vorgehen wählt *Robert Hädicke* in seinem Aufsatz. Er verwendet Frequenz- und Kookkurrenzanalysen zur Erschließung jener Zeitungstexte, die um die Anschläge des 11. Septembers 2001 herum eine Veränderung der öffentlichen Bezugnahme zu den demokratischen Grundwerten von Freiheit und Sicherheit anzeigen. *Katharina Garvert-Huijnen* und *Pim Huijnen* untersuchen das deutsche Niederlandebild, indem sie neben Frequenz- und Kookkurrenzanalysen auch Topic Modelle und Named Entities in ihre Analyse einbeziehen. Auf diesem Wege wollen sie herausfinden, ob und inwiefern die seit dem Ende des Zweiten Weltkrieges erfolgte europäische Integration auch in der wechselseitigen Wahrnehmung der

ungleichen Nachbarn ihren Niederschlag gefunden hat. *Martina Maas* widmet sich am Beispiel des PISA-Diskurses der Frage, wie in einer demokratischen Ordnung – trotz des Anspruchs auf Chancengleichheit unabhängig von sozialer Herkunft und basierend auf dem Leistungsprinzip – Bildungsungleichheiten aufgrund sozialer Herkunft in den öffentlichen Medien thematisiert werden. Anhand von Frequenz- und Kookkurrenzanalysen und durch qualitative Textanalysen deckt sie dominante Deutungsmuster im Rahmen des PISA-Diskurses auf, die darauf hindeuten, dass in der politischen Öffentlichkeit die Legitimation sozialer Ungleichheit weniger anhand des meritokratischen Prinzips erfolgt, sondern Bildungsungleichheiten stärker vor dem Hintergrund von Gerechtigkeitskriterien diskutiert werden. Der Zusammenhang von wissenschaftlicher Expertise und politischem Entscheiden steht im Zentrum des Beitrages von *Daniela Russ* und *Julia Schubert*. Sie gehen angesichts der Krisendiagnose der modernen Demokratie, wonach politische Entscheidungen zunehmend von ExpertInnen und WissenschaftlerInnen getroffen werden, der Frage nach, ob deutsche Printmedien politische Entscheidungen tatsächlich in einem zunehmend verwissenschaftlichten Kontext präsentieren. Die Anwendung von close reading, von Frequenz- und Kookkurrenzanalysen sowie die Erstellung eines eigenen Wörterbuches zur Operationalisierung dreier Typen von Verwissenschaftlichung münden jedoch in eine Ablehnung der Hypothese der zunehmenden Verwissenschaftlichung politischer Entscheidungen. Weder zeige sich eine eindeutig zu- oder abnehmende Tendenz, noch lassen sich signifikante Hoch- oder Tiefpunkte allgemeiner Verwissenschaftlichung beobachten. *Christian Rauh und Sebastian Bödeker* untersuchen das Auftreten von Nicht-Regierungsorganisationen im Zusammenhang mit der Berichterstattung über internationale Organisationen in der politischen Öffentlichkeit. Angesichts von Globalisierung und Internationalisierung untersuchen sie die Frage, ob die institutionelle Auslagerung politischer Autorität auch von angemessenen öffentlichen Willensbildungsprozessen begleitet wird. Auf Basis der mit Hilfe des LCM erzeugten Dictionary-Analysen stellen sie eigene Berechnungen an, die in der Summe entsprechend der theoretischen Erwartungen bestätigen, dass ein höheres Ausmaß der kommunizierten Autorität internationaler Organisationen – hier insbesondere der Europäischen Union und der Vereinten Nationen – mit einer stärkeren Politisierung des öffentlichen Diskurses einhergeht. *Maximilian Hösl und Abel Reiberg* nähern sich der Entstehung des Politikfeldes zur Netzpolitik an, wobei sie aus der Theorie strategischer Handlungsfelder zwei Bedingungen der Politikfeldgenese ableiten: Die Existenz eines gemeinsamen Verständnisses vom Feld und einer engen Interfeldbeziehung zwischen komplementären Handlungsfeldern. Anhand einer empirischen Analyse in Form einer Kombination von Topic Modellen, Frequenz- und Kookkurrenzanalysen sowie durch die eigene Berechnung des Politisierungsgrades eines Themas auf der Basis von

Dictionary-Zählungen gelingt es ihnen, die Kernthemen des politischen Diskurses um das Internet zu identifizieren und die Bedeutung feldbezeichnender Begriffe, wie etwa dem der ‚Netzpolitik', näher zu beschreiben. Seit einiger Zeit werde, so konstatiert *Maximilian Overbeck* in seinem Artikel, in sozialwissenschaftlichen Diskursen vermehrt eine ‚Rückkehr der Religion' konzediert. Empirisch kann er durch Frequenzanalysen, Topic Modelle und Klassifikationen zeigen, dass eine Veränderung des Themas Religion in der deutschen Presse nur insofern vorliegt, als insbesondere für den Islam und für die als zunehmend wichtiger eingeschätzte Rolle von Religion im Kontext internationaler Themen eine breitere und häufigere Verwendung festzustellen ist. *Alexander Petring* schließlich untersucht in seinem Beitrag, ob sich parallel zur Liberalisierung der Güter- und Dienstleistungsmärkte sowie der Sozialversicherungssysteme auch eine Veränderung des öffentlichen Diskurses über soziale Gerechtigkeit, Armut und Ungleichheit in Deutschland beobachten lässt. Durch die Verwendung von Frequenz- und Kookkurrenzanalysen und durch die Hinzuziehung externer Daten – hier der Gini-Koeffizient und statistische Daten über Armutsquoten – kann er mit Hilfe von Kreuz-Korrelationen auf Zeitreihen in unterschiedlichen Tageszeitungen zeigen, dass sowohl hinsichtlich der Häufigkeit wie auch des begrifflichen Kontextes eine Ausdifferenzierung des Diskurses im Zusammenhang mit tatsächlichen Vermögensverteilungen vorliegt. Ein alle drei untersuchten Tageszeitungen erfassender neoliberaler Diskurswandel sei allerdings, so seine Einschätzung, nicht zu erkennen.

Im dritten Teil schließlich fassen die Herausgeber die im Rahmen der Fallstudien gewonnenen Ergebnisse und Erfahrungen zusammen. Dabei versuchen wir aus einer methodologisch akzentuierten Meta-Perspektive heraus aufzuzeigen, welche Analysewege eher, und welche weniger gut beziehungsweise nur mit gewissen Einschränkungen gangbar sind. Gerade diese forschungspragmatische Perspektive, die anhand konkreter Einzelfallanalysen reflektiert, wo welche Erkenntnisgrenzen liegen und wie sie gegebenenfalls umgangen oder erweitert werden können, scheint uns angesichts der bisher nur geringen Erfahrungswerte zur sozialwissenschaftlichen Anwendung von Text Mining-Verfahren überaus gewinnbringend. Ohne dass damit schon das letzte Wort gesprochen wäre, können die in diesem Teil angestellten Überlegungen zur Orientierung über eine allgemeine methodische Integration von qualitativen Analysen mit Text Mining-Verfahren dienen, die notwendig immer auch eine quantitative Analysedimension umfassen.

Literatur

BMBF, 2011: Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities; http://www.bmbf.de/foerderungen/16466.php, 16.01.2012.

Cecire, Natalia, 2011: When Digital Humanities Was in Vogue. In: Journal of Digital Humanities 1 (1); http://journalofdigitalhumanities.org/1-1/when-digital-humanities-was-in-vogue-by-natalia-cecire/, 18.05.2015.

Crouch, Colin, 2011: Das befremdliche Überleben des Neoliberalismus. Postdemokratie II, Berlin.

Crouch, Colin, 2008: Postdemokratie, Frankfurt (Main).

Diaz-Bone, Rainer / Weischer, Christoph, 2015 (Hg.): Methoden-Lexikon für die Sozialwissenschaften. Lexikalischer Zugang zu allen Aspekten sozialwissenschaftlicher Methoden, Wiesbaden.

Firth, John R., 1957: A synopsis of linguistic theory 1930-1955. In: Studies in Linguistic Analysis, 1–32.

Heyer, Gerhard; Quasthoff, Uwe; Wittig, Thomas, 2006: Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse, Bochum.

Kelle, Udo, 2008: Computergestützte Analyse qualitativer Daten. In: Uwe Flick (Hg.): Qualitative Forschung. Ein Handbuch. 6. Aufl., Reinbek bei Hamburg, 485–502.

Lemke, Matthias / Niekler, Andreas / Schaal, Gary S. / Wiedemann, Gregor, 2015: Content Analysis between Quality and Quantity. Fulfilling Blended-Reading-Requirements for the Social Sciences with a scalable Text Mining Infrastructure In: Datenbank-Spektrum. Zeitschrift für Datenbanktechnologien und Information Retrieval, 15(1), 7–14.

Lemke, Matthias / Stulpe, Alexander, 2015: Text und soziale Wirklichkeit. Theoretische Grundlagen und empirische Anwendung durch Text-Mining-Verfahren in sozialwissenschaftlicher Perspektive. In: Zeitschrift für Germanistische Linguistik (ZGL), 43(1), 52–83.

McNamara, Danielle S., 2011: Computational methods to extract meaning from text and advance theories of human cognition. In: Topics in Cognitive Science, 3 (1), 3–17.

Mehler, Alexander; Wolff, Christian, 2005: Einleitung: Perspektiven und Positionen des Text Mining [Einführung in das Themenheft Text Mining des LDV-Forum]. In: LDV-Forum, 20 (1), 1–18.

Nowviskie, Bethany, 2014,: On the Origin of "Hack" and "Yack". In: Journal of Digital Humanities, 3 (2); http://journalofdigitalhumanities.org/3-2/on-the-origin-of-hack-and-yack-by-bethany-nowviskie/, 18.05.2015.

Wiedemann, Gregor / Lemke, Matthias / Niekler, Andreas, 2013: Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentation in der Bundesrepublik Deutschland 1949 – 2011. Ein Werkstattbericht. In: Zeitschrift für Politische Theorie (ZPTh) 3(1), 99–115.

Teil 1
Grundlagen

Blended Reading

Theoretische und praktische Dimensionen der Analyse von Text und sozialer Wirklichkeit im Zeitalter der Digitalisierung

Alexander Stulpe und Matthias Lemke

Zusammenfassung

Die Analyse von Text ist für die Sozialwissenschaften, von der klassischen Hermeneutik über die Wissenssoziologie bis zur Diskursanalyse, von eminenter Bedeutung, um Aussagen über die soziale Wirklichkeit zu treffen. Im Zeitalter der Digitalisierung wird der klassische Dualismus von Interpret und Text durch die immense Menge digital verfügbarer Daten und neue, elektronische Analyseverfahren aufgebrochen. Für das nunmehr ungleich komplexere und weniger transparente Verhältnis bedarf es – um die neuen Daten angemessen nutzen zu können – einer Verständigung über eine geeignete Analysestrategie. Blended reading als modularer Analyseprozess ist ein auf drei Verfahrensebenen angelegter Entwurf eines Prozesses zur Erschließung großer Textdatenmengen in den Digital Humanities.

Abstract

For the Social Sciences the analysis of text, ranging from classical hermeneutics, the sociology of knowledge to discourse analysis, is of the utmost importance in order to draw conclusions about social reality. In the digital age the classic duality of interpreter and text has changed due to the immense amount of digitally available data and new electronic methods of analysis. For today's more complex and less transparent situation, an appropriate use of the new data requires a common agreement on a suitable strategy for text mining analysis. Blended Reading as a modular analysis process offers a three-level method design, serving as a developing process for the exploration of large text-data volumes in the field of Digital Humanities.

M. Lemke, G. Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften,*
DOI 10.1007/978-3-658-07224-7_2,

1 Textanalyse und sozialwissenschaftliche Erkenntnis

Im Verlaufe eines Gesprächs mit Jean-Claude Carrière über die „große Zukunft des Buches" empfiehlt Umberto Eco für den Umgang mit der Masse an neuen Buchveröffentlichungen eine Methode, die er als „Theorie der Dezimierung" bezeichnet:

> „Es genügt, eins von zehn Büchern zu lesen. Bei den anderen reicht ein Blick in die Bibliographie und die Fußnoten, um zu erkennen, ob die angegebenen Referenzen ernst zu nehmen sind oder nicht. Ist das Werk interessant, braucht man es nicht zu lesen, weil es mit Sicherheit besprochen, zitiert und in anderen Werken kommentiert wird, einschließlich dem, das man zu lesen beschlossen hat." (Eco/Carrière 2011: 64).

Als Antwort auf die von Gregory Crane (2006) aufgeworfene Frage: „What do you do with a million books?", wäre die dezimierungstheoretische Auskunft, man müsse ja nur 100.000 davon lesen, allerdings wenig überzeugend. Dagegen verspricht die von Franco Moretti (2000) auf dem Feld der Literaturwissenschaften vorgeschlagene Methode des „Distant Reading" eine Form von Selektivität, die Zeit- mit Erkenntnisgewinn verbindet, indem sie „die reale Vielfalt an literarischen Texten bewusst reduziert und auf einer abstrakteren Ebene verhandelt" (Moretti 2009: 7). Gerade angesichts großer Textkorpora, die im Modus traditioneller Lektüre – *close reading* – in keinem vernünftigen Zeitrahmen zu bewältigen wären, macht ein mithilfe von Text Mining-Verfahren vollzogenes *distant reading* aus der Not des *close reading* eine erkenntnistheoretische Tugend: je größer die Textmenge, desto breiter und sicherer die empirische Basis für allgemeine Aussagen über Strukturen, wiederkehrende Muster und Entwicklungstendenzen in der Literatur.[1]

Eine andere Frage ist, ob und inwiefern auch die Politikwissenschaft und die Sozialwissenschaften generell die Methode des *distant reading* durch den Einsatz von Text Mining-Verfahren erkenntnisgewinnbringend anwenden können. Trotz aller strukturellen Ähnlichkeiten und interdisziplinären Affinitäten zwischen Literatur- und Sozialwissenschaften[2] besteht schließlich zwischen beiden eine

---

1 Moretti (2009: 8) selbst begründet die Vorzüge dieses Umgangs mit Texten nicht nur mit der bloßen Quantität des Analysematerials, sondern auch mit einer methodologischen Nähe zu den „objektivistischen Verfahrensweisen" der Natur- und Sozialwissenschaften, denen er ein gegenüber der traditionellen Literaturwissenschaft „größeres Erkenntnispotential" konzediert.

2 Die interdisziplinären Affinitäten variieren je nach Fachrichtung und Subdisziplin. Innerhalb der – per se in ihrem integrationswissenschaftlichen Selbstverständnis ja schon verschiedene fachdisziplinäre und damit assoziierte epistemologische und methodologische Traditionen verbindenden – Politikwissenschaft sind literaturwissenschaftliche Anschlussmöglichkeiten auf theoretischer und methodologischer Ebene vor allem im

prinzipielle Differenz, die ihre jeweilige disziplinäre Eigen- und Zuständigkeit begründet: die Differenz von Literatur und Gesellschaft. Diese Differenz rechtfertigt die Existenz beider Disziplinen (bzw. Disziplingruppen) im Hinblick auf ihre unterschiedlichen Erkenntnisgegenstände – und hat Konsequenzen für ihr jeweiliges Verhältnis zu Texten. Dass Texte Erkenntnisgegenstände literaturwissenschaftlicher Forschung sind, ist unbestreitbar. Anders verhält es sich mit den Sozialwissenschaften: Ihrem disziplinären Auftrag gemäß haben sie es mit der Beschreibung, Analyse, Diagnostik der sozialen Wirklichkeit zu tun, mit der Gesellschaft insgesamt oder in ihren Elementen oder Teilbereichen, was auch immer je nach theoretischem Standpunkt darunter zu verstehen ist. Dass eine Aussage über einen Text bzw. ein Textkorpus, wie sie mit den Mitteln der Digital Humanities (DH) ermöglicht wird, von sozialwissenschaftlicher Relevanz sei, bedarf daher einer theoretischen Begründung, die das Verhältnis von ‚Text' und ‚sozialer Wirklichkeit' betrifft. Während ein Text als legitimer Gegenstand literaturwissenschaftlicher Erkenntnis gelten kann und daher beispielsweise durch Text Mining-Verfahren generierte Aussagen prinzipiell literaturwissenschaftliche Relevanz beanspruchen können, befindet sich der sozialwissenschaftliche Erkenntnisgegenstand grundsätzlich jenseits des Textes und man muss für den Anspruch auf sozialwissenschaftliche Relevanz textanalytischer Verfahren die Erwartung begründen, dass und inwiefern sich in Texten etwas über diesen textjenseitigen Erkenntnisgegenstand finden lässt.[3]

Textanalysen können sozialwissenschaftliche Relevanz nur unter der Voraussetzung beanspruchen, dass sie sich nicht in Aussagen über den jeweiligen Text erschöpfen, sondern einen Zugang zur sozialen Wirklichkeit bieten, ein Mittel

---

Bereich der am stärksten geisteswissenschaftlich geprägten Subdisziplin, der Politischen Theorie und Ideengeschichte, zu finden.

3 Damit soll nicht behauptet werden, dass in den Literaturwissenschaften bei der Analyse literarischer Werke soziale Kontexte und Wirkungszusammenhänge nicht reflektiert würden, sondern nur, dass diese es mit Texten als primären Erkenntnisobjekten zu tun haben und daher Aussagen über Texte keiner besonderen Rechtfertigung ihrer disziplinären Zugehörigkeit bedürfen. Umgekehrt kommen wiederum auch in den Sozialwissenschaften reine Textinterpretationen vor, beispielsweise in der politikwissenschaftlichen Subdisziplin der Politischen Theorie und Ideengeschichte, die sich dann über die subdisziplinäre Programmatik rechtfertigen: Wenn Machiavelli oder Hobbes politikwissenschaftlich relevante Autoren sind, lässt sich auch für eine Interpretation des *Principe* oder des *Leviathan* politikwissenschaftliche Relevanz beanspruchen; ob dies auch für textimmanente Interpretationen gilt, ist methodologisch umstritten: Es gibt gute Gründe, die fachdisziplinäre Relevanz und Validität von Textinterpretationen an die Berücksichtigung kontextualistischer Gesichtspunkte zu koppeln (vgl. Söllner et al. 2015).

zum Zweck der Erkenntnis von ‚Gesellschaft' bzw. Aspekten des Sozialen sind. Die Sozialwissenschaften können auf zwei theoretisch-methodologische Traditionen zurückgreifen, die das Verhältnis von Texten zur sozialen Wirklichkeit reflektieren und damit einen textanalytischen Zugang zu ihrem Erkenntnisgegenstand zu rechtfertigen geeignet sind: die das Verhältnis von Text und Kontext fokussierende hermeneutische Perspektive und die wissenssoziologische Perspektive mit ihrer Leitidee von der sozialen Bedingtheit des Wissens. Um einen Text angemessen zu verstehen, ist aus hermeneutischer Perspektive die „Reflexion auf die Verstehensbedingungen" (Gadamer 1997 [1968]: 50) erforderlich, also auf die Distanz zwischen den historischen und soziokulturellen Horizonten von Interpret und Interpretand, aus wissenssoziologischer Perspektive die Berücksichtigung der „Seinsverbundenheit" (Mannheim 1995: 227 ff.) des im Text objektivierten Denkens. Beide Perspektiven nehmen in unterschiedlicher Weise die Entstehungs- und Geltungskontexte von Texten in den Blick, die den in diesen artikulierten Sinn prägen. Die Vorstellung einer solchen Kontextbedingtheit von Texten ermöglich prinzipiell zwei Erkenntnisrichtungen: vom Kontext auf den Text, den es zu interpretieren gilt; und vom Text auf den Kontext, dessen Spuren der Text in erkennbarer Weise enthalten muss. Dieser zweiten Erkenntnisrichtung gemäß können Texte also als Manifestation oder Ausdruck des sie bedingenden sozialen Seins oder Kontexts gedeutet werden, wie auch immer Letztere jeweils theoretisch konzeptualisiert werden. Aus diesem Grund können sie prinzipiell als Medium für die den Sozialwissenschaften aufgetragene Erkenntnis der sozialen Wirklichkeit fungieren. Was dabei unter ‚sozialer Wirklichkeit' verstanden bzw. wie diese konzeptualisiert und was von ihr erkennbar wird, variiert mit den jeweiligen theoretischen Ansätzen innerhalb dieser beiden perspektivischen Traditionszusammenhänge.

Wir wollen im Folgenden zunächst zeigen, wie sich die Wissenssoziologie gegen das geisteswissenschaftliche Erbe der Hermeneutik formierte und damit einen spezifisch sozialwissenschaftlichen Erkenntnisanspruch und Zugang zur sozialen Wirklichkeit im analytischen Umgang mit Texten begründete, um dann darzulegen, wie diese beiden anfänglich sich konfrontativ zueinander verhaltenden perspektivischen Traditionen in Folge des *linguistic turn* im Hinblick auf ihre Konzeptualisierung des Verhältnisses von Text und sozialer Wirklichkeit konvergieren und infolgedessen sozialwissenschaftliche Textanalysen nicht nur prinzipiell rechtfertigen, sondern auch einen erheblichen textanalytischen Empiriebedarf entwickeln (2.).[4] Anschließend wollen wir reflektieren, wie die Anwendung und Leistungsfähigkeit von Text Mining in den *Digital Humanities* zur konstruktiven

4 Einige dieser Überlegungen sowie eine exemplarische empirische Anwendung von *blended reading* finden sich bereits in kürzerer Fassung in Lemke/Stulpe (2015).

Entfaltung kommen kann. Unter dem Begriff des ‚blended reading' schlagen wir eine Strategie im Sinne einer Best Practise vor, die semiautomatische Analyseverfahren mit klassischer Textlektüre so integriert, dass sozialwissenschaftliche Erkenntnispotenziale, die sich auf die Auswertung großer Textdatenmengen stützen, optimal ausgeschöpft werden. Gerade in einer bestimmten Kombination verschiedener Verfahren des *close* wie des *distant reading* liegt, so unsere Überlegung, der Schlüssel für eine produktive Verwendung von Text Mining in den Sozialwissenschaften (3.).

2 Text und soziale Wirklichkeit in hermeneutischer und wissenssoziologischer Perspektive vor und nach dem *linguistic turn*

Der Text und das Problem seines Verstehens sind der primäre Bezugspunkt der hermeneutischen Tradition in ihren historischen Variationen und Transformationen, von der richtigen Auslegung autoritativer Texte zum Zwecke der Wahrheitsfindung bis zur Interpretation literarischer Werke unter dem Aspekt auktorialer Intentionen und zeitgenössischer Kontexte. Innerhalb des hermeneutischen Dreiecks von Text, Intention und Kontext (vgl. Luhmann 1997: 889) ist es der Kontext-Begriff, der den analytischen Bezug auf die soziale Wirklichkeit ermöglicht. Denn der Aufschluss über einen Text versprechende Kontext kann seinerseits aus Texten bestehen, beispielsweise das Gesamtwerk eines Autors oder die Werke und Debattenbeiträge anderer Autoren, aber ebenso beispielsweise aus einem zeitgenössischen Ereignishorizont und kulturellen Hintergrundgewissheiten, die auf eine konkrete historisch-soziale Realität verweisen. Reflexiv gewendet, betrifft die Kontextfrage nicht nur den zu interpretierenden Text, sondern auch den Standort des Interpreten. Durch die Universalisierung hermeneutischer Reflexion auf das Problem des Verstehens und seiner Bedingungen schlechthin, historisch vollzogen in der idealistisch-geisteswissenschaftlichen Tradition von Schleiermacher bis Dilthey, tritt die historisch-soziale Welt insgesamt, verstanden als geistiger Sinn-Zusammenhang, in den Fokus hermeneutischer Betrachtung (vgl. Gadamer 1997 [1968]: 37 ff.; Frank 1995: 7 f.). Aus dieser Perspektive kann die Hermeneutik eine prinzipielle Zuständigkeit für die Analyse sozialer Wirklichkeit geltend machen: „Denn sofern Gesellschaft immer ein sprachlich verständigtes Dasein hat, ist das eigene Gegenstandsfeld der Sozialwissenschaften selber (und nicht nur ihre Theoriebildung) durch die hermeneutische Dimension beherrscht." (Gadamer 1997 [1968]: 54)

Die Anerkennung der von Gadamer angesprochenen hermeneutischen Dimension, die den Sozialwissenschaftler bei der Analyse seines Gegenstandes einerseits zur Reflexion der eigenen Perspektivität nötigt und ihn andererseits auf einen verstehenden Zugang zu seinem Objekt verpflichtet, gehört zur geisteswissenschaftlichen Erbmasse der Sozialwissenschaften. Hermeneutik als spezifisch geisteswissenschaftliches, im Unterschied zur naturwissenschaftlichen Kausalerklärung auf Verstehen und Sinndeutung angelegtes Verfahren kommt nicht nur typischerweise in den textzentrierten sozialwissenschaftlichen Subdisziplinen zur Anwendung, innerhalb der Politikwissenschaft etwa im Bereich der Politischen Theorie und Ideengeschichte. Das hermeneutische Leitmotiv findet sich beispielsweise auch, im sozialwissenschaftlichen Selbstverständnis ebenso prominent wie paradigmatisch, in Max Webers (1972 [1921]) Programm einer „Verstehenden Soziologie". Die Hermeneutik steht den Sozialwissenschaften als methodologisches und epistemologisches Konzept zur Verfügung, das es ihnen erlaubt, in systematischer Weise Erkenntnisse über die Bedeutung von Handlungen wie von Texten zu gewinnen.

Der klassische Weg, die Bedeutung eines Textes oder einer Handlung[5] als den subjektiv gemeinten Sinn bzw. die Intention des Autors bzw. Akteurs aus dem jeweiligen Kontext des Textes oder der Handlung zu erschließen, konstituiert dabei einen Verweisungszusammenhang zwischen Text (oder Handlung) und Kontext, der prinzipiell eine Umkehrung der Blickrichtung erlaubt. Für die Frage nach der sozialwissenschaftlichen Relevanz von Textanalysen für die Erkenntnis sozialer Wirklichkeit ist dies entscheidend. Wenn ein Text nur unter Berücksichtigung seines Kontextes adäquat verstanden werden kann, dann muss er gewissermaßen Spuren dieses Kontextes enthalten, die es erlauben, beide, Text und Kontext, aufeinander zu beziehen: beispielsweise Anspielungen auf zeitgenössische Ereignisse oder bestimmte Namen oder Begriffe, von denen der Autor des Textes unterstellen konnte, dass sie seinem Publikum bekannt sind. In diesem Sinne ist der Text durch seinen Kontext geprägt und erscheint als dessen Ausdruck. Für den Beobachter bedeutet dies, dass dem Text ein Informationswert bezüglich seines Kontextes zukommt. In diesem Sinne kehrt sich die Blickrichtung um, als Schluss vom Text auf dem Kontext.

Auf dieser allgemeinen Ebene bleibt bezüglich der Frage nach dem Verhältnis von Text und sozialer Wirklichkeit noch offen, inwiefern und in welchem Ausmaß der Kontext als soziale Wirklichkeit zu konzeptualisieren ist, welche Reichweite

5 Die unter dem Topos des *linguistic turn* (siehe unten, 2.2) angesprochenen theoretischen Entwicklungen ermöglichen es dann, Texte selbst als Handlungen, nämlich als Sprechakte (Austin 1962), zu konzeptualisieren.

die Textanalyse für die Erkenntnis einer solchen sozialen Wirklichkeit haben kann und welche Rolle hierbei die Quantität (und möglicherweise auch Qualität) der hierfür heranzuziehenden Texte spielt. Grundsätzlich festzuhalten ist aber, dass die – disziplingeschichtlich in die Sozialwissenschaften eingegangene – genuin geisteswissenschaftliche Perspektive der Hermeneutik neben ihrer traditionellen textexegetischen und verstehend-soziologischen Anwendung auch einen analytischen Umgang mit Texten ermöglicht, der durch seine kontexterschließende Blickrichtung sozialwissenschaftliche Erkenntnis der sozialen Wirklichkeit verspricht. Um diese Überlegung weiter zu konkretisieren, ist eine der geisteswissenschaftlich-hermeneutischen Tradition zunächst entgegengesetzte, originär sozialwissenschaftliche Perspektive in den Blick zu nehmen: die Wissenssoziologie.

2.1 Inkongruente Perspektiven: Distanzierungen der Wissenssoziologie

Die wissenssoziologische Tradition beginnt im Ausgang von der sozialen Wirklichkeit und mit dem Nachweis, dass das ‚Wissen', im umfassenden Sinne von Ideen, Theorien und anderen Realitätsbeschreibungen, normativen Vorstellungen und Selbstverständnissen, durch diese Wirklichkeit geprägt ist. Soweit dieses Wissen sprachlich artikuliert wird, liegt es primär in Form von Texten vor, und insofern bestimmt aus wissenssoziologischer Perspektive das Erkenntnisinteresse an Weltdeutungen und ihrer sozialen Bedingtheit das analytische Interesse an Texten. Während in der Hermeneutik die Kontextualisierung des Textes die soziale Wirklichkeit in den Blick bringt, bringt in der Wissenssoziologie die Relationierung von Wissen und sozialer Wirklichkeit den Text in den Blick.

Neben der von den Geistwissenschaften geerbten bzw. entlehnten Hermeneutik verfügen die Sozialwissenschaften also mit der Wissenssoziologie über eine eigene, genuin sozialwissenschaftliche Tradition des analytischen Umgangs mit Texten und der Reflexion der Perspektivität und Kontextgebundenheit von Weltdeutungen. Diese beginnt aus Sicht der klassischen Wissenssoziologie Karl Mannheims (1995: 266) mit der soziologischen Thematisierung der ‚Seinsverbundenheit des Denkens' beziehungsweise des ‚Wissens' im Historischen Materialismus von Marx und Engels. Einen weiteren entscheidenden Ausgangspunkt hat sie für Mannheim in Nietzsches, vor allem in dessen *Genealogie der Moral* entfalteter, perspektivistischer Rückführung von Idealen, normativen Vorstellungen und Selbstverständnissen auf die psychophysiologische Dynamik von Triebentladungen und -umleitungen in den Überwältigungs- und Ermächtigungskämpfen typologisch unterschiedener Menschengruppen (vgl. Nietzsche 1993: v. a. 262 ff., 309 ff., 361 ff.). Was Nietzsche