18 Bamberger Beiträge zur Soziologie

Methode und Methodologie quantitativer Textanalyse

Jan Rasmus Riebling

University of Bamberg Press

Bamberger Beiträge zur Soziologie

Band 18

University
of Bamberg
Press
2019

Methode und Methodologie quantitativer Textanalyse

von Jan Rasmus Riebling

University
of Bamberg
Press
2019

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Informationen sind im Internet über http://dnb.d-nb.de/ abrufbar.

Diese Arbeit hat der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg als Dissertation vorgelegen.
1. Gutachter: Prof. Dr. Richard Münch,
2. Gutachter: Prof. Dr. Gerhard Schulz
Tag der mündlichen Prüfung: 20.07.2017

ISSN: 1867-8416
ISBN: 978-3-86309-640-3 (Druckausgabe)
eISBN: 978-3-86309-641-0 (Online-Ausgabe)
URN: urn:nbn:de:bvb:473-opus4-540132
DOI: http://dx.doi.org/10.20378/irbo-54013

für Anna und Elfriede

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Liste mathematischer Symbole

$\sum$	Summenzeichen;
$\prod$	Produktzeichen;
$\log$	Logarithmus;
$\ln$	Natürlicher Logarithmus zur Basis e;
e	Eulersche Zahl;
$\propto$	Proportionalität;
$\cap$	Schnittmenge;
$\cup$	Vereinigungsmenge;
$\setminus$	Differenzmenge;
$\triangle$	Symmetrische Differenz;
$P(A)$	Wahrscheinlichkeit für Ereignis A;
$P(A \cap B)$	Wahrscheinlichkeit für Ereignisse A und B;
$P(A \mid B)$	Wahrscheinlichkeit für Ereignis A gegeben B;
$\binom{n}{k}$	n über k;
χ^2	Chi-Quadrat Verteilung;
$B(k; p; n)$	Binomialverteilung für k Resultate bei n Versuchen mit Wahrscheinlichkeit p;
$L(p; k; n)$	Likelihoodfunktion;
$\mathrm{Hypergeom}(x, r, n, N)$	Hypergeometrische Verteilung;
$Dir(\eta)$	Dirichletverteilung;
$\mathcal{L}$	Eine bestimmte Sprache;

Σ	Ein endliches Alphabet; Im Kontext der SVD: die Diagonalmatrix der Singulärwerte;
w	Ein Wort (Type) für das gilt $w \in \Sigma$;
W_T	Anzahl der Worte (Typen) eines Textes T;
n_T	Anzahl der Token im Text T;
c_T	Anzahl der Zeichen im Text T;
s_T	Anzahl der Sätze im Text T;
$\bar{s}_T$	Durchschnittliche Satzlänge im Text T;
syl_T	Anzahl der Silben im Text T;
$\mathrm{polysyl}_T$	Anzahl der Worte mit drei oder mehr Silben (Mehrsilber) im Text T;
$\cos(\theta)$	Kosinus am Winkel θ;
$\vec{a}$	Vektor;
$J(\vec{a}, \vec{b})$	Jaccard-Ähnlichkeit zweier Vektoren;
$\vec{e}$	Eigenvektor;
λ	Eigenwert;
V^H	Adjungierte (hermitesch transponierte) Matrix von V;
Σ^{-1}	Invertierte Matrix von Σ;
$\hat{M}$	Approximation von M;
$\mathcal{G}(\mathscr{N}, \mathscr{L})$	Ein Graph mit $\mathscr{N}$ Knoten und $\mathscr{L}$ Kantenzügen;

Abkürzungsverzeichnis

CRA	Centering Resonance Analysis;
HDP	Hierarchical Dirichlet Process;
LDA	Latent Dirichlet Allocation;
LSA	Latente Semantikanalyse, Latent Semantic Analysis;
NTA	Netzwerk-Text-Analyse, Network-Text Analysis;
NP	Noun phrase;
POS	Part-of-speech;
SNA	Soziale Netzwerkanalyse, Social Network Analysis;
SVD	Singular Value Decomposition;
TFiDF	Term Frequency inverse Document Frequency;
VSM	Vector space model;

Danksagung

Ohne die persönliche und fachliche Unterstützung einer Vielzahl von Menschen wäre diese Arbeit niemals möglich gewesen. Ihnen allen zu danken würde ein Buch ebensolchen Ausmaßes erfordern. Mein besonderer Dank gebührt der großen Anzahl all jener Kollegen und Studenten der Universität Bamberg, die nie um eine Diskussion und eine kritische Bemerkung verlegen waren. Herausragender Dank gebührt Richard Münch, Gerhard Schulze und Kai Fischbach, für ihre Unterstützung, ihr Vertrauen und nicht zuletzt die Bereitschaft diese Arbeit in ihrer Gänze zu lesen.

Einer Reihe von Personen müssen besonders hervorgehoben werden. Was immer man an dieser Arbeit gut finden mag, es ist ihnen geschuldet. Daher danke ich Andreas, für all die Dinge zwischen erotischem Kapital und Zombie-Apokalypsen; Christian, für Unbestimmbarkeit, insbesondere die von Vorzeichen; Stephanie, für ihre Wut und ihren Mut; Oliver, für die Grenzen von Text und deren Überschreitung; Anna, für besseres Wissen, Hurenkinder und Schusterjungen; Brigitte, für Imperial Mints and practical hints; Mehmo, für die kleinen und großen Freiräume; Chetan, Katja, Julia und Shrenika, für ein Zuhause. Vor allem aber Heiko, für π und alles was dazu gehört.

Abschließend möchte ich auch meiner Familie danken, die mir all dies erst ermöglicht hat. Insbesondere meinem Vater, der mir die Wissenschaften nahe brachte und meiner Mutter, der ich die Liebe zur Literatur verdanke, leider aber nicht die zur Rechtschreibung.

Zusammenfassung

Ziel dieser Arbeit ist die Entwicklung einer sozialwissenschaftlichen Methodologie zur Analyse symbolischer Ordnungen. Darunter werden hier die Strukturen und Systeme von sozial standardisierten Zeichen (Symbolen) verstanden, die sich im Prozess der Kommunikation bilden. Im Vergleich mit soziologischen Theorien und praktischen Verfahren aus dem Bereich der Linguistik und dem Natural Language Processing, werden Nutzen und Anwendungsbereich einer solchen Methodologie diskutiert. Auf dieser Basis werden die notwendigen Voraussetzungen, Möglichkeiten und Problemstellungen der quantitativen Textanalyse eingehend erörtert. Das praktische Vorgehen wird anhand der Darstellung der konkreten Analyseschritte, unter Verwendung der Programmiersprache Python, von den rohen Textdaten bis hin zur Modellierung von Bedeutungen und Themen illustriert und kritisch diskutiert. Ein besonderes Augenmerk gilt dabei der Erläuterung der unterschiedlichen Verfahren zur Modellierung des semantischen Gehalts, insbesondere dem Topic Modelling, der Netzwerk-Text-Analyse und dem maschinellen Lernen.

1 Einleitung

Begriffe wie Kultur, Wissen und Kommunikation nehmen eine zentrale Stellung als Gegenstand soziologischer Analysen ein und verweisen damit auf die Bedeutung der sprachlichen Ebene sozialer Phänomene. Auch auf methodischer Ebene ist die Sprache ein unersetzliches Werkzeug der Soziologie. Von den Fragebögen der quantitativen Soziologie hin zu den Transkripten der qualitativen Sozialforschung stellen sprachliche Zeichen den überwiegenden Großteil soziologischer Daten. Wenngleich wir diese oft als Repräsentationen innerer Zustände auffassen, so werden sie doch mittels sozial standardisierter, sprachlicher Zeichen festgehalten und durch diese interpretiert, wie man insbesondere am Beispiel der Codierungsverfahren in der qualitativen Forschung sieht. Somit scheint Sprache aufs Engste mit den sozialen Prozessen, die hier von Interesse sind, verknüpft. Gleichzeitig wird das Fundament der Sprache und der Zeichen selten explizit in den Forschungsprozess und in die Theoriebildung integriert. Von vereinzelten Ausnahmen, insbesondere der foucaultschen Diskursanalyse (vgl. Foucault 1981) und der Symboltheorie von Norbert Elias (1991), einmal abgesehen, wird die sprachliche Vermittlung sozialer Prozesse meist als gegeben hingenommen. Wenngleich Phänomene wie Bedeutung, Sinn, Kultur und dergleichen unweigerlich eine sprachliche Basis haben, so werden sie in der Soziologie oft nicht auf diese Art und Weise untersucht.

Ob und wie sich die Besonderheiten sprachlicher Zeichen auf soziale Prozesse auswirken, oder wie sie sinnvollerweise erforscht werden könnten sind Fragen, die in den meisten soziologischen Theorien nicht adäquat formuliert werden können. Dies ist umso bedauerlicher, da eine Vielzahl von Disziplinen, wie z.B. die Semiotik, Informationstheorie und Linguistik, eine breite Wissensbasis geschaffen haben, die interdisziplinär zugänglich ist und eine Vielzahl neuer Impulse für die soziologische Theoriebildung beinhalten könnte. Letztlich werden dadurch enorme Möglichkeiten zur Reflexion und zur Erweiterung bestehender Theorien verschenkt.

In methodischer Hinsicht stellt sich das Problem, wie sprachliche Äußerungen sinnvoll untersucht werden können. Während die quantitative Sozialforschung solche Herausforderungen bisher nur im Rahmen des Fragebogendesigns diskutiert, scheint die qualitative Methodenlehre an diesem Punkt schon weiter zu sein. Dies ist auch darauf zurückzuführen, dass die Interpretation von Texten zentraler Bestandteil der qualitativen Forschung ist. Naturgemäß sind die Textmengen, die auf diese Art und Weise untersucht werden können, begrenzt. Des Weiteren gehen qualitative Verfahren meist von kognitiven Modellen der Interpretation aus und berücksichtigen daher die genuin sozialen Eigenschaften von sprachlichen Zeichen nur selten.

Demgegenüber haben computerlinguistische Verfahren und Methoden die Analyse großer Textmengen nicht nur möglich, sondern zu einem alltäglichen Phänomen werden lassen. Egal ob man das Internet mittels Suchmaschinen durchsucht, sich in sozialen Online-Netzwerken bewegt oder in Online-Kaufhäusern einkauft, es sind stets Algorithmen aus dem Bereich des *Topic Modelling* und des *Semantic Mapping*, die dafür sorgen, dass wir nicht das finden, was wir, z.B. im Eingabefeld einer Suchmaschine, geschrieben haben, sondern das was wir damit höchstwahrscheinlich meinten. Ein Großteil des menschlichen Wissens wird von eben diesen Algorithmen geordnet, kartiert und verwaltet. Die methodischen Entwicklungen in diesem Bereich sind im Rahmen der Soziologie jedoch bisher kaum genutzt worden. Von einigen nennenswerten Ausnahmen (vgl. Leydesdorff und Welbers 2011; DiMaggio, Nag und Blei 2013) einmal abgesehen sind es hauptsächlich die Politikwissenschaften (vgl. Lemke und Wiedemann 2016), der relativ junge, interdisziplinäre Bereich der Computational Social Science (vgl. Heiberger und Riebling 2016) sowie die Vertreter der Netzwerkanalyse (vgl. Carley 1997), die sich den Möglichkeiten der quantitativen Analyse von Texten geöffnet haben.

Die digitale Revolution problematisiert die soziologische Vernachlässigung von quantitativen Textanalysen darüber hinaus auch in forschungspragmatischer Hinsicht. Die weltweite *Zunahme von Online-Daten* ist nicht zuletzt auch eine Zunahme von Texten, Dokumenten und anderen verschriftlichten Sprachzeugnissen. Seien es ausführliche Diskussionen in Internet-Foren, soziale Schlammschlachten in Microblogging-Diensten oder die allgemeine mediale Berichterstattung, immer handelt es sich um soziale Prozesse, die durch sprachliche Zeichen vermittelt und als solche gespeichert werden. Dies ermöglicht es fast schon sozialen Prozessen in ihrem Ablauf zuzusehen. Ohne Zweifel handelt es sich bei den im Zu-

ge der Digitalisierung aufgehäuften Datenbergen um einen soziologisch bisher nicht gehobenen Schatz.

Ziel dieser Arbeit ist es diesen Herausforderung der soziologischen Analyse von Texten durch die Entwicklung einer *Methodologie sozialer Symbole* zu begegnen. Unter Methodologie wird dabei eine allgemeine Anleitung zur Modellierung eines bestimmten Phänomens verstanden, welches sich in eine Vielzahl konkreter Verfahren übersetzen lässt. Eine solche Methodologie ist notwendig um den bestehenden Methodenkorpus, der in anderen Forschungsbereichen zur Untersuchung sprachlicher Zeichen und der damit verbundenen Phänomene entwickelt wurde, in die soziologische Forschung und Theorietradition integrieren zu können. Zu diesem Zweck werden zwei theoretische Konzepte aus der vorhandenen Literatur abgeleitet. Zeichen, die durch Prozesse der sozialen Standardisierung erzeugt werden, die potentiell Bedeutungen kodieren können und die als physikalische Objekte existieren, werden als *Symbole* aufgefasst. Die einer bestimmten Struktur folgende, regelmäßige und relativ dauerhafte Verknüpfung dieser Symbole wird als *symbolische Ordnung* bezeichnet. Vor dem Hintergrund dieser theoretischen Konzeption geht es um die methodisch-praktische Frage, wie symbolische Ordnungen als sprachlich vermittelte, soziale Phänomene modelliert und analysiert werden können.

Die resultierende Methodologie soll dabei keinen Ersatz für unterschiedliche, theoretische Perspektiven bereitstellen, sondern die Möglichkeit bieten eine Vielzahl von Konzepten, welche auf Sprache beruhen zu operationalisieren und zu messen. Dabei wird, entgegen der Tradition, nicht von einem Primat der Theorien ausgehend, sondern von den methodischen und praktischen Möglichkeiten her kommend nach passenden Leerstellen in der soziologischen Theorie gesucht. Dieses Vorgehen erscheint sinnvoll, da einer Vielzahl von ausgearbeiteten Methoden und Verfahren eine gewisse Armut der soziologischen Theoriebildung in diesem Bereich gegenübersteht.

Dieses Buch lässt sich grob in zwei Bereiche untergliedern. Im ersten Teil der Arbeit soll eine methodologische Grundlage entwickelt werden. Dies wird wiederum in drei Kapiteln geschehen. Im ersten Kapitel geht es um einen Überblick der relevanten Definitionen aus den Bereichen der Semiotik, Linguistik und Informationswissenschaft. Im Mittelpunkt stehen dabei die Begrifflichkeiten von Zeichen, Symbol und Signal. Daran anschließend wird es um die Frage gehen, ob Symbole und deren Aggregate (Kultur, Wissen, etc.) in Form von symbolischen Ordnungen in den theoretischen Rahmen der Soziologie eingefügt werden können. Das

dritte Kapitel setzt sich mit den bestehenden, qualitativen Forschungsprogrammen zu Symbolen auseinander und versucht Gemeinsamkeiten und Unterschiede herauszuarbeiten.

Um der Frage nachzugehen, wie eine Methodologie symbolisch vermittelter, sozialer Phänomene aussehen könnte ist es zunächst notwendig sich mit den Formulierungen derjenigen Disziplinen zu befassen für die Zeichen und Sprache, den zentralen Gegenstandsbereich stellen. Die Semiotik, die allgemeine Sprachphilosophie, die Linguistik und die Informationstheorie haben in ihrer Beschäftigung mit diesen Phänomenen nicht nur eine allgemeine Wissensbasis geschaffen, sondern auch den Grundstein für die darauf aufbauenden Methoden geliefert. Eine Betrachtung dieser Grundlagen enthüllt ein fundamentales Dilemma, nämlich die Frage nach der Beziehung von Zeichen und dem was sie repräsentieren. Da Zeichen zwischen der äußeren Welt und gedanklichen Vorgängen vermitteln, steht die Frage im Raum, ob eine dieser beiden Seiten dominiert. Wenn Zeichen *Ideen* darstellen, dann kann weder erklärt werden, woher diese Ideen kommen, noch warum sie in irgendeiner Weise zu empirischen Beobachtungen passen sollten. Schlägt man sie hingegen der *Welt* zu, so ist es nur schwer begreiflich, warum der überwiegende Großteil von Zeichen keine Dinge der äußeren Welt repräsentiert, sondern Abstraktionen und Fiktionen.

Es findet sich jedoch eine dritte Position, die von einer *selbstreferenziellen Repräsentation* von Zeichen ausgeht. Demnach repräsentieren Zeichen andere Zeichen und erzeugen somit Bedeutungen als relativ dauerhafte Verknüpfungen, die sowohl auf geistige Zustände, als auch auf empirische Gegebenheiten verweisen können. Die konkreten symbolischen Ordnungen, die aus solchen sozial standardisierten Zeichenkomplexen bestehen, können dementsprechend und je nach Auflösungsgrad als Bedeutung, Wissen, Kultur oder Diskurs gedeutet werden. Ein praktischer Vorteil dieser Definition besteht darin, dass sie es ermöglicht die Eigengesetzlichkeit von Zeichensystemen in den Blick zu nehmen.

Die daran anschließende Betrachtung der soziologischen Perspektive enthüllt eine strukturell sehr ähnliche Spaltung. In der Soziologie verläuft die Frage nach der Stellung von Symbolen im sozialen Gefüge oft entlang der zentralen Unterscheidung von Individuum und Gesellschaft. Demzufolge sind soziale Symbole entweder Produkt und Gegenstand von individuellen Entscheidungen oder sie sind extern gegebene, kulturelle Normen und Imperative. Anschließend an die von Norbert Elias (1991) in *The Symbol Theory* angestellten Überlegungen wird eine Synthese der individualis-

tischen und der gesellschaftlichen Perspektive im Sinne einer Prozesssoziologie der Symbole vorgeschlagen. Grundlegend ist dabei die Idee, dass die gesellschaftliche Ebene den Rahmen einer symbolischen Ordnung bereitstellt, während auf der individuellen Ebene die Symbole eingesetzt und in Interaktionen aktualisiert werden. Das Resultat sind kommunikative Prozesse, die zwar eine gewisse Unschärfe aufweisen, sich aber im Generellen nach erwartbaren Regeln und Mustern vollziehen.

Im dritten Kapitel wird das bisherige, qualitative Programm zur Analyse von Kultur und Wissen sowie anderen symbolischen Phänomenen, eingehender untersucht. Eine solche Betrachtung macht sowohl die Bedeutung der bestehenden Methoden als auch deren relative Voreingenommenheit, gegen eine Anreicherung dieses Methodenspektrums mittels quantitativer Verfahren deutlich. Daher widmet sich dieses Kapitel insbesondere der Entkräftung solcher methodischer Vorbehalte und damit einhergehend dem Versuch das grundlegende Potenzial eines quantitativen Vorgehens im Bereich der Analyse von sprachlich vermittelten, sozialen Prozessen aufzuzeigen. Dies lässt sich auf vier Ebenen konkretisieren. Erstens, ermöglichen quantitative Verfahren eine Operationalisierung von symbolischen Prozessen, wie Kultur, Wissen und Diskurs, als objektive und eigengesetzliche Phänomene. Zweitens, sind die bestehenden Verfahren, wie sie insbesondere im Bereich der Computerlinguistik und des Topic Modellings entwickelt wurden, für die Analyse sehr großer Datenmengen ausgelegt. In Zeiten stetig wachsender Datenmengen und der zunehmenden Digitalisierung von Kommunikation, eröffnet dies eine Vielzahl von neuen Möglichkeiten für die soziologische Forschung. Drittens, erlaubt ein quantifizierendes Vorgehen ein sehr viel höheres Ausmaß an Formalisierung und Nachprüfbarkeit der Analysen. Viertens, erlaubt es der Einsatz quantitativer Methoden Anschluss an den methodischen und inhaltlichen Diskurs anderer Disziplinen, wie der Informatik oder der Linguistik, zu finden.

Die rein abstrakte Diskussion der methodologischen Prinzipien wäre aber keinesfalls zielführend. Eine Methodologie muss sich immer auch in ihrer praktischen Umsetzung bewähren. Daher ist der zweite Teil dieses Buches den konkreten Herausforderungen und Verfahren der quantitativen Textanalyse gewidmet. Dabei wird besonderes Augenmerk auf die Erläuterung praktischer Gegebenheiten gelegt, um die Besonderheiten von Textdaten und deren Analyse herauszuarbeiten.

Das erste Methodenkapitel geht der Frage nach, wie soziale Symbole operationalisiert und als sprachbasierte Daten gehandhabt werden kön-

nen. Des Weiteren wird auch der Einsatz lexikalischer Ressourcen diskutiert sowie Analysemöglichkeiten vorgestellt, die bereits auf der Ebene von einzelnen Symbolen und deren Verteilung in Texten eingesetzt werden können. Grundlegend für die Operationalisierung ist die Zerlegung von Texten in ihre einzelnen symbolischen Einheiten, welche als „Token" bezeichnet werden. Dieser Prozess der Tokenisierung erfordert vom Forscher die Festlegung, was mit Hinblick auf das jeweilige Forschungsinteresse als Token konstruiert wird (z.B.: einzelne Worte, Sätze, feste Phrasen, n-Gramme, etc.).

In praktischer Hinsicht ist hierzu das Modell einer Meta-Sprache notwendig, wie zum Beispiel das der regulären Grammatik, welche die Formulierung von Tokenisierungsregeln für beliebige Zeichenfolgen ermöglicht. Darauf aufbauend wird die Identifikation von maßgeblichen Eigenschaften der Symbole und der aus Ihnen zusammengesetzten Texte behandelt. Dazu zählt insbesondere die Bestimmung von Häufigkeiten, grammatikalischen Formen und n-Grammen (feste Verknüpfungen von n Token). Dies ist wiederum die Grundlage für weitere Verfahren der Datenaufbereitung, insbesondere des Stemmings (Rückführung auf grammatikalische Grundformen) und der Zusammenfassung von n-Grammen zu einem Token. Darauf aufbauend werden Techniken der Analyse von Häufigkeitsverteilungen und anderen textuellen Merkmalen behandelt, wie die Komplexität eines Textes, die Differenz zur „Umgangssprache" und die Identifikation mit spezifischen Symbolen. Grundlegend für das gesamte Kapitel ist außerdem die Diskussion der besonderen Anforderungen, die Textdaten aufgrund ihrer besonderen statistischen und informationstheoretischen Eigenschaften mit sich bringen.

Im zweiten Methodenkapitel werden Verfahren zur Analyse der übergreifenden symbolischen Ordnung vorgestellt. Das heißt es geht um die Bestimmung der Bedeutungen, die in der spezifischen, regelmäßigen Anordnung von Symbolen kodiert sind. Entsprechend der hier entwickelten Methodologie kann die Struktur symbolischer Ordnungen sehr unterschiedlich modelliert werden. Ähnlichkeiten zwischen Texten können sowohl als direkte Beziehungen zwischen Symbolen oder Texten aufgefasst werden, wie auch als Vektorräume oder latente Dimensionen. Auch eine Kombination beider Sichtweisen ist möglich.

Fünf unterschiedliche Ansätze werden dabei genauer betrachtet. Zum einen die Vergabe von Kategorien, welche Symbole oder Symbolgruppen direkt eine spezifische Bedeutung zuweisen. Dabei kommen Techniken des maschinellen Lernens zum Einsatz, welche Klassifikationsheuristiken

erstellen können, die anhand von bekannten Eigenschaften mittels statistischer Verfahren „trainiert“ werden. Zweitens werden die Vektorraummodelle des Information Retrievals diskutiert, die eine Analyse der Ähnlichkeit von Texten aufgrund deren Anordnung in einem multidimensionalen Raum ermöglichen. Als drittes Verfahren wird die allgemeine Logik von Netzwerk-Text-Analysen diskutiert. Dabei werden Texte nach bestimmten Regeln in eine Netzwerkstruktur überführt, um sie daran anschließend mit den etablierten Verfahren der Netzwerkanalyse zu untersuchen. Den vierten Ansatz stellen dimensionsanalytische Verfahren dar. Diese gehen ebenfalls von den Wortverteilungen in textuellen Einheiten aus, interpretieren sie jedoch als manifeste Ausprägungen von latenten Dimensionen in denen Bedeutungen kodiert werden. Dieser Bereich wird oft auch als Topic Modelling bezeichnet und umfasst eine Vielzahl von Verfahren, von denen zwei ausführlich diskutiert werden: die Latent Semantic Analysis (LSA) und die Latent Dirichlet Allocation (LDA). Die letzte Gruppe der Methoden besteht in der Kombination von dimensions- und netzwerkanalytischen Fragestellungen. Diese relativ neuen Verfahren werden unter dem Sammelbegriff der sozio-semantischen Ansätze gefasst.

2 Linguistik und Semiotik

Um zu einer soziologischen Perspektive auf Symbole zu gelangen ist es ratsam sich zunächst einmal mit dem weiteren Gebrauchs dieses Begriffes im Rahmen anderer Wissenschaft auseinanderzusetzen. Nur so kann das Spezifische an einer soziologischen Betrachtungsweise herausgearbeitet werden. Gleichzeitig ermöglicht eine solche Auseinandersetzung einen Rückgriff auf bereits bestehende Erkenntnisse, Methoden und Problemlösungen die in anderen Disziplinen erarbeitet wurden. Es ist das explizite Ziel dieser Arbeit das Rad nicht neu zu erfinden, sondern die Frage zu stellen, wie wir die bereits bestehende Technologie für die spezifischen Zwecke unseres Fahrzeugs nützlich machen können. Demzufolge sollten alle Festlegungen, die im Folgenden hinsichtlich des ontologischen Status oder der Definition von Symbolen getroffen werden, immer nur als Zugeständnisse an den soziologischen Forschungsgegenstand verstanden werden. *Symbole und symbolische Ordnungen* erlangen in dieser Hinsicht nur Relevanz insofern sie als soziale Phänomene an sich oder mit Bezug auf eben diese aufgefasst werden können. Insofern berührt diese Arbeit nicht die Falschheit oder Richtigkeit von Aussagen die über Symbole in einem anderen Kontext getroffen wurden.

Dieser Warnhinweis ist auch deswegen zu beherzigen, weil die wissenschaftliche Beschäftigung mit Symbolen ein weitverzweigtes und Disziplinen übergreifendes Netzwerk von Theorien und Begrifflichkeiten darstellt. Aufgrund unterschiedlichster Fragestellungen und disziplinärer Traditionen ist hier kein Konsens zu erwarten. Um dennoch einen prinzipiellen Überblick zu gewinnen, ist es sinnvoll sich an grundlegenden Begrifflichkeiten zu orientieren und deren unterschiedlichen Verwendungen zu diskutieren. Bereits eine oberflächliche Betrachtung zeigt, dass der Begriff des Symbols in unterschiedlichem Ausmaße mit zwei anderen zentralen Ausdrücken zusammenfällt, nämlich dem *Zeichen* und dem *Signal*. Der Gebrauch aller drei Bezeichnungen schwankt je nach theoretischem Rahmen und Präferenz des Autors zwischen einer Verwendung als Synonyme bis hin zu vollkommen unterschiedlichen Kategorien.

2.1 Zeichen

Der Begriff des Zeichens (*sign*) findet sich vor allem in der Sprachphilosophie und der Semiotik, für die er den zentralen Forschungsgegenstand darstellt. Auch hier herrscht jedoch alles andere als Einigkeit hinsichtlich der begrifflichen Definition. Trotz dieses Pluralismus der Zeichentheorien lässt sich doch so etwas wie ein kleinster gemeinsamer Nenner erkennen. Winfried Nöth (1985: 87) verweist in diesem Zusammenhang auf die „Minimalbedingung *aller* Zeichentheorien", nämlich das Vorhandensein einer „Zeichenrelation". Gemeint ist damit, dass das Zeichen auf etwas anderes verweist, also eine Beziehung zwischen zwei wie auch immer gearteten Objekten besteht. Dieser Verweis wird auch oft als die Bedeutung des Zeichens angesehen. Die eigentlichen Probleme der Begriffsbestimmung beginnen jenseits dieses minimalem Grundkonsenses. Fragen nach der Art der in Beziehung gesetzten Objekte und wie man sich die Relation vorzustellen hat, bzw. ob es sich um eine dyadische oder triadische Beziehung handelt, werden im Rahmen der einzelnen Denkschulen sehr unterschiedlich beantwortet.

Die Semiotik gründet sich im Besonderen auf die von Charles Saunders Peirce entwickelte Zeichentheorie.[1] Peirce war hauptsächlich an Fragestellungen der formalen Logik und damit einhergehend an einer möglichst exakten Definition der verwendeten Zeichen interessiert. Seiner Auffassung nach muss das Phänomen Zeichen als eine triadische Relation begriffen werden (vgl. Peirce 1931–0035: § 2.274). Die drei Teile sind das Zeichen selbst, welches auch als *Representamen* bezeichnet wird, das *Objekt* auf welches das Zeichen verweist und letztlich der *Interpretant*. In Peirce's eigenen Worten:

> The sign or *representamen* is something which stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the *interpretant* of the first sign. The sign stands for something, its *object*. It stands for the object, not in all re-

[1] Da ein Großteil des Peirces'schen Werkes nur als posthume Gesamtausgabe veröffentlicht wurde, sind seine Schriften zur Zeichentheorie stark fragmentarisch und bilden kein kohärentes Gesamtwerk. Ein Überblick zu diesen Schwierigkeiten sowie ein ausführliches Verzeichnis der Literatur, die sich mit der Rekonstruktion beschäftigt, findet sich bei Nöth (1985: 35).

> spects, but in reference to a sort of idea, which I have sometimes called the ground of the representamen. (ebd.: § 2.228)

Das prozesshafte Zusammenspiel dieser drei Elemente nannte Peirce Semiose (*semiosis*).

Als Objekt kann man dabei die Gesamtheit des Dinges verstehen, welches mittels des Zeichens dargestellt wird, also sozusagen dessen *Entfaltung als Konzept*. So steht zum Beispiel das geschriebene Zeichen „Pferd" entweder für einen bestimmten Vertreter oder die gesamte Gattung „Equus". In diesem Beispiel werden bereits eine Reihe von begrifflichen Problemen deutlich. Zum einen kann ein und dasselbe Zeichen sehr unterschiedliche Objekte repräsentieren, nämlich ein konkretes Pferd oder eben eine ganze Gattung. Zeichen sind demzufolge nicht immer eindeutig in ihrem Objektbezug.

Zum anderen zeigt sich hier auch warum Peirce das Zeichen als eine Referenz zu einer abstrakten Idee auffasste. Ein spezifisches Pferd kann der sinnlichen Erfahrung zugänglich gemacht werden, zum Beispiel indem man darauf zeigt. Im Falle eines abstrakten Konzeptes ist dies nicht mehr möglich. Die Gattung „Pferd" kann man nur bezeichnen, aber nicht zeigen. Das gleiche gilt für fiktionale Dinge wie zum Beispiel „Einhörner", welche ein sehr „reales", sprich häufig genutztes, didaktisches Hilfsmittel in der Sprachphilosophie darstellen. Dies führt jedoch zum Problem der Bestimmung des ontologischen Status der Peirces'schen Objekte, genauer gesagt zu der Frage, ob sie nicht alle fiktiv wären, was Peirce (ebd.: § 8.314) selbst schon angedeutet hatte. Demzufolge wäre das Objekt eines Zeichens besser als die „Idee vom Objekt dieses Zeichens" beziehungsweise als dessen Bedeutung umschrieben.

Aus dem Gleichsetzen des Interpretant mit dem Objekt eines Zeichens ergibt sich jedoch ein weiteres Problem. Unter dem Interpretant verstand Peirce die Folge der Interpretation des Representamen im Interpreten (vgl. ebd.: § 8.315). Ausschlaggebend ist hierbei die Vorstellung, dass sowohl Denken als auch Interpretieren als Zeichenprozesse aufgefasst werden, weshalb sich das äußere Zeichen nur in einem (oder mehreren) inneren Zeichen niederschlagen kann. Dies beinhaltet die Konsequenz, dass der Interpretant damit zum Representamen eines neuen Zeichenprozesses (einer neuen Semiose) wird. Wenn aber jeder gedankliche Vorgang aus Zeichen besteht und das Objekt eines Zeichens dessen Idee ist, kann nicht mehr klar zwischen dem Objekt und dem Interpretant unterschieden werden.

Dieses *semiotische Dilemma*[2] kann als das grundlegende Problem des Zeichenbegriffs angesehen werden. Wenn die Zeichen von den Objekten her kommen, also aus deren sinnlicher Erfahrung, dann kann es nur Zeichen geben, die ein korrespondierendes Objekt aufweisen. Damit wären viele, wenn nicht sogar fast alle Zeichen als relativ sinnlos anzusehen, denn ganz offensichtlich enthalten menschliche Sprachen eine Vielzahl von Zeichen, die wegen ihres Abstraktionsgehalts nicht direkt der sinnlichen Erfahrung zugänglich sind. Das Gleiche gilt für Zeichen, die eine rein grammatikalische Funktion haben. Setzt man das Primat hingegen auf die Zeichen, so könnten wir nur Objekte unterscheiden für die wir unterschiedliche Zeichen haben. Hier stellt sich jedoch die Frage, woher diese Zeichen dann kämen und warum sie in irgendeiner Art und Weise zu den Gegebenheiten der Welt passen sollten.

Aus dieser Problematik heraus entwickelte Umberto Eco seinen Vorschlag einer monadischen Interpretation des Zeichenprozesses. Demzufolge geht das ursprüngliche Zeichen (Representamen) im Prozess der Semiosis über in eine Reihe anderer Zeichen (die Interpretanten), welche im Rahmen einer bestimmten Kultur eine fest verknüpfte Einheit darstellen. Gleichzeitig wird der infinite Regress des Zeichenprozesses damit zu einem selbstreferenziellen System umgedeutet.

> Interpretants are the testable and describable correspondents associated by public agreement to another sign. In this way the analysis of content becomes a cultural operation that works only on physically testable cultural products, that is, on other signs and their reciprocal correlations. Therefore, the process of unlimited semiosis shows us how signification by means of continual shiftings that refer a sign to another sign or string of signs, circumscribes cultural units in an asymptotic fashion, without even allowing one to touch them directly, though making them accessible through other units. Thus a cultural unit never obliges one to replace it by means of something that is not a semiotic entity, and never asks to be explained by some Platonic, psychic, or object entity. Semiosis explains itself by itself: this continual circularity is the normal condition of signification and even allows communication processes to use

[2] Wie später noch ausführlicher gezeigt werden wird, ist dies nur der Spezialfall eines grundlegenderen Problems. Die Bezeichnung als „semiotisches Dilemma“ dient nur dazu den spezifischen Kontext und die sich darauf beziehenden Argumente einordnen zu können.

> signs in order to mention things and states of the world. (Eco 1976: 1471)

Diese Sichtweise stellt aufgrund ihres Bezuges zu einer sprachbasierten, objektiven und zumindest *potentiell empirisch erforschbaren Kultur* eine wichtige Grundlage für die Verknüpfung semiotischer und soziologischer Theorien dar. Darüber hinaus löst sie zumindest die Seite des semiotischen Dilemmas, welche auf der Annahme einer Vorrangigkeit der Objekte basiert, auf eine sehr elegante Art und Weise. Durch die Annahme, dass Zeichen nur im Verweis auf andere Zeichen existieren, wird das Problem des Objektbezugs negiert. Es steht nun zum Beispiel nicht mehr die Frage im Raum, auf welches konkrete und offensichtlich nicht-existente Objekt das Zeichen „Einhorn" verweist, sondern durch welche anderen Zeichen („Körperbau eines Pferdes", „spitzes Horn in der Mitte des Kopfes", „mythologisch", etc.) ein Einhorn in einer bestimmten Kultur gekennzeichnet ist und durch die es potentiell einer kulturfremden Person beschrieben werden könnte. Kultur meint in dieser Perspektive ein *selbstreferentielles Netzwerk von Zeichen*, in dem sich *Objekte/Bedeutungen als relativ feste Verknüpfungen* niederschlagen.

Die Selbstreferenz stellt auch eine partielle Lösung des zweiten Teils des semiotischen Dilemmas dar. Partiell deshalb, weil sie sich nur auf die Frage nach der objektunabhängigen Herkunft der Zeichen anwenden lässt. Zeichen haben demnach ihren Ursprung in anderen Zeichen. Wie kann es aber sein, dass unter der Bedingung der Selbstreferenzialität irgendeine Form von Übereinstimmung mit der sinnlichen Erfahrung, bzw. der äußeren Welt, möglich ist? Die Antwort wird bei Eco nur angedeutet und beruht auf der Festlegung der Semiosis als einem fortlaufenden Prozess, in dem Zeichen und deren Verknüpfungen ständig neu verhandelt werden. Somit ist eine Korrespondenz der Zeichen mit den Objekten stets gradueller Art und muss sich, wie zum Beispiel im wissenschaftlichen Diskurs, stets an den Gegebenheiten beweisen. Dadurch ist aber überhaupt erst die prinzipielle Möglichkeit einer Anpassung an die Welt gegeben. Diese Lösung ähnelt sehr der schrittweisen Annäherung an die Wirklichkeit, wie sie Karl Popper (2005) in seinem kritischen Rationalismus beschrieben hat.

Eine fundamental anders geartete Vorstellung von Zeichen und dem Prozess der Zeichenkonstruktion findet sich in den Arbeiten von Ferdinand de Saussure, die insbesondere für die Linguistik – als einer deren Gründungsväter er gilt – von herausragender Bedeutung war. Sein grundlegendstes Werk *Cours de linguistique générale* (Titel der deutschen Aus-

gabe: *Grundfragen der allgemeinen Sprachwissenschaft*) erschien posthum 1916. Darin spricht Saussure die Grundzüge einer allgemeinen Sprachwissenschaft an, welche er auf einem zweiseitigen Zeichenmodell aufbaut (vgl. Saussure 2001: 76ff). Demnach besteht das Zeichen einerseits aus der psychischen Vorstellung eines Objekts, dem *Bezeichneten*, und andererseits dem *Bezeichnenden*, dem sprachlichen Zeichen, welches diese Vorstellung repräsentiert. Die Festlegung der Beziehung von sprachlichem Zeichen und Vorstellung ist auch bei Saussure ein grundsätzlich soziales Phänomen, deren Träger die Sprachgemeinschaft ist:

> Die Sprache besteht in der Sprachgemeinschaft in Gestalt einer Summe von Eindrücken, die in jedem Gehirn niedergelegt sind, ungefähr so wie ein Wörterbuch, von dem alle Exemplare, unter sich völlig gleich, unter den Individuen verteilt wären. Sie ist also etwas das in jedem Einzelnen von ihnen vorhanden, zugleich aber auch allen gemeinsam ist und unabhängig von dem Willen der Aufbewahrer. Insofern kann das Vorhandensein der Sprache dargestellt werden durch die Formel 1 + 1 + 1 ... = I (gemeinsames Vorbild). (ebd.: 23)

Hieran lässt sich jedoch auch bereits ablesen, dass es Saussure und in seinem Gefolge auch der Linguistik eher um die allgemeinen Regeln des richtigen Sprechens geht und nicht so sehr um die Inhalte oder die symbolische Ordnung einer spezifischen Sprache. Dies wird noch deutlicher, wenn er den konkreten Akt des Sprechens als eine rein individuelle Angelegenheit darstellt (ebd.: 23).

Trotz des unzweifelhaft geringen Konsens in der Soziologie bezüglich ihres Gegenstandsbereichs, widerspricht die Vorstellung eines rein additiven Charakters des Sozialen allen gängigen Theorierichtungen und Schulen. Dies gilt sogar für diejenigen Richtungen der Soziologie, die das von Durkheim (1984) begründete Prinzip der Emergenz und des *sui generis* Charakters des Sozialen ablehnen oder kritisch betrachten. Selbst der *methodologische Individualismus* musste dem nicht-additiven Charakter des Sozialen Rechnung tragen, wie man in der intensiven Suche nach einer Logik der Aggregation von individuellen Handlungen zu Makrophänomenen deutlich sieht (vgl. z.B.: Kroneberg 2009). Die Vorstellung einer Ordnung der Zeichen, in die eine eigene Logik eingeschrieben ist und die vom konkreten Sprachgebrauch aktualisiert wird, wie sie in der Semiotik zumindest grundsätzlich zu finden ist, verträgt sich besser mit einer

soziologischen Perspektive, als eine kollektiv gleichverteilte Ansammlung von Zeichenpaaren (*Signifikat* / *Signifikant*).

Auf Seiten der Semiotik wird Saussures Zeichentheorie entsprechend kritisch beäugt. Ihr genereller Einfluss auf die frühe Semiotik ist teilweise heftig umstritten und unter modernen Vertreter spielt sie nur eine geringe Rolle (vgl. Nöth 1985: 65f). Interessanterweise ist es gerade die Saussure'sche Auffassung von Zeichen als sozialen Konventionen die ihren Niederschlag in inneren Vorstellungen finden, die von den Vertretern der Semiotik zurückgewiesen wird. Eine solchen Perspektive würde eine ganze Reihe von Zeichenphänomenen, wie zum Beispiel Zeichenprozesse zwischen Maschinen, im menschlichen Körper oder im Tierreich (Zoosemiotik), die in der Semiotik eine wichtige Rolle spielen nicht adäquat berücksichtigen (vgl. Eco 1972: 28f). Folgt man Saussures Auffassung, so ist Sprache als Zeichensystem an klare und arbiträre Konventionen gebunden, während der individuelle Sprechakt sich zwar mittels dieses Zeichensystems ausdrückt, in seinen Inhalten aber vollkommen unabhängig ist. Damit wäre der Auftrag an eine Wissenschaft von den Zeichen aber nur deskriptiver Natur, nämlich die *Rekonstruktion und Beschreibung des allgemeinen Wörterbuchs* (natürlicher Sprachen).

2.2 Symbol

Nicht zuletzt aufgrund der sehr weiten Fassung des Zeichenbegriffs und der oben genannten Probleme mit einer hauptsächlich sozialen Bestimmung von Zeichen, wird in vielen Sprachtheorien oft der Begriff des Symbols für sozial konstruierte Zeichen verwendet. Allerdings herrscht in diesem Bereich noch weniger einheitliche Begriffsverwendung, als im Falle des übergeordneten Zeichenbegriffs (vgl. Nöth 1985: 96ff). Obwohl die soziale Dimension von Symbolen oft hervorgehoben wird, geschieht dies unter sehr unterschiedlichen Prämissen und theoretischen Perspektiven. Zentrales Problem in diesem Bereich sind die Fragen nach den sozialen Eigenschaften von Zeichen im Allgemeinen und Symbolen im Speziellen. Im Diskurs der Zeichentheorie werden hauptsächlich zwei Eigenschaften als kennzeichnend für soziale Zeichen angesehen, nämlich *Konventionalität* und *Arbitrarität*, letzteres wird oft auch über den gegensätzlichen Begriff der *Motiviertheit* angesprochen (vgl. ebd.: 101ff).

Zeichen können als konventionell angesehen werden, wenn sie durch *soziale Übereinkunft* zustande kommen und von einer Mehrzahl von Personen (mindestens zwei) anerkannt werden. Die Konventionalität von Zei-

chen impliziert somit immer die Existenz einer Gemeinschaft in der diese Zeichen verbreitet sind. Wie oben schon erwähnt trifft dies bei Saussure und auch im Großteil der modernen Linguistik immer schon deshalb zu, weil Sprache grundsätzlich als soziale Übereinkunft aufgefasst wird. Auch im Rahmen der Semiotik werden *Sprachzeichen als ein soziales Regelwerk* angesehen. Darüber hinaus können aber auch verschiedene nicht-sprachliche Zeichen als konventionell gelten, wenn sie innerhalb einer Gruppe von Menschen Verwendung finden (vgl. Nöth 1985: 106f).

Arbitrarität bezeichnet die Kontingenz von Zeichen. Das Zeichen steht dabei nicht in einer festen Beziehung zum Objekt welches es repräsentiert. Peirce zufolge sind Symbole eine Klasse von Zeichen, die sich spezifisch durch Konventionalität und Arbitrarität auszeichnen (vgl. Peirce 1931–0035: § 2.292ff). Er bezieht sich dabei im Wesentlichen auf Warnsignale und nicht-sprachliche Zeichen die aufgrund von sozialer Übereinkunft eine bestimmte Information mitteilen, wie zum Beispiel Rangabzeichen beim Militär, die Farbkodierung von Ampeln und dergleichen. Vom Symbol unterscheidet Peirce (ebd.: § 2.82ff) das *Ikon*, ein konventionelles Zeichen, das sich durch geringe Arbitrarität auszeichnet. Ein Ikon kennzeichnet eine sehr hohe Ähnlichkeit mit dem Objekt, auf welches es verweist. Die grafische Darstellung von Feuer wäre hierfür ein Beispiel, ebenso wie Wegweiser in Form von Pfeilen und dergleichen. Im Rahmen der Aufarbeitung des Peirces'schen Werkes ist darauf hingewiesen worden, dass die Unterscheidung von Symbol und Ikon idealtypisch verstanden werden muss (vgl. Nöth 1985: 111). Es ist leicht einsichtig, dass es sehr unterschiedliche Grade der Übereinstimmung zwischen dem Symbol und dem Objekt, das es repräsentieren soll, geben kann. So kann zum Beispiel ein Warnhinweis für Feuer sehr realistisch gestaltet sein oder eher abstrakt.

Demgegenüber versteht Saussure (2001: 80) unter Symbolen gerade eine Klasse von Zeichen, die sich durch eine *geringere Arbitrarität* auszeichnen:

> Man hat auch das Wort Symbol für das sprachliche Zeichen gebraucht, genauer für das, was wir die Bezeichnung nennen. Aber dieser Ausdruck hat seine Nachteile, und zwar gerade wegen unseres ersten Grundsatzes. Beim Symbol ist es nämlich wesentlich, daß[sic] es niemals ganz beliebig ist; es ist nicht inhaltslos, sondern bei ihm besteht bis zu einem gewissen Grade eine natürliche Beziehung zwischen Bezeichnung und Bezeichnetem. Das Symbol der Gerechtigkeit, die Waage, könn-

> te demnach nicht etwa durch irgendetwas anderes, z.B. einen Pferdewagen, ersetzt werden.

Dieser Begriff ist dem Peirces'schen Ikon relativ nahe. Auch in dieser Auffasung ist die Arbitrarität von Zeichen eher ein graduelles Phänomen und trifft nicht auf alle Sprachzeichen gleichermaßen zu.

Wenn man einmal von Widersprüchen in der verwendeten Terminologie absieht, sind sich beide Autoren hinsichtlich der Konventionalität und der Arbitrarität zumindest prinzipiell einig. Aus einer soziologischen Perspektive erscheinen diese beiden Kriterien jedoch etwas widersprüchlich und problematisch. So wären konventionelle Zeichen im Rahmen einer Sprachgemeinschaft immer nur potentiell arbiträr, d.h. bis zum Zeitpunkt ihrer *Festschreibung als soziale Konventionen*. Das würde aber auch bedeuten, dass neue Zeichen keinen erkennbaren Bezug zum bereits bestehenden Zeichensystem hätten. Wenn es auch Fälle geben mag, wie zum Beispiel bestimmte Modewörter, die scheinbar spontan auftauchen, so kann man in der überwältigenden Mehrzahl der Fälle einen klaren Bezug zwischen den bestehenden sozialen Regelungen und der Einführung neuer Zeichen erkennen. Würde man heute eine digitale Waage als Symbol/Ikon für Gerechtigkeit einführen können? Wohl eher nicht, denn die Vorstellung von Ausgleich und Abwägung, welche mittels der Waagschalen symbolisiert wird, verträgt sich nicht mit modernen Instrumenten zur Messung des Gewichts. Noch grundsätzlicher ist aber der Hinweis, dass eben nur eine Gesellschaft, die unter Gerechtigkeit einen Ausgleich von Interessen verstand, ein solches Symbol dafür hervorbringen konnte.

Arbitrarität und Konvention würden somit auf ein festgeschriebenes und sich nur durch Zufallsprozesse veränderndes *Zeichensystem* hindeuten, deren einzelne Elemente keinerlei Verbindung zueinander aufweisen. Eine solche Vorstellung wäre für soziologische Fragestellungen jedoch kaum dienlich, da es gerade die sozialen Prozesse hinter der Entstehung und Reproduktion von Zeichensystemen sind, die hier von Interesse sind. Anders ausgedrückt, gerade die Frage was die Verwendung der „Waage" als Symbol der Gerechtigkeit über eine entsprechende Kultur ausssagt, ist das soziologisch relevante Phänomen. Daher erscheint es sinnvoll zu den vorherigen Überlegungen zurückzukehren und *Zeichen* sowie deren *Relationen* als die *konstituierenden Elemente eines Zeichensystems* aufzufassen. Damit wird auch das Problem der Genese neuer Zeichen gelöst, da diese aus dem bereits bestehenden Relationen heraus entstehen, beziehungsweise in diese eingepasst werden können.

Jedoch lohnt es sich das Konzept der Konventionalität nicht ganz aufzugeben, da es daran erinnert, dass ein spezifisches Zeichensystem immer mit einer Zeichengemeinschaft einhergeht. Aus soziologischer Sicht ist dies ein weiterer Hinweis auf die Notwendigkeit, Zeichen als Elemente eines übergreifenden Zeichensystems aufzufassen. Schließlich ist der Fall, in dem verschiedene Gemeinschaften dasselbe Zeichen verwenden, dieses aber in unterschiedliche Zeichensysteme (Kulturen) eingebettet ist, der wahrscheinlich soziologisch interessanteste. Der Kampf um die Deutungshoheit zentraler Konzepte und Vorstellungen ist ohne Zweifel ein zentraler Gegenstandsbereich der sozialwissenschaftlichen Forschung. Würde man aber nur die Zeichen an sich betrachten, so käme man wohl nicht darauf, dass zum Beispiel das Wort „Gott" für unterschiedliche Religionsgemeinschaften sehr differente Bedeutungen haben kann und das die Tatsache, dass Menschen dieselben Wörter verwenden noch lange nicht heißen muss, dass sie sich einig sind.

Insofern dient das Konzept der Konventionalität als wichtiger Hinweis auf den Prozess der Aushandlung, in dem Zeichen entstehen und dem sie auch fortlaufend ausgesetzt sind. Ein und dasselbe Zeichen kann zu verschiedenen Zeitpunkten einen unterschiedlich hohen Grad an Konventionalität aufweisen. Dies kann aber immer nur relativ zu den anderen Zeichen eines Zeichensystems bestimmt werden. Daher kann man Konventionalität eher als das Ausmaß der Eindeutigkeit und des Konsens bezüglich der relativen Position eines Zeichens sehen. Diese und die vorangegangenen Überlegungen machen deutlich, dass Zeichen (und damit auch Symbole) nur als Elemente eines Zeichensystems sinnvoll betrachtet werden können.

Im weiteren Verlauf der Arbeit wird der Begriff des Symbols ausschließlich für soziale Zeichen verwendet werden. Gemeint sind damit Zeichen, die durch ein System von Zeichen bestimmt sind, welches einer Mehrzahl von Menschen bekannt ist und in sozialen Interaktionen genutzt und aktualisiert werden kann. Dieses symbolische System ist ein Synonym für Kultur entsprechend der oben vorgestellten Definition, also ein selbstreferentielles Netzwerk oder System von Zeichen.

2.3 Signal

Ein weiterer Kontext, der eine genauere Begriffsbestimmung sozialer Symbole erlaubt, ist die allgemeine Informationstheorie oder Kommunikationstheorie. Hierbei geht es vor allem um die mathematische For-

malisierung von Symbolen und die Frage, wie der Informationsgehalt und die strukturellen Eigenschaften von informationsverarbeitenden und -übertragenden Systemen beschrieben werden können. Um die begriffliche Verwirrung gering zu halten wird im Folgenden der Ausdruck „Signale“ für die Gegenstände der Informationstheorie verwendet. Deren Formulierung geht vor allem auf die Arbeit von Claude Shannon (1948) zurück, der ein wahrscheinlichkeitstheoretisches Modell der Signalübertragung entwickelte.[3]

Ausgangspunkt seiner Überlegungen war dabei das Problem des Einflusses von Störungen auf die Übertragung von Signalen mittels beliebiger Kanäle. Er konnte zeigen, dass die Möglichkeit die einzelnen Impulse voneinander zu unterscheiden und somit Information von bloßem „Rauschen“ zu trennen, als eine Wahrscheinlichkeitsverteilung beschrieben werden kann. Dadurch konnten erstmals die formalen Eigenschaften von Zeichen und Zeichensystemen bestimmt werden, die sie als Träger von Informationen auszeichnen. Diese informationstheoretischen Formalisierungen sind in zweierlei Hinsicht von Bedeutung für diese Arbeit. Zum einen stellen sie in vielen Bereichen die Grundlage für die computergestützte Analyse von Zeichen und Zeichensystemen dar. Desweiteren waren sie auch zentral für die Entwicklung sozial- und kommunikationswissenschaftlicher Theorien, insbesondere der Kybernetik und der Systemtheorie (vgl. Bertalanffy 1969; Luhmann 1981).

Der von Shannon verwendete Informationsbegriff geht auf Überlegungen von Ralph Hartley (1928) zurück, der als erster versuchte Signale und deren Übertragung formal zu beschreiben. Signale sind dabei für ihn materielle Prozesse (Schallwellen, chemische Reaktionen, elektrische Impulse, etc.), die von anderen Prozessen der gleichen Art potentiell unterschieden werden können. Um den Informationsgehalt einer Sequenz von Signalen zu quantifizieren, ging Hartley von einem idealen Sender und einem ebensolchen Empfänger aus, die nicht am Inhalt interessiert sind. Die Kommunikation zwischen diesen ereignet sich in einem spezifischen Code, der aus einer Anzahl von s verschiedenen Zeichen besteht, aus denen eine Sequenz von der Länge n gebildet wird. Der hier beschriebene Informationsgehalt, besser als Komplexität verstanden, misst wie viele Kombinationen von s möglich sind und damit die Schwierigkeit eine

[3] Der ausschlaggebende Artikel trägt im Englischen den Titel „A Mathematical Theory of Communication“ was in der deutschen Übersetzung von Helmut Dreßler als „Mathematische Grundlagen der Informationstheorie“ angegeben wurde. Mittlerweile scheint sich der Begriff der Informationstheorie durchgesetzt zu haben.

spezifische Sequenz zu rekonstruieren. In der Kombinatorik spricht man hierbei von einer *Variation mit Wiederholung*. Somit entspricht die Anzahl der möglichen Kombinationen s hoch n. Aus mathematischen und praktischen Gründen verwendete Hartley den Logarithmus um eine vergleichbare Maßzahl zu erhalten:

$$H = n \log s.$$

Je nachdem welche Basis verwendet wird, unterscheiden sich die Namen der resultierenden Einheiten: *Bits* für den Logarithmus zur Basis Zwei, *Nats* für den natürlichen Logarithmus und *Hartleys* für die Basis Zehn.

Die Unterscheidung zwischen dem Kode als einer Menge von Zeichen und deren Realisation in einer spezifischen Sequenz von Zeichen ist auch in dieser Arbeit von zentraler Bedeutung, so dass ein kurzer Exkurs gerechtfertigt erscheint. In den meisten Wissenschaften, die sich mit Zeichen und Zeichensystemen beschäftigen, hat sich unter Rückbezug auf Charles S. Peirce' (1931–0035: § 4.537) Formulierung das Begriffspaares *Typ* und *Token* eingebürgert.

> A common mode of estimating the amount of matter in a MS. or printed book is to count the number of words. There will ordinarily be about twenty *the*s on a page, and of course they count as twenty words. In another sense of the word "word", however, there is but one word "the" in the English language; and it is impossible that this word should lie visibly on a page or be heard in any voice, for the reason that it is not a Single thing or Single event. It does not exist; it only determines things that do exist. Such a definitely significant Form, I propose to term a *Type*. A Single event which happens once and whose identity is limited to that one happening or a Single object or thing which is in some single place at any one instant of time, such event or thing being significant only as occurring just when and where it does, such as this or that word on a single line of a single page of a single copy of a book, I will venture to call a *Token*. (Peirce 1906: 505f, Hervorhebungen im Original)

Ausgearbeitet und formalisiert wurden diese Begriffe in der Typentheorie von Bertrand Russell (1908), die heute als eine wichtige Grundlage der Mengenlehre gilt. Unter einem Typ wird dabei eine spezifische Einheit einer Menge verstanden. Die Menge aller Typen kann im Rahmen von Zei-

chensystemen, je nach Betrachtungsebene, auch als Kode, Alphabet, Wortschatz, etc. bezeichnet werden. Ein Token ist hingegen das Resultat einer Zerlegung einer konkreten Zeichenfolge in die einzelnen Zeichen aus denen sie besteht. Zum Beispiel besteht die Zeichenfolge „papapapa“ aus acht einzelnen Tokens, aber nur aus zwei Buchstaben, sprich Typen („p“ und „a“). Dies gilt allerdings nur, wenn man als Kode das lateinische Alphabet zugrunde legt. Wählt man eine andere Menge von Typen aus, z.B. den Wortschatz der deutschen Sprache, so würde es ebenso Sinn machen die Sequenz in zwei Token des Type „papa“ zu zerlegen. Die Beziehung zwischen Typen und Token sowie die technischen Implikationen dieses Unterschieds werden im Abschnitt 5.2 ausführlich diskutiert.

Auf der Vorstellung aufbauend, dass Information als eine Auswahl einer bestimmten Sequenz von Signalen aus dem Raum aller möglichen Sequenzen beschrieben werden kann, erweiterte Claude Shannon (1948) das Modell von Hartley. Dabei fasste er den Sender als einen stochastischen Prozess auf, der die einzelnen Signale z nun mit spezifischen Wahrscheinlichkeiten aus einem endlichen Kode $Z = \{z_1,\ z_2, \ldots z_n\}$ auswählt. Daher ergibt sich der Informationswert eines Zeichens aus der Wahrscheinlichkeit seines Auftretens $P\left(z_i\right)$:

$$I(z_i) = \ -\log_b P\left(z_i\right) \ = \log_b \left(\frac{1}{P\left(z_i\right)}\right)$$

Auch hier wählt man normalerweise die Basis Zwei für den Logarithmus, was zur Folge hat, dass die *Information eines Zeichens* in Bit angegeben wird.[4] Der Zusammenhang zwischen der Auftretenswahrscheinlichkeit und dem Informationsgehalt eines Signals kann als die Anzahl von Bits beschrieben werden, die benötigt werden, um dieses Zeichen darzustellen und es von anderen Zeichen unterscheidbar zu machen. Da sich der Informationsgehalt invers zur Wahrscheinlichkeit des Zeichens verhält, wurde für die Information auch der Ausdruck *Überraschung* eingeführt. Dahinter steht die Annahme, dass die Realisation eines unwahrscheinlichen Zeichens informativer ist (vgl. Tribus 1961).

Ausgehend von der spezifischen Selbst-Information eines jeden einzelnen Zeichens kann der Informationsgehalt einer Menge von Zeichen als der Erwartungswert beschrieben werden:

[4] Das negative Vorzeichen bzw. die Restdivision dienen nur dazu den Term positiv werden zu lassen.

$$H = -\sum_{i=1}^{n} P(z_i) \log_b P(z_i)$$

Diese Formel wird auch als Shannon-Entropie bezeichnet, da sie bis auf eine fehlende Konstante der aus der Thermodynamik und der statistischen Mechanik bekannten Entropie (oder auch H-Theorem)[5] entspricht. Die Entropie ist ein Maß für die (Un)-Sicherheit und den Informationsgehalt der aus einer bestimmten Quelle, d.h. Zufallsverteilung, zu erwarten ist. Sie erreicht ihren maximalen Wert für eine bestimmte Verteilung, wenn alle Ereignisse gleich wahrscheinlich sind, sprich die größtmögliche Unsicherheit darüber herrscht, welche Sequenz vom Sender zu erwarten ist.

Obwohl die Shannon-Entropie ursprünglich nur dazu gedacht war eine formale Grundlage für die Signalübertragung und damit zusammenhängende technische Probleme (z.B.: Übertragungsraten, Störungen, Kompression, etc.) zu stellen, verbreitete sie sich in diverse wissenschaftliche Bereiche und wurde zu einem Grundbaustein der modernen Kommunikations- und Informationstheorien. Dabei waren es sicherlich nicht zuletzt auch die Begrifflichkeiten wie „Information", „Unsicherheit" und „Entropie", die eine große Anziehungskraft auf andere Disziplinen ausübten. Durch die Kooperation von Claude Shannon und Warren Weaver (1976) wurde die Informationstheorie in einen größeren theoretischen Zusammenhang gestellt, vor allem indem sie mit Überlegungen zur Kybernetik und der Systemtheorie in Verbindung gebracht wurde, die auf die Arbeiten von Norbert Wiener zurück gingen. Von da aus war es nur ein kleiner Schritt in die sozialwissenschaftliche Kommunikations- und Systemtheorie. Vor allem Niklas Luhmann (vgl. 1984: 195f) popularisierte den Shannonschen Informationsbegriff, indem er ihn zu einem zentralen Bestandteil seiner Auffassung von Kommunikation als der Einheit von *Mitteilung*, *Information* und *Verstehen* machte. In diesem Zusammenhang wurde auch die Unwahrscheinlichkeit von einer mathematischen Beschreibung des Informationsgehalts zu einem essentiellen Charakteristikum der Kommunikation (vgl. Luhmann 1981).

Ungeachtet des praktischen und theoretischen Nutzens des Informationsbegriffs von Shannon, barg dessen Popularität jedoch auch das Potential zu Missverständnissen. Sowohl Shannon (1948: 379) als auch Hartley (1928: 536) hatten explizit darauf hingewiesen, dass ihre Begrifflichkeiten nur im spezifischen Kontext der Signalübertragung Sinn machten und

[5] Das H steht in diesem Fall für den griechischen Buchstaben Eta.

daher nicht mit der Bedeutung, die in diesen Signalen kodiert sein mag, gleichgesetzt werden dürfen. Allerdings ging bereits Weaver als Koautor über Shannons ursprünglichen Ansatz hinaus, indem er feststellte, dass es vielleicht noch ein weiter Weg zu einer „Theorie der Bedeutung" sei, aber die „Theorie der Information" der erste Schritt in die richtige Richtung wäre (vgl. Shannon und Weaver 1976: 38). Dies wurde vor allem von Seiten der Semantiker als eine unzulässige Verallgemeinerung und Übertragung dieses spezifischen Informationsbegriffs in die Sphäre der Bedeutungen angesehen, die mit einer eigenständigen Formalisierung der semantischen Information dagegen hielten (vgl. Bar-Hillel 1955; und Carnap und Bar-Hillel 1952).[6] Von da an multiplizierte sich das Problem, indem verschiedenste Disziplinen sehr unterschiedliche Auffassungen von Information entwickelten, die sich zum Teil nur noch sehr lose an der ursprünglichen Formel von Shannon orientieren.[7]

Das Problem der sauberen Begriffstrennung wird nicht zuletzt dadurch erschwert, dass eine Beziehung zwischen Information als Eigenschaft von Verteilungen und als Bedeutung naheliegend ist. So gibt es zum Beispiel Verfahren der Textklassifikation, die auf dem Konzept der Entropie aufbauen (Jaynes (1957); beispielsweise die Maximum Entropie Klassifikation auf Seite 222) und auch für die Identifikation von charakteristischen Wörtern in Texten hat sie sich als brauchbar erwiesen (vgl. Montemurro und Zanette 2013). Grundsätzlich ist es für diese und ähnliche Verfahren jedoch egal, welche konkrete Bedeutung die Texte für eine spezifische Gruppe von Menschen haben können. Vielmehr schließen sie von der charakteristischen Verteilung der Zeichen auf die grundsätzliche Möglichkeit, dass in diesen Zeichenketten Bedeutungen codiert sein können.

Der Fall des von Marcelo Montemurro und Damián Zanette (ebd.) untersuchten Voynich Manuskript kann helfen den Zusammenhang von Information und Bedeutung näher zu erläutern. Beim sogenannten Voynich Manuskript handelt es sich um ein linguistisches Rätsel in Form eines

[6] Die Berechnung der semantischen Information bezieht sich jedoch auf die Aussagenlogik und beansprucht daher nur für logische Sätze Gültigkeit. Für eine alltagsweltliche und sozio-kulturelle Rekonstruktion von Bedeutung, wie sie die Sozialwissenschaften anstreben, scheinen diese Formalisierungen eher weniger geeignet zu sein. Damit soll aber nicht jene Auffassung unterstützt werden, wonach sich die Soziologie selbst sich nicht der allgemeinen Aussagenlogik beugen müsste, nur weil ihr Gegenstandsbereich es nicht tut.

[7] Ein umfassenderer Literaturüberblick findet sich bei Kullback (1997: 1ff).

Buches, welches nach Wilfrid Voynich benannt ist, der es 1912 erwarb.[8] Obwohl sich der Besitz des Buches bis ins 17. Jahrhundert zurückverfolgen lässt, ist der eigentliche Entstehungszeitpunkt ebenso wie der Inhalt vollkommen unbekannt. Sowohl die Sprache als auch das Zeichensystem in welchem es verfasst ist, entsprechen keinen bekannten, natürlichen Sprachen. Konsequenterweise ist es eine zentrale Frage, der sich damit beschäftigenden Forschung, ob es sich um eine simple Fälschung handelt, d.h. eine bedeutungslose Menge an Zeichen, oder um eine Chiffre, also die Transliteration eines Textes mit Bedeutungen in ein unbekanntes Zeichensystem. Montemurro und Zanette gelang es nachzuweisen, dass die bedingte Entropie einzelner Textteile charakteristisch für natürliche Sprachen ist. Zudem konnten sie Token identifizieren, die aufgrund ihrer Verteilung im Text besonders kennzeichnend für einzelne Textteile sind. Diese statistischen Merkmale des Textes zu fälschen wäre so gut wie unmöglich, da sie einer Person des 17ten Jahrhunderts nicht bekannt gewesen sein dürften. Dies kann als ein Hinweis aufgefasst werden, dass es sich um einen Text handelt in dem Bedeutungen kodiert sind, allerdings besteht nach wie vor keinerlei Wissen darüber, welche Inhalte dies sein könnten. Die statistische Verteilung der Tokens kann also Aufschluss über das Potential eines Textes zur Kodierung von Bedeutungen geben, ohne jedoch diese Bedeutungen notwendigerweise entziffern zu können. Dementsprechend stellen die Gesetzmäßigkeiten denen Informationsträger unterliegen, die Rahmenbedingungen für die Kodierung von Bedeutungen in Texten dar, ohne den konkreten Inhalt zu determinieren.

Berücksichtigt man diese spezifischen Eigenschaften von Signalen als Signalträger und bezieht sie auf die bereits getroffenen Feststellungen, so gelangt man zu folgender Erweiterung der Definition: Soziale Symbole sind informationstragende Zeichen mittels derer Bedeutungen kodiert werden können und die einem spezifischen System von Zeichen entstammen, welches per Konvention als die Grundlage der Verständigung einer angebbaren Gruppe von Menschen dient (oder diente). Diese Definition umfasst somit auch „tote“ Sprachen, die nicht mehr aktiv genutzt werden, schließt jedoch diejenige Kommunikation aus die nur zwischen Maschinen stattfindet. Ebenso finden natürliche Zeichenprozesse, wie zum Beispiel die Kommunikation von Nervenbahnen, hier keine Berücksichtigung. Wichtig ist zudem der Hinweis, dass der Begriff der Information

[8] Gegenwärtig befindet es sich im Besitz der Beinecke Rare Book and Manuscript Library der Yale University. Eine eingescannte Version ist als Teil der Wikimedia Commons öffentlich zugänglich: http://commons.wikimedia.org/wiki/Voynich_manuscript.

sich auf die Shannon-Information bezieht, d.h. auf eine geordnete, vom Zufall unterscheidbare Sequenz von Zeichen.

2.4 Symbole und Soziales

Abschließend lässt sich feststellen, dass trotz unterschiedlicher Perspektiven und disziplinärer Traditionen eine grundlegende Übereinkunft in der Betrachtung von Zeichen festgestellt werden kann. Diese Übereinkunft kommt vor allem im *Dilemma der Semiotik* zum Ausdruck, welches besagt, dass Zeichen weder als Repräsentation von Objekten, noch als subjektive Zustände ausreichend beschrieben werden können. Zeichen dienen fundamental als Vermittler zwischen Entitäten, egal ob es sich dabei um Maschinen, kognitive Prozesse oder menschliche Individuen handelt. Auch in der informationstheoretischen Auffassung, dass Information letztlich nur als Signalübertragung fassbar ist, kommt diese primäre Vermittlungsfunktion zum Ausdruck. Gleichzeitig war diese Zwischenposition auch immer wieder Grund zu Ausschlägen der Debatte in die eine oder andere Richtung. Dies ist insbesondere im Hinblick auf die in Zeichen kodierten Bedeutungen ein nicht zu unterschätzendes Problem. Unterschiedliche Perspektiven betonen diesbezüglich oft entweder den subjektiven oder objektiven Charakter von Zeichen. Im ersten Fall erscheinen Bedeutungen als Aktionen und Reaktionen innerhalb von Einheiten, die Zeichen verarbeiten. Andererseits kann auch die Objektivität von Zeichen hervorgehoben werden, die ihre Vermittlungsfunktion nur dann erfüllen können, wenn sie gleichermaßen von allen beteiligten Einheiten verarbeitet werden können.

Diese Zwischen-Position rückt den Begriff des Zeichens aber auch in die Nähe sozialer Prozesse. Blendet man einmal die Zeichen und Zeichensysteme aus, die nicht von Menschen in sozialen Interaktionen genutzt werden, so erscheinen sozial standardisierte Zeichen (Symbole) als ein grundlegendes, soziales Phänomen. Da sich auch die soziale Interaktion zwischen Menschen vollzieht, liegt es nahe anzunehmen, das Symbole das geeignete Medium darstellen. Das alleine begründet aber noch keine soziologische Auseinandersetzung mit diesem Thema. Schließlich ließe sich auch eine Position wie die von Saussure einnehmen, indem man sagt, dass Bedeutung allein in der Auffassung der Sprechenden liegt und die Symbole selbst nur inhaltsleere Konventionen sind. Eine solche Position kann vielleicht noch auf die Sprache angewendet werden, vernachlässigt jedoch, dass jegliches Wissen notwendigerweise in Symbolen kodiert

ist und Bedeutungen nur als Verweisungszusammenhänge in diesem intersubjektiv geteilten Wissen kodiert und dekodiert werden können. Eine Kommunikation über ein bestimmtes Objekt setzt voraus, dass dessen Position relativ zu anderen Objekten bestmmt werden kann. Eine Rede über Einhörner macht eben nur dann Sinn, wenn man diese von Pferden und dergleichen abgrenzen kann. Auch im Bereich der Sprache wird es schwierig diese Argumentation aufrechtzuerhalten, wenn man metaphorische Ausdrucksweisen, Wortspiele und Fachsprachen in die Diskussion miteinbezieht. Schließlich hat auch das Einhorn seinen Namen nicht von ungefähr.

Symbole sind genuin soziale Phänomene, da sie das Medium darstellen in dem sich soziale Interaktionen vollziehen und sie diesen unweigerlich ihren Stempel aufdrücken. Dies kann entweder in Form eines geteilten oder nicht-geteilten individuellen Wissens geschehen, welches sich auf den weiteren Gang der Interaktion auswirkt, oder in einer übergreifenden Kultur, deren Relationen von Bedeutungen den Rahmen abstecken, in dem jeglicher sinnhafte, soziale Prozess verläuft. Gleichzeitig wirken sich auch die sozialen Interaktionen auf die Welt der Symbole aus. Wissen kann nur im Wechselspiel mit bestehendem Wissen gelernt werden. Dies setzt jedoch voraus, dass es in irgendeiner Form durch ein Handeln zugänglich gemacht wurde. Bücher schreiben sich nicht von selbst und Geschichten können nur in der Erzählung existieren. Gleiches gilt auch für die Kultur, die einem ständigen Wandel unterliegt, welcher durch soziales Handeln am Laufen gehalten wird. Insofern lässt sich sagen, dass Symbole prinzipiell als ein eigenständiges, soziales Phänomen aufgefasst werden können. Daran anschließend stellt sich jedoch die Frage, wie sich eine solche Konzeption in die bestehenden Theorien der Soziologie einpflegen lässt und welche methodologischen und methodischen Schlüsse daraus zu ziehen sind.

3 Grundlagen einer Soziologie der Symbole

Schlägt man die Stichwortverzeichnisse soziologischer Klassiker des letzten Jahrhunderts auf, so findet man in der Mehrzahl der Fälle einen Verweis auf „Symbole“ oder verwandte Begrifflichkeiten. Die Verbreitung des Begriffs ist ein Hinweis auf dessen zentrale Bedeutung für die Disziplin als Ganzes. Gleichzeitig bedeutet dies nicht, dass Symbole oder symbolische Ordnungen einen zentralen Stellenwert im jeweiligen Werk oder in spezifischen soziologischen Theorien einnehmen. Im Gegenteil, so die These des folgenden Kapitels, scheinen die meisten soziologischen Theorien doch ein eher angespanntes und unbequemes Verhältnis zu diesem Begriff zu pflegen. Es soll im Folgenden gezeigt werden, dass die Gründe für dieses Spannungsverhältnis in der fundamentalen Konzeption der Soziologie zu finden sind. Dieser Problemaufriss deutet aber auch das Potential an, welches eine explizite und methodologisch fundierte Betrachtung von Symbolen für die soziologische Theorie haben kann.

Der wahrscheinlich weitreichendste Versuch sich einen Weg durch das Dickicht soziologischer Theoriebildung mit Bezug auf Symbole zu bahnen, wurde von Norbert Elias (1991) unternommen. Sein Buch *The Symbol Theory* befasst sich mit der umfassenden Frage nach dem Stellenwert und der Konzeption von Symbolen im Rahmen der Soziologie. Ursprünglich ein dreiteiliger Artikel, der 1989 in *Theory, Culture & Society* veröffentlicht wurde, gehört es zu den letzten Werken, die vor seinem Tod fertiggestellt und publiziert wurden. Von allen größeren Büchern, die Elias im Laufe seines überaus produktiven, akademischen Lebens verfasste, handelt es sich hier um das wahrscheinlich am wenigsten rezipierte. In der soziologischen Aufarbeitung von Elias Werk wird es meist nur erwähnt, aber nicht eingehender besprochen (z.B.: Smith 2001). Die bisherige fachliche Rezeption und Bezugnahme findet hauptsächlich im Rahmen der Mediatisierungsdebatte statt und wird dort meist der Vollständigkeit halber und mit Bezug zu Elias bekannteren Arbeiten genannt (z.B.: Krotz 2003).

Vergleicht man den Einfluss von *The Symbol Theory* mit *Die Gesellschaft der Individuen* (Elias 1988), welches in dieselbe Schaffensperiode fällt, so wird die mangelnde Resonanz noch deutlicher. In der Datenbank von Google Scholar finden sich für *The Symbol Theory* ca. 377 Zitation (für die drei Artikel und das Buch), wohingegen *Die Gesellschaft der Individuen* auf ca. 885 Zitationen für die deutsche Version und ca. 751 für die posthum erschienene englischsprachige Ausgabe kommt.[1] Diese Differenz ist vielsagend, da beide Bücher im Kern dasselbe Thema bearbeiten, nämlich die grundsätzliche Konzeption des Sozialen und das Verhältnis von Individuum und Gesellschaft. Zudem sind beide Arbeiten für Norbert Elias Verhältnisse ungewöhnlich abstrakt und allgemein gehalten. Elias hat diese zögerliche Rezeption in gewisser Weise vorhergesehen, denn neben der Ausformulierung der Grundzüge einer soziologischen Theorie der Symbole handelt es sich auch um einen Versuch die bisherige Zurückhaltung der Soziologie hinsichtlich dieses Themas zu erklären.

Als zentrales Problem thematisiert Elias (1991: 10f) das Festhalten der Humanwissenschaften an *problematischen Dualismen*, welche einer Erforschung des sozialen Gehalts von Symbolen im Wege stehen. Hierunter fasst er im Besonderen die Unterscheidung von „Individuum“ und „Gesellschaft“ sowie die Trennung von „Natur“ und „Kultur“. Dabei geht es ihm nicht darum Partei zu ergreifen, sondern darzulegen, warum eine Perspektive, die dem Gegenstand der Symbole gerecht werden will, nur in einer Synthese bestehen kann, die diese Gegensätze überwindet.[2] Letztlich, so das Plädoyer der Symbol Theory, kann man Symbolen nur dann gerecht werden, wenn man sie als Prozesse auffasst.

3.1 Evolution eines sozialen Phänomens

Ausgehend von Überlegungen zum Verhältnis von biologischer und sozialer Entwicklung des Menschen, verweist Elias (ebd.: 24ff) auf die Bedeutung der Nutzung von abstrakten Symbolen als dem charakteristischen Merkmal unserer Spezies. Im Gegensatz zu unseren nächsten Verwandten im Tierreich sind wir nicht nur in der Lage auf äußere Gegenstände und innere Zustände zu verweisen, die nicht direkt unserer sinnlichen Erfahrung zugänglich sind, sondern auch das Ausmaß in dem wir dies prak-

[1] Datum der Suchanfrage: 11.03.2014

[2] Die Würdigung der Synthese als zentralem Element im Elias‘schen Denken ist auch deswegen wichtig, weil sie von der bisherigen Rezeption der Symbol Theory weitestgehend ignoriert wurde.

tizieren ist einzigartig. Der Austausch über abstrakte Dinge, zeitlich und örtlich entfernte Begebenheiten sowie fiktive Geschichten machen ohne Frage den Großteil unserer alltäglichen Kommunikationen aus. Weil diese Eigenschaften so gut wie universell menschlich sind, liegt es nahe sie als eine biologische Eigenschaft des Menschen anzusehen. Während die Fähigkeit sich symbolisch Auszudrücken ein allgemeines Merkmal unserer Spezies ist, sind die konkreten Sprachen das Produkt sozialer Prozesse. Sprache ohne eine Gemeinschaft der Sprechenden ist sinnlos. Die Form und Fumktion standardisierter Symbole zum Zwecke der Kommunikation kann somit als ein rein soziales Phänomen aufgefasst werden.

Diese Sichtweise wird jedoch problematisch, sobald man die beiden Ebenen ins Verhältnis setzt und nach den Ursachen fragt. Dann muss man feststellen, dass es nie einen Sprecher gab, ohne dass die Sprache schon vor ihm dagewesen wäre. Zugleich kann eine Sprache aber nur existieren, wenn es Sprecher eben dieser gibt. Es bleiben dann zwei Möglichkeiten damit umzugehen, man kann das Problem ignorieren und als nicht erklärbar einstufen, oder man findet eine methodologische Lösung dafür. Im Falle des Gegensatzes von *Natur* und *Kultur* kann man dieses Dilemma durch eine Trennung der Disziplinen von Soziologie und Biologie erreichen. In gewisser Weise ist diese Lösung historisch wohl die bevorzugte Variante gewesen. Dies wird nicht nur im Bereich der Sprache deutlich, sondern in einer Vielzahl von Grenzbereichen. Man denke zum Beispiel an die unterschiedlichen Ansätze zur Partnerwahl, als einer genetisch prädisponierten Maximierungsstrategie, im Gegensatz zur soziologisches Auffassung von sozialen Mechanismen der Reproduktion.

Es ist kein Zufall, dass das grundlegende Problem an das bereits thematisierte *semiotische Dilemma* erinnert. Vielmehr ist diese Unterscheidung wohl in gewissem Maße in allen Wissenschaften zu finden. Der Biologe Rupert Riedl (1985) beschreibt sie gar als die fundamentale Spaltung unseres Weltbildes und dehnt ihren Geltungsbereich auf die Geistes- und Kulturgeschichte aus. Insbesondere fasst er darunter den Gegensatz von *Empirismus und Rationalismus* sowie *Materialismus und Idealismus*. Spezifisch geht es ihm um das Verhältnis von *a priori und a posteriori*. Wir kommen nur zu allgemeinen Aussagen, indem wir die Erfahrung des speziellen Falles verallgemeinern. Dieser induktive Schluss ist, wie Karl Popper (2005) gezeigt hat, jedoch im logischen Sinne unmöglich, da er eine Erweiterung des Wahrheitswertes des ursprünglich Satzes darstellen würde. Nur der Weg von den allgemeinen Sätzen zu den elementaren Sätzen ist als Deduktion möglich. Dabei bleibt der empirische Wahrheitswert unabhängig

von der logischen Bewertung.[3] Somit stellen sich zweierlei Fragen: was machen wir, wenn unsere Theorien falsifiziert werden und noch dringender, woher kommen überhaupt die allgemeinen Sätze? Wieder begegnet uns das bereits angesprochene Henne-und-Ei-Problem.

Ausgehend von einer evolutionsbiologischen und systemtheoretischen Perspektive, erscheint Riedl das Paradoxon des Verhältnisses von induktiver Praxis und deduktiver Logik, die so gegensätzlich erscheinen und doch untrennbar verknüpft sind, in der evolutionären Entwicklung allen Lebens angelegt. Dabei wird das Problem der Erkenntnis in das allgemeinere Problem der Anpassung an die prinzipiell lebensfeindliche Realität überführt:

> Wir betrachten die Evolution selbst als einen kenntnisgewinnenden Prozeß. Die Ketten der Generationen überleben dabei unter der Voraussetzung, daß sie auf die Bedingungen ihres Milieus richtig reagieren. Und dies setzt, weit vor jedem bewußten Handeln, Programme voraus, die etwas wie Kenntnis von den relevanten Gesetzen der Umgebung enthalten. (Riedl 1985: 45)

Auch hier nimmt das Problem seinen Ausgang in der mittlerweile vertrauten Form: Wie ist Anpassung an neue, individuelle Situationen möglich, wenn dies immer schon voraussetzt hinreichend angepasst zu sein? In der Stammesgeschichte des Lebens wird das Problem der Anpassung mittels der Tradierung und Inkorporation des per Versuch und Irrtums erworbenen Wissens gelöst. Erkenntnis ist demnach das Ergebnis eines Selektionsvorganges, eines Zusammenstoßes von Möglichkeiten mit der Realität. Was dabei im Zeitverlauf herauskommt sind wiederum a priori, sind *Isomorphien*, also eine Anpassung von Strukturen an die Erkenntnis in den für die Selektion relevanten Bereichen. Perspektiven der Beobachtung können daher als Verzeitlichung oder Inkorporation von momentaner Erkenntnis gesehen werden.

Die Evolution des Wissens enthält damit aber nicht nur eine mehr oder minder gelungene Anpassung an die Lebensumstände, sondern auch eine interne Steigerung der Möglichkeiten des Wissenserwerbs selbst, indem sie neue Informationsträger (Medien) hervorbrachte, die eine schnellere und flexiblere Anpassung ermöglichten. Die ursprünglichste Form der Kodierung des erworbenen Wissens waren höchstwahrscheinlich Gene, bzw.

[3] In dem Sinne, dass empirische Sätze nicht schon allein aus logischen Gründen wahr oder falsch sein dürfen.

deren direkten Vorläufer. In dieser Phase war das Lernen auf die Abfolge von Generationen beschränkt und die Selektion betraf somit noch das Individuum als Ganzes. Die darauf folgende Entwicklung der Nervenbahnen erlaubte es den einzelnen Lebewesen hingegen im Falle einer „Falsifikation" mit der blanken Haut davonzukommen, indem der *Realitätskontakt über Sinnesorgane* erfolgte. Diese erlaubten irgendwann auch das Kopieren der Verhaltensweisen von Artgenossen, womit die Verstetigung des Wissens nicht mehr zwangsläufig mit dem Ende desjenigen, der die Beobachtung machte, einherging. Irgendwann erreichte auch diese Entwicklung ihre kritische Masse und brachte Lebewesen hervor die nicht nur in der Lage waren, externe Reize zu verarbeiten, sondern diese auch kognitiv simulieren konnten. Eine solche *Simulation* erlaubte es „Hypothesen" über einen möglichen Verlauf der Dinge zu formulieren und diese anhand gemachter Erfahrungen zu evaluieren. Auch diese Entwicklung erwies sich als äußerst nutzbringend, da nun die Selektion auf die Hypothesen wirken konnte und das Individuum somit relativ geschützt war.

Mit dem Aufkommen der Simulation eng verknüpft, war die Fähigkeit abstrakte Konzepte zu handhaben. Ab diesem Zeitpunkt waren Zeichen nicht mehr direkt an die sinnliche Erfahrung geknüpft, sondern konnten zu Kategorien zusammengefasst werden. Damit begann die oben schon festgestellte Selbstreferenz von Zeichen. Der nächste Schritt bestand dann wohl im Austausch und der Standardisierung dieser Zeichen innerhalb *sozialer Gruppen*. Hier besteht der Überlebensvorteil in der Weitergabe und auch der systematischen Konstruktion von Wissen, welches jenseits eines spezifischen Kontextes anwendbar ist. Erst ab diesem Zeitpunkt, den wir entwicklungsgeschichtlich wahrscheinlich nie konkret bestimmen können werden, sind wir bei den Symbolen als sozial standardisierten Zeichen angelangt.

Diese grobe Skizze[4], die weder historische Korrektheit noch Vollständigkeit für sich beanspruchen kann, soll vor allem zwei Punkte illustrieren. Zum einen die *inhärente Dynamik* des Prozesses, in der die Steigerung der Komplexität mit der Entstehung immer neuer Verarbeitungsmechanismen einhergeht. Was wiederrum mehr Komplexität zulässt und damit einen sich selbst verstärkenden Prozess aufrechterhält. Zweitens wird in der Entwicklungsgeschichte der Symbole deren Tragweite für das harte

[4] Diese Überlegungen sind angelehnt an die sehr viel ausführlicheren Darstellungen von Rupert Riedl (1981) zu den biologischen Grundlagen dieser Entwicklung und den Arbeiten von Dietrich Dörner (2001), welcher sich mit der Genese des abstrakten Denkens in der Psyche auseinandersetzt. Auch in diesen Fällen geht es nicht um den konkreten, historischen Ablauf, sondern um das generelle Verlaufsmuster.

Geschäft des Überlebenskampfes deutlich. Mit der Zunahme der Möglichkeit die Welt in Symbolen zu fassen ging immer auch eine *Verbesserung der Anpassungs- und Kontrollmöglichkeiten* der Individuen in Bezug auf ihre jeweilige Umwelt einher.[5]

Das beste Beispiel für die Richtigkeit der letztgenannten Behauptung liefert die evolutionäre Erfolgsgeschichte der Gattung Homo. Zwar gibt es *Kommunikation mit abstrakten Symbolen* auch bei anderen Tieren und sicherlich ist der Übergang fließend, dennoch kann man feststellen, dass die menschliche Neigung zu Symbolen ihresgleichen sucht. Dies wird insbesondere an der Vielzahl von Techniken zur Speicherung (z.B.: Schrift, Bilder, etc.) und Verbesserung (z.B.: Diskussion, Experimente) von symbolischen Ordnungen, wie zum Beispiel unserem Wissen über die Welt, deutlich. Wie man auch zum menschlichen Wesen stehen mag, es sind diese Techniken und die einzigartige, menschliche Prädisposition zu Symbolen, die zu der heutigen, umfassenden Dominanz unserer Spezies geführt hat. Dank ihrer symbolischen Ordnungen waren Menschen in der Lage Kooperation und technologisches Wissen in einem nie gekannten Ausmaß einzusetzen und sich somit immer weiter vom Selektionsdruck der Evolution zu emanzipieren. Statt genetischer Anpassung an die Umwelt wurde die Anpassung der Umwelt an menschliche Bedürfnisse zum Programm.

Die inhärente Steigerungsdynamik dieses Prozesses wird damit jedoch keineswegs aufgehoben. In dem Maße in dem wir uns von der genetischen Selektion unabhängig machten, wurden wir zum Gegenstand der sozialen Selektion. Die symbolische Ordnung selbst ist damit der Rahmen unserer Möglichkeiten geworden. Was einem Einzelnen ermöglicht oder verwehrt wird hängt in sehr hohem Maße von der sozialen Zuschreibung ab. Status und Stellung des Einzelnen sind in menschlichen Gesellschaften in erster Linie sozialen Mechanismen geschuldet. Somit geht auf der individuellen Ebene eine gesteigerte Fähigkeit mit Symbolen umzugehen auch mit gesteigerten Möglichkeiten einher, wie zum Beispiel verstehbar

[5] Dieser Zugewinn an Wissen und Erkenntnis im Laufe der Menschheitsgeschichte wird von einigen Denkern der „Postmoderne“ in Zweifel gezogen. Meistens unter Verweis auf den Charakter der Wirklichkeit als einer „Konstruktion“. Hier scheint jedoch eine Verwirrung der Begrifflichkeiten vorzuliegen. Die Tatsache, dass etwas konstruiert wurde impliziert noch lange nicht dessen Korrespondenz mit der Realität. Vielmehr kann es bessere und schlechtere Konstruktionen geben. Am ontologischen Gehalt einer handwerklich gut konstruierten Pistole mag man Zweifeln, ihren Realitätsgehalt möchte selbst ein überzeugter „Konstruktivist“ wohl nur ungern zu spüren bekommen. Ebenso verhält es sich mit unserem Wissen über die Welt, dessen Realitätsgehalt ebenfalls von den Bedingungen seiner Konstruktion, d.h. den Methoden, abhängt.

kommunizieren, sich mit anderen koordinieren oder auch nur das tradierte Wissen für seine eigenen Zwecke einsetzen zu können. Des Weiteren enthält nur der individuelle Gebrauch von Symbolen das grundsätzliche Potential Symbole und deren Bedeutungsgehalt, d.h. deren Verknüpfung mit anderen Symbolen, zu verändern. Gleichzeitig stellt jedoch die symbolische Ordnung als Ganzes eine externe Beschränkung des Möglichen, Erlaubten oder Kommunizierbaren dar. Es ist der Anschluss an die bestehende Ordnung, der aus den Möglichkeiten und den individuellen Ereignissen auswählt.

3.2 Die Fraktalisierung des Problems

All die bisher untersuchten theoretischen und epistemischen Perspektiven teilen im Kern dieselbe Problembeschreibung, nämlich, wie der wechselseitige Bezug von „Teilen“ und „Ganzem“ zu fassen wäre. Da sich hierfür jedoch eine Reihe von sehr verschiedenen Namen und Bezeichnungen herausgebildet hat, ist es nicht immer einfach das grundlegende Wesen dieses Problems sowie dessen Tragweite zu erkennen. Es ist gerade diese Janusköpfigkeit des Symbolischen, so die hier vertretene These, die sich in den Differenzen der soziologischen Perspektiven widerspiegelt und damit einer nutzbringenden Analyse dieser Tatbestände im Wege steht. Deswegen erscheint es zweckmäßig die Folgen dieses Bruchs für die soziologische Beschäftigung mit Symbolen zu verdeutlichen, bevor der Versuch unternommen werden kann das Problem zu überwinden.

So schwierig das Streben nach Interdisziplinarität und übergreifenden Erklärungen sein mag, das eigentliche Problem beginnt nicht erst an den Grenzbereichen zu anderen Disziplinen, vielmehr ist es tief eingeschrieben in die Forschungspraktiken und theoretischen Perspektiven der Soziologie selbst. Hier nimmt es die bereits angesprochene Form der Differenz von „Individuum“ und „Gesellschaft“ an. Diese Unterscheidung tritt unter einer Vielzahl von Namen und mit sehr spezifischen Akzenten auf. Die Unterscheidung von Mikro- und Makroebene, die gegensätzlichen Positionierungen von methodologischem Individualismus und Holismus sowie auch die Trennlinie von qualitativer und quantitativer Soziologie können als Verlaufslinien dieses Bruchs aufgefasst werden. Man kann dementsprechend von einem epistemologischen Bruch sprechen, da er sich bis in die erkenntnistheoretischen Grundlagen der Soziologie erstreckt. Eine erschöpfende Aufzählung der Ausdrucksformen dieser Differenz ist sicherlich nicht möglich, aber zumindest kann sie eingehender analysiert

und verstanden werden. Dafür zumindest scheint die grobe analytische Unterscheidung zwischen „Individuum“ und „Gesellschaft“ als dem jeweiligen Schwerpunkt, den eine theoretische Perspektive im Hinblick auf Symbole und symbolische Ordnung setzt, hilfreich. Jenseits dieses Nutzens als Heuristik und Werkzeug der Analyse ist sie jedoch nicht dazu gedacht den Wert oder den ontologischen Gehalt spezifischer Theorien zu beschreiben.

Neben der mangelnden Geradlinigkeit des Bruchs muss auch zugestanden werden, dass es sich hier sicherlich nicht um eine absolute Differenz handelt und deshalb auch nicht davon auszugehen ist, dass sich die jeweiligen theoretischen Perspektiven einem der beiden Pole klar zuordnen lassen. Vielmehr muss ein Kontinuum angenommen werden, dessen Extrempunkte idealtypische Fassungen der gesellschafts- und individuumszentrierten Perspektiven auf Symbole sind. Hiermit wird auch dem Umstand Rechnung getragen, dass es sicherlich auch soziologische Theorien gibt, die das Phänomen sozialer Symbole nicht oder kaum berücksichtigen.

Diese Spaltung hat jedoch nicht nur den Charakter eines Kontinuums, sondern auch den einer „fractal distinction“ (Abbott 2001: 9). Mit diesem Ausdruck bezeichnet Andrew Abbott die in der Soziologie häufig anzutreffende Tendenz grundlegende Spaltungen innerhalb bereits bestehender Spaltungen zu wiederholen. In seinem so treffend betitelten Buch *Chaos of the Disciplines* macht er dies an einer Mehrzahl verschiedener Brüche innerhalb der Soziologie fest, wie z.B. qualitative vs. quantitative Sozialforschung, Realismus vs. Konstruktionismus[6] oder Geschichtswissenschaft vs. historische Soziologie (vgl. ebd.: 10ff).

Die *Fraktalisierung* kann am Beispiel des inhärenten, soziologischen Methodenstreits gut verdeutlicht werden. Hierbei führte die ursprüngliche Unterscheidung zwischen interpretativem Vorgehen und statistischer Analyse zu zwei relativ eigenständigen Bereichen des soziologischen Methodendiskurses, in denen dieselbe Unterscheidung (Analyse vs. Interpretation) wieder auftaucht (ebd.: 11). Im Rahmen der quantitativen Methode manifestiert sich dies in der methodischen Auseinandersetzung zwischen regressionsanalytisch, auf Kausalität ausgerichteten Verfahren

[6] In der deutschsprachigen Soziologie wäre hier der Begriff des „Konstruktivismus“ üblicher. Das von Abbott (2001: 17ff) verwendete „Constructionism“ erscheint jedoch an disere Stelle besser geeignet das Forschungsprogramm der soziologischen Konstruktivisten, also die Frage nach der Konstruktionsleistung der Subjekte, von ähnlich lautenden mathematikphilosophischen („Erlanger Konstruktivismus“) und epistemologischen („Radikaler Konstruktivismus“) Formulierungen abzugrenzen.

(z.B.: OLS Regressionen, Event History Analysis) und dimensionsanalytischen Klassifikationsverfahren (z.B.: Korrespondenzanalyse, Clusteranalyse, etc.). Auf der Seite der qualitativen Methoden findet sich dasselbe Phänomen in Form der Unterscheidung von analytischer Interpretation (z.B. Inhaltsanalysen mit quantitativen Elementen) und interpretativer Interpretation (z.B.: Grounded Theory, Objektive Hermeneutik).

Im soziologischen Diskurs werden diese Brüche und ihre Tendenz zur Fraktalisierung vor allem durch den Versuch der Reintegration durch „Wiederentdeckung" deutlich:

> The centrality of rediscovery is also evident in the litany of articles entiteled „Bringing the Something-or-other Back In". Some ninety-one articles and books have brought something back in since George Homans first used the phrase in the title of his 1964 ASA presidential address, a virulent attack on Parsons for ignoring purposive action. And the things brought back in have included both sides of most of the important social scientific dichotomies. Some writers have brought people back in, others behavior. Some have brought social structure back in, others culture. Some have brought ourselves, others the context. Some circulations, others structure. Some capitalists, other workers. Some firms, other unions. (ebd.: 16)

Damit verwandt sind auch die zahlreichen „Turns", welche die Soziologie in den letzten Jahren erlebt hat. So zum Beispiel den „Visual Turn", den „Semantic Turn", den „Cultural Turn" und einige andere. Obwohl diese Versuche das Ruder herumzureißen, bzw. altes Wissen neu zugänglich zu machen, stets einen integrativen Anstrich hatten, war das Ergebnis in den meisten Fällen eine Vertiefung des Bruchs oder seine Wiederholung unter anderen Vorzeichen. Es ist leicht ersichtlich, dass genau dies unter der Annahme einer Fraktalisierung zu erwarten ist. Jeder Versuch der Reintegration wird mit hoher Wahrscheinlichkeit zu einer weiteren Spaltung führen, wenn die vorangegangene Unterscheidung bereits länger besteht und sich dementsprechend als zu gut institutionalisiert erweist.

Auch für die folgenden Erläuterungen hat sich die Annahme einer fraktalisierten Soziologie als nützlich erwiesen. So zeigt sich relativ klar, dass das grundlegende Problem einer Verortung sozialer Symbole aufgrund deren Doppelcharakters stets eine Wiederholung dieses Bruches im nachfolgenden Diskurs zur Folge hat. Dies unterstreicht die Wichtigkeit diesem Bruch und seinen Verlaufsformen zu folgen, wenn man zu einer dem

Gegenstandsbereich angemessenen Sichtweise gelangen will. Gleichzeitig ist es als eine Warnung zu verstehen, nicht ebenfalls in das fraktale Denkmuster zu verfallen und den Unterschied in sich selbst zu wiederholen. Wie schon angedeutet, wird hier der Versuch unternommen den Bruch durch eine Synthese der einzelnen Ebenen zu überwinden.

3.2.1 Individualistische Perspektive

Kennzeichnend für die individualistische Perspektive ist, dass Symbole und symbolische Ordnungen als *Merkmale von Individuen* aufgefasst werden. Wissen, Kultur und Sprache erscheinen als Funktionen des geistigen Innenlebens des Menschen. Dementsprechend ist *Sinn* hier der zentrale Untersuchungsgegenstand, also der innere Entschluss eines Individuums, welches aus persönlichen Gründe heraus handelt. Dieser Entschluss kann dabei sowohl bewusst als auch unbewusst, rational oder irrational gefällt werden. Die grundlegende Methodologie dieser Perspektive ist dementsprechend die Rekonstruktion und Modellierung des subjektiven Sinns hinter den Handlungen.

Die aufgeführten Begrifflichkeiten machen deutlich, dass Max Webers Auffassung von „Erklären“ und „Verstehen“ als methodologischen Grundprinzipien für die individualistische Auffassung Pate gestanden haben. In soziologischen Lehrbüchern und Einführungskursen findet man mit hoher Wahrscheinlichkeit Webers berühmte Definition der Soziologie:

> Soziologie (im hier verstandenen Sinne dieses sehr vieldeutig gebrauchten Wortes) soll heißen: eine Wissenschaft, welche soziales Handeln deutend verstehen und dadurch in seinem Ablauf und seinen Wirkungen ursächlich erklären will. „Handeln“ soll dabei ein menschliches Verhalten (einerlei ob äußeres oder innerliches Tun, Unterlassen oder Dulden) heißen, wenn und insofern als der oder die Handelnden mit ihm einen subjektiven Sinn verbinden. „Soziales“ Handeln aber soll ein solches Handeln heißen, welches seinem von dem oder den Handelnden gemeinten Sinn nach auf das Verhalten *anderer* bezogen wird und daran in seinem Ablauf orientiert ist. (Weber 2006: 11f)

Den subjektiven Sinn ins Zentrum der Soziologie zu stellen war sicherlich nicht alleine Webers Verdienst. Dennoch war diese Perspektive ein

Grundstein für spätere Ausführungen, die man gemeinhin als *methodologischen Individualismus* bezeichnet. Jedoch verbirgt sich hinter diesem Oberbegriff eine Vielzahl von zum Teil sehr unterschiedlichen Konzeptionen (Udehn 2002). Sie reichen von der Annahme das Individuen ein prinzipielles Primat in der Erklärung sozialer Phänomene zukommt bis hin zur kompletten Reduktion aller sozialer Phänomene auf psychische Vorgänge. Was hier interessiert ist jedoch keine lückenlose Aufarbeitung dieser Unterschiede, sondern ein Herauspräparieren der spezifischen Perspektive auf soziale Symbole entsprechend der oben eingeführten Definition.

Betrachtet man den methodologischen Individualismus durch die Brille der Fraktalisierung bezüglich der Frage, welche Stellung soziale Symbole in der jeweiligen Theorie einnehmen, so lässt sich im Anschluss an Weber eine Differenzierung in Handlungs- und Wissenstheorie feststellen. Erstere bezog sich vor allem auf die methodologischen Formulierungen Webers bezüglich der Handlungen und tendierte dementsprechend dazu die „Kausaladäquanz" ins Zentrum zu stellen und erklärend vorzugehen. Als namenhafte Vertreter sind hier vor allem Talcott Parsons (1968), George Homans (1964)[7] und James Coleman (1994) zu nennen. Während Talcott Parsons den Versuch einer handlungstheoretischen Fundierung später wieder aufgab wurde dieses Theorieprogramm von Homans und Coleman entscheidend vorangetrieben und führte letztlich zur gegenwärtigen Form der Handlungstheorie, als einer Abwandlung des Rational Choice. Die andere Seite dieser Unterscheidung ging von Webers Auffassung des Sinns als zentraler Kategorie der Soziologie aus. Folglich wurde in diesem Bereich die „Sinnadäquanz", bzw. das Verstehen zur Grundlage des methodologischen Programms. Federführend waren hier vor allem die Arbeiten von Alfred Schütz (1974) sowie, daran anschließend, Peter Berger und Thomas Luckmann (1980). Daraus resultierte schließlich die wissenssoziologische Richtung der Sozialphänomenologie bzw. des sozialen Konstruktivismus.

Handlungstheorien

In der klassischen, ökonomisch orientierten Handlungstheorie tauchte die symbolisch, normative Ordnung, wenn überhaupt, fast ausschließlich als Randbedingung eines rationalen Kalküls auf (vgl. Münch 1998).

[7] Hier handelt es sich um eine der oben bereits erwähnten „Bringing x back in" Publikationen. Dies verdeutlicht noch einmal den fraktalen Charakter der fundamentale Differenz von Individuum und Gesellschaft.

Normen werden nur relevant sofern sie mit Kosten der Sanktion bewehrt sind. Die einzige Form des Wissens, die in den Begrifflichkeiten einer solchen Theorie gefasst werden konnte, ist das instrumentelle Wissen, welches jedoch zumeist hinter der Annahme der „complete information" zurücktreten musste. Unter der Annahme einer streng *rationalen Wahl* erschien der Sinn einer Handlung als eben dieses Kalkül: Die Begründung der subjektiven Auswahl aus einer Mehrzahl von Handlungsalternativen entsprechend deren subjektiv empfundenem Nutzen und der subjektiv eingeschätzten Chance der Realisierung des Ziels der jeweiligen Handlung, besser bekannt als „von-Neumann-Morgenstern-Nutzensfunktion" (vgl. von Neumann und Morgenstern 1953: 15ff).

Dementsprechend ist es nicht verwunderlich, dass die methodologische Folge der Auftrag zur Rekonstruktion des Sinns aus der konkreten Handlung war. Deutlich wird dies vor allem im Versuch Handlungen zu erklären, die auf den ersten Blick kontra-intuitiv oder irrational erscheinen. So zum Beispiel Gary Beckers (1985) Erklärung der Arbeitsteilung im Haushalt als eine rationale Strategie der Spezialisierung. Dadurch dass komparative Vorteile zwischen den Geschlechtern existieren (z.B.: Höhere Kosten des Zugangs zum Arbeitsmarkt für Frauen oder geringere Bildungschancen), ist die Strategie sich auf einen spezifischen „Markt" zu spezialisieren, also entweder den Haushaltsmarkt oder den Arbeitsmarkt, die Folge eines rationalen Kalküls. Das grundsätzliche Muster der geschlechtsspezifischen Arbeitsteilung in einer bestimmten Gesellschaft entsteht demnach spontan aus den rationalen Überlegungen der Individuen. Allerdings ist die Spezialisierung nur unter der Bedingung jenes ursprünglichen, komparativen Vorteils rational. Woher dieser gesellschaftliche Unterschied zwischen den Geschlechter kommt und worin genau diese komparativen Vorteile bestehen, wie also die institutionellen Regeln aussehen, die ein spezifisches Nutzenkalkül rational werden lassen, wird von Becker hingegen explizit ausgeklammert (vgl. ebd.: 41f).

Demgegenüber konzentrierte sich die phänomenologische Schule auf de Gegenstand des *intersubjektiv geteilten Wissens*. Der Fokus lag insbesondere auf den Alltagsstrukturen und den Wissensbeständen der jeweiligen Individuen. Als dominante Handlungsform wurde hier das traditionale Handeln angenommen. Wissen entsteht demnach durch einen Prozess der wechselseitigen Typisierung von Handlungen (Institutionalisierung), welche durch *Sozialisation* weitergegeben werden und somit dem Individuum als objektives Wissen erscheinen (vgl. Berger und Luckmann 1980: 49ff). Die methodologische Konsequenz war die Erforschung des Sinns als

subjektivem Wissen, um die Handlungen von Personen dadurch verstehen zu können. Die Erforschung des subjektiven Wissens geschieht dabei konsequenterweise an den Personen selbst. Dementsprechend erscheint Wissen als auf einen sehr kleinen, spezifischen Personenkreis ausgerichtet und als eine grundsätzlich sehr fragile Konstruktion. Dies wird zum Beispiel in der Beschreibung der Familie als einem Mikrokosmos der sozialen Konstruktion von Wirklichkeit deutlich:

> Also, coming from broadly similar sectors of the larger society (in terms of region, class, ethnic and religious affiliations), the two individuals will have organized their stock of experience in similar fashion. In other words, the two individuals have internalized the same overall world, including the general definitions and expectations of the marriage relationship itself. Their society has provided them with a taken-for-granted image of marriage and has socialized them into an anticipation of stepping into the taken-for-granted roles of marriage. All the same, these relatively empty projections now have to be actualized, lived through and filled with experiential content by the protagonists. This will require a dramatic change in their definitions of reality and of themselves. (Berger und Kellner 1964: 10)

Hier zeigt sich die Verwandtschaft zur handlungstheoretischen Perspektive. Der Dreh- und Angelpunkt sind in beiden Fällen die Individuen. Darüber hinaus wird das konkrete Handeln hier jedoch als eine Funktion des im Laufe der Biographie erworbenen Wissens und der grundsätzlichen Schwierigkeit der Synchronisierung von Bewusstseinsinhalten aufgefasst. Interessanterweise gibt es ebenfalls einen Verweis auf ein gesellschaftlich verbreitetes Wissen, dass jedoch erst in seiner konkreten, individuellen Ausprägung zum Gegenstand der phänomenologischen Forschung wird.

Das Modell der Fraktalisierung legt die Vermutung nahe, dass eine individualistische Annäherung an intersubjektiv geteilte Inhalte, selbst in der individualistischen Fassung der Phänomenologie, mit einer Gegenbewegung in die entgegengesetzte Richtung einhergeht. Und in der Tat kann man die von Harold Garfinkel (1967) begründeten *Studien zur Ethnomethodologie* als eine solche fraktale Bewegung auffassen. In Abgrenzung zur damals in der us-amerikanischen Soziologie insbesondere von Talcott Parson vertretenen Annahme, dass Handeln von einem kulturel-

len Reservoir von Wissen angeleitet wird, ging Garfinkel von einer situativen und kontextabhängigen Konstruktion von Wissen aus. „Consistent with the emphasis in Garfinkel's early work, ethnomethodologists reject the idea that persons make sufficient sense of each others' actions by attaching culturally encoded meanings to particular words and gestures" (Lynch und Peyrot 1992: 114). Ethnomethodologie ist dabei die umfassende Bezeichnung für die konkreten Praktiken, Riten und Heuristiken, mit denen Menschen zu einer sinnhaften Interpretation der konkreten Situation gelangen. Die Erforschung dieser Formen der „common-sense rationalities" (Garfinkel 1967: 68) ist der Gegenstand der ethnomethodologischen Forschung. Dementsprechend ist das Vorgehen dieser Forschungsrichtung von akribischen Beobachtungen und Versuch der Rekonstruktion der konkreten Logiken des Verstehens *in situ* gekennzeichnet, was sowohl mit klassisch ethnographischen Verfahren als auch mit Krisenexperimenten (vgl. ebd.) und Konversationsanalysen (vgl. Schegloff 1989) erreicht werden soll.

Die Ethnomethodologie hat in der Soziologie lange Zeit ein Nischendasein geführt. Eine gewisse Tendenz zu eigenwilliger Begriffsbildung und sehr kleinteiligen Studien hat, selbst nach Einschätzung der Vertreter dieser Forschungsrichtung, zu einem Ruf der Obskurität und der Sektiererei beigetragen (vgl. Lynch und Peyrot 1992: 114). Seit Garfinkels ursprünglichem Forschungsprogramm hat die Ethnomethodologie vor allem im Bereich der Wissenschaftssoziologie Fuß gefasst (Knorr-Cetina 2002; Lynch 1997). Obwohl neuere Ansätze in diesem Bereich oft weniger von kognitiven Prozessen aus argumentieren, ist die grundsätzliche Ausrichtung auf die individuelle Interpretation und Konstruktionsleistung erhalten geblieben (vgl. Lynch und Peyrot 1992: 114f). Mittlerweile bezieht die ethnomethodologische Forschung auch kulturelle Artefakte, wie Bilder (z.B. Lynch 1985), Zeichnungen und Berichte (z.B. Knorr-Cetina 2002), stärker mit ein. Dabei gilt das eigentliche Forschungsinteresse jedoch nach wie vor dem *subjektiven Konstruktionsprozess* dieser Objekte und deren *Interpretation* durch konkrete Personen.

Diese Interpretationslogik wird im Rahmen der *Konversationsanalyse* auch auf sprachliche Zeichen angewandt. Hierbei wird untersucht wie wechselseitiges Verstehen im Zuge der Konversation aufwendig konstruiert werden muss (Schegloff 1989). Diese Auffassung ist derjenigen der Linguistik und Semiotik so diametral entgegengesetzt, dass man hier von inkommensurablen Axiomatiken sprechen kann. Jedwede Beschäftigung mit Sprache oder Symbolen als zentralem Gegenstand der Forschung setzt

die Annahme eines gewissen Objektcharakters für diese Phänomene voraus, da es sonst schlichtweg nichts zu forschen gäbe. Dementsprechend ist diese spezifische, ethnomethodologische Position auch nur schwer in die hier geführte Debatte zu integrieren.

Neben dieser strikt interaktionstheoretischen Auslegung ist in der jüngeren ethnomethodologischen Forschung eine stärkere Akzeptanz von situationsübergreifenden und kulturell vermittelten Bedeutungsmustern festzustellen. Dies geht im Wesentlichen auf die Arbeiten von Karin Knorr-Cetina (2002, 2005) zurück. In Abgrenzung von einer praxisorientierten, ethnomethodologischen Wissenschaftssoziologie, die sich hauptsächlich auf „performative Agency" (Pickering 2010), also auf die Praxis der Handelnden konzentriert, schlägt Knorr-Cetina eine Ergänzung des Praxisbegriffs um den Begriff der Kultur vor, die den Übergreifenden Rahmen von Symbolen und Bedeutungen darstellt und damit die Voraussetzung jeglichen individuellen Handelns darstellt, welches ansonsten nur unkoordiniert und erratisch wäre.

Auch in anderen Bereichen lässt sich in den letzten Jahrzehnten eine verstärkte Tendenz zur Integration der individualistischen Theorien beobachten. Dies geschieht sehr oft unter der Schirmherrschaft der aus dem Rational Choice hervorgegangenen Handlungstheorien. Da sowohl die Phänomenologie als auch die Handlungstheorie von einer *zentralen Stellung des Individuums* ausgehen, ist es nicht verwunderlich, dass Sinn und Wissen in beiden Fällen als rein subjektive Kategorien erscheinen. Es gibt in dieser Sichtweise kein Wissen, außer dem, welches von Menschen gewusst wird und keinen Sinn, außer dem, den konkrete Individuen ersinnen. Es besteht eine grundsätzliche Komplementarität dieser individualistischen Perspektiven. Die Phänomenologie bietet eine Sozialisationstheorie des Wissenstransfers an, während die Rational Choice Theorie über ein Handlungsmodell verfügt, welches die Gründe für die Auswahl konkreter Handlungen anzugeben vermag. Diese Passgenauigkeit wurde insbesondere von Hartmut Esser (1996) erkannt, der sie in sein Modell der soziologischen Erklärung integrierte. Im Rahmen der Frame-Selektionstheorie sollen Normen, Wissen und Regeln handlungstheoretisch fundiert und gedeutet werden (Esser 1990; Kroneberg 2007).

Ausgangspunkt ist dabei die Überlegung, dass Handlungen in zwei verschiedenen Modi stattfinden können, nämlich einem affektiv-spontanen und einem rational-kalkulierenden. Diese Selektion des Modus findet sukzessive auf drei unterschiedlichen Ebenen statt (vgl. Kroneberg 2005). Erstens, auf der Ebene der Frames. Hier fällt die Entscheidung, ob man ei-

ne bestimmte *Definition der Situation* akzeptiert oder rationale Kalkulation einsetzt um Handlungsalternativen in Betracht zu ziehen. Zweitens, auf der Ebene der Skripte. Hier wird, wiederrum in einem der beiden Modi, zwischen verschiedenen „Programmen des Handelns" unterschieden. Drittens, auf der Ebene des Handelns selbst. Letztlich findet sich hier die klassische Entscheidung zwischen verschiedenen Handlungsalternativen, welche um die Möglichkeit einer Entscheidung zugunsten traditionaler oder affektiver Handlungsmodi erweitert wurde. Demzufolge kann man auf der Ebene der möglichen Definitionen der Situation (*Frames*) von einem intersubjektiv geteilten Wissensbestand sprechen, der einer Mehrzahl von Individuen erlaubt praktisch übereinstimmende Situationsdefinitionen zu finden. Im Falle der Skripte handelt es sich hingegen um geteilte Normen richtigen Verhaltens.

Symbolisches Wissen

Man kann gegenwärtig von einer starken Hinwendung der individualistischen Theorien zur Betrachtung kultureller und normativer Verhaltensregeln sprechen. Dabei wird das grundlegende Prinzip des methodologischen Individualismus jedoch nicht aufgegeben, sondern bildet den übergreifenden Rahmen für die Interpretation von Kultur. Somit wird es möglich die grundlegenden Eigenschaften und Axiome zu bestimmen die allen bisher aufgeführten Ansätzen in gewisser Weise gemein sind. Diese zentralen Grundannahmen der individualistischen Perspektive lassen sich entlang dreier analytischer Dimensionen herausarbeiten: welche Repräsentationsfunktion haben Symbole in diesen Theorien, wie entsteht und verändert sich die symbolische Ordnung und in welchem Verhältnis steht sie zu anderen sozialen Phänomenen. Im Rahmen der individualistischen Perspektive erscheinen soziale Symbole und deren Bedeutung, also die Kultur, Normen, und andere Informationen die in ihnen kodiert sind, stets als Phänomene im Inneren von Individuen. Des Weiteren ist der Wandel und die Genese von soziale Symbolen eine Funktion der auf individuellen Entscheidungen basierenden Handlungen und der daran anschließenden Interpretation durch andere Individuen. Dies läuft auf die Feststellung hinaus, dass die Auswahl sozial standardisierter Symbole sowie deren Umdeutung und Änderung durch entscheidungsfähige Individuen keine symbolisch vermittelten Prozesse sind.

Alle individualistischen Ansätze enthalten einen zumindest impliziten Bezug auf eine intersubjektiv geteilte, symbolische Ordnung, seien es die generellen Kategorien in denen die praktische Handlung abläuft, die

handlungsleitenden Skripte oder das Wissen von Personen. Diese Phänomene sind jedoch grundsätzlich im Individuum angesiedelt. Das Wissen einer sozialen Gruppe oder einer ganzen Gesellschaft erscheint somit als das, was entweder im konkreten Einzelfall, im Durchschnitt oder typischerweise von denjenigen Personen gewusst wird, die Mitglieder dieser Gruppe sind. Im Prinzip entspricht dies der bereits dargestellten Perspektive von Sassure, der ebenfalls den Sprachgebrauch strikt von der Sprache selbst unterschied. Für die Betrachtung von Normen, Werten und Institutionen hat dies zur Folge, dass diese nur dann als wirkmächtig angesehen werden, wenn die betreffenden Personen entweder keine Alternative sehen oder Ihnen aus guten Gründen folgen. Letztgenannte stellen, zumindest in der Rational Choice Perspektive, oft Sanktionsmächte dar, die eine Befolgung der Spielregeln rational werden lassen.

Die Annahme eines rein individuellen Wissens ist jedoch nicht unproblematisch. Ein entscheidender Schritt in der menschlichen Entwicklungsgeschichte war schließlich die Herausbildung der Sprache, als ein kodifizierter und standardisierter Korpus von Symbolen, mittels derer Wissen überindividuell gespeichert und individuell abrufbar wurde. Im Prinzip stellt jedoch bereits die orale Weitergabe von Geschichten eine technologische Abkehr von der individuellen Verankerung von Wissen dar. Dass die Eigenschaften und Befindlichkeiten von Individuen aber auch stets ein praktisches Problem für die Weitergabe von Wissen darstellten, wird an den zahllosen Versuchen deutlich dieses Problem in den Griff zu bekommen. Die Entwicklung von Versschemata und Erzähltechniken können als solche frühen Versuche verstanden werden, dem Wissen Persistenz und überindividuelle Gültigkeit zu verleihen. Spätestens aber mit der Erfindung der Schrift und ähnlicher Techniken, die eine vom Menschen unabhängige Aufzeichnung von Informationen möglich machten, kann man von einer gewissen Unabhängigkeit der symbolischen Ordnung von den auf dieser Basis handelnden Individuen ausgehen. Solange wir ein ausreichendes Wissen über die jeweilige Grammatik und die verwendete Sprache haben, sind wir dazu in der Lage Wissen zu entschlüsseln und uns anzueignen, das lange vor unserer Zeit geschaffen wurde und das so gut wie keinerlei Bezug zu den Regeln unseres Alltags oder den praktischen Anforderungen konkreter Situationen hat.[8]

[8] Die relative Einfachheit, mit der wir komplexe, symbolische Probleme lösen, wie zum Beispiel die Entschlüsselung von in Texten kodierten Bedeutungen oder das Erkennen einer Person aufgrund vager Beschreibungen, ist umso erstaunlicher wenn man die Probleme betrachtet, die Maschinen auch heute noch bei der Lösung dieser für uns so trivialen Aufgaben haben.

Gegen die Annahme eines die Individuen überschattenden Wissens wird von Seiten der individualistischen Perspektive oft der Einwand vorgebracht, dass es ja zumindest noch Individuen bräuchte, um dieses Wissen in Handlungen zu übersetzen. Diese Feststellung ist prinzipiell richtig und kann um den Zusatz erweitert werden, dass es auch keine Schriftzeugnisse ohne Schreibzeug und keine Lautsprache ohne Schallwellen geben kann. Allerdings ändert dies nichts an der Tatsache, dass symbolische Ordnungen relativ unabhängig von ihrem materiellen Substrat sind. Die Existenz eines Sprechers und einer Situation in der gesprochen wird sind eben noch keine hinreichenden Gründe für das was gesagt werden wird. Vor allem dann nicht, wenn es darum geht festzustellen, was welche Sprecher unter welchen Umständen sagen werden.

Methodologisch besteht hier zusätzlich das Problem, dass zwar die Existenz von Traditionen, Skripten und Wissen angenommen wird, diese jedoch von konkreten Handlungen und Individuen ausgehend rekonstruiert werden. Die Feststellung, dass Menschen in ihren Handlungen Skripten folgen und zwischen diesen wählen bleibt relativ inhaltsleer, wenn wir nicht angeben können wie genau die Menge an Skripten beschaffen ist, die einem Akteur typischerweise zur Verfügung steht. Ansonsten erscheint jegliche Abweichung von einem konkreten Handlungsmodell (zum Beispiel dem Modell der rationalen Wahl) post facto als die Folge eines nicht näher bestimmbaren Skriptes. Es erscheint problematisch von der Handlung selbst auf das Vorhandensein eines spezifischen Skriptes oder Wissens zu schließen. Vor allem deshalb, weil gleichartige Handlungen durchaus von unterschiedlichen Handlungslogiken angeleitet werden können. Eine Mehrzahl von Gründen, Verpflichtungen oder erwartetem Nutzen kann letztlich dieselbe Handlung nach sich ziehen. Jemanden zu begrüßen kann zum Beispiel aus Schmeichelei, Höflichkeit oder als Ausdruck der Zuneigung geschehen. Zugleich ändern diese unterschiedlichen Motive aber nichts an der allgemeinen Form der Begrüßung, da es den meisten Mitgliedern eines bestimmten Kulturkreises bekannt ist, welche sozialen Funktionen eine Begrüßung erfüllt und welche Bedingungen und Konsequenzen daraus hervorgehen.

Des Weiteren ist es auch schwierig handlungsleitende Skripte und individuelles Wissen, als isolierte Einheiten aufzufassen. Im obigen Beispiel impliziert das Wissen um das Skript der Begrüßung immer auch ein Wissen über eine Vielzahl anderer allgemeiner Regeln und sozialer Institutionen, wie eben Schmeichelei, Höflichkeit oder Zuneigung. Dies verweist, wie in der Betrachtung der allgemeinen Zeichentheorie schon festgestellt,

auf ein System von Zeichen in dem Bedeutungen aus den Relationen zwischen diesen Zeichen hervorgehen.

Die zweite Grundannahme der individualistischen Perspektive besteht in der Feststellung, dass die Entstehung und die Veränderung von Bedeutungen, Wissen und Institutionen entweder auf individuelle Entscheidungen oder spezifische Interaktionssituationen zurückzuführen ist. Damit geht oft eine Anerkennung von allgemeinen sozialen Institutionen einher, zum Beispiel im Sinne des *Mikro-Makro-Modells*, d.h. der „Colemannschen Badewanne". Die Bezeichnung erscheint dabei etwas irreführend, da die Makro-Ebene hier normalerweise nur als eine analytische Ebene vorkommt, die in der Sichtweise der Handlungstheorien vollständig aus der Mikro-Ebene heraus erklärt werden kann. Selbst in der moderaten Fassung gilt, „während das analytische Primat, also das letztlich interessierende Erklärungsobjekt, in der Soziologie auf der kollektiven Makro-Ebene liegt, findet sich das theoretische Primat, das erklärende theoretische Gesetz, auf der individuellen Mikro-Ebene" (Albert 2009: 22). Jegliche Auswirkungen der Makro-Ebene auf die Mikro-Ebene bestehen in dieser Perspektive normalerweise nur in der Form von strukturellen Rahmenbedingungen, die jedoch ebenfalls nur durch soziales Handeln vermittelt und wirkmächtig werden.

Eine solche prinzipielle Verankerung von Kausalität und Entwicklungsdynamik in der Interaktionssituation findet sich ebenfalls in der Phänomenologie und, wahrscheinlich am eindeutigsten, im Forschungsprogramm der Ethnomethodologie. Somit entfalten symbolische Ordnungen in dieser Perspektive keine eigene Dynamik jenseits der konkreten Handlungslogik der Individuen, welche sie in der Interaktion anwenden. Auch diese Konzeption erinnert stark an Sassures Vorstellung einer objektiven Sprache die von der Dynamik und den kausalen Effekten eines individuellen Sprachgebrauchs strikt unterschieden werden muss.

Zunächst einmal kann man dieser strukturierenden Wirkung symbolischer Ordnungen auf menschliches Handeln sicher zustimmen, wenn man unter den internalisierten Skripten bzw. dem Wissen einer Person das Vorhandensein einer kognitiven Vorstellung vom potentiellen Ablauf und den Konsequenzen einer Handlung versteht. Auch Institutionen funktionieren in dieser Sichtweise als eine innere Repräsentation von externen Begrenzungen, die sich auf die Einschätzung der Wahrscheinlichkeit auswirken mit der eine bestimmten Handlungsalternative zum Erfolg führt. Das gleiche gilt für Normen und Werte, die nur dann als wirkmächtig angesehen werden, wenn bei den Individuen ein entsprechender

„Glaube“ vorliegt. Kroneberg bedient sich zur Illustration dieses Sachverhaltes der Verfassungstreue:

> Das Grundgesetz wirkt immer noch ausschließlich über ›die kausale Kraft‹ der Akteure und ihres Handelns. Einerseits wirkt das Grundgesetzt als *subjektive* Vorstellung von und Glauben an eine legitime Ordnung. Andererseits ist dies gleichzeitig eine indirekte Wirkung anderer Akteure, insoweit dieser Glauben auf *Sozialisation*seinflüsse zurückzuführen ist und in sozialen Interaktionen aktualisiert (und dabei u.U. modifiziert oder gar aufgegeben) wird. (Kroneberg 2009: 230)

Dreh- und Angelpunkt ist demnach die individuelle Überzeugung von Individuen, die entweder auf deren eigenen Glaubenssätzen basiert oder auf der Annahme, dass diese Glaubenssätze bei anderen vorhanden sind. Diese Sichtweise führt in letzter Konsequenz zu einer sozialkonstruktivistischen Position, in der die Annahme vertreten wird, dass alles Soziale auch immer auf andere Art und Weise konstruiert sein könnte. Wie auch im Falle des philosophischen Relativismus, ist diese Position nicht grundsätzlich widerlegbar und zweifelsohne analytisch spannend. Jedoch muss die Frage erlaubt sein, wie plausibel eine solche Sichtweise vor dem Hintergrund der relativen Persistenz von Glaubenssätzen und Überzeugungen auch unter widrigsten Umständen ist. Ein kurzer Blick in den Alltag offenbart recht schnell, dass es zum einen die unterschiedlichsten und wechselseitig ausschließlichsten Weltbilder gibt, dass die Vielfalt dieser standardisierten, symbolischen Ordnungen jedoch in keinem Fall auch nur annähernd so groß wäre wie die Menge der Individuen die sich mit ihnen identifizieren. Tatsächlich scheint die Anzahl der dominanten Weltbilder so begrenzt, dass Forscher wie John Meyer (2005) sogar von einer „Weltkultur“ im Singular sprechen. Ein solches Ausmaß an Standardisierung überrascht nicht, wenn man bedenkt, dass es sich dabei um die grundlegende Funktion von informationstragenden Zeichen in sozialer Kommunikation handelt.

Die Annahme einer rein passiv strukturierenden Wirkung von Symbolen und symbolischen Ordnungen weist jedoch noch einen weiteren blinden Fleck auf. So wie individualistische Theorien das Soziale als in individuelles Wissen und Handlungen zerlegt betrachten, so zerfällt die spezifische symbolische Ordnung – die wir auch als *Kultur* bezeichnen können – in dieser Sichtweise ebenfalls in einzelne Skripte und getrennte Wissensbestände deren einziger Bezugspunkt die gemeinsame Koexis-

tenz in einem spezifischen Individuum ist. Dabei gerät der grundsätzlich, relationale Charakter des Wissens aus dem Blickfeld. So ist zum Beispiel die Verfassung kein in sich geschlossener Wissensbestand, sondern weist eine Vielzahl von Bezügen zu anderen symbolischen Konstrukten auf, wie zum Beispiel Normen, abstrakten Werten, rechtlichen Prozeduren, offenen Verfassungsfragen, Witzen, Kritiken und so weiter und so fort. All diese Relationen sind im Rahmen einer spezifischen, symbolischen Ordnung verankert und somit keineswegs zufällig oder beliebiger Natur. Wissen entsteht nicht spontan in einem Individuum, sondern aus bereits bestehendem Wissen und mittels standardisierter Zeichen und sozialer Symbolen. Die vorausgehende symbolische Ordnung gibt den Rahmen vor, benennt Leerstellen und verweist auf kritische Punkte. Das Grundgesetz der Bundesrepublik Deutschland steht nicht für sich. Es ist vielmehr die Folge eines historisch gewachsenen Wissens um Verfassungen, Menschenrechte, politische Prozeduren und dergleichen. Deswegen ist es nicht verwunderlich, dass eine solche überwältigende Ähnlichkeit zwischen den Verfassungen der unterschiedlichsten Staaten existiert (vgl. Beck, Drori und Meyer 2012), die durch die Intensität des kulturellen Austauschs zwischen Ländern erklärt werden kann (vgl. Goderis und Versteeg 2011).

Im Prinzip geht es hier um die Frage ob bestimmte Symbole oder symbolische Ordnungen einen größeren Anreiz auf Menschen ausüben als andere. Die Erforschung einer solchen Eigendynamik von Symbolen ist der Gegenstand der *Memetik*, einem relativ jungen Forschungszweig der Biologie in dem versucht wird die Modelle der genetischen Evolution auf die kulturelle Entwicklung zu übertragen. Ausschlaggebend hierfür sind insbesondere die Arbeiten von Susan Blackmore (2000), die den von Richard Dawkins im Jahre 1976 in seinem Buch „The Selfish Gene" geprägten Begriff übernahm und zu einem Forschungsprogramm ausbaute. Demnach ist ein Mem eine „Einheit der kulturellen Vererbung" (Dawkins 1996: 309), welche durch den Prozess der Imitation weitergegeben wird und dadurch in Analogie zu Genen die Eigenschaften eines „Replikators" (ebd.: 44) erhält, d.h. einer Einheit von Information, die das Potential (im stochastischen Sinne) besitzt Kopien seiner selbst herzustellen. Drei Kriterien müssen hierbei erfüllt sein: Variation, Selektion und Vererbung, die in ihrem Zusammenspiel einen evolutionären Prozess ergeben bei dem Meme um die Vorherrschaft im Mempool konkurrieren (vgl. Blackmore 2001).

Die Verbreitung der Meme geschieht mittels Imitation durch Menschen, die in gleicher Art als Kopiermaschinen für die Meme fungieren wie Organismen für genetische Kopien. Die individuellen Gründe für die Imitation bestimmter Meme können dabei vielfältiger Art sein. Kognitive Präferenzen, praktische Gründe und soziale Zwänge, um nur ein paar zu nennen. Für die memetische Betrachtung des evolutionären Prozesses ist dies jedoch prinzipiell nachrangig, da es ausreichend ist festzustellen, welche Überlebenswahrscheinlichkeiten ein Mem im Durchschnitt innerhalb einer bestimmten Population hat. In Anlehnung an die Theorie des „selfish gene" wird angenommen, dass Meme über eine klar definierbare Nutzensfunktion verfügen. Dies erlaubt die Analyse memetischer Populationen mittels Modellen der evolutionären Spieltheorie (vgl. Maynard-Smith und Price 1973).

Die Begrenzung der Memetik auf imitierbare Information ermöglicht nicht nur den Anschluss an allgemeine Formulierungen der Evolutionstheorie, sondern macht auch deutlich, dass Meme ein genuin soziales Phänomen sind:

> This means we can immediately exclude many things that a few authors have confusingly included as memes, such as perceptions, emotional states, cognitive maps, experiences in general, or '*anything that can be the subject of an instant of experience*'. Furthermore we can build on the long history of research in animal behaviour to distinguish imitation from contagion, and from individual and social learning, and so to eliminate from memetics the catching of yawns or all the many things we each learn for ourselves, by ourselves. (Blackmore 1998: Kap. 7)[9]

Die Definition eines Mems als eine imitierbare Einheit der Information entspricht somit sehr genau der hier verwendeten Definition eines Symbols als einer sozial standardisierten Bedeutungskodierung. Dies ist wenig verwunderlich, wenn man bedenkt, dass viele der bisher vorgestellten Überlegungen aus den Bereichen der Informations- und Zeichentheorie, ebenso wie die Elias'sche Symboltheorie in einem wechselseitigen Bezug zu Ideen der Kybernetik und der Evolutionstheorie stehen. Dennoch

[9] Der Begriff „social learning" umfasst in der Biologie nicht nur Lernen durch Imitation und Nachahmung, sondern auch durch Prozesse der Konditionierung in Gruppen, welche von Blackmore nicht als memetische Vorgänge im engeren Sinne angesehen werden, da hier keine Information weitergegeben wird (vgl. Blackmore 1998: Kap. 4).

muss man festhalten, dass es auch einige Unterschiede gibt. Hierzu zählt insbesondere die starke Ausrichtung auf einzelne Informationseinheiten als Erkenntnisgegenstand. Die Eigendynamik der Meme basiert letztlich auf den Eigenschaften der Meme ihr eigenes Überleben und ihre Kopierwahrscheinlichkeit zu erhöhen. Dies blendet jedoch bis zu einem gewissen Grad die Einbettung in ein Netzwerk von Symbolen – die symbolische Ordnung – aus. An dieser Stelle ist die Memetik hauptsächlich wegen ihres Verweises auf die grundsätzliche Eigendynamik von Symbolen (Memen) erwähnenswert. Es kann jedoch auch festgehalten werden, dass es sich im Prinzip um eine „Rational Choice" Theorie handelt, bei der es jedoch um die Bestimmung der Nutzensfunktionen von Memen geht, also keine Betrachtung der kognitiven Gründe stattfindet.

Individualistisches Wissen und Dynamik durch individuelle Handlungen, die bisher dargestellten Grundannahmen des individualistischen Modells, verweisen bereits auf die dritte perspektivische Festlegung, nämlich die zumindest analytisch praktizierte *Trennung von Denken und Zeichenprozessen*. Diese Tendenz ist vor allem bei denjenigen individualistischen Theorien zu beobachten, die eine Entscheidungsregel oder ein wie auch immer geartetes Prozedere der Auswahl zwischen verschiedenen symbolischen Alternativen als zentralen Gegenstandsbereich auffassen. Ein Paradebeispiel hierfür sind die bereits vorgestellten Modelle der Frameselektion, die zwar die Existenz von allgemeinen Skripten und Wissen anerkennen, aber diesen immer eine rationale Wahl voranstellen. In der Ethnomethodologie nimmt die praktische Rationalität eine ähnliche Rolle ein. Demgegenüber zeichnet sich das Verhältnis von Entscheidung und Wissen in der Wissenssoziologie eher durch eine Dominanz des Wissens aus, die als Lebenswelt den Horizont aller Entscheidungen darstellt. Dennoch bleibt innerhalb dieser Grenzen reichlich Raum für ein individuelles Kalkül.

Das bedeutet auch, dass das Entscheidungskalkül in dieser Perspektive relativ unabhängig von der symbolischen Ordnung ist. Demnach können wir uns Symbole, Wissen, Skripte und dergleichen aneignen und sie auf vielfältige Arten und Weisen einsetzen, sind jedoch immer auch dazu in der Lage uns dagegen zu entscheiden. Dies ist nur dann möglich, wenn die Maxime der Entscheidung nicht ebenfalls von der symbolischen Ordnung vorgegeben wird. Damit entspricht diese Sichtweise der *triadischen Zeichenrelation* der Semiotik, in der das Representamen von den Objekten und den Zeichen analytisch getrennt erscheint. Je nach Theorie werden die einzelnen Teile und damit das Verhältnis von Subjektivität zu Objektivität

sowie das Ausmaß der erreichbaren Intersubjektivität anders bestimmt. Tendenziell neigen individualistische Theorien jedoch zu einer stärkeren Betonung der Subjektivität und einer zentraleren Stellung des Representamen, bzw. der Bedeutungsebene.

Damit geht ein geschärftes Verständnis für die Rolle individueller Entscheidungen ein her. Menschen sind somit keine Automaten, die auf vorprogrammierten Bahnen ihre Kreise ziehen, sondern relativ eigenständige Entscheider. Die gesellschaftliche Entwicklung erscheint dann als Funktion von individuellen oder in Interaktionen getroffenen Entscheidungen über die Richtigkeit oder Falschheit sozialer Regeln und vordefinierten Wissens. Wandel und Kontingenz werden gegenüber Strukturen und Pfadabhängigkeiten stärker in den Blick genommen.

Spiegelbildlich zum semiotischen Dilemma und dem Problem der a priori, bleibt jedoch auch hier die Frage offen woher die Regeln stammen nach denen die Entscheidungen getroffen werden und wie sie im konkreten Fall aussehen. Zum Beispiel die Präferenzordnung, die einer rationalen Wahl zugrunde liegt oder die Kategorien nach denen über die Gültigkeit von Aussagen entschieden wird. Zwar werden diese Probleme auch im Rahmen der Handlungstheorie diskutiert, jedoch geschieht dies meistens mit Hinblick auf die Form in der sich die Entscheidung vollzieht (z.B. rational oder nicht-rational). Dabei ist es oft von vornherein ausgeschlossen, dass die Kategorien selbst sozial standardisiert sind. Stattdessen werden deren Ursachen in unbewussten kognitiven Prozessen (vgl. Kroneberg 2005: 347), persönlichen Biografien (vgl. Etzrodt 2007: 272) oder in der biologischen Grundausstattung des Menschen (vgl. Riedl 1981) vermutet.[10]

Damit geht die Vorstellung eines sehr basalen, affektuellen und spontanen Entscheidungskalküls einher, welches keinen Bezug zu symbolischen Prozessen aufweist. Menschliche Diskurse – beispielsweise über Fragen der Gerechtigkeit – basieren jedoch normalerweise nicht auf emotionalen Affekten, sondern auf dem Austausch von Argumenten auf der Grundlage abstrakter Symbole über einen längeren Zeitraum hinweg. Das ist jedoch sicherlich nicht immer und grundsätzlich der Fall. Vielmehr muss man annehmen, dass ein symbolisch vermitteltes und sozial standardisiertes Entscheidungskalkül umso wahrscheinlicher ist, je abstrakter der

[10] Für die letztgenannte Sichtweise spricht die Tatsache, dass es a priori Kategorien zu geben scheint, die nicht nur für Menschen universell sind. So kann man bereits bei Kapuzineräffchen einen grundlegenden Gerechtigkeitssinn finden, der sich gegen ungleiche Verteilungen richtet (vgl. de Waal und Berger 2000).

Gegenstand der Entscheidung ist. Somit würde man zum Beispiel von einer wissenschaftlichen Auseinandersetzung eine sehr viel präzisere, logisch stringentere Argumentation nach vorher feststehenden Regeln erwarten dürfen, von jenem sprichwörtlichen Streitgespräch in einer Paarbeziehung während eines Einkaufs in einem Möbelhaus.[11] Der wichtige Hinweis besteht hier in der Tatsache, dass die Erzeugung und Aufrechterhaltung symbolisch abstrakter Interaktionen über einen längeren Zeitraum nur durch Institutionen und soziale Strukturen möglich ist.

Zusammenfassend kann gesagt werden, dass die individualistische Position im Bezug auf Symbole tendenziell mittels dreier Merkmale beschrieben werden kann. Erstens existiert die symbolische Ordnung und erlangt ihre soziale Relevanz als Wissen, d.h. als eine persönliche, kognitive Repräsentation der Welt durch Zeichen. Zweitens sind der Wandel und die Genese dieses Wissens auf die Entscheidungen und Handlungen von Individuen zurückzuführen. Drittens sind diese Entscheidungen selbst nicht symbolischer Art, bzw. nicht durch Symbole vermittelt, stattdessen basieren sie auf intra-personellen Prozessen.

3.2.2 Gesellschaftliche Perspektive

Während die individualistische Perspektive hauptsächlich auf Max Weber zurückgeführt werden kann, ist die kollektivistische Sichtweise bezüglich sozialer Symbole eher mit den Arbeiten Emil Durkheims verknüpft. Insbesondere mit seinen Feststellungen zu den „sozialen Tatbeständen“, als sozial geteilte, objektive Gegebenheiten, die einen Zwangscharakter auf die Individuen ausüben und prinzipiell unabhängig von diesen existieren (vgl. Durkheim 1984: 101ff). Diese Externalität des Sozialen sieht er nicht zuletzt in deren Festlegung durch Zeichen und Sprache begründet:

> Die kollektive Gewohnheit existiert nicht nur im Zustand der Immanenz in den sukzesiven Akten, die sie bestimmt; vermöge einer Besonderheit, für die es im Bereich des Biologischen kein Beispiel gibt, wird sie ein für allemal in einer Form ausgedrückt, die von Mund zu Mund geht, durch Erziehung sich

[11] Die Metapher bezieht sich auf die Befolgung von Coding-Styles, welche die Lesbarkeit von formalen Sprachen durch Menschen sicherstellen sollen. Der unter Programmierern beliebte Webcomic *xkcd.com* verwendete sie zur Charakterisierung eines extremen Negativbeispiels: „It's like someone took a transcript of a couple arguing at IKEA and made random edits until it compiled without errors.“ Zu finden unter: `http://www.xkcd.com/1513/`.

> fortpflanzt und sogar durch die Schrift festgehalten wird. Solcherart sind der Ursprung und die Natur rechtlicher und sittlicher Gebote, der Sprichwörter und volkstümlicher Wendungen, der Dogmen, in denen religiöse oder politische Sekten ihr Glaubensbekenntnis festlegen, der Regeln des Geschmacks, die von literarischen Schulen festgehalten werden, usw. Keine dieser Normen geht vollkommen in den Anwendungen auf, die die Einzelnen von ihr machen, da sie ja vorhanden sein können, ohne wirklich angewendet zu werden. (Durkheim 1984: 109f)

Hier findet sich demnach eine Vorstellung einer symbolischen Ordnung, welche hauptsächlich aus Regel und Normen besteht, aber daneben auch noch grundlegende Kategorien, wie Geschmäcker und Stilistik oder moralische Gefühle umfasst. Diese Auffassung entspricht relativ genau der Konzeption sozialer Symbole als sozial standardisierte Zeichen, die in einer überindividuellen symbolischen Ordnung festgelegt sind. Allerdings ist es wichtig darauf hinzuweisen, dass der Begriff der sozialen Tatbestände bei Durkheim auch eine Reihe von anderen sozialen Faktoren umfasst. Hierzu zählen vor allem die strukturellen Bedingungen, wie zum Beispiel soziale Ungleichheit, die auch dann als soziale Tatbestände auf Menschen einwirken, wenn sie nicht in Symbolen kodiert werden. Die soziale Ungleichheit in der man lebt, die Ressourcen bereitstellt und vorenthält, muss unterschieden werden von der Bewertung sozialer Ungleichheit im Diskurs. Ersteres ist ein strukturelles Phänomen, letzteres ein symbolisches.

Des Weiteren muss präzisiert werden, dass einzelne, soziale Symbole nur in einigen wenigen Fällen eigenständige, soziale Tatbestände darstellen. Ein Gesetzestext besteht ja auch nicht nur aus einem Wort. Demzufolge ist es sinnvoller den sozialen Tatbestand mit den Bedeutungen gleichzusetzen, die durch eine Mehrzahl von Symbolen in nachvollziehbarer Weise kodiert werden. In der Perspektive Durkheims sind Symbole letztlich nur das *Medium*, in dem sich die sozialen Tatbestände ausdrücken und durch die sie Zwang auf die Individuen ausüben. Somit erscheinen die Symbole bei Durkheim als „*Modelle von* Wirklichkeit und als *Modell für* Wirklichkeit" (Hülst 1999: 132). Sie sind Abbilder der gesellschaftlichen Tatbestände und formen so die Sichtweisen der Individuen. Er weist darauf hin, dass selbst die grundlegendsten Kategorien, die a priori, wie sie von Kant, Hume und anderen Vertretern des Rationalismus untersucht worden sind, mitnichten transzendentalen Ursprungs sind. Vielmehr sind

es die gesellschaftlichen Kräfte und Tatbestände, welche die Kategorien des Denkens hervorbringen. Für ihn hat das einzelne Individuum schlicht „weder einen Grund, noch die Mittel, sie zu erlernen, zu reflektieren, auszudrücken und sie in distinkte Begriffe zu fassen“ (Durkheim 2007: 647). Tiere, so seine weitere Ausführung, sind schließlich auch dazu in der Lage, sich räumlich und zeitlich zu orientieren, Unterscheidungen zu treffen und zu lernen, sprich kausale Verknüpfungen herzustellen. Sie bewältigen all diese Lebensaufgaben problemlos ohne dabei über abstrakte Kategorien zu verfügen. Vielmehr machen diese nur als eine spezifische Lösung für das Probleme der Koordination von Menschengruppen Sinn:

> [Die Gesellschaft] ist nur möglich, wenn die Individuen und Dinge, die sie zusammensetzen in verschiedene Gruppen aufgeteilt, d.h. klassifiziert sind, und wenn diese Gruppe selbst in Bezug aufeinander, in Klassen eingeteilt sind. Die Gesellschaft setzt also eine bewusste Organisation ihrer selbst voraus, die nichts anderes ist als eine Klassifizierung. Diese Klassifizierung teilt sich natürlich dem Raum mit, den sie einnimmt. Um jeden Zusammenstoß zu vermeiden, braucht jede einzelne Gruppe einen bestimmten Raumanteil. Mit anderen Worten: der Gesamtraum muss aufgeteilt, unterschieden und ausgerichtet werden, und diese Einteilung und diese Ausrichtung müssen allen Menschen bewußt sein. Andererseits setzt jede Einberufung zu einem Fest, zu einer Jagd, zu einem Kriegszug voraus, daß man die jeweiligen Zeitpunkte fixiert und festgelegt hat und das folglich eine gemeinsame Zeit ausgebildet wurde, die alle Welt auf die gleiche Art und Weise versteht. Schließlich ist die Zusammenarbeit mehrerer zur Erreichung eines gemeinsamen Ziels nur möglich, wenn man sich über die Beziehung einig ist, die zwischen diesem Ziel und den Mitteln besteht, die es zu erreichen erlauben, d.h. wenn ein und dieselbe Kausalbeziehung von allen Beteiligten an der gleichen Aufgabe unterstellt wird. Es ist also nicht weiter erstaunlich, wenn die soziale Zeit, der soziale Raum, die soziale Klasse und die kollektive Kausalität den entsprechenden Kategorien zugrunde liegen, da die verschiedenen Relationen vom menschlichen Bewußtsein zunächst nur in ihren sozialen Formen mit einiger Klarheit erfasst worden sind. (ebd.: 648f)

Somit führt Durkheim das Vorhandensein dieser grundlegenden Kategorien in allen Menschen auf die Tatsache zurück, dass alle menschlichen Gesellschaften bestimmte fundamentale Koordinationsprobleme teilen und deshalb ähnliche Kategorien zu deren Lösung hervorbringen. Dies impliziert jedoch nicht, dass die konkreten Inhalte dieser Kategorien vollkommen gleich sein müssten. So ist z.B. das Raumverständnis der modernen Physik, die Raum und Zeit nicht länger als getrennte Qualitäten auffasst, ein fundamental anderes als dasjenige der alltäglichen Erfahrung. Auf den ersten Blick erscheint diese Formulierung sowohl materialistische, als auch funktionalistische Züge aufzuweisen. Materialistisch insofern, als dass die Kategorien aus den sozialen Lebensumständen, also aus objektiven Gegebenheiten hervorgehen und funktionalistisch, da sie als eine zweckmäßige Lösung eines spezifischen Problems aufgefasst werden.

Die konkrete methodische Umsetzung erfolgt bei Durkheim vor allem auf zwei Arten. Zum einen werden die grundlegenden sozialen Tatbestände durch die statistische Betrachtung kollektiver Merkmale rekonstruiert, dies geschieht vor allem in *Der Selbstmord* (Durkheim 1973). Zur Erfassung der grundlegenden symbolischen Ordnung – in der sich die sozialen Tatbestände ausdrücken – bedient sich Durkheim der Analyse von Vorschriften, Geschichten und Normen, die in Ritualen, Texten oder Überlieferungen sozial standardisiert und tradiert werden. Besonders deutlich wird diese Vorgehensweise in *Die elementaren Formen des religiösen Lebens* (Durkheim 2007). Dabei stellte er jedoch keine formalisierte Interpretationslehre, bzw. Methodologie der Textinterpretation auf, wie dies insbesondere in der qualitativen Sozialforschung heute üblich wäre. Einerseits mag dies einfach eine Folge der damaligen Auffassung von wissenschaftlicher Praxis gewesen sein. Andererseits folgt es auch aus Durkheims Auffassung des Symbolischen, als einem externen sozialen Tatbestand und seiner klaren Ablehnung des Rationalismus. Da für ihn soziale Tatbestände und nicht abstrakte Ideen die soziale Wirklichkeit bestimmen, muss ihm das Problem, welches sich aus der überindividuellen Interpretation individueller, geistiger Gehalte ergibt, als relativ gering erschienen sein. Soziale Tatbestände zeichnen sich ja dadurch aus, dass es sich um kollektiv geteilte Vorstellungen handelt, die genau deswegen auch einfach zu teilen und nachzuvollziehen sein müssen. In moderner, memetischer Perspektive: Symbole, deren Weitergabe von Generation zu Generation mit hohen Kosten der Interpretation verbunden wäre, hätten eine größere Chance nicht tradiert zu werden und würden somit schneller in Vergessenheit geraten.

Die Tradition der Betrachtung des Sozialen, als einem gesellschaftlichen Phänomen, kann man auch als *methodologischen Holismus* bezeichnen. Im Gegensatz zum methodologischen Individualismus, der die Ursache sozialer Phänomene primär in den Individuen, bzw. einzelnen Handlungen sieht, wird hier eine kausale Wirkung der Gesellschaft, beziehungsweise größerer Kollektive und Gruppen, auf die Individuen angenommen. Hierfür sind vor allem zwei Annahmen ausschlaggebend. Zum einen wird das Soziale als etwas den Individuen externes, unabhängiges und damit letztlich objektives angesehen. Zweitens, übt diese externe Vorgabe einen Zwangscharakter auf die Individuen aus. Damit sind jedoch keineswegs nur Sanktionen gemeint, sondern auch der Mangel an sozial legitimierten Alternativen zu bestimmten Handlungen oder auch strukturelle Rahmenbedingungen, die alternative Handlungen ausschließen. Wie schon im Falle des methodologischen Individualismus diskutiert, gibt es hier natürlich auch eine Vielzahl von Abstufungen und auch eine weiterführende Fraktalisierung kann ausgemacht werden. Letztlich handelt es sich bei der Beschreibung dieses erkenntnistheoretischen Bruchs eben auch nur um ein perspektivisches Hilfsmittel, dass die Unterschiede zwischen verschiedenen Theorierichtungen hervorheben soll, um die hier relevante Frage nach möglichen Anschlüssen für die Analyse sozialer Symbole im Allgemeinen und der quantitativen Textanalyse im Besonderen zu beantworten.

Gesellschaftstheorien

Den wahrscheinlich exzessivsten Gebrauch des Begriffs des Symbols oder symbolischer Eigenschaften im Rahmen einer Gesellschaftstheorie findet man in den Arbeiten von Pierre Bourdieu. Dabei scheint er Symbole jedoch nicht als einen eigenständigen Gegenstandsbereich zu betrachten, da er den Begriff hauptsächlich als eine Eigenschaft grundsätzlicher Phänomene auffasst. So spricht er zum Beispiel des Öfteren von „symbolischer Herrschaft“ und „symbolischem Kapital“, aber verhältnismäßig selten von den „Symbolen“ selbst. Zudem verwendet er diese Begrifflichkeiten nicht immer auf dieselbe Art und Weise. In *Die feinen Unterschiede* erscheint symbolisches Kapital als eine Art der Repräsentation anderer Kapitalformen, wie zum Beispiel ökonomischem oder kulturellem Kapital (vgl. Bourdieu 2003: 438ff). Es sind demnach Handlungen, Artefakte und Kunstwerke, die man sich aneignen kann und durch die man seine zur Verfügung stehenden Kapitalien ausdrücken kann. Eine „stilvolle“ Inneneinrichtung oder „ausgesuchte“ Malereien, ebenso wie „hausgemachtes“

Essen werden damit als Ausdrücke einer bestimmten sozialen Position aufgefasst, die den „Besitzer" charakterisieren und sein Kapitalvermögen symbolisieren. Demgegenüber stellt symbolisches Kapital in *Die Regeln der Kunst* eine eigenständige Kapitalsorte im Feld der Kulturproduktion dar (vgl. Bourdieu 2001: 228f). Die Trennung vom ökonomischen Kapital ist in dieser Fassung absolut und vollständig, da das Feld der Kunst durch die Unterscheidung von ökonomischer („Mainstream") und symbolischer Logik („reine Kunst") erst konstituiert wird.

Trotz dieser unterschiedlichen Fassungen des Symbolbegriffs lässt sich ein gemeinsamer Kern ausmachen. In noch viel stärkerem Ausmaß als Durkheim, fasst Bourdieu *Symbole als Repräsentationen sozialer Strukturen* auf. Dies wird insbesondere in *Zur Soziologie der symbolischen Formen* (Bourdieu 1974) deutlich. Der Titel des Buches legt eine Beziehung zu Ernst Cassirers Hauptwerk *Philosophie der symbolischen Formen* nahe. Allerdings finden sich im Verlaufe der Argumentation so gut wie keinerlei Verweise auf die Arbeiten von Cassirer.[12] Die wesentliche Gemeinsamkeit welche die beiden Werke verbindet besteht in der Frage nach den grundsätzlichen Strukturierungsprinzipien mittels derer wir die Welt wahrnehmen. Im Falle Cassirers sind dies die symbolischen Formen, die sinnliche Zeichen mit inneren Geisteszuständen verknüpfen (vgl. Cassirer und Schmücker 2001: 161). Demgegenüber stellt Bourdieu symbolische Formen als eine Verknüpfung von individuellen und kollektiven Bewertungsschemas dar, die auf einer Übersetzung der Sozialstruktur in ein System von Symbolen basiert:

> So läßt die Logik dieses Systems der Äußerungen und Signalements sich nicht unabhängig von seiner Funktion begreifen, d.h. als eine symbolische Übersetzung des sozialen Systems, eines Systems von ‚Einschluß und Ausschluß' [...], als Zeichen von Gemeinschaft und Unterscheidung, das ökonomische Güter in Symole und auf ökonomische Ziele gerichtete Handlungen in kommunikative Akte (die auch Ausdruck einer Verweigerung der Kommunikation sein können) verwandelt. Nichts wäre in der Tat irriger als die Annahme, die symbolischen Handlungen (bzw. deren symbolischer Aspekt) bedeuten nichts außer sich selbst: Sie verleihen stets der der sozia-

[12] Wie auch schon im Falle des Untertitels von „Die feinen Unterschiede. Eine Kritik der gesellschaftlichen Urteilskraft" scheint es Bourdieu eher um eine soziologische Aufarbeitung eines philosophisch besetzten Themas gegangen zu sein und in keinster Weise um eine Auseinandersetzung mit der entsprechenden Philosophie.

> len Stellung Ausdruck und zwar gemäß einer Logik, die eben die der Sozialstruktur selbst ist, d.h. die der Unterscheidung. (Bourdieu 1974: 62)

Als Bindeglied zwischen dem kollektiven System der Unterscheidungen und deren individueller Reproduktion sieht Bourdieu den *Habitus*, den er in Anlehnung an Noam Chomsky (1964) als eine „generative Grammatik" beschreibt (vgl. Bourdieu 1974: 143). Gemeint ist damit ein den Individuen nicht bewusst zugängliches Regelsystem, welches aus einer endlichen Anzahl von Symbolen (im Falle der Sprache: Lexeme) eine potentiell unendliche Menge an Sätzen hervorbringen kann, die in einer spezifischen Sprache sinnhaft formuliert werden können. Bourdieu stellt sich analog dazu den Habitus als ein System von Regeln vor, welches in der Lage ist „alle typischen Gedanken, Wahrnehmungen und Handlungen einer Kultur zu erzeugen - und nur diese" (ebd.: 143). Es wird hier jedoch auch deutlich, dass er in einem entscheidenden Punkt von Chomskys Vorstellungen abweicht, da er das Regelsystem des Habitus in erster Linie als eine Begrenzung der Möglichkeiten der Kommunikation auf diejenigen Unterscheidungen ansieht, die bereits in der Sozialstruktur vorgezeichnet sind.

Demgegenüber betont Chomsky, dass es sich gerade nicht um ein Modell der Satzkonstruktion im Rahmen kultureller Vorstellungen handelt:

> To avoid what has been a continuing misunderstanding, it is perhaps worth while to reiterate that a generative grammar is not a model for a speaker or a hearer. It attempts to characterize in the most neutral possible terms the knowledge of the language that provides the basis for actual use of language by a speaker-hearer. When we speak of a grammar as generating a sentence with a certain structural description, we mean simply that the grammar assigns this structural description to the sentence. When we say that a sentence has a certain derivation with respect to a particular generative grammar, we say nothing about how the speaker or hearer might proceed, in some practical or efficient way, to construct such a derivation. These questions belong to the theory of language use — the theory of performance. No doubt, a reasonable model of language use will incorporate, as a basic component, the generative grammar that expresses the speaker-hearer's knowledge of the language; but this generative grammar does not, in itself,

> prescribe the character or functioning of a perceptual model or a model of speech production. (Chomsky 1964: 9)

Da die konkreten symbolischen Ordnungen, in denen Menschen leben und durch die sie interagieren, bei Bourdieu durch die Sozialstruktur bestimmt werden, nimmt deren direkte Untersuchung in seinen Arbeiten relativ wenig Raum ein. Wenn überhaupt, so dienen sie nur als Indikatoren für die zugrundeliegenden Prinzipien der Sozialstruktur und des sie ausdrückenden Habitus. Hierin zeigt sich die materialistische Grundausrichtung der Bourdieuschen Theorie am deutlichsten.[13] Die symbolischen Ordnungen sind somit unbewusste Handlungs- und Bewertungsprogramme, die wiederum ein Produkt der gesellschaftlichen Verteilung wertvoller Güter sind. Dabei betont er, dass die Strategien und Regeln, die im Habitus inkorporiert sind, weder auf kollektiver noch auf individueller Ebene bewusst sind (Bourdieu 1974: 139). Vielmehr basiert der Erhalt der gesellschaftlichen Zustände gerade auf der fortlaufenden Verkennung der eigentlichen Gründe für die Spielregeln. Diesen Mechanismus bezeichnet er als die „Illusio" des Spiels (vgl. Bourdieu 2001: 363).

Neben Bourdieus feldtheoretischem Ansatz gibt es gerade in jüngerer Zeit eine Vielzahl weiterer Theorien, die sich auf das Konzept des Feldes beziehen, wie zum Beispiel die „Strategic Action Field Theory" von Fligstein und MacAdam (2011) oder der „Structurisation" Ansatz von Anthony Giddens (1984). John Levi Martin (2003) verdanken wir in diesem Kontext die Herausarbeitung der grundsätzlichen Charakteristika einer allgemeinen Feldtheorie. Als bezeichnend sieht er vor allem die Annahme, dass es ein grundlegendes Verteilungsprinzip (das Feld) gibt, welches eine Funktion der Elemente ist, die das Feld konstituieren und gleichzeitig von diesem erzeugt werden (vgl. ebd.: 3ff). Beobachtbar ist das Feld jedoch nur indirekt, durch die Verteilung der einzelnen Elemente. Diese Vorstellung der *wechselseitigen Hervorbringung von Feld und Feldeffekten* (d.h. Elemen-

[13] Dies wird in Bourdieus Nachwort zur französischen Übersetzung von Erwin Panofskys „Gothic Architecture and Scholasticism", welches als viertes Kapitel in Bourdieus (1974: 125ff) Arbeit *Zur Soziologie der symbolischen Formen* wiedergegeben ist, besonders deutlich. Hierin geht Bourdieu der Frage nach, wie und auf welche Weise sich ein einheitlicher gotischer Stil im Kathedralenbau herausgebildet hat. Ausschlaggebend ist seiner Meinung nach die Veränderung in den Schreibpraktiken der Scholastiker, welche sich weg von den rigiden und redundanten Formen der *Summa* und hin zu den präziseren und eleganteren Formen der *Lectio* und *Disputatio* entwickelte. Die analoge Verschiebung der Stilistik hin zu eleganteren und gleichzeitig klareren Linien erklärt er mit dem Habitus als dem entscheidenden Bindeglied zwischen der sich verändernden scholastischen Praxis und der Herausbildung einer neuen kollektiven Ästhetik.

ten) hat der Feldtheorie den Vorwurf der Tautologie oder zumindest eines unklaren ontologischen Status eingebracht (vgl. ebd.: 8ff).

Hinsichtlich ihrer Einschätzung der Rolle symbolischer Ordnungen ähneln sich diese neueren feldtheoretischen Ansätze sehr. Sie nehmen die Form von „Spielregeln", „Legitimationen" und „symbolische Ressourcen" an, die im Feld produziert und konsumiert werden. Im Unterschied zu Bourdieus Auffassung wird der Genese und Existenz symbolischer Ordnungen bei anderen Autoren jedoch weit weniger Aufmerksamkeit geschenkt. So sind die kulturellen Legitmationen und Spielregeln zwar auch bei Fligstein und McAdam (2011: 12) essentieller Bestandteil und umkämpfte Ressource des Spiels, jedoch findet sich dort wenig über die Herkunft dieser Regeln. Letztlich deuten aber auch diese Fassungen der Feldtheorie an, dass die symbolischen Ordnungen ein *Ergebnis der zugrundeliegenden Feldstrukturen* sind und dies sowohl in ihrer kollektiven Variante, als Kultur und übergreifende Spielregeln, wie auch in individueller Form, als inkorporiertes (Orientierungs-)Wissen der Akteure.

Die Annahme, dass symbolische Ordnungen durch das Feld bestimmt werden, wie sie insbesondere in Bourdieus Feldtheorie vertreten wird, bringt ein tiefgreifendes Problem mit sich. Empirisch ist des Öfteren zu beobachten, dass der Wandel der Kapitalverteilungen, also die Verschiebung der Ungleichheitsstruktur des Feldes, nicht zwangsläufig zu einer Veränderung der Spielregeln oder der symbolischen Bewertungen führen muss. Das Problem einer solchen *Entkopplung von symbolischem und sozialem Wandel* versucht Bordieu durch den „Hysteresis Effekt des Habitus, demzufolge auch einem veränderten Stand des Titel-Marktes noch die Wahrnehmungs- und Bewertungskategorien appliziert werden, die einem früheren Stand der objektiven Chancen der Einschätzung entsprachen" (Bourdieu 2003: 238), zu begegnen. Dahinter steht die Auffassung, dass der Habitus durch die Sozialisation festgelegt wird und dem Bewusstsein nur bedingt zugänglich ist. Dies hat zur Folge, dass der Wandel der symbolischen Ordnung aus dieser Perspektive heraus eigentlich nur im Wechsel der Generationen denkbar ist.

Aus dieser Konzeption ergeben sich zwei Probleme. Zum einen müsste eine solche Zeitverschiebung deterministisch sein, d.h. der symbolische Wandel müsste sich in bestimmbaren, regelmäßigen Abständen als Reaktion auf einen sozialen Wandel vollziehen.[14] Des Weiteren dürfte es

[14] Gegen dieses Argument ließe sich einwenden, dass die Geschwindigkeit mit der die objektiven Verteilungsstrukturen in den Habitus inkorporiert werden ebenfalls einem Wandel unterliegen kann. Allerdings wäre es in diesem Fall notwendig eine genauere und empi-

keine Fälle geben, in denen ein sozialer Wandel auf lange Sicht nicht zu einem entsprechenden symbolischen Wandel geführt hätte oder gar eine Verschiebung der symbolischen Ordnung, die einer Veränderung der Feldstruktur vorausgegangen wäre. Insbesondere gegen letzteres wird oft der Einwand vorgebracht, dass Bourdieus Feldtheorie als Prozess angelegt sei und man daher gar nicht von einem richtigen „vorher" und „nachher" sprechen könne. Hierzu muss man festhalten, dass dies zwar für die Feldtheorie im Allgemeinen gelten mag, aber eben nicht für den Bereich des Wissens und der Kultur, da diese als bloße Repräsentationen der objektiven Feldstruktur aufgefasst werden. Letztlich befinden wir uns somit wieder im semiotischen Dilemma und zwar auf der Seite des Materialismus in dem die Annahme vertreten wird, dass die Dinge ursächlich für die Zeichen sind.

Die zweite, gesellschaftstheoretische Perspektive auf Symbole sind die Systemtheorien, die mit der Durkheimschen Theorie vor allem der Fokus auf die Erklärung gesellschaftlicher bzw. systemischer Phänomene sowie die Verwurzelung in der biologischen Theorie, insbesondere der Evolutionsbiologie, verbindet. Die Abgrenzung zwischen der Feld- und der Systemtheorie ist nicht immer sehr eindeutig und der generelle Fokus auf die Ebene emergenter, sozialer Phänomene hat anscheinend dazu geführt, dass manche Autoren Feld und System wie Synonyme behandeln (z.B.: Martin 2003: 9). Da ein ausführlicher Vergleich dieser Theorierichtungen hier nicht notwendig oder sinnvoll erscheint, soll im folgenden vor allem auf zwei zentrale Unterschiede eingegangen werden, die im Hinblick auf die Rolle symbolischer Ordnungen besonders bedeutsam sind. Dies sind zum einen die zentralere Stellung von Kommunikation und Information im theoretischen Gebäude der Systemtheorien und zum anderen die systemtheoretische Vorstellung der Eigenlogik und der geschichtlichen Entwicklung sozialer Phänomene, die sich bei genauerer Betrachtung stark von ihren feldtheoretischen Gegenstücken unterscheiden.

Generell verstehen Systemtheoretiker unter einem System ein emergentes Phänomen, welches durch die *Relationen seiner Elemente* hervorgebracht wird und auch dadurch gekennzeichnet ist:

> Since a social system is a system of processes of interaction between actors, it is the structure of the relations between the actors as involved in the interactive process which is essentially

risch überprüfbare Sozialisationstheorie zu entwickeln, da sonst nur eine Verwässerung der ursprünglichen Aussage erreicht wäre.

> the structure of the social system. The system is a network of such relationships. (Parsons 1964b: 25)

In ähnlicher Weise werden Systeme auch von anderen Autoren der systemtheoretischen Tradition aufgefasst. So geht Richard Münch im grundlegendsten Fall von „Beziehungen zwischen Elementen" verschiedener Mengen von Ereignissen aus (z.B.: Münch 1986: 15) und Niklas Luhmanns Theorie behandelt den Sonderfall von autopoietischen „Systemen, die nicht nur ihre Strukturen, sondern auch die Elemente, aus denen sie bestehen, im Netzwerk eben dieser Elemente selbst erzeugen" (Luhmann 1997: 65). Auch wenn sich systemtheoretische Ansätze in einer Vielzahl von Punkten unterscheiden, so treffen sie sich doch in diesen basalen Grundverständnis. Dies gilt für die biologischen Theorien zu Systemen ebenso wie für die soziologischen oder diejenigen der Physik. Für Handlungssysteme in gleichem Maße wie für Funktionssysteme.

Die grundlegende Idee, dass die konkreten Beziehungen zwischen den Elementen ausschlaggebend für die Strukturen sind, ist der Grund für die zentrale Stellung, die Begriffe wie Komplexität, Emergenz und Funktion in der allgemeinen Systemtheorie einnehmen. Sie alle können als Versuch aufgefasst werden das Zusammenspiel zwischen den Eigenschaften eines Systems und dem spezifischen Beziehungsgeflecht aus dem es besteht fassbar zu machen. In diesem Punkt unterscheiden sich Feld- und Systemtheorie wahrscheinlich am deutlichsten. Für die Feldtheorie ist die Beziehung zwischen Elementen und Feld ausschlaggebend, während im Fall der Systemtheorie die spezifischen Relationen zwischen den Elementen das System bestimmen. Im erstgenannten Fall versucht man die Eigenschaften des Feldes mittels einer Beobachtung der Elemente zu erfassen, zum Beispiel indem man aus der Anordnung von Eisenspänen auf die Struktur eines spezifischen Magnetfeldes schließt. In der Systemtheorie wird demgegenüber versucht aus dem Zusammenspiel spezifischer Elemente Aussagen über die Konstitution des Systems zu treffen. Dies ist auch der Grund für die systemtheoretische Faszination mit Regelkreisläufen, Programmen, Mechanismen und nicht zuletzt Funktionen. Der Versuch solcher Erklärungen durch Mechanismen, wie sie typisch sind für die aus der Kybernetik hervorgegangene Systemtheorie, ist jedoch grundsätzlich inkommensurabel mit einer feldtheoretischen Perspektive: „field theories are proposed, whether reluctantly or not, when no such mechanistic explanations currently offer promise: if there were a mechanism, there would be no need for a field theory" (Martin 2003: 12).

Neben den Beziehungen der Elemente zueinander ist das *Wechselspiel von System und Umwelt* eine weitere grundlegende Gemeinsamkeit der systemtheoretischen Perspektiven. Wenngleich die Details des Verhältnisses von System und Umwelt, welche auch immer andere Systeme umfasst, sehr unterschiedlich aufgefasst werden kann. Die gängigen Konzeptionen reichen dabei von Input/Output Schemata über Interdependenzen und Interpenetration hin zu autopoietischer Geschlossenheit.

Soziale Symbole und symbolische Ordnungen scheinen vor allem an solchen Schnittstellen und in der Form von Interdependenzen ihren Platz in der Systemtheorie zu haben. Dies wird insbesondere im Rahmen einer handlungstheoretisch ausgelegten Betrachtung deutlich, wie sie insbesondere von Talcott Parsons und im deutschsprachigen Raum von Richard Münch vorgeschlagen wurde. Die Grundidee ist hierbei die Verortung des Sozialsystems am Schnittpunkt individueller Handlungen und abstrakter Kultur:

> Reduced to the simplest possible terms, then, a social system consists in a plurality of individual actors interacting with each other in a situation which has at least a physical or environmental aspect, actors who are motivated in terms of a tendency to the "optimization of gratification" and whose relation to their situations, including each other, is defined and mediated in terms of a system of culturally structured and shared symbols. (Parsons 1964b: 5f)

Die „shared order of symbolic meanings" (ebd.: 11) stellt dabei die Grundlage für jegliche Kommunikation und Interaktion dar. Gleichzeitig legt Parsons sehr großen Wert auf die Unterscheidung des kulturellen Systems vom sozialen System. Er begründet dies auf zweierlei Arten. Zum einen sind kulturell definierte Symbole und Bedeutungen nicht an den Rahmen konkreter, sozialer Systeme gebunden (vgl. ebd.: 15). Vielmehr sind sie transmissiv, d.h. sie können die Grenzen sozialer Systeme überwinden und in sehr verschiedenen Kontexten auftreten. Im Grunde genommen verweist dies auf eine grundsätzliche Eigenlogik der übergreifenden symbolischen Ordnung. Zweitens, hat das kulturelle System keine Funktion außerhalb eines konkreten Handlungssystems, „it just ‚is"' (ebd.: 17). Hier wird betont, dass sich die Konstitution des symbolischen Systems nicht aus den funktionalen Erfordernissen sozialer Systeme ableiten lässt. Der Grund ist vor allem die historische Genese des kulturellen Systems, das den konkreten Individuen und Handlungen vorausgeht.

Damit stellt das kulturelle System stets den alleinigen Rahmen seiner eigenen Bewertung dar.

In Richard Münchs Arbeiten wird diese Differenz noch analytisch vertieft. Er unterscheidet zwischen der Symbolkomplexität, also der Menge der zu einem konkreten historischen Zeitpunkt vorhandenen Symbole, die potentiell in einer Handlungssituation gewählt werden könnten, und der Handlungskontingenz, d.h. der Menge der Handlungsalternativen, die einem Akteur in einer konkreten Situation offen stehen (vgl. Münch 1986: 15).

Des Weiteren wird Kultur hier ebenfalls als ein gesellschaftliches System begriffen, welches den Individuen übergeordnet ist. Insofern ähnelt der grundlegende Ansatz dem Modell von Parsons. Jedoch stellt Richard Münch die Bedeutung der symbolischen Ordnung sehr viel stärker in das Zentrum seiner Überlegungen, als dies bei anderen Systemtheoretikern geschieht. Auch in anderer Hinsicht ist seine Auffassung der Funktionsweise von Kultur näher an der hier angestrebten Synthese der individualistischen und der gesellschaftlichen Perspektive auf soziale Symbole. Aus diesem Grund werden seine Überlegungen in Abschnitt 3.3.1, in Zusammenhang mit der prozesssoziologischen Perspektive, näher ausgeführt und mit den hier getroffenen Feststellungen kontrastiert. Für den Moment reicht es festzuhalten, dass es sich auch hier primär um eine gesellschaftliche Perspektive handelt, welche Kultur, bzw. die symbolische Ordnung, als einen umfassenden sozialen Raum auffasst, der den Rahmen für das soziale Handeln darstellt.

Wenngleich Niklas Luhmanns Systemtheorie um eine Abgrenzung zu den gerade dargestellten Ansätzen bemüht ist, werden auch hier Symbole, Zeichen und Signale als grundlegende Elemente des Sozialen aufgefasst. Er fasst sie unter dem Begriff der „sinnhaften Identitäten (empirische Objekte, Symbole, Zeichen, Zahlen, Sätze usw.)“, welche in *rekursiven Operationen des Mediums Sinn* erzeugt und überliefert werden:

> Sinn ist demnach eine durch und durch historische Operationsform, und nur ihr Gebrauch bündelt kontingente Entstehung und Unbestimmtheit künftiger Verwendung. [...] In der rekursiven Erzeugung von Sinn wird diese Rekursivität vor allem durch die Worte der Sprache geleistet, die in einer Vielzahl von Situationen als dieselben verwendet werden können. (Luhmann 1997: 47f)

Sinn nimmt im Rahmen der Luhmannschen Systemtheorie eine besondere Stellung ein, da sich hierin sowohl die Operationen psychischer als auch sozialer Systeme vollziehen und dadurch die strukturelle Kopplung beider Systeme ermöglicht wird. Dennoch bleiben beide Ebenen aufgrund ihrer Autopoiesis grundsätzlich getrennt (vgl. Luhmann 1997: 100ff). Strukturelle Kopplung bedeutet demnach nur, dass in beiden Systemen Strukturen bestehen, die eine wechselseitige Bearbeitung von Irritationen ermöglichen. Aufgrund der unterschiedlichen Geschwindigkeiten mit denen soziale und psychische Systeme operieren und des überindividuellen Charakters von Sinn, kann jedoch geschlossen werden, dass die Entstehung von sinnhaften Identitäten auf der Ebene des psychischen Systems anzusiedeln wäre.

Da sich Kommunikation, der zentrale Baustein der Luhmannschen Systeme, im Medium Sinn vollzieht ist diese fundamental auf die Bereitstellung von Anschlussfähigkeit durch eine symbolische Ordnung angewiesen. Nur durch standardisierte Formen und relativ feste Unterscheidungen kann gewährleistet werden, dass in irgendeiner Art und Weise auf vergangene Kommunikationen aufgebaut werden kann. Somit erlauben Symbole nicht nur eine Kopplung von psychischen und sozialen Systemen, sondern auch die Verkettung der jeweiligen, systeminternen Operationen des Denkens und Kommunizierens.

Luhmann geht davon aus, dass Symbole zwei Arten von Kopplungsmechanismen bereitstellen. Zum einen durch eine Sprache, bzw. ein Zeichensystem, welches „psychisch unreflektiert und sozial unkommentiert funktioniert“ (ebd.: 110). Hierunter fallen die jeweiligen Typen eines Zeichensystems und die im weitesten Sinne grammatikalischen Regeln der Komposition dieser Typen. Dies entspricht relativ genau den Vorstellungen einer „generativen Transformationsgrammatik“ (vgl. Chomsky 2002), d.h. einem Regelsystem, welches *unabhängig vom semantischen Gehalt* konkreter Sätze die Regeln zur Kodierung von Information bereitstellt. Dabei ist auch nicht davon auszugehen, dass es sich hierbei um eine reine Schriftgrammatik handeln müsste. Auch die Kompositionsregeln von Bildern, Musik und andere symbolische Ordnungen sind hierbei denkbar. Im Unterschied dazu beschreibt er „Schemata“ als den zweiten symbolischen Kopplungsmechanismus:

> In einem schlecht koordinierten Forschungsgebiet hat er auch viele andere Namen, zum Beispiel „frames“, „scripts“, „prototypes“, „stereotypes“, „cognitive maps“, „implicit theories“ - um nur einige zu nennen. Diese Begriffe bezeichnen Sinn-

> kombinationen, die der Gesellschaft und den psychischen Systemen dazu dienen ein Gedächtnis zu bilden, das fast alle eigenen Operationen vergessen, aber einiges in schematischer Form doch behalten kann und wiederverwenden kann. (Luhmann 1997: 110f)

Hier handelt es sich demnach um eine Anreicherung von sprachlichen Zeichen und Grammatik mit Bedeutungen. Zum Beispiel die Assoziation einer Waage mit Gerechtigkeit oder bestimmte Bewertungen, die auch eine Handlungsaufforderung beinhalten, zum Beispiel „gerecht" oder „ungerecht".

Diese Schemata können jedoch nicht mit Kultur oder Wissen gleichgesetzt werden. Vielmehr handelt es sich, übersetzt in die hier verwendeten Begrifflichkeiten, um relativ feste Verknüpfungen von Symbolen die in das Netzwerk einer bestimmten symbolische Ordnung eingeschrieben sind. Aufgrund Luhmanns Annahme eines Gesellschaftssystems als übergreifendem, sozialen System, kann es für ihn jedoch nur eine einzige symbolische Ordnung geben. Kulturen und Wissensordnungen existieren jedoch im Plural, dies kann man allein schon daran erkennen, wie sehr sie sich widersprechen. Da dieser Widerspruch in unterschiedlichen Graden und Spezifität vorliegen kann, ist er auch nicht durch eine Dichotomisierung Luhmannscher Prägung aufzulösen. Zumindest nicht ohne einen fundamentalen Verlust an Information. Des Weiteren nimmt er an, dass Schemata wie auch Sprache unabhängig und unbemerkt von den Operationen psychischer und sozialer Systeme existieren. Dies ist, wie er selbst betont, eine notwendige Folge aus der Annahme der Autopoiesis. Wäre es psychischen Systemen möglich, die Inhalte der Kommunikation direkt nachzuvollziehen und direkt auf sie einzuwirken, so käme das einer Verletzung der System-Umwelt Differenz gleich. Daraus folgt für ihn, dass es sich bei Sprache und Schemata nicht um Systeme handeln kann, sondern vielmehr um „symbolische Generalisierungen" im Sinne Parsons (vgl. ebd.: 112). Wie dieser setzt Luhmann ebenfalls die symbolische Ordnung als gegeben voraus. In gewisser Weise gilt auch hier: „It just is".

Neben den System- und Feldtheorien gibt es noch eine Reihe weiterer Ansätze die Kultur und symbolische Ordnung als ein Kollektivmerkmal auffassen. So zum Beispiel der „World-Polity" Ansatz von John Meyer (Meyer 2005). In der deutschen Übersetzung wird dieser Ansatz sogar unter der Bezeichnung *Weltkultur* geführt. Dies ist jedoch in gewisser Weise irreführend, da es in diesem Forschungsansatz hauptsächlich um die Analyse der leitenden Annahmen geht, die dem Handeln poli-

tischer Akteure (Staaten, NGOs und dergleichen) zugrundeliegen. Von dieser Einschränkung einmal abgesehen weist auch dieses Forschungsprogramm Merkmale der Sichtweise auf die hier als gesellschaftliche Perspektive bezeichnet wurde und hätte es somit, neben einer Vielzahl anderer Ansätze, ebenfalls verdient näher erläutert zu werden. Da es hier jedoch vorrangig um die Illustration der allgemeinen Merkmale geht erscheint es nicht zweckmäßig alle Fälle einer Kategorie abzubilden. Stattdessen können die Feld- und Systemtheorie im Sinne der bereits besprochenen Fraktalisierung als die jeweiligen Endpunkte eines Spektrums gelten, welches vom individualistisch-gesellschaftlichen Blickwinkel (Habitus-Feldtheorie) bis zu einer gesellschaftlich-gesellschaftlichen Perspektive (Systemtheorie nach Luhmann) reicht.

Symbolische Kultur

Was diese gesellschaftszentrierten Ansätze eint, kann entlang der drei analytischen Dimensionen festgestellt werden, die bereits zur Illustration der Grundannahmen der individualistischen Perspektive verwendet wurden. Erstens, die Frage nach den Trägern der symbolischen Ordnung, oder anders ausgedrückt, was sind die Grenzen in denen eine solche Ordnung existiert. Dies waren im Falle der individualistischen Perspektive die Individuen und ist hier wie zu erwarten die Gesellschaft. Zweitens, der Mechanismus der Entstehung und Aktualisierung spezifischer Ausprägungen der symbolischen Ordnung. Im Rahmen der gesellschaftlichen Perspektive wird dies als ein historischer Vorgang aufgefasst, der im weitesten Sinne aus Kommunikationen besteht. Drittens, was steuert den symbolischen Wandel? Hier findet sich die Annahme eines gesellschaftlichen Vorgangs, der für die Individuen in den meisten Fällen nicht nachvollziehbar oder steuerbar ist.

In der gesellschaftlichen Perspektive erscheinen symbolische Ordnungen als Merkmale von Kollektiven. Genauer gesagt werden Symbole als Repräsentationen einer überindividuellen Sozialität dargestellt. Symbole und deren Relationen sind somit eine Funktion des Feldes, der Gesellschaft als Ganzem, spezifischer Gruppen oder eines spezifischen Systems (z.B.: des kulturellen Systems) und nicht individueller Sozialisation und Erfahrung. Um die hier verwendeten Begriffe von denen der individualistischen Sichtweise abzugrenzen, wird im Folgenden die Bezeichnung *Kultur* für eine symbolische Ordnung verwendet, die über eine Vielzahl von Individuen hinweg existiert.

Dabei gehen die meisten Theorien dieser Richtung durchaus davon aus, dass sich die Kultur auch im Inneren des Individuums niederschlägt. Diese *Verinnerlichung einer externen symbolischen Ordnung* wird am stärksten in der Bourdieuschen Feldtheorie betont, in der der Habitus als eine vorbewusste Inkorporation des Feldes aufgefasst wird. Den Gegenpol bildet die Annahme einer strikten Trennung von symbolischen Vorgängen der Psyche und denen des Sozialen, wie sie in Niklas Luhmanns Version der Systemtheorie zum Ausdruck kommt. Allerdings verfügt auch diese mit dem Begriff des Schemas über eine Vorstellung von einer Verinnerlichung überindividueller Wissensbestände. Ungeachtet dessen ist die Annahme der Verankerung überindividueller Vorgänge in Individuen jedoch mitnichten als Annäherung an die individualistische Position anzusehen. Regeln, Normen, Werte, Geschichten, Fakten und dergleichen, kurzum alles was hier als Kultur bezeichnet wurde, muss dem Wissen der Individuen in dieser Perspektive immer vorausgehen. Um es in den Begrifflichkeiten der Feldtheorie auszudrücken: Es kann Regeln für Spiele geben, die niemand mehr spielt oder je gespielt hat, aber ein Spiel ohne Regeln ist nicht möglich. In allen Ausprägungen der gesellschaftlichen Perspektive stellen Symbole eine notwendige Bedingung des Sozialen dar, da symbolische Kommunikation erst durch den Prozess der sozialen Standardisierung möglich wird. Damit ist nicht nur eine Basis für individuelle Orientierung geschaffen, sondern auch für die Anschlussfähigkeit von Kommunikationen.

Weil Symbole und symbolische Ordnungen als *Repräsentationen* bzw. als *Funktionen des Kollektivs* gesehen werden, sind sie so sehr mit diesem verbunden, dass eine Unterscheidung zwischen diesen zwei Bereichen fast zwangsläufig einen analytischen Charakter annehmen muss. In der Feldtheorie erkennt man dies an der Annahme, dass die Kultur eine symbolische Repräsentation des Feldes ist. Die (Spiel-)Regeln und das worum gespielt wird ergeben sich letztlich aus der Konstitution und der Geschichte des Feldes. Dies trifft nicht nur auf die materialistische Fassung von Bourdieus Feldtheorie zu. Auch in der Strategic Action Field Theory sind symbolische Repräsentationen und Feldeffekte gleichgesetzt. Deutlich wird dies an der Ableitung der meist impliziten Spielregeln aus der Beobachtung der Verteilungslogiken des Feldes, zum Beispiel indem die Häufigkeit und Verteilung unterschiedlicher Arten von Finanzmarkttransaktionen als Indikatoren (und Folgen) der auf dem amerikanischen Finanzmarkt vorherrschenden Strategien angesehen werden (vgl. Fligstein und Goldstein 2010). Im Falle der Systemtheorie stellen Symbole die

Grundlage der sozialen Interaktionen bei Parsons und auch der Kommunikation bei Luhmann. In beiden Fällen werden sie in gewisser Weise mit den sozialen Phänomenen die sie repräsentieren gleichgesetzt. Insofern lässt sich die Trennung von symbolischer Ordnung und sozialen Phänomenen in beiden Fällen als analytisch kategorisieren, da das Verhältnis der beiden bereits in der Axiomatik der Theorie festgeschrieben wird und deshalb keiner empirischen Bestimmung mehr zugänglich scheint.

Kultur wird dabei zu einem externen und relativ statischen Phänomen. In gewisser Weise stellt sie nur die Repräsentation der sozialen Felder und Systeme selbst dar, bzw. ist kaum von diesen zu unterscheiden. Eine Betrachtung und Analyse von Kultur in ihrem empirischen Verhältnis zum Handeln, zu unterschiedlichen Feldeffekten oder systemischen Operationen wird dadurch erschwert. So reduziert die Feldtheorie den Kampf um und mit Spielregeln auf seine „materiellen" Komponenten, entweder auf die Verteilung der jeweiligen Kapitalia oder auf die strategischen Positionen der Individuen. Fragen nach der Legitimation oder der Delegitimation von Spielregeln, Normen und Werten können so nicht gestellt werden. Warum sich bestimmte Legitimationen durchsetzen oder es überhaupt zu solchen Konflikten kommt, erscheint in dieser Perspektive immer als ein Ergebnis der bestehenden Strukturen. Die konkreten Inhalte und mit welchen Symbolen und Argumenten die Kämpfe um Legitimation und Deutungshoheit gefochten wurden, spielen dementsprechend nur eine untergeordnete Rolle.

Damit eng verwandt ist ein zweites Problem. Die Auffassung, dass Kultur eine Funktion der jeweiligen Gesellschaft, des spezifischen Feldes oder die Selbstbeschreibung eines Funktionssystems sei, führt zur Annahme, dass der *soziale und symbolische Raum deckungsgleich* wären. Natürlich ist es nicht von der Hand zu weisen, dass es Fälle gibt in denen soziale Strukturen an eine spezifische Kultur gebunden sind. Der wahrscheinlich bekannteste Fall einer solchen Deckungsgleichheit sind die von Erving Goffman beschriebenen totalen Institutionen (Goffman 1997). Im Falle von Gefängnissen, Klöstern, Altenheimen und dergleichen kann sinnvoll von einer Übereinstimmung der sozialen Struktur und der symbolischen Ordnung gesprochen werden. Allerdings ist von solchen Einrichtungen auch bekannt, dass sich die Normen, Werte, Geschichten und Riten, sich also die Kultur der Insassen stark von derjenigen des Personals unterscheidet. Diese inneren Differenzen und symbolischen Spannungen könnten jedoch noch als eine Repräsentation der sozialen Aufteilung innerhalb der Institution gelten.

Die Beurteilung des Einflusses von externen Kulturen stellt hingegen bereits ein größeres Problem dar. Innnerhalb totaler Institutionen gibt es symbolische Ordnungen, die auch außerhalb dieser existieren und mit den Strukturen der Institution in einer komplexen Wechselbeziehung stehen. So zum Beispiel die unterschiedlichen Subkulturen des organisierten Verbrechens, die normalerweise sowohl in der totalen Institution des Gefängnisses vorkommen als auch außerhalb. Ein Forschungsvorhaben, welches die spezifischen, symbolischen Praktiken verschiedener Gruppen in solchen Institutionen untersuchen zum Ziel hätte, käme demnach nicht umhin jene weiter gefassten kulturellen Räume ebenfalls zu berücksichtigen.

Unterschiedliche Kulturen können im selben, sozialen Raum koexistieren oder miteinander konkurrieren. Symbolische Ordnungen können weit über den Rahmen spezifischer Institutionen hinausgehen oder ein exklusiver und relativ kleiner Bestandteil davon sein. Letztlich bleibt nichts anderes übrig, als das Verhältnis empirisch zu bestimmen. Dies ist im Fall der gesellschaftlichen Perspektive jedoch genauso problematisch, wie aus einer individualistischen Sichtweise heraus. Nimmt man Individuen als Träger von sozialen Symbolen an, so erweist es sich ebenfalls als schwierig die Reichweite symbolischer Ordnungen zu beschreiben, da individuelles Wissen immer in einen sehr spezifischen psychischen und biografischen Kontext eingebettet ist. Dementsprechend schwierig ist es die Kultur einer Gruppe als das durchschnittliche Wissen ihrer Mitglieder zu beschreiben. In der gesellschaftlichen Perspektive sind symbolische Ordnungen hingegen meist deckungsgleich mit den sozialen Strukturen die sie repräsentieren, weswegen auch hier das empirische Verhältnis bereits theoretisch vorherbestimmt ist.

Die einzige Möglichkeit die Grenzen symbolischer Ordnungen empirisch festzustellen ist sie zunächst als eigenständige Phänomene zu betrachten. Sind ihre Reichweite und sonstigen Eigenheiten erst einmal bestimmt, kann man sie in einem nächsten Schritt mit anderen Phänomenen des sozialen Raums und der sozialen Struktur in Beziehung setzen. Um die Eigenschaften und die Grenzen konkreter, symbolischer Ordnungen empirisch festlegen zu können sind Verfahren notwendig die Symbole als eigenständige Phänomene auffassen.

Im strikten Gegensatz zur individualistischen Perspektive, sehen die auf die Gesellschaft orientierten Ansätze die Entstehung einer spezifischen symbolischen Ordnung als eine historische Entwicklung. *Kultur als Produkt der Geschichte* aufzufassen bedeutet jedoch auch anzunehmen,

dass die gegenwärtigen Ausprägungen als das Resultat einer singulären Entwicklungslinie verstanden werden können. Dies wirft jedoch das Problem auf, ob es eine Richtung in dieser Entwicklung geben kann und wie diese festzustellen wäre. Talcott Parsons (vgl. z.B.: 1964a) war wahrscheinlich der letzte Vertreter der großen Gesellschaftstheorien, der der Meinung war eine solch allgemeine Richtung in der Evolution von Gesellschaften erkennen zu können. Seitdem hat die Soziologie dem Fortschrittsglauben größtenteils abgeschworen und eine eher historische Haltung an den Tag gelegt. Es finden sich zwar grundlegende Annahmen über bestimmte Entwicklungstendenzen, wie zum Beispiel Arbeitsteilung (vgl. Durkheim 2012) oder Differenzierung gesellschaftlicher Teilsysteme (vgl. Luhmann 2009: 205-227), diese beziehen sich jedoch fast ausschließlich auf die sozialen Strukturen und werden daher selten als Theorien allgemeinen, kulturhistorischen Fortschritts gelesen.

Dadurch lassen sich zwar die Fallstricke überzogener Fortschrittsgläubigkeit und des Kulturzentrismus umgehen, gleichzeitig begrenzt dies die Erkenntnismöglichkeiten auf eine reine Deskription der dominanten, symbolischen Ordnung zu verschiedenen Zeitpunkten. Sowohl in den Feldtheorien, als auch in den systemtheoretischen Ansätzen wird davon ausgegangen, dass die Entwicklung einer symbolischen Ordnung den strukturellen Gegebenheiten folgt. Dadurch wird das bereits beschriebene Problem der empirischen Abgrenzung kultureller Räume auch zu einem zeitlichen, da angenommen wird, dass die strukturelle und die kulturelle Entwicklung deckungsgleich oder zumindest von denselben Phänomenen getragen wäre. Des Weiteren bedeutet dies auch, dass die Entwicklung verschiedener Kulturen nur schwer miteinander verglichen werden kann und somit keine allgemeingültigen Aussagen darüber möglich sind. Es ist schwierig nach den generellen Rahmenbedingungen kultureller Entwicklungen zu fragen, wenn diese als ein Kapitel in der Geschichte des jeweiligen Systems bzw. Feldes aufgefasst werden.

Aus dieser Annahme einer geschichtlichen Entwicklung ergibt sich auch, dass sich die Steuerung dieser Vorgänge der menschlichen Schaffenskraft (in handlungstheoretischen Begriffen: Agency) entzieht. Die Eigenlogik der symbolischen Entwicklung wird als ein System jenseits der menschlichen Einflusssphäre angesehen, da es den individuellen Menschen zeitlich vorausgeht. Dies wiederum hat zur Folge, dass es als ein objektiver, äußerer sozialer Tatbestand erscheint, der sich nicht einfach ändern lässt. Hieraus folgt der von Durkheim festgestellte grundsätzliche Zwangscharakter des Sozialen:

> [W]enn es allgemein ist, so ist es das, weil es kollektiv (d.h. mehr oder weniger obligatorisch) ist; und nicht umgekehrt ist es kollektiv, weil es allgemein ist. Es ist ein Zustand der Gruppe, der sich bei den Einzelnen wiederholt, weil er sich ihnen aufdrängt. Er ist in jedem Teil, weil er im Ganzen ist, und er ist nicht im Ganzen, weil er in den Teilen ist. (Durkheim 1984: 111)

Während die individualistische Perspektive von einer Steuerung der Entwicklung symbolischer Ordnungen durch individuelle Entscheidungen oder deren Aggregate ausgeht, ist es im Falle der Kultur eher ein blinder, historischer Prozess, der über die Köpfe der Individuen hinweg den Lauf der Dinge entscheidet. Gerade weil Kultur als etwas den Individuen Vorausgehendes begriffen wird, ist der Wandel und Genese nicht vorhersehbar oder direkt regulierbar. Eine Steuerung der Entwicklungsrichtung der symbolischen Ordnung scheint wenn, dann überhaupt nur durch Macht – im Sinne einer Veränderung der objektiven Strukturen und Spielregeln – möglich. Dieser feldtheoretischen Sichtweise kann von systemtheoretischer Seite noch hinzugefügt werden, dass nur bestimmte hochspezialisierte Teilsysteme einer Gesellschaft dazu in der Lage wären solche Veränderungen einzuleiten.

Allerdings ist die Annahme einer solchen „Blindheit" der Entwicklung einer symbolischen Ordnung in gewisser Weise eine überspitzte Darstellung, die aber helfen kann das grundsätzliche Problem zu sehen. Da die gesellschaftstheoretischen Ansätze von der symbolischen Ordnung als einer Repräsentation fundamentalerer Prozesse ausgehen, tendieren sie dazu sowohl deren Eigengesetzlichkeit als auch die Rolle von Individuen hinsichtlich der Konstruktion, Aktualisierung und Entwicklung zu unterschätzen. Diese beiden Probleme können auf eine gemeinsame Ursache zurückgeführt werden, nämlich das Unvermögen die symbolischen Ordnungen menschlicher Gesellschaften als eine Form von Technologie aufzufassen, genauer gesagt der *sozialen Technologie* schlechthin.

Sprache und Zeichensysteme sind aber auch in anderen Spezies beobachtet worden. Laborversuche in denen großen Menschenaffen eine Zeichensprache beigebracht wurde, haben gezeigt, dass diese von ihren kognitiven Fähigkeiten und ihrer Neugierde her mit zweieinhalbjährigen Kindern vergleichbar sind (vgl. Premack und Premack 1994). In diesem Rahmen wird auch von den beeindruckenden Fähigkeiten unserer nächsten Verwandten berichtet, komplexe Fragen nachzuvollziehen und beantworten zu können (vgl. z.B.: Rumbaugh, Gill und Glasersfeld 1973; Patter-

son 1980; Wallman 1992). Unabhängig davon wurde jedoch nie eine Frage von seitens der Menschenaffen gestellt und das obwohl ihnen entsprechende Ausdrücke beigebracht wurden und sie ganz offensichtlich in der Lage waren Fragen zu verstehen und adäquat zu antworten. Demgegenüber ist die Frageform (Interrogativ) nicht nur Bestandteil aller menschlichen Sprachen, sondern tritt auch in Form einer spezifischen Intonation für „ja/nein" Fragen auf, die bereits in Kleinkindern beobachtet werden kann (vgl. Cruttenden 1997: 162ff). Die erste Frage, die ein Mitglied unserer Spezies stellte, ist demzufolge sehr viel mehr, als nur eine Fußnote unserer Entwicklungsgeschichte.

> Accordingly I would suggest that it is not the recognition of ourselves as individuals that makes us humans (we know that apes, at least chimpanzees and orangutans, are as good as humans at recognizing themselves in the mirror). It is, rather, **recognition of other members of the society as individuals with equal cognitive abilities and the employment of their cognitive abilities as a source of information (asking questions), that makes us human, and our language – human language**. (Jordania 2006: 336, Hervorhebung im Original)

Es ist anzunehmen, dass der Gewinn an Koordinationsmöglichkeiten und individuellem Wissen ein enormer Selektionsvorteil für die frühen Hominiden gewesen sein muss. Es greift jedoch zu kurz nur den funktionalen Aspekt der menschlichen Prävalenz für Fragen-Antwortspielchen zu betrachten. Eine Frage und die darauf gegebene Antwort, sind auch immer die Grundlage weiterer Fragen und Antworten. Die sich daraus entspannende Konversation ist zugleich auch Verortung der beteiligten Personen relativ zueinander. Es bedarf nicht viel Fantasie um sich vorzustellen, wie sich aus dem beständigen Austausch von Informationen auch die ersten sozialen Gebilde und Unterscheidungen entwickelten, die sich zu einer Vielzahl sozialer Formen (Arbeitsteilung, Hierarchien, Gruppenunterscheidungen, etc.) verdichteten. Mit den ständigen Fragen wurden auch immer neue Antworten geschaffen. Es bedurfte Erklärungen für die Phänomene der natürlichen Welt und für die soziale Organisation. Die Kodifizierung und Standardisierung dieser Erklärungen dürfte der Grund für die Entwicklung der ersten abstrakten Symbolsysteme gewesen sein. Gleichzeitig lernten Menschen offensichtlich auch, dass Manipulation und Kontrolle des Informationsaustausches entscheidende Vorteile mit sich bringen konnten.

Die angeborene Tendenz Fragen zu stellen ist aber auch deswegen so relevant, weil sie Rückschlüsse auf das grundsätzliche Appetenzverhalten des Menschen zulässt. Neben Schlaf-, Nahrungs- und Fortpflanzungstrieben haben wir auch einen natürlichen Hang zur Kommunikation und zum Spiel mit abstrakten Symbolsystemen, der im gesamten Tierreich seinesgleichen sucht. Dieses Spiel mit Symbolen ist ein aktiver Prozess, den man bereits bei sehr kleinen Kindern beobachten kann. Sie geben sich Spielregeln, konstruieren Fantasiewelten und denken sich eigene Sprachen aus, um nur ein paar Varianten dieses Verhaltens zu beschreiben.

Auch wenn sich der genaue Hergang nicht rekonstruieren lässt, so kann man doch Rückschlüsse aus den heutigen menschlichen Gesellschaften ziehen. Irgendwann in der Vergangenheit muss die Idee aufgekommen sein, dass Symbole nicht nur Respekt verlangende heilige Objekte sind, sondern auch bewusst konstruiert werden können. Da sie aber schon immer auch eine soziale Seite hatten, d.h. die grundlegenden Verhältnisse zwischen Personen wiederspiegelten, war es wohl nur ein kleiner Schritt zur gezielten Erschaffung sozialer Ordnungen. Dies geschah indem man Positionen und Verhältnisse markierte, Abläufe und Tätigkeiten kodifizierte, Normen und Werte abstrahierte und festhielt. Die Folge waren von den Personen unabhängige, soziale Prozesse.

Die ältesten Kulturzeugnisse die uns überliefert sind versuchen entweder bestimmte Umstände oder Gegebenheiten im sozialen Gedächtnis zu verankern (vgl. Assmann 2007: 66-86) oder sie stellen Regelwerke und technische Anleitungen – insbesondere Gesetzestexte (Neumann 1989) – dar. Solche normativen Regelwerke sind die eindeutigste Form sozialer Technologie im hier verstandenen Sinne. Jedoch erschöpfen sich diese expliziten und impliziten Regelwerke nicht nur in der Kodifizierung der sozialen Ordnung, sie werfen auch Fragen nach den Konstruktionsprinzipien auf. Es ist bezeichnend für die menschliche Spezies, dass ein „weil-es-eben-so-ist“ nie vollkommend ausreichend war. Jede Legitimation der sozialen Ordnung war immer auch Ausgangspunkt für Zweifel und Revisionen. Vor allem aber führt dies auch immer zu einer Veränderung der Kriterien der Bewertung dieser Regeln. Dies hat weitreichende Konsequenzen, weil man sich symbolische Ordnungen als ein komplexes Wechselspiel vorstellen muss, in dem Veränderungen einzelner Elemente immer auch Wirkungen auf das gesamte Ensemble haben.

Insofern ähnelt die hier dargestellte Version der sozialen Entwicklung den Ausführungen von Jürgen Habermas (1995), der ebenfalls davon ausgeht, dass eine Zunahme an Diskurs und kritischer Reflexion zur Arbeits-

teilung und zur Entwicklung moderner Gesellschaften geführt hat. Jedoch weicht die hier vertretene Auffassung an zwei entscheidenden Punkt von der Habermas'schen Position ab. Zum einen wird dieser Prozess nicht als eine zwingende Folge der Vernunft mit einem vorherbestimmten Endziel betrachtet, sondern als eine evolutionäre Entwicklung und zum anderen werden hier symbolische Ordnungen immer auch als soziale Technologien aufgefasst. Letzteres bedeutet, dass sie Kooperation und Verständnis erzeugen können aber auch genauso gut das Gegenteil. Denn auch Streit und Wettbewerb setzen einen Rahmen voraus an dem sich die beteiligten Personen orientieren und den sie letztlich auch manipulieren können.

Zusammenfassend lässt sich der an der Gesellschaft orientierte Zugang zu sozialen Symbolen mit der *materialistischen Interpretation von Symbolen* vergleichen. Wie schon in den Ausführungen über das Dilemma der Semiotik dargelegt, können Zeichen auch als eine direkte Repräsentation von Objekten aufgefasst werden. Das Problem besteht dann allerdings darin, dass es keinen eigenständigen Wandel dieser Zeichen geben kann und dass der Vorrat der Zeichen und der daraus konstruierbaren Sätze und Aussagen begrenzt wäre auf die Repräsentation von Objekten. Es ist jedoch evident das die gesellschaftliche Kommunikation sich auch auf Entitäten und Konzepte bezieht die es außerhalb dieser Diskurse nicht gibt oder geben kann. Vorstellungen wie Freiheit, Gerechtigkeit, Wahrheit und andere „praktische" Fiktionen sind keine Objekte außerhalb der symbolischen Ordnung in der sie existieren. Aber es sind genau solche Kriterien, welche die Entwicklung menschlichen Wissens und Kultur steuern. Dies ist nur möglich weil es sich dabei selbst um symbolische Ordnungen handelt.

Auf den hier untersuchten drei Ebenen kann die gesellschaftliche Perspektive folgendermaßen charakterisiert werden. Die symbolische Ordnung wird hier als eine von den Individuen größtenteils unabhängige Struktur aufgefasst, welche eine Repräsentation der sozialen Verhältnisse darstellt. Entsprechend wird der Wandel dieser symbolischen Ordnung als ein historischer Vorgang dargestellt, bei dem die konkreten Eigenschaften der gegenwärtigen Struktur entweder eine Abfolge sozialer Veränderungen wiederspiegeln oder einen bestimmten vergangenen Zustand der Gesellschaft. In beiden Fällen herrscht die Vorstellung einer direkten Repräsentation sozialer Phänomene durch Symbole vor. Dementsprechend gehen von den Individuen in dieser Vorstellung so gut wie keine Steuerungsimpulse aus. Wahlentscheidungen, Rationalitäten und funktionale Erfordernisse (zumindest von Personen) vollziehen sich im Rahmen und

unter den Imperativen der symbolischen Ordnung, die in manchen dieser Theorien den Individuen nicht einmal bewusst ist.

3.3 Prozesssoziologische Perspektive

Die bisher betrachteten Perspektiven auf den Gegenstand der symbolischen Ordnungen wirken zunächst relativ unvereinbar, da sie von gegensätzlichen Prämissen ausgehen. Sieht man Symbole als Funktionen von Individuen und deren Entscheidungsmaximen, beziehungsweise als Folge von Mikrointeraktionen an, so ist nicht zu erklären woher die Zeichensysteme kommen und wieso sie so gut zur sozialen Realität passen. Eine gesellschaftliche Perspektive löst dieses Problem. Allerdings um den Preis einer Reduktion auf eine monolithische, symbolische Ordnung, die den Individuen als eine externe Macht entgegen tritt. Damit kann allerdings nicht erklärt werden, wieso die bewusste Konstruktion und der Wandel sozialer Strukturen immer symbolisch vorbereitet und in vielen Fällen auch auf diese Weise durchgeführt wurden.

Wie schon angedeutet stützt sich der Vorschlag einer *Synthese der individualistischen und der gesellschaftlichen Perspektive*, der im Folgenden erläutert werden soll, in weiten Teilen auf Norbert Elias Ausführungen zu einer eigenständigen Theorie sozialer Symbole. Darüber hinaus findet sich aber noch eine Vielzahl anderer Ansätze, die in eine ähnliche Richtung deuten. Grundlegend ist dabei die Annahme, dass Symbole ein eigenständiges Phänomen darstellen, welches sowohl im Innenleben von Individuen auftritt, wo es die Basis des Denkens darstellt, als auch in der Kultur einer Gesellschaft, deren grundlegenden Elemente ebenfalls Symbole sind. Es wird hierbei die Annahme vertreten, dass es zwischen den unterschiedlichen Ebenen auf denen soziale Symbole eingesetzt werden keinen wesentlichen Unterschied gibt. Anders ausgedrückt, Denken, Sprechen und Diskutieren sind in dieser Perspektive ein und dasselbe.

3.3.1 Prozesstheorien

Es wurde bereits angedeutet, dass eine Synthese von individualistischen und gesellschaftlichen Ansätzen in der Dimension der Zeit zu suchen wäre. Wenn wir die beiden Ebenen des individuellen Symbolgebrauchs und der umfassenden Kultur in die sie eingebettet sind aufeinander beziehen, erhalten wir ein Modell welches dem „Schraubenprozess der Erkenntnis“ von Rupert Riedl (1985: 55ff) entspricht. In dieser Vorstellung ist es das

iterative Aufeinandertreffen von empirischen Beobachtungen und kategorischen Vorannahmen, welches Erkenntnisse über die Welt generiert. Die Einzelbeobachtungen und deren induktive (aber nicht logisch zwingende) Verallgemeinerung schaffen Spekulationen und Hypothesen, die dann systematisch und deduktiv getestet werden können. In der Sprache der allgemeinen Systemtheorie ausgedrückt lässt sich sagen,dass die einzelnen Elemente den Raum der Möglichkeiten öffnen, also Komplexität bereitstellen, welche wiederum die Grundlage für eine Selektion durch das System als Ganzes stellt. Was dabei im Zeitverlauf herauskommt sind nicht nur Erkenntnisse, Wissen und Kultur, sondern auch neue Methoden und Kategorien der Beobachtung. Da jede dieser beiden Seiten des Prozesses die jeweils andere voraussetzt, macht es in dieser Perspektive wenig Sinn nach einem absoluten Anfang oder einem ebensolchen Ende zu fragen. Des Weiteren bedeutet die Annahme eines Prozesses, dass die symbolischen Ordnungen, die durch ihn generiert werden, nicht transzendentale Objekte sind, sondern selbst dem Wandel unterliegen, dessen fester Bestandteil sie sind. Im Gegensatz zu den Befürchtungen der Rationalisten, führt dies jedoch nicht zu Beliebigkeit und Chaos, sondern ist gerade erst die Bedingung der Konstruktion einer immer realitätskongruenteren Beschreibung der Welt:

> Human beings have developed *within* a world. Their cognitive functions evolved in continuous contact with objects to be recognized. The symbol emancipation in the course of which socially acquired means of communication gained dominance over those which were genetically fixated enabled humans to adjust their judgement and their actions to an almost infinite variety of situations. Humans did not enter the world as aliens. Subject and object form part of the same world. (Elias 1991: 98)

Die wahrscheinlich weitreichendste Konsequenz dieser Überlegungen besteht darin, dass zwischen Natur und Kultur kein nennenswerter Unterschied festgestellt werden kann. Da Symbole genau wie Menschen Teil der Welt sind, unterliegen sie den gleichen Gesetzmäßigkeiten, wie alle anderen Träger von Informationen auch. Hiergegen wird gerne eingewendet, dass es ja immer noch des Menschen bedürfte um soziale Symbole auszutauschen und sie überhaupt erst zu erfinden. Dies mag als eine Beschreibung der gegenwärtigen Situation korrekt sein, ist aber kein Hinweis auf eine fundamentale Sonderstellung des Menschen. Vielmehr zei-

gen sowohl die oben erwähnten Sprachexperimente mit Affen, als auch die rasante Entwicklung von Sprachen, die entweder der Kommunikation mit Maschinen oder zwischen diesen dienen, dass Symbole von allen nur erdenklichen Systemen genutzt werden können, die über die Möglichkeiten verfügen diese zu verarbeiten.[15]

In der bisherigen Diskussion wurde der Begriff der symbolischen Ordnung relativ unterschiedslos und recht allgemein verwendet, um einen wie auch immer gearteten Raum von Symbolen[16] zu beschreiben, in dem die wechselseitigen Beziehungen zwischen diesen so etwas wie Bedeutungen hervorbringt. In seiner Gesamtheit wurde dieser Raum auch als Kultur oder gesellschaftliches Wissen bezeichnet. Der hier vertretenen Auffassung nach stellen Prozesse jedoch immer ein Mehrebenenphänomen dar. Dies ist letztlich auch, was Norbert Elias mit einer Entwicklung „within" andeutet, nämlich das Herausbilden von mehreren Schichten von symbolischen Ordnungen, die sich zwischen den Extrempunkten relativ ungeordneter Kommunikationen und der umfassenden Ordnung von Symbolen auftuen. Aus einer prozesstheoretischen Perspektive spricht Elias (ebd.: 44f) hier von „levels of synthesis", während Rupert Riedl (1985: 66ff) ein solches Schichtmodell in Form des „hierarchischen Bau[s] der Welt" für die Erklärung einer Vielzahl natürlicher Prozesse heranzieht.

Geht man von einem hierarchischen Aufbau symbolischer Ordnungen aus, dann umfasst dieser Begriff eine Vielzahl von Phänomenen mit zum Teil sehr unterschiedlichen Reichweiten. Die dominante Kultur einer Gesellschaft wäre demnach genauso eine symbolische Ordnung, wie die dominanten Themen eines Gesprächs unter Freunden. Diese Phänomene konnten als gleichartig betrachtet werden, weil es in beiden Fällen um spezifische, regelmäßige Verknüpfung von Symbolen geht. Der Unterschied besteht jedoch in der Reichweite der jeweiligen symbolischen Ordnung.

[15] Diese Feststellung berührt nicht die Definition von Symbolen als sozial standardisierten Zeichen, die der Kommunikation zwischen Menschen dienen. Da es sich hier immer noch um menschliche Sprachen handelt, die nicht-menschlichen „Kommunikanden" beigebracht wurden.

[16] Die Formulierung „wie auch immer geartet" soll anzeigen, dass zu diesem Zeitpunkt jede Festlegung auf eine bestimmte Struktur oder Form dieses Raums nicht zielführend oder sinnvoll möglich ist. Wir können uns den Raum der Symbole zum Beispiel als einen euklidischen Raum vorstellen. Aber auch ein geodätischer Raum (ein Netzwerk-Graph) wäre ebenso denkbar, wie auch gänzlich andere Konzeptionen. Wie noch zu zeigen sein wird, setzt die Anwendung unterschiedlicher Verfahren eine Entscheidung des Forschers hinsichtlich der Konstitution dieses Raumes voraus. Eine Entscheidung, die im Hinblick auf die zu verfolgende Fragestellung und das zu untersuchende Zeichensystem getroffen werden muss.

Gemeint ist damit der Kreis der potentiellen Adressaten symbolisch kodierter Äußerungen, von denen erwartet wird, dass sie in der Lage sind, nicht nur die verwendeten Symbole in der intendierten Art und Weise zu verstehen, sondern vor allem sich adäquat daran zu orientieren. Sicherlich ist dies keine sonderlich trennscharfe Unterscheidung. Schließlich muss man davon ausgehen, dass es unterschiedlich schwierig sein kann, sich im jeweiligen Diskurs zurechtzufinden und darin wirkmächtig zu werden. Bei einem Kneipengespräch würde es wahrscheinlich schon reichen mit der allgemeinen Kultur der Gegenwart vertraut zu sein, um nach relativ kurzer Zeit zumindest dem Gespräch folgen zu können. Der Fall wäre jedoch ganz anders gelagert, wenn man es mit einer sehr spezifischen „Subkultur" zu tun hätte.

Eine solche Hierarchie symbolischer Ordnungen wird von den meisten Prozesstheorien als das Ergebnis des wechselseitigen Bezugs zweier Ebenen aufgefasst. Zum einen die der Möglichkeiten, welche durch den konkreten Gebrauch von sozial standardisierten Symbolen entstehen und zum anderen die der Selektion durch die übergreifende symbolische Ordnung, die wir hier als Kultur bezeichnet haben. Mit jeder Handlung, jeder Interaktion und jeder Kommunikation, die auf kulturelle Skripte zurückgeht und sich sozialer Symbole bedient, wird die Möglichkeit geschaffen die symbolische Ordnung tiefgreifend zu verändern. Setzen sich bestimmte symbolische Konfigurationen gegenüber konkurrierenden Modellen durch so verändert sich die symbolische Ordnung, was wiederum zu neuen Möglichkeiten des Gebrauchs sozialer Symbole durch individuelle oder kollektive Akteure führt und so weiter und so fort. Richard Münch (1986: 16f) skizziert diesen Zusammenhang in seiner Synthese der handlungs- und gesellschaftstheoretischen Perspektiven wie folgt: „Die symbolische Komplexität und die Kontingenz des Handelns bilden zwei unabhängige Achsen eines Koordinatensystems, das einen Raum aufspannt, innerhalb dessen sich das konkrete, symbolisch gesteuerte und faktisch ausgeführte Handeln bewegt."

Symbolkomplexität bezieht sich dabei auf die Komplexität des Wissensbestandes einer Person und ist damit eine Eigenschaft symbolischer Ordnungen. Je allgemeingültiger, situationsunabhängiger und kontextfreier dieses Wissen ist, umso höher ist seine Komplexität. Auch wenn dieses Wissen bei Münch zunächst einmal als individuelles Wissen beschrieben wird, so erhält es seinen Komplexitätsgrad nur im Vergleich mit der Gesamtheit der symbolisch kodierten Wissensbestände. Anders ausgedrückt, eine symbolische Ordnung ist dann komplexer als eine andere, wenn sie

eine höhere Synthese im Raum der symbolischen Ordnungen darstellt. Dies ist dann der Fall, wenn sich möglichst viele symbolische Ordnungen darunter subsumieren lassen, bzw. daraus ableiten lassen. Damit ergibt sich für das Individuum eine größere Auswahl und ein flexiblerer Umgang mit symbolischen Ordnungen auf einem niedrigeren Syntheseniveau. Am Beispiel der Wissenschaft lässt sich dies wie folgt illustrieren: „Die allgemeine Idee der Wahrheit gewährt dem Wissenschaftler eine größere Auswahl zwischen alternativen Methoden der Wahrheitssuche als die Korrespondenztheorie der Wahrheit oder gar eine positivistische oder hermeneutische Methodologie der Prüfung von Hypothesen" (ebd.: 16). Es ist der relationale (räumliche) Charakter sozialer Wissensbestände, der so etwas wie den Transfer zwischen verschiedenen Wissensbereichen und auch die Erweiterung des Wissens möglich macht. Andererseits erklärt diese Konzeption auch, warum es selbst in alltagspraktischer Hinsicht leichter ist aus allgemeinen Sätzen und Regeln partikulare Erklärungen abzuleiten als anders herum.

Auch bei Münch verändert sich das symbolische Wissen im Laufe der gesellschaftlichen Entwicklung. Hier ist es die Beziehung zwischen der Kultur einer Gesellschaft und der Welt, die den Prozess in Gang hält:

> Auf der abstraktesten Ebene von Sinnkonstruktionen wird die Interpenetration von Kultur und Welt durch eine allgemeine sinnhafte Haltung zur welt repräsentiert. In dieser Haltung zur Welt liegt die Dynamik der Moderne begründet. Die Verflechtung mit der Welt prägt die Kultur. Sie besitzt keinen rein kontemplativen Charakter, sondern nimmt Elemente der Welt in sich auf, um sie kulturell zu bearbeiten. Wo die Kultur nicht mit der Welt verflochten ist, bleibt sie der rein kontemplativen Sinnsuche und der Formulierung abstrakter Ideen verhaftet. Wo sie sich der Welt anpasst, wird sie von deren Utilitarismus, Pragmatik und Partikularismus aufgefressen. (ebd.: 23f)

Zwar beschreibt Münch an dieser Stelle hauptsächlich die besondere Dynamik der Kultur der westlichen Moderne, er macht jedoch auch deutlich, dass der Unterschied zu anderen Zeiten und Gesellschaftsformen im Grad der Interpenetration liegt. Die Extreme einer „rein kontemplativen" sowie einer „absolut weltangepassten" Gesellschaft, stellen Idealtypen dar, für die sich keine historische Beispiele finden lassen. Allerdings finden sich Annäherungen. So würde diese Theorie erwarten lassen, dass man in einer auf Kontemplation ausgerichteten Gesellschaft eine symbolische

Ordnung von niedriger Komplexität erwarten dürfte, bzw. das es nur einen sehr rudimentären Wissenbestand gäbe, der sich durch ständige Reproduktion und Affirmation seines Kernbestandes auszeichnen würde.

Eine ähnliche Vorstellung findet sich in der Prozesstheorie von Norbert Elias, der ebenfalls in der symbolisch kodifzierten Auseinandersetzung mit der Welt die den Prozess antreibende Kraft vermutet (vgl. Elias 1991: pp. 74-78). Als Ursache hierfür sieht er den menschlichen Drang zum Wissen, der zu in einem sich selbst antreibenden Wechselspiel von „fantasy knowledge" und „reality congruent knowledge" führt, was wiederum die ständige Weiterentwicklung menschlichen Wissens zur Folge hat. Dabei ist das Fantasie-Wissen, also die Spekulation und das Spiel mit Symbolen, ebenso bedeutsam wie das Testen dieses Wissens und die damit einhergehende Feststellung der Realitätskongruenz. „The human capacity for producing fantasies in answer to questions which present themselves has no less survival value than their capacity for discovering what used to be called the truth" (ebd.: 75). Dies zeigt sich nicht zuletzt in den sozialen Institutionen die Menschen geschaffen haben, um zur Spekulation und zum Fantasieren anzuregen. Unsere enorme Unterhaltungsindustrien zählen ebenso dazu, wie unser Bedürfnis nach Streitgesprächen und sozialem Austausch im Alltag.

Die Unterscheidung zwischen Fantasie und Realität ist, ebenso wie diejenige von Kultur und Welt, selbst eine symbolische Unterscheidung. Die Modelle und Beschreibungen der Realität sind notwendigerweise immer eine verkürzte Darstellung der Wirklichkeit, über deren Wahrheitsgehalt nur ein Vergleich mit Daten Aufschluss geben kann. Da Daten aber ebenfalls nur symbolische Repräsentationen von Messungen darstellen, handelt es sich letztlich um einen rein symbolischen Prozess. Wie oben schon vorgestellt, kann dieses Problem durch eine inkrementelle Annäherung der Modelle an die Realität gelöst werden. Die sinnvolle Ausgestaltung dieses Prozesses fällt in den Bereich der Erkenntnistheorie und muss uns an dieser Stelle nicht weiter interessieren. Offensichtlich war es in praktischer Hinsicht nie ein großes Problem für Menschen alle möglichen symbolischen Ordnungen als gültig zu akzeptieren, wenn auch nur temporär.

Die Frage nach den Konstruktionsregeln symbolischer Ordnungen deutet jedoch auf eine Unterscheidung hin, die für die Untersuchung symbolischer Phänomene von Bedeutung ist. Zwei grundlegende Arten von symbolischen Ordnungen lassen sich hier unterscheiden. Zum einen der statische Gehalt des Wissensbestandes einer spezifischen Kultur und zum anderen die symbolisch kodierten Regeln und Verfahren, die über die rich-

tige Zusammensetzung von symbolischen Ordnungen entscheiden. Bei ersterem handelt es sich um Aussagen über spezifische Sachverhalte und auch über andere symbolische Ordnungen, die bestimmte Zustände zum Ausdruck bringen. Die zweite Art symbolischer Ordnungen umfasst die Normen und Regeln des richtigen Verhaltens, der richtigen Argumentationsweise und auch der Zulassung zum Diskurs. Es ist jedoch wichtig klarzustellen, dass es sich dabei auch nur um symbolische Ordnungen handelt und sie ebenfalls durch andere Regelwerke kritisiert werden können. Im Folgenden wird die Bezeichnung *symbolisches Programm* für ein symbolisch kodifiziertes Regelwerk verwendet.

Die Untersuchung solcher symbolischer Programme ist mit dem grundlegenden Forschungsprogramm der Diskursanalyse von Michel Foucault verwandt. Der symbolische Prozess, bzw. der Diskurs, wird in Foucaults Auffassung durch den Gegensatz von Subjekt und Macht erzeugt (Foucault 1981: 23). Wie auch im Falle anderer Prozesstheorien, umfasst dieser Wechselbezug seine beiden Extrempunkte, was nichts anderes sagt, als dass sowohl Subjekt als auch die Macht Ursache und Wirkung des Diskurses sind. Auch in diesem Fall wird symbolisch kodiertes Wissen als ein sich selbst erzeugendes Steigerungsspiel begriffen, in dem Wissbegierde und der Versuch dikursiver Kontrolle sich gegenseitig bedingen (vgl. Foucault 1991: 41ff). Grundsätzlich gleicht Foucaults Beschreibung des Prozesses symbolischer Ordnungen den Vorangegangenen. Er legt jedoch ein sehr viel größeres Augenmerk auf die symbolischen Konflikte um die Deutungshoheit über die umfassende, symbolische Ordnung. Nicht zuletzt deshalb verwendet er eine stark an militärisches Fachjargon angelehnte Terminologie. So beschreibt er eine strategisch günstige, argumentative Position die innerhalb eines Diskurses eingenommen werden kann als „Dispositiv" (ebd.: 29). Symbolische Ordnungen werden in dieser Sichtweise zu „diskursiven Formationen" (Foucault 1981: 48ff) und Regelmäßigkeiten im Ablauf werden zu „dikursiven Strategien" (ebd.: 94ff).

Konsequenterweise sieht Foucault die Aufgabe der Diskursanalyse in der Beschreibung dieser spezifischen Strukturmerkmale von konkreten symbolischen Prozessen (Diskursen). Es geht ihm also nicht um die allgemeinen Regeln des Sprechens oder der impliziten sozialen Regeln der Sprache. Das eigentliche Ziel der Diskursanalyse ist vielmehr die historische Beschreibung eines Diskurses:

> [S]ogar wenn sie seit langem verschwunden ist, wenn niemand sie mehr spricht und man sie aufgrund seltener Fragmente restauriert hat, bildet eine Sprache stets ein System

> für mögliche Aussagen: es ist eine endliche Menge von Regeln, die eine unendliche Zahl von Performanzen gestattet. Das Feld der diskursiven Ereignisse ist die stets endliche und zur Zeit begrenzte Menge von allein den linguistischen Sequenzen, die formuliert worden sind;[...] Die Beschreibung der diskursiven Ereignisse stellt eine völlig andere Frage: wie kommt es daß eine bestimmte Aussage erschienen ist und keine andere an ihrer Stelle? (Foucault 1981: 42)

Im Gegensatz zu den Arbeiten von Elias und Münch liegt das Augenmerk hier sehr viel mehr auf dem konkreten historischen Prozess und dessen innerer Logik. Daraus resultiert auch eine ausführlichere Terminologie zur Beschreibung der Merkmale von Diskursen und deren inneren Mechanismen. Gleichzeitig bleibt die Perspektive der Diskursanalyse dem spezifischen Merkmalen und der historischen Entwicklung so stark verhaftet, dass es hier nur sehr schwer möglich ist Aussagen zu generieren die einen allgemeineren Gültigkeitsgrad aufweisen.

3.3.2 Symbolische Prozesse

Aus der Perspektive einer prozessorientierten Soziologie erscheinen soziale Symbole und deren spezifische Ordnung als ein Prozess, der sich gemäß einer Eigenlogik entwickelt. Der Unterschied dieser Perspektive zu den vorangegangenen Konzeptionen, die symbolische Ordnungen als Merkmale von Individuen und Gesellschaften auffassen, kann ebenfalls entlang dreier Ebenen verdeutlicht werden. Zunächst wäre dies die Frage nach dem konkreten Untersuchungsgegenstand, der mit einer solchen prozesstheoretischen Herangehensweise einhergeht. Zweitens, wie Wandel und die Genese symbolischer Prozesse adäquat erfasst werden können. Drittens, soll geklärt werden, welche Beziehungen zu anderen sozialen Phänomenen existieren, insbesondere dem sozialen Handeln und der Gesellschaft.

Wie im Falle der individualistischen und gesellschaftlichen Perspektiven auch, stellen Symbole und symbolische Ordnungen nicht den zentralen Untersuchungsgegenstand dar. Dies liegt nicht zuletzt daran, dass die Soziologie in erster Linie an genuin sozialen Phänomenen interessiert ist. Die prozessorientierten Soziologien sehen soziale Symbole daher in erster Linie als Repräsentation von Mustern der gesellschaftlichen Entwicklungen. Gemeint sind damit wiederkehrende soziale Konfigurationen, die

auf eine vergleichbare Art und Weise verlaufen. Ein Paradebeispiele dafür ist der Zivilisationsprozess bei Norbert Elias (1939), der sich durch symbolische und mediale Prozesse ausdrückt und von diesen ebenfalls getragen wird (vgl. Krotz 2003). Damit gehen sie über den Rahmen einer reinen historischen Beschreibung der Entwicklung hinaus und versuchen zeitlich ausgedehnte, allgemeine Muster sozialer Interaktionen zu finden. Selbst im Falle Foucaults, der sich eher einer deskriptiven Vorgehensweise verpflichtet sieht, ist das Endergebnis seiner Arbeit doch fast immer eine Erklärung in Form eines sozialen Prozesses.

Unter sozialen Prozessen werden somit wiederkehrende Muster des Handelns und Bewertens verstanden, die nach der Art eines „offenen Systems" (Bertalanffy 1969: 102ff) über ein bestimmtes Maß an Selbstregulierung verfügen. Im Unterschied zur Konzeption eines autopoietischen, sozialen Systems (vgl. z.B.: Luhmann 1997), stehen offene Systeme in einer fortlaufenden Austauschbeziehung mit ihrer Umwelt. Ebenso wie autopoietische Systeme produzieren sich offene Systeme tendenziell aus sich selbst heraus, die Beeinflussung ihrer internen Prozesse durch die Umwelt ist jedoch eine graduelle Frage. Nach Bertalanffys (1969: 150) Auffassung bewegen sich biologische Organismen im Laufe ihrer Entwicklung von offenen, selbstregulierten Organismen immer mehr hin zu mechanisierten, auf Rückkopplung basierenden Prozessen. Darin zeigt sich, dass es sich grundsätzlich um ein Kontinuum zwischen absoluter Geschlossenheit und absoluter Offenheit handelt.

Um diese Betrachtungsweise auf soziale Symbole zu übertragen, ist es hilfreich sich die Gene eines Organismus als eine Form der Information vorzustellen. Das Modell der Autopoiesis würde in diesem Fall besagen, dass sich ein Organismus aus der ständigen Reproduktion seiner genetischen Struktur ergibt und das dieser Zusammenhang auch innerhalb der Evolution gilt. Die genetische Information innerhalb einer bestimmten Entwicklungslinie kann demnach als ein autopoietischer Prozess gesehen werden, denn auch jede auftretende Mutation ist nichts anderes als die Abwandlung der bereits bestehenden Information.[17]

Demgegenüber erlaubt das Modell eines offenen Systems prinzipiell einen Transfer von Informationen zwischen Systemen. Diese Art des *horizontalen Gentransfers* mag in der biologischen Welt recht selten vorkommen, im Bereich der Informationsübertragung zwischen Individuen dürf-

[17] Das Beispiel der genetischen Evolution wurde hier hauptsächlich deshalb gewählt, weil sich die ursprüngliche Formulierung des Begriffs der Autopoiesis darauf bezieht (vgl. Maturana und Varela 1987).

te es doch eher die Norm sein. Übertragen auf soziale Symbole bedeutet dies, dass sich Personen in ihrem Denken und dem daraus resultierenden Handeln direkt und wechselseitig beeinflussen können. Auch die Übertragung, Inkorporation und Veränderung von symbolischen Ordnungen auf der Ebene von ganzen Kulturkreisen ist in dieser Konzeption möglich. Symbolischer Austausch und Kämpfe um Legitimation und Deutungshoheit können so prinzipiell erforscht werden, anstatt sie durch die Behauptung eines umfassenden sozialen Systems auszuklammern.

Sowohl in der Perspektive der autopoietischen Systemtheorie, als auch in der handlungstheoretischen Perspektive der Memetik ist die Entstehung und der Wandel symbolischer Ordnungen ein evolutionärer Prozess der Rückkopplung von Variation und Selektion. Dabei ist es letztlich unerheblich, ob die Auswahl aufgrund bewusster Entscheidungen oder durch bloßes Nachahmen geschieht. Die zusätzliche Annahme, dass dieser Prozess autopoietisch geschlossen ist, löst einige theoretische Probleme, insbesondere hinsichtlich der Interaktion von Systemen, die als strukturell gekoppelt aufgefasst werden. Da strukturelle Kopplung annimmt, dass es keinen direkten Austausch von Information geben kann, bleibt nur die Anpassung der Kommunikation und Verhaltensweisen im Laufe einer gemeinsamen Koevolution der beteiligten Systeme. Wie in der Diskussion um die systemtheoretische Auffassung von Sprache schon gezeigt, führt dies letztlich aber auch dazu, dass Sprache selbst nicht als System aufgefasst werden kann und das erfolgreiche Kommunikation ein unwahrscheinliches Phänomen ist. Folglich liegt die Hauptfunktion der Sprache und des Sprechens in der sozialen Koordination und der Weitergabe tradierten Wissens. Diejenigen Informationen, die sich aus welchen Gründen auch immer (z.B.: Nachahmung, rationale Wahl, Affekte) durchsetzen, werden weitergegeben und formen die symbolischen Ordnungen.

Diese Sichtweise beinhaltet ein grundsätzliches Problem. Der Prozess der Kommunikation wird darin zu einem Rückkopplungsprozess, der zwar zu einer strukturellen, wechselseitigen Anpassung der Systeme führen kann, dafür aber auf lange Sicht den Preis der beständig sinkenden Komplexität zu zahlen hätte. Dies ergibt sich aus der Arbeitsweise geschlossener Systeme und dem zweiten Hauptsatz der Thermodynamik:

> In an open system increase of order and decrease of entropy is thermodynamically possible. The magnitude, „information," is defined by an expression formally identical with negative entropy. However in a closed feedback mechanism information can only decrease, never increase, i.e., information can be

> transformed into „noise," but not vice versa. (Bertalanffy 1969: 50)

Die Phänomene Sprache, Kommunikation und symbolische Ordnung weisen jedoch zweifelsfrei eine Tendenz zu steigender Komplexität auf. Das wahrscheinlich beeindruckendste Beispiel dafür ist der enorme Zuwachs des Internetverkehrs, der sich selbst nach vorsichtigen Schätzungen jedes Jahr verdoppelt (vgl. Coffman und Odlyzko 2002). Das eigentliche Argument ist jedoch grundsätzlicher Natur, ein geschlossener Rückkopplungsprozess kann für sich genommen den Prozess der *Variation* und der *Selektion* erklären, daraus geht jedoch noch lange keine zwangsläufige Tendenz zu Systemen mit komplexerer Organisation hervor. Die von Elias beschriebene Entwicklung hin zu höheren Ebenen der Synthese, ebenso wie Modernisierung als ein Prozess immer komplexerer Modellierungen der Wirklichkeit sind im Konzept autopoietisch geschlossener Systeme nur schwer zu fassen. Allerdings muss man klar sagen, dass auch die Theorie offener Systeme hier keinen eindeutigen Lösungsmechanismus anbietet, sondern nur die prinzipielle Möglichkeit zulässt.

Bei der Betrachtung symbolischer Programme, also derjenigen symbolischen Ordnungen, die regelmäßige Abläufe spezifizieren wird deutlich, dass die Dynamik der symbolischen Entwicklung ein Prozess ist, der in der Lage ist seine eigene *Komplexität* zu steigern. Symbolische Programme unterscheiden sich von anderen symbolischen Ordnungen durch ihre Fähigkeit zur Selbstreproduktion aus den in ihnen kodierten Informationen heraus. Das kodierte Information prinzipiell dazu in der Lage ist, sich selbst zu reproduzieren und komplexe Maschinen hervorzubringen, wissen wir vor allem aus dem Bereich der Biologie und der Informatik. DNA beispielsweise, ist nicht nur in der Lage Kopien ihrer selbst zu erzeugen, sondern komplexe Organismen zu generieren, die in der Lage sind auf Umwelteinflüsse zu reagieren und so zu einer erfolgreichen Weitergabe beizutragen. Was dabei im Zeitverlauf herauskommt, sind immer komplexere Programme der Beobachtung und Interaktion mit der Umwelt. Der formale Beweis, das bereits recht einfache Regeln solche komplexitätssteigernden Systeme hervorbringen können, die turing-vollständig sind, d.h. auf denen jeder uns bekannte Algorithmus ausführbar wäre, wurde von John Conway mit seinem „Game of Life" erbracht (vgl. Gardner 1970).

Die Erkenntnis, dass Information in Form von Regelwerken in der Lage ist prinzipiell endlose Komplexität hervorzubringen, kann jedoch nicht über das Problem hinwegtäuschen, dass bis jetzt keine ausreichende Erklärung für die grundlegenden „organizational laws" (Bertalanffy 1969:

155) vorliegt, die solche Prozesse hervorbringen könnten. Für die Frage nach einer grundsätzlichen Methodologie sozialer Symbole ist dies auch nicht notwendig. Vielmehr reicht an dieser Stelle die Feststellung aus, dass eine Erklärung für die Genese und den Wandel symbolischer Ordnungen unvollständig ist, wenn sie nicht anerkennt, dass dieser Wandel durch symbolische Programme gesteuert und aufrechterhalten wird, die aber ebenfalls Gegenstand dieses Wandels sein können.

Dies hat auch Auswirkungen auf die Konzeption der grundsätzlichen Steuerungsfähigkeit und die Rollen von Individuen und Gesellschaften vis-a-vis symbolischer Prozesse. Im Falle der individualistischen Perspektive ist es das unabhängige Enscheidungskalkül der Individuen das über Gültigkeit, Legitimität und Anwendung eines in symbolischen Ordnungen kodierten Wissens entscheidet. Demgegenüber stellt in der gesellschaftliche Perspektive der externe Selektionsdruck von Normen und Institutionen die Verbindung zwischen sozialen Phänomenen und symbolischen Ordnungen her. Symbolische Prozesse, in der hier vertretenen Auffassung, sind hingegen das Resultat beständiger sozialer Interaktion zwischen den Ebenen von individuellen psychischen Vorgängen und den sozialen Institutionen. Die aktive menschliche Kommunikation produziert die notwendige Dynamik um diesen Prozess am Leben zu erhalten, indem neue Kombinationen von Zeichen hervorgebracht werden. Auf der anderen Seite stellt die Rahmung der Kommunikation durch den bestehenden Raum sozialer Symbole die notwendige soziale Standardisierung und die Regeln der Konstruktion von symbolischen Ordnungen bereit. Somit sind Menschen keine einsamen Entscheidungkünstler, die den Gang der Dinge entsprechend ihres akkumulierten Willens beherrschen, da ihre Entscheidungen nur auf das Zurückgreifen können, was innerhalb einer bestimmten übergreifenden symbolischen Ordnung als Wissen verfügbar ist. Gleichzeitig sind wir aber auch keine Sklaven anonymer Kräfte der Kultur und Geschichte.

Aus soziologischer Sichtweise kommt noch eine weitere wichtige Funktion symbolischer Ordnungen hinzu, nämlich die Koordination von sozialen Handlungen mittels symbolischem Austausch und insbesondere symbolischen Programmen. Symbolische Ordnungen werden damit zu *sozialen Technologien*, die es sozialen Organisation erlauben relativ unabhängig von den beteiligten Personen zu werden. Die Errichtung von politischen Institutionen mittels Verfassungen und Rechtskodizes ist dabei wohl das beeindruckendste Beispiel der Errichtung einer sozialen „Maschine" mittels eines symbolischen „Quellcodes". Sehr viel subtiler, aber nicht weni-

ger wirkmächtig, ist die Wissenschaft, welche trotz enormer institutioneller Umwälzungen ihre zentrale Ausrichtung auf die Fragen des Schönen, Guten und Wahren, ebenso wie die Verfolgung dieser Ziele durch kritische und vernünftige Methoden nie aus den Augen verloren hat.

Mit der Definition symbolischer Prozesse, als zeitlich ausgedehnte, offene Systeme, gehen auch zwei grundlegende, methodische Probleme einher. Zum einen ist dies die Abgrenzung der Prozesse und zum anderen deren Messung in Abhängigkeit von der Zeit. Ersteres ist sowohl aufgrund der potentiellen Übertragung von Informationen zwischen Systemen, als auch wegen der Konzeption von symbolischen Ordnungen als spezifische Anordnungen von Symbolen problematisch. Die Analyse solcher relationaler Strukturen leidet in besonderem Ausmaß unter fehlenden Werten und Verschiebungen zwischen einzelnen Beobachtungszeitpunkten, eil jedes Element in Relation zu allen anderen steht. Daher kann ein einzelner Ausfall zu einer kompletten Rekonfiguration der Struktur führen. Dieses Problem ist insbesondere hinsichtlich der Stichprobenziehung aus Netzwerken bekannt (vgl. z.B.: Wasserman und Faust 1994: 33f). Das zweite Problem sind die generellen Probleme bei der Messung zeitlich ausgedehnter Vorgänge, die noch dadurch verstärkt werden, dass sich Veränderungen auf mehreren Ebenen und in unterschiedlichen Geschwindigkeiten ergeben. Daraus resultieren methodische Herausforderungen und Probleme des Umgangs mit entsprechenden Datenstrukturen (vgl. z.B.: Hedeker und Gibbons 2006: 2f; Blossfeld und Rohwer 2002: 5ff).

Die Annahme eines objektiven und eigenlogischen Charakters von sozialen Symbolen birgt aber auch die Gefahr von immensen Missverständnissen, wie sie insbesondere in der kritischen oder kultursoziologischen Tradition zu finden sind. Das Problem liegt in der Verwechselung von symbolischen Ordnungen und ihren intersubjektiv nachvollziehbaren Bedeutungen mit menschlichen Intentionen und Handlungslogiken. Gerade weil Menschen sich und ihre Umwelt fasst ausschließlich in Symbolen begreifen, liegt es nahe dass es auch eine Tendenz in die andere Richtung geben kann, sprich das sozialen Symbolen menschliche Eigenschaften zugeschrieben werden. Dies sieht man gerade in der heutigen Zeit an einer Vielzahl von Diskursen, die alle möglichen Texte, Filme und andere kulturellen Artefakte durchforsten und diese als „sexistisch", „rassistisch" und dergleichen einstufen. Hierbei handelt es sich um eine grobe Verwechselung von Kategorien. Die Abwertung von Personen und Gruppen aufgrund bestimmter Merkmale, setzt die Fähigkeit zu Werturteilen voraus, etwas über das symbolische Ordnungen definitiv nicht verfügen.

Ein eindrucksvolles Beispiel ist der von Anna Hickey-Moody verfasste Artikel *Carbon Fibre Masculinity*, der im *Journal of the Theoretical Humanities* veröffentlicht wurde. Darin stellt die Autorin fest: „carbon fibre can be a homosocial surface; that is, carbon fibre becomes both a surface extension of the self and a third-party mediator in homosocial relationships, a surface that facilitates intimacy between men in ways that devalue femininity in both male and female bodies" (Hickey-Moody 2015: 139). Hier wird die metaphorische Beschreibung eines Gegenstands als „männlich" und „effizienzsteigernd" zur wirkmächtigen, symbolischen Gewalt hochstilisiert. Das eigentliche Problem liegt aber im Wort „devalue". Um es klar und unmissverständlich auszudrücken: Symbole können zur Bewertung eingesetzt werden, aber der Akt der Bewertung setzt Intention und Handlungsfähigkeit voraus, Eigenschaften welche symbolische Ordnungen schlicht nicht besitzen. Damit soll nicht gesagt werden, dass es keine erkennbaren Muster in der Art und Weise gibt, mit denen Symbole zur Grenzziehung und Unterscheidung von allen Möglichen Dingen und Personen eingesetzt werden. Aber die damit verbundenen Werturteile sind diesen symbolischen Ordnungen nicht inhärent.

3.4 Methodologische Schlussfolgerungen

Die vorangegangenen Ausführungen haben gezeigt, dass sozial standardisierte Zeichen, welche hier als Symbole bezeichnet werden, und die aus ihnen gebildeten, relationalen Strukturen eines der grundlegenden Phänomene der Soziologie darstellen. Folglich nehmen sie eine zentrale Stellung in einer Vielzahl soziologischer Theorien und Forschungsprogramme ein. Gleichzeitig findet aber fast immer eine Unterordnung unter die jeweilige theoretische Perspektive statt, die sich meistens entlang der Unterscheidung von individualistischer und gesellschaftlicher Ausrichtung vollzieht. Dies scheint sich wie ein fraktaler, roter Faden durch die Theoriegeschichte der Soziologie zu ziehen. Als Versuch einer Synthese und Überwindung dieses Gegensatzes, wurde hier ein Modell symbolischer Prozesse vorgestellt und diskutiert. Gerade weil es sich hier um axiomatische Voreinstellungen handelt, ist diese Synthese nicht dazu gedacht die bestehen Sichtweisen zu ersetzen. Vielmehr soll damit gezeigt werden, dass diese Perspektive unterschiedliche Fragestellungen ermöglichen, die jedoch in vielen Fällen komplementärer Art sind. Eine Methodologie der Erforschung sozialer Symbole kann zu zentralen Fragestellungen aller drei Perspektiven einen entscheidenden Beitrag leisten.

In jeder der drei hier untersuchten Perspektiven füllen soziale Symbole und symbolische Ordnungen eine wichtige theoretische Leerstelle ohne die das restliche Theoriegebäude nicht stehen könnte. Sie erfüllen stets eine spezifische Funktion, ohne die soziale Interaktion und Handeln nicht vorstellbar wären. Diese Korrespondenz unterschiedlicher Perspektiven auf das Soziale mit einzelnen Funktionsbereichen von Symbolen ist auch Norbert Elias aufgefallen, der in der Aufklärung dieses Zusammenhangs eine zentrale Aufgabe einer Theorie sozialer Symbole sah:

> The present tendancy to treat knowledge, language and thought as independent, perhaps even as separately existing items without bothering much about their relationship, can itself serve as example. In traditional language one might say, it is symptomatic of an imbalance in the relationship between analysis and synthesis in favour of the former. To correct this imbalance is one of the aims of a theory of symbols. It would lead to far here to indicate in greater detail the complex relationship between analysis an synthesis. It must be enough to remain within the confines of this example and to indicate briefly that all three activities or products refer to perspectives of symbols: knowledge mainly to the function of symbols as means of orientation, language mainly to their function as means of communication, thought mainly to their function as means of exploration, usually at a high level of synthesis and without any action at a lower level. (Elias 1991: 70f)

Eine von den Individuen ausgehende Soziologie, sieht symbolische Ordnungen als Wissen an, welches das *Entscheidungskalküls* von Personen informiert. Unabhängig von der genauen Konzeption von Entscheidung in den unterschiedlichen Theorierichtungen, sehen sie die Notwendigkeit individuellen Wissens um den zielgerichteten und sinnhaften Charakter menschlicher Handlungen zu erklären. In einer gesellschaftlichen Perspektive gilt das Gleiche für die *Koordinierungsfunktion* symbolischer Ordnungen. Ohne die Existenz sozial standardisierter Zeichen und Bedeutungen wäre keine weitreichende Koordination sozialer Handlungen möglich und damit auch keine Gesellschaft. Die prozessorientierte Perspektive erweitert diese Betrachtungen um die *Explorationsfunktion* des Denkens mittels symbolischer Ordnungen. Dies entspricht in dem hier vertretenen Prozessmodell dem iterativen Wechselspiel von individuellen Sprechakten mit einer übergreifenden symbolischen Ordnung. Anders

ausgedrückt, Denken ist ein Gespräch, gleichgültig ob dies im Inneren der Person stattfindet oder mit externen Personen.

Obwohl unterschiedliche Perspektiven andere Aspekte symbolischer Ordnungen betonen, handelt es sich doch stets um ein und dasselbe Phänomen. Egal ob von Wissen, Kultur, Skripten oder praktischen Rationalitäten die Rede ist, gemeint sind damit symbolische Ordnungen, d.h. die Kodierung von Bedeutungen in spezifischen Relationen von sozial standardisierten Zeichen. Es ist diese Auffassung aus der die Forderung resultiert Symbole als ein eigenständiges, genuin soziales Phänomen ernst zu nehmen. Gleichzeitig deutet es darauf hin, dass die theoretische Perspektive eine untergeordnete Rolle spielt. Vielmehr erscheint eine Methodologie zur Erforschung sozialer Symbole unabhängig von der fraktalen Spaltung der Soziologie möglich und auch lohnenswert.

Allerdings wurde in der Erörterung der einzelnen Ansätze auch deutlich, dass der Stellenwert sozialer Symbole innerhalb der einzelnen Theorierichtungen nicht das eigentliche Problem ist. Es ist nicht die Axiomatik, die zu wünschen übrig lässt. Vielmehr ist es die Messung und Operationalisierung symbolischer Ordnungen, die bisher vernachlässigt worden ist. Zum Beispiel nehmen Frames eine wichtige Rolle in der modernen Handlungstheorie ein, werden aber nicht direkt gemessen, sondern post facto als Erklärung der geringen Varianz von beobachteten Handlungenstrategien angeboten. Ebenso wird in der gesellschaftlichen Perspektive oft die Wichtigkeit einer übergreifenden Kultur hervorgehoben, diese jedoch ebenfalls meist nur durch deren Artefakte, wie etwa bestimmte Institutionenordnungen und Kulturgüter, bestimmt. Dies ist vor allem deshalb problematisch, weil es ein reichhaltiges Methodenwissen zur Analyse von Zeichensystemen gibt, welches über eine Vielzahl von Disziplinen hinweg mit großem Erfolg eingesetzt wird.

4 Das Verhältnis qualitativer und quantitativer Textanalyse

Die sozialwissenschaftliche Analyse von Symbolen, Texten und Bedeutungen wird oft als eine mehr oder minder ausschließliche Domäne der qualitativen Forschung gesehen. Demgegenüber hat allerdings auch die quantitative Analyse dieses Gegenstandsbereichs eine Vielzahl von Forschungserfolgen hervorgebracht, die jedoch fast ausschließlich in anderen Disziplinen, wie der Linguistik, der Informatik oder der Narratologie, stattfanden. Dementsprechend wurden die dort entwickelten Verfahren und Methoden in den Sozialwissenschaften, bis auf wenige Ausnahmen, bisher kaum aufgegriffen. Allerdings kann ein zunehmender Trend zum Einsatz von Verfahren der quantitativen Textanalyse auf Forschungsfragen festgestellt werden, die normalerweise dem qualitativen Forschungsprogramm zugerechnet werden würden. Darunter fällt beispielsweise der Vergleich medialer Diskurse (vgl. Bail 2012) oder die Analyse des Wechselspiels von semantischen Ähnlichkeiten und Koautorenschaft (vgl. Leydesdorff und Welbers 2011).

Dies verweist auf die Notwendigkeit die Gemeinsamkeiten, Unterschiede sowie Kombinationsmöglichkeiten beider Perspektiven kritisch zu würdigen. Es kann und soll dabei nicht um ein entweder oder gehen, sondern um die Frage was ein Methodologie der quantitativen Textanalyse von ihrem qualitativen Gegenstück lernen kann und wo Möglichkeiten zu einer Kombination beider Vorgehensweisen bestehen.

Die qualitative Sozialforschung hat ihren Ursprung in der ethnographischen Untersuchung primitiver Gesellschaften (vgl. z.B.: Malinowski 1979; Mead 1965). Diese Ansätze wurde in den fünfziger Jahren in die Soziologie übernommen und auf die Untersuchung von Alltagshandeln und Subkulturen moderner, westlicher Gesellschaften übertragen (vgl. z.B.: Goffman 1961; Whyte 1973). Im Laufe der Zeit wurde die ethnographische Beobachtung immer mehr von der Analyse schriftlicher Zeugnisse, insbesondere den Transkripten von Interviews, abgelöst. Spätestens seit den Konversationsanalysen (vgl. Schegloff 1989) der Ethnomethodologen

kann man davon sprechen, dass geschriebene Texte die dominante Datengrundlage der qualitativen Sozialforschung geworden sind. Das Erbe der Ethnographie wird dennoch deutlich, vor allem in der grundsätzlichen Ausrichtung der qualitativen Sozialforschung. Dazu zählen im Besonderen die Tendenz zur „dichten Beschreibung" (Geertz 1999) der sozialen Wirklichkeit und der Versuch der Herstellung subjektiver Nachvollziehbarkeit durch *Verstehen*. Im deutschsprachigen Raum kommt außerdem eine lange Tradition des interpretativen Verstehens im Sinne der „Hermeneutik" Friedrich Schleiermachers hinzu.

Die daraus hervorgegangene qualitative Sozialforschung zeichnet sich durch eine große Vielfalt an spezifischen Methoden und methodologischen Festlegungen aus. Neben sehr unterschiedlichen praktischen Auffassungen gibt es auch eine Reihe von Strömungen, die in der qualitativen Sozialforschung sehr viel mehr als nur eine Methode sehen, sondern beispielsweise eine eigenständige „Interpretative Soziologie" (Keller 2012). Generell lässt sich beobachten, dass die meisten qualitativen Ansätze mit sehr weitreichenden und sich gegenseitig ausschließenden Annahmen über den Gegenstandsbereich einhergehen.

Dennoch lassen sich grundsätzliche Merkmale der qualitativen Forschung feststellen. Insbesondere wenn man Lehrbücher (vgl. Brüsemeister 2008; Flick 1995; Lamnek 2005; Mayring 2002) zur qualitativen Sozialforschung betrachtet, lassen sich trotz geringfügiger Unterschiede eine Reihe von allgemeinen Grundannahmen angeben, die charakteristisch für den Großteil der qualitativen Forschung sind:

1. Analyse von Text.
2. Interpretatives Verstehen.
3. Subjekte als zentrale Kategorien.
4. Induktiv-exploratives Vorgehen.

Die Vielfalt der unterschiedlichen Methoden ist in dieser Liste bewusst nicht enthalten. Obwohl Methodenvielfalt in den Lehrbüchern für gewöhnlich als ein zentrales Merkmal der qualitativen Sozialforschung genannt wird, soll hier die Auffassung vertreten werden, dass es sich dabei eher um eine Beschreibung der disziplinären Beschaffenheit und des mühsam errungenen Burgfriedens handelt als um eine eigenständige, methodische Grundannahme.

4.1 Qualitative Textanalyse

Die verborgene Einheit der qualitativen Sozialforschung lässt sich am einfachsten an der Debatte über qualitative Daten illustrieren. In den meisten Fällen werden qualitative Daten als „Interviewdaten" aufgefasst (vgl. Breuer et al. 2014; Helfferich 2005). Zusätzlich wird auch von „Visual Data" im Sinne von Filmen, Videos und Bildern (vgl. Ball und Smith 1992; Denzin 2000; Harper 2000), ethnographischen „Beobachtungsdaten" (vgl. Lüders 2000), sowie „elektronischen Prozessdaten" (vgl. Bergmann und Meier 2000) gesprochen. Die ausschlaggebende Unterscheidung ist hier anscheinend die Art der Erhebung und nicht das Format in dem die Daten gefasst werden. Es ist sicherlich wichtig, die Erhebungsinstrumente zu reflektieren und ihre Auswirkungen auf die generierten Daten zu berücksichtigen. Allerdings führt dies nicht zu unterschiedlichen Arten von Daten. Auch die Autoren fassen dies nicht so auf. Vielmehr scheint diese Unterscheidung der Konzentration des Forschenden auf ein spezifisches Verfahren geschuldet zu sein, die wohl ein Spezifikum des Diskurses der qualitativen Sozialforschung ist. Aufs Ganze genommen führt diese Spezialisierung der individuellen Forscher auch zur beschriebenen Methodenvielfalt des Forschungsprogramms.

Von einem analytisch, technischen Standpunkt aus ließe sich argumentieren, dass es nur eine Form von Daten gibt, da alle schematischen Repräsentationen von Messungen (welcher Art auch immer) unabhängig vom Schema als Daten aufgefasst werden können. Nach dieser Definition ließe sich nur das Schema unterscheiden. Für Interviewdaten wäre zum Beispiel die Abfolge von Frage-Antwort Blöcken das gängige Schema in dem ein Ausschnitt der konkreten Interaktion repräsentiert wird. Allerdings wären auch andere Schemata für die Repräsentation denkbar, zum Beispiel könnten die einzelnen Frage-Antwort Blöcke thematisch und nicht zeitlich geordnet werden oder die Repräsentation könnte in Form eines gemalten Bildes geschehen, welches die Eindrücke des Interviewers auffängt. Zugegebenermaßen ist die letztgenannte Variante relativ unwahrscheinlich[1], aber sie hilft die Bandbreite der Möglichkeiten und die relative Unabhängigkeit des Schemas von der Messung heraus zu stellen. Der Vorteil dieser Definition liegt auch darin das Augenmerk von der Genese der Daten auf deren Struktur zu legen und dadurch technisch-praktische

[1] Tatsächlich werden „productive methods" als eigenständoge Erhebungsinstrumente genutzt, deren Endergebnis auch handgemalte Bilder sein können (McDonnell 2014: 249).

Fragen, zum Beispiel nach der Konvertierung, Aufbewahrung und kooperativen Bearbeitung, zu ermöglichen.

Abgesehen von dieser formalen Äquivalenz zeigt sich auch eine inhaltliche Gemeinsamkeit bezüglich der Tatbestände, welche durch qualitativen Datensätze normalerweise repräsentiert werden. Egal ob es sich um detaillierte Interaktionen, im Diskurs getroffene Äußerungen oder ethnographische Beobachtungen handelt, letztlich werden sozial standardisierte Symbole festgehalten. Dies ergibt sich zwangsläufig aus dem grundlegenden Selbstverständnis der qualitativen Forschung als „interpretativer Sozialwissenschaft" (vgl. Soeffner 2014) die am Verstehen orientiert ist. Im Anschluss an Max Weber (2006) wird dabei Verstehen als die „Erfassung des Sinnzusammenhangs" gedeutet, in dem eine bestimmte Handlung steht. Lassen wir die kontroverse Frage nach der Lokalisation dieses Sinnzusammenhangs (Individuum oder Gesellschaft; siehe Kapitel 3.2) einmal außen vor und betrachten nur die sozialen Handlungen, so lässt sich die Beziehung von konkretem Sinn einer Handlung und Sinnzusammenhang in den hier verwendeten Begriffen als der von symbolischer Äußerung und symbolischer Ordnung wiedergeben. Solange es der qualitativen Sozialforschung um das interpretative Verstehen geht, sind ihre Datensätze stets Repräsentationen von Symbolen, d.h. Texte. Dies gilt ebenso für Bilder, Videos und Tonbandmitschnitte, die mit dem Ziel erhoben wurden die Bedeutungen und Sinnzusammenhänge zu rekonstruieren. Als Text lässt sich im Prinzip jede Repräsentation von sozial standardisierten Zeichen auffassen. Auch wenn man von einer so breiten Definition absieht, lässt sich immer noch feststellen, dass schriftliche Texte (z.B.: Transkripte, Beobachtungsprotokolle, Feldtagebücher, etc.) eine zentrale Rolle in der qualitativen Sozialforschung einnehmen.

Obwohl jedes Verstehen ein grundsätzliches Wissen über die symbolische Ordnung und die darin kodierten Bedeutungen und Wissensbestände erfordert, wird im Rahmen der qualitativen Sozialforschung vor allem der subjektive Anteil an der Interpretationsleistung hervorgehoben. Die *ganzheitliche Subjektorientierung* wird als zentraler Grundsatz der qualitativen Forschung angesehen (z.B.: Flick, Kardorff und Steinke 2004: 21; Mayring 2002: 20). Gemeint ist damit einerseits die Subjektivität im Gegenstandsbereich, d.h. die möglichst umfassende Berücksichtigung der Personen in ihren Einstellungen, Interessen und Handlungsprinzipien. Diese Orientierung am Forschungssubjekt kommt auch in Ausdrücken wie „Lebensweltorientierung", „sozialer Kontext" und dergleichen zum Ausdruck. Des Weiteren erstreckt sich die Bedeutung des Subjektiven auch

auf den Forscher und dient als Erklärung und methodologische Begründung seiner Interpretationsleistung (vgl. Soeffner 2014). Dass sich auf beiden Seiten Subjekte gegenüberstehen, die prinzipiell zur Reflexion fähig sind, stellt in dieser Auffassung die Grundlage jeglicher Interpretation dar. Somit erscheint Sinn prinzipiell immer als eine Funktion von inneren Vorgängen im Bewusstsein. „Objektiver" Sinn ist dann bloße, äußere Repräsentation der Subjektivität und selbst eine objektive Betrachtung von Wissen und Kultur muss zwangsläufig wie Magie erscheinen:

> Nur Wesen mit höheren Einsichtsfähigkeiten als „wir alle" können den anderen ins Herz, ins Bewusstsein oder gar in das hineinschauen, was sich unter, hinter oder über deren Bewusstsein abspielt. Erleuchtete und begnadete Menschen – und vielleicht auch Strukturalistinnen und Strukturalisten und Objektivistinnen und Objektivisten – mögen das können. Interpretativ arbeitende Sozialwissenschaftlerinnen und Sozialwissenschaftler vermögen es jedenfalls nicht. Wie für jeden von uns, so ist auch für die Sozialwissenschaftlerin und Sozialwissenschaftler fremder, subjektiv gemeinter Sinn unabdingbar nur über „bezeichnende Indizien" – von „einfachen" körperlichen Appräsentationen bis hin zu komplexen kulturellen Objektivationen – rekonstruierbar und trivialer Weise eben keineswegs unmittelbar erfassbar. (Hitzler 2014: 64f).

Darin kommt ein zentrales Dilemma der qualitativ-interpretativen Schule zum Ausdruck. Sinn wird als inneres, subjektives Phänomen begriffen, kann aber offensichtlich immer nur durch objektive, äußerliche Symbole erfasst werden. Erkenntnistheoretisch problematisch scheinen hier vor allem diejenigen Lösungsansätze zu sein, welche sich tendenziell in Mystifikation und praxisorientierter Kunstlehre ergehen. Man sieht dies an der Bedeutung des Alltagswissens und der praktischen Erfahrung, die in allen qualitativen Methodenlehrbüchern eine zentrale Stellung einnimmt. Interpretative Methoden rekurrieren deshalb oft auf die „Lebendigkeit und Dynamik sozialer Prozesse" (Keller 2012: 316), welcher mit der „Kunst der Interpretation" (Bude 2004: 516) Rechenschaft getragen werden soll.

Das Resultat ist dann auch ein gelinde gesagt *zwiespältiges Verhältnis zur quantitativen Forschung* und der in diesem Bereich vertretenen methodischen Standards der Validität, Repräsentativität und Objektivität. Diese Abgrenzung wird vor allem durch den Kampfbegriff des „Positivismus"

markiert, welcher von verschiedenen Autoren in sehr unterschiedsloser Art und Weise gebraucht wird. Gemeint ist damit wohl am häufigsten der Positivismusbegriff der kritischen Theorie in Anschluss an Theodor Adorno (1976) und die Frankfurter Schule, der ebenfalls den Vorrang des Subjektes vor objektiver Erkenntnis anmahnte. Verwirrend an diesem Wortgebrauch ist die Tatsache, dass er an die Adresse des kritischen Rationalismus – insbesondere vertreten durch Karl Popper und Hans Albert – gerichtet war, der explizit gegen den logischen Positivismus (auch logischer Empirizismus) des *Wiener Kreises* angetreten war. Daher lässt sich hier wohl mit recht von einem ideologisch motivierten Begriff sprechen, der als solcher auch in neueren Debatten zu beobachten ist. So werden zum Beispiel Bemühung hin zu einer stärkeren Formalisierung und Qualitätskontrolle qualitativer Methoden als „neo-positivistische[r] Objektivismus" (Keller 2012: 177) bezeichnet.[2] Allerdings scheint es dabei mittlerweile fast nur noch um eine Abgrenzung gegenüber dem Anspruch der Formalisierung und objektiver Gültigkeitskriterien zu gehen. Die ursprüngliche Argumentation der kritischen Theorie, welche die Subjektzentrierung noch normativ zu begründen suchte, kommt jedenfalls kaum noch zur Anwendung.

Der starke Fokus auf subjektive Interpretation, bzw. interpretative Subjekte, erklärt auch die Tendenz zum induktiven Vorgehen. Wegen der Orientierung an Komplexität und Subjektivität sind die Fallzahlen meist relativ klein. Da Zeit die wichtigste Ressource einer tiefergehenden Interpretation ist, können bestimmte Analyseverfahren zudem nur für eine sehr begrenzte Anzahl an Fällen durchgeführt werden. Dem daraus resultierenden Problem mangelnder Repräsentativität und dem Vorwurf der Beliebigkeit versucht die qualitative Forschung mittels einer stärkere Einbettung in sozialwissenschaftliche Theorien zu begegnen (vgl. Diaz-Bone 2006; Keller 2012; Soeffner 2014). Unabhängig vom Erfolg dieses Vorgehens kann man es als Ironie der Philosophiegeschichte auffassen, dass die daraus resultierende induktiv-explorative Ausrichtung und die damit einhergehende Tendez zur Bestätigung theoretischer Vorrausannahmen, wahrscheinlich näher am logischen Positivismus des Wiener Kreises ist, als irgendeine andere gegenwärtige Forschungsperspektive.[3]

[2] Auch der logische Positivismus des Wiener Kreises wurde schon als „Neopositivismus" bezeichnet, da er sich in der Tradition des Positivismus von Auguste Comte sah.

[3] 'Eine solche Parallele hier im Detail aufzuzeigen, würde den Rahmen dieser Arbeit übersteigen. Dies ist vor allem auf die recht unterschiedlichen und weitreichenden Arbeiten der Mitglieder des Wiener Kreises zurück zu führen. Dennoch kann man sagen, dass die Vorstellung von der Begründung in Einzelbeobachtung (Induktion) und deren an-

4.2 Warum quantitative Textanalyse?

Die Feststellung von Gemeinsamkeiten und Unterschieden reicht natürlich nicht aus, um die praktische Entscheidung für oder gegen ein quantitatives Vorgehen zu treffen. Dementsprechend wird es im folgenden Kapitel um die Frage gehen, welche Möglichkeiten eine standardisierten Auswertung von Texten bietet und wie deren Bezug zum qualitativen Forschungsprogramm aussehen könnte.

Prinzipiell sprechen fünf Eigenschaften der quantitativen Textanalyse für deren Anwendung auf die Analyse symbolischer Ordnungen. Erstens, die *Angemessenheit* dieser Analyse an den Gegenstandsbereich. Zweitens, die Möglichkeit die im gegenwärtigen technologischen Wandels des Alltagslebens entstehenden Mengen an Daten bearbeiten zu können, sowie deren neuen Strukturen gerecht zu werden (*Machbarkeit*). Drittens, die Chance einer objektiveren und formaleren Analyse aufgrund expliziter Algorithmen, welche *Reproduzierbarkeit und Nachvollziehbarkeit* gewährleistet. Viertens, wird durch die Nutzung von Analyseverfahren anderer Disziplinen (z.B.: Linguistik, machine learning) die *Anschlussfähigkeit* an einen größeren, interdisziplinären Fachdiskurs ermöglicht und damit die Möglichkeit für neue Erkenntnisse geschaffen. Zu guter Letzt muss es auch nicht auf eine entweder oder Entscheidung zwischen quantitativer und qualitativer Textanalyse hinauslaufen, da es vielversprechende Möglichkeiten der Kombination beider Vorgehensweisen gibt (*Integrierbarkeit*).

4.2.1 Angemessenheit

Qualitative Verfahren rühmen sich insbesondere der Subjektivität und Komplexität von Symbolen und den darin kodierten Bedeutungen gerecht zu werden. Demgegenüber erscheint ein quantitatives Vorgehen problematisch, da Zeichen hier auf zählbare Einheiten verkürzt werden, die dann in Häufigkeiten zerlegt werden. Der gängige Vorwurf lautet, damit gehe gerade das Einzigartige und das Subjektive jeder symbolischen Äußerung verloren.

schließende Verifikation mit logischen Verfahren zentraler Bestandteil des Positivismus waren (vgl. Creath 2004). Ähnliche Annahmen scheinen auch das qualitative Forschungshandeln anzuleiten, welches ebenfalls von einzelnen Beobachtungen zu theoretischen Sätzen gelangt. Dass das Bindeglied dabei „Sinn“ und nicht „Logik“ darstellt ist nur ein minimaler Unterschied in der Terminologie. Auch im soziologischen Sinnbegriff wird eine deterministische Beziehung von Gründen und daraus resultierenden Handlungen angenommen, da jegliche Interpretation sonst schlicht unmöglich wäre.

Wie in den vorangegangen Kapitel schon ausführlich besprochen geht die hier vertretene Perspektive davon aus, dass Symbole und symbolische Ordnungen sowie die darin kodierten Bedeutungen keine rein subjektiven Phänomene sind. Sie müssen stets auch einen sehr objektiven Charakter haben, da sie ansonsten ihren zentralen Zweck nicht erfüllen könnten, wegen dem sie seit Jahrtausenden von Menschen kultiviert und verfeinert werden, nämlich intersubjektiv nachvollziehbare Kommunikation zu ermöglichen. Es mag eine letztlich nicht weiter reduzierbare Subjektivität in der Entschlüsselung und Konstruktion von symbolisch übermittelten Informationen zwischen Menschen geben, aber dieser Bereich scheint letztlich nicht sonderlich groß zu sein und kann aufs Ganze der symbolischen Ordnung gerechnet wohl eher vernachlässigt werden. Nimmt man das Gegenteil an, so müsste man zugestehen, dass sich der Großteil der menschlichen Kommunikation und des menschlichen Wissens um Missverständnisse dreht. Und zwar spezifisch um eine Störung der Übertragung vom Sender zum Empfänger einer Botschaft und nicht um bewusste Täuschung, die ebenfalls nur dann funktionieren kann wenn die Informationsübertragung in berechenbarer Weise gelingt. Auch faktisch falsche Information ist hierbei nicht gemeint, da diese ebenfalls nur mittels relativ objektiver und intersubjektiv nachvollziehbarer Symbole ausgetauscht werden kann.

Sicherlich gibt es solche „richtigen" Missverständnisse, die sich aus unterschiedlichen subjektiven Einschätzungen der Symbole ergeben, aber sie werden durch zwei sehr einfache Mechanismen in Schach gehalten. Zum einen durch die Bestimmung einzelner Symbole als Verweise in einem komplexen System aus Symbolen. Anders ausgedrückt, Bedeutungen existieren nicht an und für sich, sondern nur als eine Struktur von Verweisen. Diese Struktur macht es, je formaler die eingesetzte Sprache und je länger der produzierten Text ist, umso schwieriger die Bedeutungen nicht in ihrer intendierten Weise zu erfassen. Man sieht dies in besonderer Weise an symbolische Techniken, wie der Poesie oder Rätseln, die vorzüglich mit verschiedenen Bedeutungsebenen arbeiten. Wären Sprache und Kommunikation tendenziell missverständlich und schwer zu entschlüsseln, so hätte man wohl kaum aufwendige Verfahren zur ästhetisch anmutenden Verschleierung von Bedeutung entwickeln müssen. Der zweite Grund für die Behandlung von Symbolen als objektiven Gegenständen ist ihre physikalische Gegebenheit. Sie existieren als Schriftzeichen auf dem Papier, elektrische Impulse auf Schaltkreisen oder als Schallwellen in der

Luft. Wie auch immer ihre Beschaffenheit, sie sind Phänomene der Welt und können damit auch als solche erfasst und beobachtet werden.

Die Notwendigkeit sich der Erforschung von Kultur – bzw. der symbolischen Ordnung in der hier verwendeten Terminologie – auf einem quantitativen und formalen Weg zu nähern, wird mittlerweile auch im soziologischen Theoriediskurs rezipiert. Im 2014 erschienen Sonderband *Measuring Culture* der Zeitschrift *Theory and Society* spielt diese Frage eine zentrale Rolle:

> The impetus for our gathering was our shared sense that although social scientific studies of 'culture' have made great strides on conceptual clarification of the concept (its many meanings), we still have made but little progress on the problem of how to measure culture (the ways of operationalizing it, determining appropriate indicators for it, breaking it down into observable analytic units, and thus studying it). How can we represent the social world and provide some observational grounding to it? How can we coordinate our theoretical concepts — something like 'culture', for instance, with our observational procedures for studying it? And how can we diversify our measurement outcomes to include theoretical, perceptual, numerical, statistical, and qualitative representations, all of which can advance our knowledge about the state of an empirical object? (Ghaziani 2014: 228)

Die im Rahmen dieses Sonderbandes erschienen Artikels enthalten neben methodologischen Äußerungen auch eine Reihe von praktischen Beispielen aus der Forschung. Diese zeigen recht eingängig, dass Kultur hier zwangsläufig als ein objektives Phänomen aufgefasst werden muss, wenn man damit empirische Forschung betreiben will. Andererseits wird auch deutlich, dass eine Messung von Kultur sowohl auf der Ebene des Individuums als auch auf der Ebene der Symbole ansetzen kann. Im erstgenannten Fall wird versucht durch die gleichartigen Reaktionstendenzen und Antwortschemata einer Mehrzahl vom Menschen auf die dahinterliegenden Kulturmuster zu schließen (vgl. z.B.: Ghaziani 2014; Vaisey und Miles 2014). Darüber hinaus gibt es auch eine Reihe von Versuchen diesen kulturellen Bedeutungsschemata direkt durch kulturelle Artefakte (Texte, Bilder, Lieder, etc.) auf die Spur zu kommen (vgl. z.B.: Ghaziani 2014; McDonnell 2014). In beiden Fällen ist das Ziel dasselbe, nämlich eine objek-

tive, d.h. von den Individuen relativ unabhängige, Struktur von Bedeutungen zu analysieren.

Trotz der Vielfalt der Datenquellen, lässt sich hier die Tendenz erkennen, die Erforschung symbolischer Ordnungen durch die direkte Beobachtung kultureller Artefakte, d.h. ohne ein Subjekt als Mittelsmann, und nicht mit quantitativen Verfahren zu betreiben. Es sind hauptsächlich diejenigen Daten deren zentrale Merkmalsträger Subjekte darstellen, die mit komplexeren, statistischen Verfahren behandelt werden. Dabei scheint es sich um eine allgemeine Tendenz in der soziologischen Methodenlehre zu handeln, da kulturelle Artefakte (z.B.: Bücher, Bilder, etc.) meist als dieser Art der Analyse nicht oder nur sehr schwer zugänglich angesehen werden.

Der schwerwiegendere Grund dürften jedoch methodische Konventionen gepaart mit fehlender, technischer Expertise sein. So verweisen einige Autoren des Sonderbandes explizit auf die Existenz solcher Methoden und deren vielversprechenden Ergebnisse ohne Sie jedoch selbst anzuwenden (vgl. Bail 2014; Vaisey und Miles 2014: 322). In seiner durchaus konstruktiven Kritik der Verfahren der quantitativen Textanalyse bringt Christopher Bail (2014: 467f) die Befürchtungen der kulturellen Soziologie auf den Punkt:

> For all the promise of big data for cultural sociology, formidable obstacles remain. First of all, the sheer volume of data can be overwhelming. Large corpora cannot be coded by hand, and automated data mining techniques are of little utility if they are not guided by theory. Second, big data is untidy. Although computer-assisted data classification and data reduction techniques have improved in the past decade, much big data analysis remains computationally intensive and therefore out of reach for many cultural sociologists—particularly those without any background in statistics or computer programming. Third—and perhaps most importantly—there is much that is of interest to cultural sociologists that is not easily reducible to text. The greatest challenge for cultural sociologists interested in big data is to develop new techniques to measure the unspoken or implicit meanings that occur in-between words.

Die ersten beiden Probleme sind vor allem technischer Natur und können dementsprechend durch bessere Kenntnisse im Umgang mit solchen Datenstrukturen behoben werden. Es ist Bail durchaus bewusst, dass es sich hierbei keinesfalls um ein Problem des Gegenstandsbereichs, son-

dern um eines der technischen Expertise der Disziplin handelt. Das dritte Problem offenbart dann auch die grundsätzliche Bewältigungsstrategie der kulturellen und interpretativen Soziologie, nämlich das Umdefinieren der Fragestellung, indem auf Komplexität und Subjektivität verwiesen wird. Dahinter verbirgt sich ein non sequitur. Denn aus der bloßen Feststellung, dass es andere Phänomene gibt, lässt sich nicht zwingend folgern, dass auf diese mehr Wert gelegt werden sollte.

Dies scheint dann auch das grundsätzlichere Problem der Auseinandersetzung zwischen qualitativer und quantitativer Sozialforschung zu sein. Natürlich ist Kritik an methodischem Vorgehen grundsätzlich angebracht, aber es ist schon sehr auffällig, dass diese Kritik meist ohne technische Kenntnisse in diesem Bereich auskommt und statt einer ernsthaften Auseinandersetzung mit den Möglichkeiten neuer Analyseverfahren lieber den Gegenstandsbereich für ungültig erklärt. Ebenso problematisch ist aber auch die Feststellung der mangelnden Expertise der Sozialwissenschaften in diesem Bereich. Egal wie sehr die Diagnose zutreffen mag, sie taugt nicht als Begründung um Verfahren im Voraus abzulehnen. Im Gegenteil, mangelnde Expertise in einem bestimmten Bereich sollte entweder Anlass genug sein sich mit Kritik zurück zu halten oder sich die notwendigen Kenntnisse anzueignen, um sich auf konstruktive Weise am Diskurs beteiligen zu können. Daher muss man die Feststellung, dass computergestützte Verfahren der Analyse großer Textmengen der „guidance of theoretically and qualitatively oriented cultural sociologists" (ebd.: 478) bedürfen, äußerst kritisch beurteilen. Wie sollte eine „guidance" aussehen, die auf der (korrekten) Einschätzung basiert, dass es innerhalb der Disziplin so gut wie kein technisches Wissen zu diesem Bereich gibt?

Dieses Gebaren ist umso problematischer, wenn man sich vor Augen führt, dass die Verfahren zur quantitativen Analyse von Texten seit Jahrzehnten erfolgreich in einer Vielzahl von Bereichen angewendet werden und mittlerweile in viele Aspekte unseres Alltags integriert sind.[4] Die Bedeutung dieser Technologien und deren Verknüpfung lassen sich insbesondere am Beispiel des „Information Retrievals" demonstrieren:

> Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). (Manning 2008: 1)

[4] Zahlreiche Beispiele finden sich in Landauer et al. (2013).

Im Unterschied zu gängigen Datenbankabfragen zielen diese Techniken nicht darauf ab vorher abgelegte Informationen nach einem festen Schema zurückzugeben. Ein Beispiel für ein solches festes Schema wäre die Abfrage des Kundenkontostandes durch die Eingabe der Kundennummer. Stattdessen sollen im Information Retrieval Informationen herausgelesen werden, die dem entsprechen was mit der Suchanfrage „gemeint" war. Dazu ist es natürlich notwendig einen Algorithmus zu haben, welcher der menschlichen Art Fragen an Texte zu stellen und diese zu verstehen möglichst ähnlich ist.

Die Angemessenheit der Verfahren an das Problem des algorithmischen Verstehens von Texten zeigt sich auf zweierlei Arten. Einerseits kann dies am beispiellosen praktischen Erfolg dieser Verfahren verdeutlicht werden. Wann immer wir Anfragen an Suchmaschinen stellen, Vorschläge in Online-Kaufhäusern bekommen oder auch nur nach Dateien auf einem Computer suchen bedienen wir uns solcher Algorithmen. Ohne sie wäre die Bewältigung der immensen und stetig wachsenden Menge an Information nicht möglich. Der zweite Grund ist theoretischer Natur. Das dem Information Retrieval zugrundeliegende Modell von „verstehenden" und „lernenden" Algorithmen stellt eine praktische Lösung von „Plato's Problem" dar, welches auch als „Poverty of the Stimulus" bekannt ist (siehe auch die Unterscheidung von Heuristiken und Algorithmen in Abschnitt 6.2). Darin geht es um die Frage, wie Menschen neue Worte und Grammatik problemlos lernen können ohne mit den formalen Regeln vertraut zu sein, d.h. schlichtweg durch Beobachtung von Sprechakten. Man hat versucht dieses Problem durch eine angeborene Universalgrammatik zu lösen.

Die Techniken des Information Retrieval, insbesondere die *Latent Semantic Analysis*, legen jedoch eine einfachere Lösung nahe (vgl. Landauer und Dumais 1997). Ein Vektorraummodell der Sprache, welches davon ausgeht, dass die in Texten zum Ausdruck gebrachten Bedeutungen latente Dimensionen darstellen, die in den Relationen der Wörter kodiert ist (siehe hierzu auch Abschnitt 6.5.1). In diesem Modell ist es die Struktur von Texten und Wissen, welches es möglich macht daraus zu lernen. Die Angemessenheit dieses Verfahrens konnte durch den Nachweis gezeigt werden, dass sich die Fähigkeit der LSA zur Erschließung neuer Bedeutungen aus Texten in etwa mit der von Schulkindern vergleichen lässt (vgl. ebd.).

Was hier für den Bereich des Information Retrievals diskutiert wurde gilt natürlich auch für eine Vielzahl anderer Verfahren der quantitativen

Textanalyse, wie z.B. Machine Learning (siehe Abschnitt 6.2.1) oder Readability Scores (siehe Abschnitt 5.7.2). Auch hier deuten sowohl der praktische Erfolg als auch die theoretische Passgenauigkeit auf eine große Angemessenheit dieser Verfahren zur Analyse von Texten. Zudem weist die ausführliche und seit Jahrzehnten betriebene Forschung in diesem Bereich darauf hin, dass es nicht die Informatik und die Linguistik sind, die ihre Hausaufgaben nicht gemacht haben.

Es wird hier aber nicht dafür plädiert unsere Erfahrung in der qualitativen Analyse von Texten aufzugeben, vielmehr sollten wir sie auf konstruktive Weise in den Diskurs mit einbringen. Die unerlässliche Voraussetzung dafür ist jedoch die Entwicklung der notwendigen, technischen Expertise in den Sozialwissenschaften.

4.2.2 Machbarkeit

In den letzten zwanzig Jahren sind neue Informationstechnologien und digitale Medien zentraler Bestandteil fast sämtlicher sozialer Prozesse in industrialisierten Gesellschaften geworden. Von der automatisierten Steuerung der industriellen Produktion, über die Abwicklung von administrativen Vorgängen hin zur Beziehungspflege und -anbahnung in sozialen Netzwerken. Diese Technologien erzeugen enorme Datenmengen, die zum ersten Mal in der Geschichte der Menschheit die detaillierte Konstruktion der zugrundeliegenden sozialen Phänomene erlauben. Es ist quasi möglich geworden soziale Interaktion und Organisation direkt und ihrem „natürlichen" Umfeld zu beobachten. Technologische, methodische und ethische Probleme scheinen die Erschließung dieses Datenlagers bisher noch gehemmt zu haben, wenn auch keineswegs aufgehalten. Dennoch gibt es keinen Zweifel daran, dass diese Entwicklung dieser Technologien in den kommenden Jahren stetig zunehmen wird.

Die Analyse solcher Datenmengen wird gemeinhin als „Big Data" bezeichnet. Man kann Burrows und Savage (2014) in ihrer Diagnose zustimmen, dass das Aufkommen von Big Data zu einer tiefgreifenden Krise der Sozialwissenschaften geführt hat. Als Grund geben die Autoren die enge Verknüpfung unseres disziplinären Selbstverständnisses mit den verwendeten Daten an. Wenn man davon ausgeht, dass die Soziologie tendenziell eine deskriptive und beobachtende Disziplin ist, so ist es folgerichtig anzunehmen, dass die Daten welche die Grundlage dieser Beobachtungen bilden auch das grundsätzliche Verständnis unseres Gegenstandsbereichs bestimmen. So kann zum Beispiel das Erstarken der sozialen Netzwerk-

analyse in den letzten Jahren sicherlich auch auf den enormen Zuwachs an Netzwerkdaten im selben Zeitraum zurückgeführt werden.

Da ein Großteil dieser neuen Daten Texte sind bietet deren Analyse die Möglichkeit zu ganz neuen Erkenntnissen zu gelangen. Prozessmodelle von Diskursen werden ebenso möglich, wie die Erforschung sprachlicher Interaktion ohne Probleme der Reaktivität. Es ist zum jetzigen Zeitpunkt noch nicht absehbar welche Potentiale diese Schwemme an kulturellen Artefakten nach sich ziehen wird. Sicher ist jedoch, dass diesen Datenmengen zunächst einmal nur mit standardisierten und computergestützten Verfahren beizukommen ist. Um sich ein Bild von den enormen Dimensionen dieser Daten machen zu können, sind ein paar Beispiele hilfreich:

- Ca. 30 Millionen digitalisierte Bücher in Google Books (vgl. Darnton 2013).[5]
- Facebook: ca. 40 petabyte gespeicherte Daten; Anstieg um ca.100 terrabyte pro Tag. (vgl. Kerzner & Maniyam 2013: 6)
- Grob geschätzte 10 exabytes aktiver Speicher der Google Searchengine.[6]

Es wird schnell deutlich, dass selbst verhältnismäßig kleinere Datenmengen, wie zum Beispiel Twitters 500 Mio. Kurznachrichten pro Tag[7], ohne Algorithmen und standardisierte Datenverarbeitung keiner Analyse zugänglich sind. Die Anwendung qualitativer Analyseverfahren ist hier zwar keineswegs ausgeschlossen, allerdings müssen dazu erst Verfahren des Text- und Data Minings angewendet werden, um handhabbarere Datenmengen zu erzeugen.

Weil sie die „processing capacity of conventional database systems" (Dumbill 2012) überschreiten, werden solche Datnemengen als Big Data bezeichnet. „Konventionell" bezieht sich hier allerdings auf die technischen Standards der Informatik bzw. der IT-Branche, von denen die Sozialwissenschaften noch sehr weit entfernt sind. Zum jetzigen Zeitpunkt ist das Problem noch relativ gering, da ein Großteil der Daten die man zu Recht als Big Data bezeichnen kann ohnehin nicht offen zugänglich ist. Systeme die solche Datenmengen erzeugen sind häufig das Privileg

[5] Diese Bücher sind über den Google Ngram Viewer auch als annotierter Korpus abrufbar und auswertbar.

[6] Kalkuliert von Randall Munroe in seinem What If Blog. Basis der Kalkulation waren die von Google veröffentlichten Zahlen zum Stromverbrauch seiner Rechenzentren.

[7] `https://about.twitter.com/company`

von großen Internet-Konzernen, staatlicher Administration oder Geheimdiensten. Daraus folgt ein gewisses Maß an Geheimhaltung. Des Weiteren stellt sich bei genauerer Betrachtung heraus, dass mit Big Data in den Sozialwissenschaften eigentlich ein viel allgemeineres Problem gemeint ist. Eine kurze Betrachtung des soziologischen Diskurses um Big Data zeigt zunächst einmal, dass es in den meisten Fällen nicht um die methodisch praktische Anwendung geht und der Begriff Big Data meist sehr unterschiedslos für eine Vielzahl von Datenstrukturen und -mengen verwendet wird (z.B.: Bail 2014; Burrows und Savage 2014; Uprichard 2013). Dies liegt wahrscheinlich in der auf „konventionellen Techniken" basierenden Definition des Begriffs begründet. Was jedoch für die Informatik in den Bereich des Konventionellen fällt, mag in der Soziologie noch nicht einmal bekannt, geschweige denn gängige Praxis sein.

Was also sind die Konventionen der Soziologie bezüglich Daten? Das dominante Modell sozialwissenschaftlicher Datensätze ist eine Tabelle mit Beobachtungen entlang der Zeilen und Variablen entlang der Spalten. Zumindest stellt dies die gängige Form in Fachbüchern und Auswertungsprogrammen (z.B. der Dataframe in R) dar. So oder so ähnlich werden die Aufbereitung, Aufbewahrung und Analyse von Daten an Studenten vermittelt. Sicherlich hat dieses Datenformat gewisse Vorteile, wenn es um die Analyse von Daten geht, da die meisten Statistikprogramme auf solche Datenstrukturen ausgelegt sind. Die Transformation und vor allem das langfristige Datenmanagement kann jedoch schnell eine Komplexität erreichen, welche von tabellarischen Daten nicht mehr verarbeitet werden kann. Das größte Problem ist jedoch die starke Orientierung am Modell des Fragebogens. Dies hat zur Folge, dass Daten die andere Sachverhalte repräsentieren nur sehr schwer in Datentabellen gefasst werden können.

Dies kann man als „Medium Data" Problem bezeichnen (vgl. Heiberger und Riebling 2016: 3f; Riebling 2018). Gemeint ist damit das Problem der adäquaten Repräsentation von Datenstrukturen, welche einerseits nicht mittels der konventionellen Verfahren der Sozialwissenschaften bearbeitet werden können, andererseits aber auch nicht in den Bereich von Big Data fallen. Beispiele für Medium Data sind Netzwerk-, Text- und unstrukturierte Online-Daten. Entscheidend für die Einordnung von Textdaten als Medium Data sind vor allem zwei Eigenschaften: Größe und Komplexität.

Um das Problem der Größe von Texten in Datentabellen näher zu erläutern wird im Folgenden das Beispiel einer Term-Document Matrix verwendet. Eine ausführlichere Darstellung dieser numerischen Repräsentation von Texten findet sich in den Abschnitten 5.3 und 5.6. Nehmen wir an, dass

	1. Satz	2. Satz	3. Satz
das	1	1	1
ein	1	1	1
für	0	0	1
ist	1	1	0
satz	1	1	1
sein	0	0	1
soll	0	0	1
was	0	0	1

Tabelle 4.1: Schematische Darstellung einer Word-Dokument Matrix.

die Häufigkeitsverteilung folgender drei Sätze dargestellt werden soll (der Einfachheit halber wird hier auf Satzzeichen und Groß-/Kleinschreibung verzichtet):

1. das ist ein satz
2. ist das ein satz
3. was für ein satz soll das sein

Repräsentiert in Form einer Word-Dokument Matrix (*Term-Document Matrix*), stellt jedes Wort[8] eine Zeile dar und jede Spalte einen Satz. Das Ergebnis ist Tabelle 4.1, deren einzelne Zellen jeweils die absoluten Häufigkeiten enthalten.

Von den insgesamt 24 Feldern der Tabelle enthalten 9 den Wert 0. Genau genommen enthalten diese keine eigene Information. Die Darstellung der Texte als Datentabelle macht es notwendig, dass jede Spalte mindestens so lang ist, wie die Menge aller verwendeten, einzelnen Worte. Dieses Problem wird manchmal als die *sparsity* einer Matrix bezeichnet. Gemeint ist damit das Verhältnis der Null-Zellen einer Matrix zur Gesamtzahl der Zellen. Bereits in diesem artifiziellen Beispiel beträgt die sparsity $37{,}5\%$. Größere Textmengen bestehen in der Praxis meist zum überwiegenden Teil aus Null-Zellen, so dass eine sparsity von annähernd hundert Prozent keine Seltenheit ist.

Das Zweite Problem ist die hohe Komplexität von Textdaten. Unter Komplexität soll im Folgenden die Menge der Relationen zwischen den

[8] Genauer gesagt handelt es sich hier um Token (siehe Abschnitt 5.2).

einzelnen Elementen verstanden werden, die notwendig ist um die Daten adäquat zu repräsentieren. Durch die Transformation der Sätze in eine Term-Document Matrix geht eine Vielzahl von Informationen unwiederbringlich verloren, vor allem die Reihenfolge der Wörter. Als Folge davon sind die Sätze 1 und 2 in der Matrixschreibweise nicht mehr zu unterscheiden. Erschwerend kommt noch hinzu, dass eine nachträgliche Transformation dieses Datensatzes, zum Beispiel durch das oft notwendige Herbeiführen der grammatikalischen Grundform (siehe Abschnitt ??: Stemming und Lemmatisierung), durch die Tabellenform enorm erschwert wird. Texte in Datentabellen zu transformieren kann unter bestimmten Umständen sehr sinnvoll sein, deren Aufbereitung und Verwaltung in diesem Format ist es jedoch so gut wie nie.

Der angemessene Umgang mit Textdaten setzt zunächst einmal voraus, dass man Texte als einen eigenständigen Datentyp begreift. Es mag für die sozialwissenschaftliche Betrachtung ungewohnt sein, aber Texte stellen bereits äußerst komplexe und relativ gut komprimierte Daten dar. Die Anordnung der einzelnen Wörter kodiert die Grammatik, Absätze strukturieren die Argumentation und je nach Herkunft des Textes können auch Metadaten (z.B.: Autorenschaft, Entstehungsdatum, etc.), Hyperlinks und Formatierungen enthalten sein. Dies ist nicht verwunderlich, wenn man sich vor Augen führt, dass die möglichst präzise Repräsentation von Sachverhalten und Fakten eine der grundlegenden Funktionen von Texten ist. Einige der komplexesten, modernen Datenstrukturen stellen im Prinzip nichts anderes dar als einen annotierten Text der in einer für diesen Zweck optimierten Sprache, wie zum Beispiel XML (eXtensive Markup Language), verfasst wurde.

Ein Großteil der Daten welche durch moderne Informationstechnologien produziert wird besteht aus solchem „Medium Data“, d.h. Textdaten die in einer bestimmten, formalen Sprache verfasst sind. Das vielleicht eingängigste Beispiel dafür sind die enormen Datenmengen der in HTML (HyperText Markup Language) verfassten Webseiten. Da ein Großteil des Internetverkehrs und der dadurch vermittelten sozialen Prozessen in dieser Form kodiert sind, handelt es sich hier um einen, für die Sozialwissenschaften, sehr bedeutsamen, neuen Datentypus. Um mit solch spezifischen Datenstrukturen umgehen zu können – vor allem aber um der Versuchung zu widerstehen, den Text einfach heraus zu kopieren und in eine Datentabelle einzutragen – sind ein grundlegendes Verständnis der sie produzierenden Sprachen sowie der darauf aufbauenden Techniken der Transformation und Aufbereitung notwendig. Der richtige Umgang

mit digitalem Text ist aber nicht nur notwendig für die quantitative Analyse. Gerade weil die qualitative Sozialforschung fast ausschließlich auf Textdaten zurückgreift besteht hier die Notwendigkeit auf angemessene Art und Weise mit diesen Daten umzugehen.

Man kann davon ausgehen, dass wir uns erst am Anfang eines goldenen Zeitalters sozialwissenschaftlicher Daten befinden, welches durch die zunehmende Digitalisierung aller Lebensbereiche weiter befeuert werden wird. Dadurch entstehen eine Vielzahl prozessgenerierter Daten, die einen bis dato nicht gekannten Einblick in die zugrundeliegenden, sozialen Phänomene erlauben. Diese Daten liegen in Form von Texten vor, die sowohl natürliche als auch formale Sprachen enthalten. Um diesen Datenmengen und deren komplexen Strukturen gerecht zu werden, ist es notwendig Texte als eigenständige, formalisierte Daten aufzufassen und sich die entsprechenden Techniken der Bearbeitung solcher Daten anzueignen. Dies gilt für Big Data ebenso wie für Medium Data, für die quantitative ebenso wie für die qualitative Sozialforschung.

4.2.3 Reproduzierbarkeit

Wissenschaft ist stets ein soziales Spiel. Daher nehmen die Reproduzierbarkeit und die intersubjektive Nachvollziehbarkeit einen zentralen Stellenwert unter den Gütekriterien der wissenschaftlichen Arbeit ein. Da sich wirklichkeitsähnlicheres Wissen nur über ständige Kritik im Prozess des „organisierten Skeptizismus“ (vgl. Merton 1993: 277f) herstellen lässt, muss sichergestellt werden, dass das Zustandekommen der Ergebnisse möglichst genau nachvollzogen werden kann. Hierin zeichnet sich jedoch ein gewisser Konflikt mit der zentralen Stellung der Subjektivität in der qualitativen Sozialforschung ab. Im Folgenden soll gezeigt werden, dass die Verfahren der quantitativen Textanalyse sowohl für sich genommen, als auch in Verbindung mit qualitativen Verfahren zu einer Steigerung der Reproduzierbarkeit und Nachvollziehbarkeit beitragen, ohne dabei die „Interpretation“ durch „Zahlengläubigkeit“ zu ersetzen.

Wie schon angesprochen, stützt sich die Interpretation im Rahmen der qualitativen Forschung sowohl auf die Einbettung in Theorien als auch auf die subjektive Interpretationsleistung des Forschers. Ersteres legt die stärkere Anleitung der Interpretation durch theoretische Begrifflichkeiten nahe. Im gegenwärtigen Diskurs um eine Neuausrichtung qualitativer Forschung hat Rainer Keller (vgl. Keller 2012) hierfür den Begriff des „Theorismus“ eingeführt. Damit soll ein Gegengewicht zu empiristische-

ren Auffassungen (z.B.: Grounded Theory) gebildet werden, sowie auch zu einer post-qualitativen Position, die man vielleicht nur noch als tiefe Verunsicherung bezüglich der eigenen Methodologie beschreiben kann:

> If we cease to privilege knowing over being; if we refuse positivist and phenomenological assumptions about the nature of lived experience and the world; if we give up representational and binary logics; if we see language, the human, and the material not as separate entities mixed together but as completely imbricated "on the surface" – if we do all that and the "more" it will open up – will qualitative inquiry as we know it be possible? Perhaps not. (Lather und Pierre 2013: 629f)

Leider versäumen die Autoren es hier, ihrer eigenen Intuition zu folgen und das „not" zu akzeptieren. Dieses Zitat, welches der Einleitung des Sonderbandes *Post-Qualitative Research* entnommen wurde, illustriert die inhärente Gefahr des Abgleitens in den Mystizismus und die damit verbundene Leugnung selbst der einfachsten, wissenschaftlichen Grundannahmen besser als es jedes systematische Argument könnte. Der Fokus einer solchen post-qualitativen „Forschung" liegt damit nicht auf intersubjektiv kritisierbaren Aussagen, sondern scheint sich um ein nicht genauer beschriebenes, spirituelles Erweckungserlebnis zu drehen. Da jegliche Logik oder empirische Demonstration von vorneherein ausgeschlossen werden, lässt sich dagegen auch nichts einwenden. Über Offenbarung kann man zwar praktisch streiten, nicht aber inhaltlich.

Der Gegenvorschlag des Theorismus ist jedoch auch mit gewissen Mängeln behaftet. Grundsätzlich ist natürlich jede Schärfung der Begrifflichkeiten zu begrüßen, da es zu Kritik und zum wissenschaftlichen Erkenntnisfortschritt beiträgt. Jedoch besteht hier die Gefahr einer so starken perspektivischen Festlegung, dass die Ergebnisse von den theoretischen Annahmen vorherbestimmt werden. Damit wäre nur eine ständige Bestätigung dessen was wir ohnehin schon wussten erreicht. Selbst wenn Falsifikation im strengen Sinne für sozialwissenschaftliche Aussagen oft nicht möglich ist, so muss man doch zugestehen, dass Widerlegung und Kritik auch in diesem Fall die einzigen Wege zu besserer Erkenntnis sind. Damit verschiebt sich das Problem von der Ausführlichkeit der Theorie zur Frage wie die systematische Kritik der Ergebnisse gewährleistet werden kann, und damit letztlich zur Güte der Interpretation.

Eine Möglichkeit könnte hier in der *Stärkung der individuellen Fähigkeit zur Interpretation* liegen. Wie schon ausgeführt, gibt es Anzeichen für ei-

ne solche Strategie in verschiedenen Lehrbüchern. Praktische Erfahrung würde damit zu einem entscheidenden Qualitätsmerkmal. Es ist sicherlich richtig anzunehmen, dass der Faktor Praxiserfahrung wichtig für die Interpretation ist. Zugleich eignet er sich aber auch nur sehr bedingt als ein Qualitätsmerkmal. Die Begründung des Wahrheitsgehaltes einer Aussage durch die angenommene Kompetenz eines Individuums läuft dem Grundverständnis der modernen Wissenschaft zuwider. Schließlich geht es ja gerade um die Steigerung der Objektivität, Überprüfbarkeit und intersubjektiven Nachvollziehbarkeit von Aussagen.

Trotz der zentralen Stellung der Interpretation in der qualitativen Sozialwissenschaft, ist es jedoch nicht klar in welchem Maße diese Methode als objektiv gelten kann. Dies ist sicherlich auch auf die zentrale Annahme der Subjektivität in der qualitativen Forschung zurückzuführen. Dennoch kann man davon ausgehen, dass der Großteil der qualitativ arbeitenden, sozialwissenschaftlichen Forscher Interpretation für einen zumindest prinzipiell nachvollziehbaren Prozess handelt, da sie sonst nicht versuchen würden ihre Aussagen durch Zitationen zu belegen. Selbst Ausnahmefälle, wie die weiter oben zitierten post-qualitativen Forscher, bemühen sich darum ihre Gedanken zu Papier zu bringen und sie einem Publikum gegenüber verständlich zu machen.

Wir können die Frage weiter vereinfachen: Ist ein einzelner Akt der Interpretation *deterministisch*? Das heißt, würde dieselbe Person unter denselben Bedingungen zu einem gleichen oder zumindest stark ähnlichen Ergebnis kommen? Beantwortet man diese Frage mit Nein, so ist jegliche Erkenntnis mittels Interpretation letztlich unmöglich und damit auch jede Forschung hinfällig. Genau genommen wäre damit jegliche Form von Sprache oder intersubjektivem Verstehen hinfällig. Entscheidet man sich dafür, so muss dies nicht nur für eine Person gelten, sondern für alle. Aber sicherlich nicht für alle Personen in gleicher Art und Weise. Worum es hier geht ist festzustellen, dass Interpretation nur entweder regelhaft und damit intersubjektiv oder erratisch und damit rein subjektiv sein kann. Es kann wohl keine letztgültigen Argumente für eine der beiden Seiten geben. Da es meine feste Überzeugung ist, dass eine rein subjektive Interpretation der Wirklichkeit die lohnenswerte Aufgabe der Kunst und nicht die der Wissenschaft ist, wird Interpretation hier als ein regelhaftes Phänomen aufgefasst.

Die für die Analyse von Text zur Verfügung stehenden Algorithmen, können die menschliche Interpretationsleistung sicherlich nicht ersetzen, aber sie stellen eine Verbesserung der Nachvollziehbarkeit in zweierlei

Hinsicht bereit. Zum einen wird uns dadurch ermöglicht Texte auf eine regelmäßige Art und Weise zu analysieren und somit eine gewisse Reproduzierbarkeit der Ergebnisse zu gewährleisten. Andererseits kann die Interpretation im Rahmen eines qualitativen Forschungsbereichs nachvollziehbarer gemacht werden, indem die Fallauswahl und die Ergebnisse mit formalen Verfahren kontrolliert werden. Da die Integration von quantitativer und qualitativer Textanalyse an anderer Stelle ausführlicher behandelt wird (siehe Abschnitt 4.2.5), soll es im Folgenden nur um die Steigerung der Nachvollziehbarkeit durch Formalisierung gehen.

Die gesteigerte Nachvollziehbarkeit durch den Einsatz eines explizit formulierten Algorithmus ergibt sich teilweise aus dessen deterministischem Charakter, aber vor allem aus der Verwendung einer formalen Sprache.[9] Während sich die intersubjektive Überprüfung im Bereich der qualitativen Sozialforschung meist nur auf Plausibilität und Vertrauen in die Fähigkeiten des Forschers stützen kann, erlaubt die Verwendung einer formalen Sprache den Weg zu den Ergebnissen explizit und damit auch kritisierbar zu machen. Gegen den oft gemachten Vorwurf der „Theorieferne“ quantitativer Verfahren lässt sich anführen, dass die eine Formalisierung ja nichts weiter als eine theoretische Aussage über den Gegenstandsbereich ist. Der Nutzen einer Formulierung mittels einer formalen Sprache liegt schlichtweg in der Eineindeutigkeit der Aussagen. Einen Text zum Beispiel als Vektor in einem geometrischen Raum aufzufassen, der durch latente Dimensionen erzeugt wird, enthält bereits eine reichhaltige Theorie mit expliziten Vorannahmen bezüglich Sprache, Bedeutungen und kognitiven Prozessen. Dadurch wird auch der Streit über die theoretische „Einordnung“ – der insbesondere in der Soziologie gerne praktiziert wird – enorm vereinfacht und man kann sogar sagen, endlich fruchtbar gemacht. Eine formale Sprache macht es möglich wesentlich präzisere Begrifflichkeiten zu formulieren und sich so nicht länger mit dem Streit über die richtige Definition von Wörtern aufzuhalten. Dieser Vorteil zeigt sich klar an der Dominanz formaler Sprachen, wie Mathematik,

[9] Dieses Argument ist bewusst sehr allgemein formuliert. Man kann sich hierunter auch eine explizite Handlungsanweisung an menschliche Interpretierende vorstellen, die ebenfalls ein Algorithmus wäre. Allerdings sind Menschen für gewöhnlich nicht direkt programmierbar. Das soll jedoch nicht heißen, dass soziale Organisationen, wie zum Beispiel die Bürokratie, nicht extrem effiziente Verhaltensprogramme für Menschen bereithalten. Von den ethischen Problemen einmal abgesehen, muss aber festgestellt werden, dass Menschen sehr ineffektiv sind, wenn es um die wiederholte und schnelle Ausführung präziser Anweisungen geht.

Computercode, chemischen Formeln und dergleichen, in allen empirisch orientierten Wissenschaften.

Es ist hier noch wichtig anzumerken, dass vollkommene Reproduzierbarkeit auch unter der Verwendung standardisierter Verfahren nicht immer möglich ist. Einige Verfahren wie zum Beispiel die Latent Semantic Analysis (siehe Abschnitt 6.5.1) sind nicht-probabilistisch und liefern daher stets dasselbe Resultat. Es gibt jedoch auch eine Reihe Verfahren die auf Zufallsverteilungen aufbauen (siehe zum Beispiel Naive-Bayes Klassifikation in Abschnitt 6.2.1). Diese Verfahren erzeugen daher nicht immer den gleichen Output und sind bis zu einem gewissen Grad abhängig von der Qualität der verwendeten Daten und Zufallsalgorithmen. Der grundsätzliche Vorteil expliziter Modellbildung bleibt jedoch auch hier erhalten.

4.2.4 Anschlussfähigkeit

Außerhalb der sozialwissenschaftlichen Forschung findet sich die Anwendung standardisierter Verfahren und Algorithmen zur Analyse von Texten, Sprache und Symbolen in einer Vielzahl unterschiedlicher Kontexte und Disziplinen. Die ausführlichste Beschäftigung mit diesem Themenfeld findet sich in der Linguistik, der theoretischen Informatik und der Informationstheorie. Diese Forschungslinien waren vor allem durch eine Abstrakte Beschäftigung mit den Regeln zeichenverarbeitender Systeme gekennzeichnet. Daraus erwuchsen später auch Forschungsfelder die auf die empirische Beobachtungen von Kultur und Zeichensystemen sowie auf praktische Anwendungen abzielten. Als Folge davon hat sich ein weit verzweigtes, interdisziplinäres Netzwerk entwickelt, welches mehr durch gemeinsame Methoden und Methodologien zusammengehalten wird als durch eine theoretische Basis.

Dies lässt sich gut am Beispiel der *Narratologie* explizieren. Im Zentrum der modernen Erzähltheorie steht der Begriff des Narrativs bzw. der Erzählung, welcher explizit als ein Bindeglied zwischen verschiedensten Disziplinen und Forschungsfeldern gesehen werden kann (vgl. Koschorke 2012: 19ff). Es geht dabei um die Frage, in welchem Verhältnis semantischer Inhalt und linguistische Struktur, hier insbesondere die Reihenfolge und die logischen Bezüge der Aussagen, zueinander stehen. Im Prinzip geht es um die Analyse von Verknüpfungen in Aussagensystem. Ein Besipiel wäre folgende, sehr simple Erzählung: „A geht nach X, weshalb B C anruft“. Narrative können in unterschiedlicher Art und Weise formali-

siert werden. Die naheliegendste Lösungen sind Netzwerkgraphen oder Sequenzen.

Das Einsatzgebiet der Narratologie umfasst so unterschiedliche Bereiche wie Literaturwissenschaften, Soziologie, Psychologie, computergestützte Linguistik oder Forschung an künstlicher Intelligenz (vgl. Herman, Jahn und Ryan 2007: 9ff). Je nach Disziplin haben sich eine Reihe von Beiträgen herausgebildet, die sowohl theoretischer Natur als auch methodischerund empirischer Art sein können. Jenseits der akademischen Beschäftigung finden sich aber auch Möglichkeiten praktischer Anwendung. Zum Beispiel im Bereich der künstlichen Intelligenz, in dem diskursive Narrative als Modelle für den Prozess mentaler Planung (DPOCL planning systems von Young und Moore (1994)) vorgeschlagen wurden oder der Einsatz in der automatischen Erkennung von Frageformen im Information Retrieval (QUEST-Modell von Graesser, Gordon und Brainerd (1992)). Auch in den Sozialwissenschaften findet eine Beschäftigung mit dem Begriff der Narration statt. Dazu zählen vor allem die Forschung bezüglich biographischer Narrative (vgl. Loch und Rosenthal 2002), Narrative in Diskursen (vgl. Bearman und Stovel 2000) und die Analyse von Narrativen als Netzwerkstrukturen (vgl. Mützel 2007). Dieser Forschungsbereich umfasst dabei sowohl qualitative als auch quantitative sowie mixed-method Ansätze.

Insgesamt illustrieren diese Beispiele die *Interdisziplinarität* und die weiten Anwendungsmöglichkeiten der Narratologie. Ein solcher Austausch über verschiedene Kontexte hinweg basiert vor allem auf der formalen Definition der zentralen Begrifflichkeiten. Ein solches Potential ist jedoch nicht nur auf die Narratologie begrenzt, sondern findet sich mehr oder minder stark in allen Bereichen, die mit formalen und computergestützten Verfahren arbeiten. Durch die Verwendung formaler Sprachen wird nicht nur ein interdisziplinärer Diskurs möglich, sondern auch eine Implementation als *praktische Anwendung*. Es ist insbesondere dieser Praxisbezug der hier hervorgehoben werden muss, da die Soziologie äußerst selten technologische Lösungen hervorbringt und sogenannten „Sozialtechnologien" normalerweise sehr misstrauisch gegenübersteht, da diese als oft unzulässige Manipulation von Menschen angesehen werden und somit dem Projekt der „soziologischen Aufklärung" zuwiderlaufen.

Die Befürchtungen, die mit den neuen Technologien der Informationsverarbeitung, insbesondere dem Text und Data Mining, verbunden sind, lassen sich nicht so einfach von der Hand weisen. Die Massenüberwachung durch staatliche Geheimdienste, die in den letzten Jahren publik ge-

macht wurde, gibt reichlich Anlass zur Besorgnis. Eine Vielzahl der in diesem Rahmen eingesetzten Techniken basieren auf semantischen Erkennungsverfahren und Techniken des Information Retrievals, wie sie auch Gegenstand dieses Buches sind. So besorgniserregend diese Entwicklungen auch sein mögen, sie demonstrieren zugleich auch die praktische Wirkmächtigkeit dieser Verfahren und damit auch ihren Realitätsbezug. Wie bei allen Technologien, ist auch hier die technologische Anwendung selbst nicht das Problem, sondern die Art und Weise in der sie eingesetzt wird. Für jeden moralisch bedenklichen Einsatz lassen sich auch zahllose Gegenbeispiele einer vernünftigen und hilfreichen Anwendung finden.[10]

Letztlich wird eine Weigerung der Soziologie an der technologischen Entwicklung teilzuhaben diese weder aufhalten noch verlangsamen. Gerade weil standardisierte Verfahren und auf Algorithmen basierende Analysen die einzigen Möglichkeiten sind mit einem technologischen Wandel schrittzuhalten, der als ein „Steigerungsspiel" (gl. Schulze 2004) der Komplexität angesehen werden kann. Dabei machen komplexere Datenbestände nicht nur neue Algorithmen der Datenverarbeitung notwendig, sondern schaffen gleichzeitig die Basis für die Entstehung neuer Komplexitäten. Allerdings birgt die Teilnahme an diesem Spiel auch gleichzeitig das Potential die soziologische Analyse von Texten und Kultur aus ihrer Obskurität und Selbstbezüglichkeit zu lösen und sie mit anderen Disziplinen zu verknüpfen. Darüber hinaus eröffnen sich damit auch Möglichkeiten zu einer Forschung die technologische Anwendungen hervorbringen kann, welche nicht nur zu einer Verbesserung menschlicher Lebensumstände beitragen können, sondern auch zu einer Stärkung des Wirklichkeitsbezuges soziologischer Theorie.

4.2.5 Integrierbarkeit

Die bisherigen Äußerungen haben die unterschiedlichen Herangehensweisen von quantitativer und qualitativer Textanalyse in den Vordergrund gestellt. Dies geschah nicht zuletzt um die bestehenden Gegenargumente der qualitativen Forschung gegenüber einer standardisierten Betrachtung von Text zu entkräften. Damit ist noch keine grundsätzliche Ablehnung qualitativer Vorgehensweisen impliziert. Gleichzeitig scheint eine strikte Trennung unterschiedlicher Auffassungen, wie sie in der Soziologie gerne praktiziert wird, ebenfalls nicht sinnvoll. Eine kritische Auseinander-

[10] Praktische Einsatzgebiete werden in der Diskussion der jeweiligen Verfahren näher beleuchtet.

setzung ist der einzig sinnvolle Weg zu besserer Erkenntnis und eröffnet auch erst die Möglichkeit diese Gegensätze zu überwinden.

Die Verbindung von qualitativer und quantitativer Sozialforschung wird unter einer Reihe verschiedener Begrifflichkeiten, wie zum Beispiel „Triangulation" (vgl. Flick 2011) oder „Mixed Methods" (vgl. Greene 2008) und mit sehr unterschiedlichen theoretischen und praktischen Akzenten diskutiert. Aus einem Vergleich 19 verschiedener Definitionen konnten Johnson et al. (2007: 123) folgende grundlegende Auffassung herausarbeiten:

> Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration.

Daran anschließend entwickeln die Autoren die Vorstellung eines Kontinuums welches von rein qualitativer über rein mixed methods zu rein quantiativer Forschung führt. Nicht zuletzt um den pragmatischen Ansatz einer solchen Forschung hervorzuheben (ebd.: 123ff). Demnach wären diejenigen Verfahren und Methoden, sowie deren Kombinationen auszuwählen, die zur Verfolgung einer spezifischen Fragestellung notwendig sind. Da auch der vorliegenden Arbeit ein pragmatischer und auf Synthesen abzielender Impetus zugrunde liegt, kann dieser Haltung voll und ganz zugestimmt werden. Allerdings ist damit nur die Feststellung gewonnen, dass man pragmatisch forschen sollte. Sehr viel wichtiger scheint jedoch die Frage zu beantworten, wie eine solche Verknüpfung im Bereich der quantitativen Textanalysen praktisch aussehen könnte. In Anlehnung an das oben beschriebene Modell eines Kontinuums soll diese Frage für die drei Extremwerte der *qualitativ*, *mixed methods* und *quantitativ* dominierten Forschung, genauer betrachtet werden.

Auch wenn ein Forschungsvorhaben strikt qualitativ orientiert zu sein scheint, so gibt es doch eine Reihe impliziter und expliziter Bezugspunkte zur computergestützten Verarbeitung und Analyse von Texten. Dies ist nicht zuletzt darauf zurückzuführen, dass Textdaten heutzutage fast ausschließlich in digitaler Form vorliegen und daher als solche auch verarbeitet und verwaltet werden müssen. Um adäquat mit solchen Daten umgehen zu können bedarf es zumindest eines grundlegenden Verständnises hinsichtlich der *digitalen Repräsentation von Textdaten* (siehe Abschnitt 5.1 und 5.3). Auch die Beschaffung dieser Daten kann in vielen Fällen durch

Text Mining und ähnliche Verfahren unterstützt werden. In dem Maße wie kulturelle Artefakte mittels sozialer Medien direkt zugänglich (z.B.: Websiten, Foren, Online-Medien) geworden sind, ist auch die qualitative Forschung auf technische Verfahren angewiesen, die solche Texte erheben und als menschenlesbares Material bereitstellen können.

Liegen die Daten dann in entsprechend aufbereiteten Formaten vor, so werden im nächsten Schritt der Analyse meistens sogenannte QDA-Programme (*Qualitative Data Analysis Programs*) eingesetzt. Die Bedienung solcher Anwendungen ist auf den ersten Blick relativ voraussetzungslos. Wenn jedoch einige grundlegende Programmiertechniken und ein Verständnis für das digitale Verarbeiten von Texten vorliegen, so kann der Spielraum für die qualitative Analyse enorm erweitert werden. Dies gilt insbesondere für den überaus zentralen Arbeitsschritt des *Text Retrievals*, bei dem markierte Textpassagen nach bestimmten Kriterien ausgewählt werden können. Erst durch den Einsatz von formalen Suchkriterien, wie zum Beispiel der Überschneidung codierter Textstellen oder die Betrachtung der Verteilung von Codes kommen die Stärken solcher Programme zur Geltung (Kuckartz 2010: 26ff). Dieses Vorgehen hat seine Entsprechung in Verfahren der Korpuslinguistik, die auf strukturellen Merkmalen des Textes basieren, wie zum Beispiel dem relativ häufigen Auftretens zweier Wörter in direkter Nachbarschaft (Bigramme; siehe auch Abschnitt 5.5.1).

Viele der Operationen und Analyseschritte in QDA-Programmen greifen zudem implizit auf Algorithmen und quantitative Verfahren zurück, wie sie auch Gegenstand dieser Arbeit sind. Deshalb kann eine tiefergehende Beschäftigung mit diesen formalen Methoden auch in rein qualitativen Forschungsprojekten zu einer Verbesserung der praktischen Abläufe und einer Erweiterung der Analysemöglichkeiten führen. Dazu ist in diesem Bereich auch keine Veränderung der grundlegenden Arbeitsweise notwendig. Es reicht aus ein besseres Verständnis für die bestehenden technologischen Lösungen und die ihnen zugrunde liegenden Verfahren zu entwickeln.

Im mittleren Bereich der Skala (mixed methods) lassen sich eine Vielzahl von Möglichkeiten der Integration finden, die es notwendig machen eine Auswahl zu treffen. Daher sollen im Folgenden zwei Beispiele herausgegriffen werden, bei denen die Integration quantitativer und qualitativer Verfahren im Rahmen eines spezifischen Forschungsvorhabens am weitesten fortgeschritten ist. Dabei handelt es sich zum einen um die Stich-

probenziehung und zum anderen um den Einsatz von qualitativer Kodierung in quantitativen Klassifikationsverfahren.

Wie im Abschnitt über die Reproduzierbarkeit schon angesprochen, ist die Stichprobengröße qualitativer Forschungsarbeiten aufgrund der aufwendigen Interpretation und langwierigen Datenerhebung notwendigerweise begrenzt. Deshalb wurden im Rahmen der qualitativen Forschung auch spezifische Stichprobentechniken entwickelt (Morse 2007). Manche dieser Stichprobentechniken, insbesondere das theoretical sampling, sind jedoch sehr voraussetzungsvoll, da sie oft davon ausgehen, dass ein grundsätzlicher Überblick über das zu erforschende Feld besteht. Hier können Verfahren der computergestützten Verarbeitung von Texten eine wichtige Hilfsfunktion übernehmen. Insbesondere durch *Techniken des Text Minings*, welche größere Textmengen bereitstellen können, aus denen anschließend eine Zufallsauswahl an Texten gezogen werden kann. Des Weiteren sind auch automatische Klassifikationsalgorithmen (siehe Abschnitt 6.2.1) und semantische Analysen (siehe Unterkapitel 6.5) eine grundlegende Möglichkeit, sich erst einmal einen Überblick bezüglich der Themen und Textarten zu verschaffen und daraus dann eine kleine Stichprobe auszuwählen. Durch dieses Vorgehen werden die Chancen maximiert, dass die resultierende Stichprobe diejenigen symbolischen Muster enthält, die einen spezifischen Textkorpus am umfassendsten beschreiben.

Am Pol der quantitativen Analyse von Text gibt es ebenfalls zahlreiche Möglichkeiten einer sinnvollen Ergänzung durch qualitative Verfahren. Dies gilt in besonderem Ausmaße für Algorithmen der Klassifikation von Texten im Bereich des maschinellen Lernens. Beim sogenannten „überwachten Lernen" werden Algorithmen dazu gebracht Klassifikationsprobleme zu lösen, indem sie von Klassifikationen lernen die von Menschen gemacht wurden. Ein klassisches Beispiel hierfür ist die Sentiment Analyse (Überblick in Pang und Lee (2008); siehe auch 6.2.2), bei der versucht wird festzustellen ob ein Text oder Satz eher positiv oder negativ über ein bestimmtes Thema spricht. Beim überwachten Lernen wird hierzu eine Stichprobe aus den zu klassifizierenden Texten ausgewählt und manuell codiert um die Richtung der Aussage zu erhalten. Diese „händische Klassifikation" ist die notwendige Vorausbedingung um den Algorithmus trainieren zu können. Mit dem daraus resultierenden Modell ist es möglich weitere Texte automatisch zu codieren. Ein solches Vorgehen kann zudem auch eingesetzt werden um qualitative Codierungen auf ihre innere Konsistenz und Reliabilität zu testen. Dabei überführt man bestehende Codie-

rungen in einen Algorithmus und prüft wie gut dieser die Kategorisierung des Ausgangsmaterials wiederherstellen kann.

Weniger technisch, aber nicht minder wichtig ist die Rolle der qualitativen Analyse bei der *Interpretation der statistischen und mathematischen Modelle*. Oft sind die Endergebnisse einer quantitativen Analyse von Text nur als Zahlen und Verteilungen von Worten zugängig. Was sich hinter diesen Verteilungen und Indikatoren verbirgt, sind die für uns theoretisch relevanten Tatsachen. Diese jedoch herauslesen zu können erfordert ein gewisses grundlegendes Verständnis der analysierten Texte und des Entstehungskontextes. Für die Bereitstellung dieses Wissens sind qualitative Methoden oft sehr hilfreich und in manchen Fällen schlichtweg unentbehrlich.

4.3 Pragmatische Methodenwahl

Aus den bisherigen Äußerungen wird deutlich, dass entscheidende Differenzen zwischen den qualitativen Verfahren der Textanalyse, welche momentan die dominante Vorgehensweise in den Sozialwissenschaften darstellt, und der computervermittelten, auf Algorithmen basierenden Analyse von Text bestehen. Unterschiedliche Konzeptionen und Vorgehensweisen finden sich insbesondere hinsichtlich der starken *Orientierung an Subjekten* und der Subjektivität der Interpretation. Auf der quantitativen Seite dominieren hingegen formale Modelle und symbolische Ordnungen werden als *objektive Gegebenheiten* aufgefasst. Aufbauend auf diesen grundlegenden Unterscheidungen wurden fünf Vorzüge der formalen Analyse von Texten näher beleuchtet sowie in Bezug zum qualitativen Methodendiskurs und dessen Einwänden gesetzt.

Ohne die subjektiven Elemente der Interpretation gänzlich leugnen zu wollen, kann doch gezeigt werden, dass die Auffassung von Symbolen und Bedeutungen als objektiven Phänomenen dem Gegenstandsbereich ebenfalls angemessen ist. Zum einen setzen Kommunikation und wechselseitiges Verstehen objektiv nachvollziehbare Symbole voraus. Zweitens müssen diese Symbole in einer physikalischen Form vorliegen, um übertragen werden zu können. Der objektive Charakter von Symbolen ist nicht nur notwendige Voraussetzung für das Gelingen der Kommunikation zwischen Menschen, sondern auch für den Austausch von Menschen und Maschinen. Beispiele, wie der weitreichende Einsatz von Algorithmen zur Suche nach Informationen, bezeugen den praktischen Nutzen symbolische Ordnungen als Objekte zu betrachten. Darüber hinaus ist die Auffassung,

dass die Interpretation von Texten nur der Subjektivität des Forschers geschuldet sei, auch vor wissenschaftstheoretischen Gesichtspunkten problematisch. Besteht hier doch die Gefahr subjektive Expertise zur Richtschnur von Ergebnissen zu machen.

Die zunehmende Kommunikation zwischen Mensch und Maschine, bzw. die digitale Vermittlung von menschlicher Interaktion stellt nicht nur die gängigen Konzeptionen der Soziologie in Frage, sondern auch die Disziplin vor neue technische Herausforderungen. Bis dato unbekannte Mengen von Daten (*Big Data*), sowie neue Schemata und Datenstrukturen (*Medium Data*) können mit den momentanen Kompetenzen der Sozialwissenschaften oft nur unzureichend behandelt werden. Da ein Großteil dieser neuen Daten aus digitalen Texten besteht, findet sich hier ein wichtiges Einsatzfeld für die Verfahren des Text Minings sowie der quantitativen Textanalyse. Um diesen Datenschatz zu bergen und ihn für die Sozialwissenschaften nutzbar zu machen, ist eine Beschäftigung mit formalen Verfahren und Algorithmen dringend notwendig geworden.

Das formale Vorgehen der quantitativen Textanalyse löst aber nicht nur technische Probleme. Gleichzeitig ermöglicht es eine Steigerung der intersubjektiven Nachvollziehbarkeit und der Reproduzierbarkeit der Ergebnisse. Dies geschieht hauptsächlich durch den formalen Charakter der Modelle und deren grundsätzliche Replizierbarkeit. Hierdurch wird es möglich Kultur und Bedeutungen bis zu einem gewissen Grad standardisiert zu erfassen und die daraus resultierenden Modelle einem kritischen Diskurs zugänglich zu machen. Gleichzeitig zwingt die Verwendung formaler Sprachen und Verfahren auch zur Präzision in der Modellbildung.

Ein weiterer Vorzug der Verwendung formaler Modelle und quantitativer Verfahren besteht in dem Anschluss an den bestehenden, interdisziplinären Diskurs. In einer Vielzahl von Disziplinen finden sich Bereiche, die sich durch einen gemeinsamen Fokus auf die formale Analyse von Texten, Zeichen und Bedeutungen auszeichnen. Die Verwendung formaler Sprachen und einer klaren Modellbildung ermöglicht eine Teilhabe an diesem Diskurs und damit auch eine Nutzung der entsprechenden Verfahren sowie einen Rückgriff auf das bestehende Wissen. Eine Besonderheit dieses interdisziplinären Bereichs der Textanalyse ist zudem der Bezug zu technologischen Anwendungsfeldern. Damit eröffnen sich neue Möglichkeiten der praktischen Überprüfung von theoretischen Erkenntnissen. Technologische Lösungen können sozialwissenschaftlichen Theorien zu neuem Gewicht verhelfen und dadurch auch zu einer Steigerung der wissenschaftlichen Legitimität beitragen.

All diese Vorzüge müssen jedoch keinen grundsätzlichen Gegensatz zur qualitativen Forschung darstellen. Vielmehr ist eine Integrierbarkeit der hier diskutierten formalen und quantitativen Verfahren auf verschiedenen Ebenen möglich und wird auch zunehmend praktiziert. Dies beginnt in der rein qualitativen Forschung bei der Verwendung von QDA-Programmen, deren unterstützende und visualisierende Funktionen auf Verfahren der quantitativen Textanalyse aufbauen. Konsequenterweise führt eine Beherrschung dieser Techniken zu besseren und gehaltvolleren Analysen. Mixed-method Ansätze erlauben die zielführende Integration von qualitativer und quantitativer Analyse von Texten auf einer Vielzahl verschiedener Ebenen (z.B.: Stichprobenziehung, Validierung). Aber auch in stark quantitativ orientierten Forschungen finden sich Anwendungen, welche qualitative Techniken voraussetzen. Das wahrscheinlich beste Beispiel ist hierbei die Interpretation der inhaltlichen Dimensionen. Egal wie konsistent diese modelliert worden sind, letztlich ist ein gewisser Grad an subjektiver Interpretation notwendig.

Insgesamt entsteht somit auch hier der Eindruck möglicher Synthesen von qualitativen und quantitativen Verfahren. Allerdings besteht der Nachholbedarf wohl eher hinsichtlich des Bereichs der computergestützten Textanalyse, da diese sich unabhängig von großen Teilen der sozialwissenschaftlichen Analyse von Kultur entwickelt und erst mit der neuen, verbesserten Datenlage für uns an Bedeutung gewonnen hat. Außerdem besteht ein nicht zu unterschätzender Mangel an praktischer, technischer Expertise in den Sozialwissenschaften, der jedoch grundsätzlich behebbar scheint. Die Entstehung neuer Fachbereiche, insbesondere der Computational Social Science, deutet zumindest darauf hin, dass sich in diesem Problemfeld einiges zum Besseren wenden könnte. Eine Überwindung der technischen Schwierigkeiten scheint deshalb zentral, weil hierin die einzige Chance zu einer praxisorientierten Überwindung der Trennung von quantitativen und qualitativen Forschungsansätzen liegt. Technische Kompetenzen ermöglichen es den Forschenden ihre Methoden entsprechend ihrer Fragestellungen zu wählen, diese in Forschungsvorhaben praktisch umzusetzen und die resultierenden Ergebnisse nachvollziehbar zu gestalten.

Da textanalytische Verfahren grundlegende Kenntnisse des Programmierens und der Verarbeitung von Textdaten erfordern, stellen sie einen möglichen Eintritt in die Welt der formalen und computergestützten Verfahren dar, der über ihren bloßen Status als Werkzeuge hinausgeht. Die Beschäftigung mit Sprache als einem Gegenstandsbereich, auch im Kon-

text der sozialen Welt, führt beinahe schon zwangsläufig in die Benutzung formaler Sprachen und die Möglichkeiten computergestützter Analysen ein. Dies liegt vor allem daran, dass die Komplexität von Texten und symbolischen Ordnungen eine Vielzahl von Anforderungen an den Umgang mit unkonventionellen Datenstrukturen, die Nutzung relativ aufwendiger statistischer und mathematischer Verfahren sowie grundlegende Programmiertechniken erfordert.

4.4 Ein Plädoyer für Digital Literacy

Die Bedeutung computergestützter Verfahren für die Geistes- und Kulturwissenschaften wird in einer Reihe gegenwärtiger Forschungsbemühungen deutlich, die unter einer Reihe von Bezeichnungen diskutiert werden, wie zum Beispiel: *Computational Social Science*, *Digital Sociology* oder auch *eHumanities*. Trotz dieses grundsätzlichen Trends kann ein gewisses Misstrauen und Festhalten an klassischen Forschungsparadigmen in weiten Teilen der Sozialwissenschaften beobachtet werden. Bei manchen Befürwortern neuer Methoden und Arbeitsweisen hat dies die Befürchtung geweckt, dass hier nur ein Lippenbekenntnis zur digitalen Welt im Dienste der Vergabe von Drittmitteln geleistet werde (vgl. Diekmann 25. September 2016, Uhr). Die Frustration über das Festhalten an überkommenen Konzepten und der unbegründeten Ablehnung digitaler Verfahren äußert sich zum Beispiel in der „more hack; less yack“ Debatte in den Digital Humannities (vgl. Nowviskie 2014).

Aus Ermangelung an Umfragedaten und dergleichen, lässt sich hier nur ein ungefähres und anekdotisches Bild des vorherrschenden Techniverständnisses in den Sozialwissenschaften zeichnen. Es ist jedoch möglich, diese Beobachtungen in gewisser Weise zu verallgmeinern. Da es sich beim Großteil der Leser dieses Buches um Sozialwissenschaftler handeln dürfte, kann von einer gewissen Expertenschaft hinischtlich der alltäglichen Arbeitsweisen ausgegangen werden. Phänomene wie das repetitive Kopieren mittels Tastenkombinationen („copy-paste“), das manuelle Abtippen von Online-Daten, nachträgliche, hektische Literaturrecherche anstelle der Verwendung einer einheitlichen Datenbank und die Beschäftigung von studentischen Hilfskräften zur manuellen Lösung von Problemen, die einfach zu automatisieren gewesen wären, sind wohl jedem Sozialwissenschaftler schon einmal begegnet.

Das es sich hier um tiefsitzenderes Problem handelt wird deutlich wenn man die gängige Argumentation betrachtet, mit der den Forderungen

nach einer größeren Technikexpertise begegnet wird. Diese Argumentationslinie, der zufolge man diese Art der technischen Expertise von einem Sozialwissenschaftler nicht erwarten könne, ist bereits im Bereich der qualitativen Forschungstradition illustriert worden. Leider beschränkt sie sich nicht hierauf, sondern findet sich auch in den Arbeiten expliziter Befürworter computergestützer Analyseverfahren. So treffen Lemke und Wiedemann in der Einleitung zu ihrem Sammelband *Text Minining für Sozialwissenschaftler* folgende Feststellung:

> 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. Anstelle der „morehack"-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. (Lemke und Wiedemann 2016: 4)

Auch hier wird jene Haltung deutlich, die den Wunsch nach einer voraussetzungsfreien Nutzung der Möglichkeiten computergestützter Verfahren zum Vater des Gedankens macht. Dies erscheint in zweierlei Hinsicht zutiefst problematisch.

Zum einen fördert es die ohnehin in den Sozialwissenschaften schon weit verbreitete Tendenz zur *Orientierung an Standardverfahren*. Dieses Problem wird wohl am deutlichsten in der gegenwärtig durch wissenschaftliche Fachzeitschriften vorangetriebenen Absage an den t-Test und die damit einhergehenden Jagd nach Signifikanzniveaus, der sich Sozialwissenschaftler in den letzten Jahren in zunehmendem Maße verschrieben haben. In einer bisher einmaligen Aktion sah sich sogar die American Statistical Association dazu genötigt einzuschreiten und klarzustellen, dass die unreflektierte Verwendung dieses Standardverfahrens ein zentrales Problem der gegenwärtigen Forschung darstellt (vgl. Wasserstein und

Lazar 2016). Das Bedeutsame an diesem Schritt ist jedoch die Tatsache, dass der ASA-Bericht explizit keine Alternative für den t-Test anbietet und damit dem naheliegenden Fehler entgeht, eine schlechte Praxis durch eine andere zu ersetzen. Um Euklid zu paraphrasieren, es führt eben auch kein Königsweg zur Statistik. Es sollte selbstverständlich sein, dass der Einsatz von Verfahren und Methoden ein Verständnis eben dieser voraussetzt, gerade weil die „Inhalte" nie sinnvoll von den Methoden, die sie hervorgebracht haben, getrennt werden können.

Der zweite Einwand ist sogar noch allgmeiner, da er sich nicht nur auf die methodischen Aspekte der Arbeit eines Sozialwissenschaftlers bezieht, sondern für die große Gesamtheit des modernen Arbeitslebens gilt. Wir verbringen einen Großteil unserer Arbeitszeit im Austausch mit digitalen Maschinen. Auch wenn der Anteil der *digitalen Arbeit* an der Gesamtarbeitszeit sicherlich abhängig vom gewählten Beruf ist, so kann man doch davon ausgehen, dass dies in den Sozialwissenschaften die wohl gängigste Form der Arbeit sein dürfte. Diese Bedeutung digitaler Maschinen, als zentrale Werkzeuge unserer alltäglichen Arbeit, lässt es angezeigt sein unser Verhältnis zu ihnen zu überdenken. Hier kommt das Konzept der *Digital Literacy* ins Spiel. Grob gesagt wird damit die Kompetenz zur Konversation und zur effektiven Interaktion mit digitalen Maschinen beschrieben. Da Interaktion wechselseitiges Verstehen impliziert, bedarf es einer gemeinsamen Sprache um diesen Austausch zu bewerkstelligen. Programmiersprachen dienen genau diesem Zweck. Wie bei jeder natürlichen Sprache auch, erwirbt man durch den aktiven Sprachgebrauch mit Muttersprachlern nicht nur vertiefte Sprachkenntnisse, sondern immer auch kulturspezifisches Wissen. Anders ausgedrückt, das praktische Handlungswissen, das man durch die Nutzung von Programmiersprachen erwirbt, ist seinem Wesen nach nicht auf einen einzelnen Kontext oder ein spezifisches Problem beschränkt.

Gerade wenn man sich im Arbeitsalltag ständig mit Maschinen herumschlagen muss, liegen die Vorteile einer Möglichkeit seine Wünsche und Bedürfnisse klar formulieren zu können auf der Hand. Um auf die Metapher der natürlichen Sprache zurückzukommen, stelle man sich vor, dass man in einem Team arbeitet, in dem alle anderen eine bestimmte Sprache sprechen und auch keine andere lernen können. In so einer Situation würde einem das Beharren auf Gesten und unkodifizierter Zeichensprache auch nicht sehr viel bringen. Vor allem aber wird es an der Zeit sich den Tatsachen zu stellen: digitale Maschinen, Informationsnetzwerke und Computer sind ein fester Bestandteil der alltäglichen Arbeitswelt gewor-

den und es gibt keinerlei Anzeichen für eine Umkehrung dieser Entwicklung. Dadurch eröffnen sich aber auch Chancen zu einer effektiveren Gestaltung der Arbeit und einem selbstbestimmteren und reflektierten Umgang mit Daten und Verfahren.

5 Text und Token

Allgemein kann man Texte als endliche Sequenzen von Zeichen auffassen. Die Zeichen und ihre Anordnung im Text sind über ihre Zugehörigkeit zu einer bestimmten Sprache gekennzeichnet. Hierbei lässt sich zwischen formalen und natürlichen Sprachen unterscheiden. Bei den natürlichen Sprachen handelt es sich um historisch gewachsene Zeichensysteme, welche dem Zweck der Kommunikation dienen. Demgegenüber sind formale Sprachen bewusst konstruierte Gebilde, wie zum Beispiel Programmiersprachen oder mathematische Notationen, für die präzise, logische Regeln gelten. Allerdings können die Definitionen und Merkmale formaler Sprachen auch auf die natürlichen Sprachen ausgeweitet werden. Dies ist hier sogar notwendig, da eine quantitative Analyse natürlicher Sprachen deren digitale Verarbeitung voraussetzt, diese also in formaler Sprache abgefasst werden muss.

Formal betrachtet, verfügt eine *Sprache* $\mathcal{L}$ sowohl über ein endliches Vokabular bzw. *Alphabet* Σ, d.h. die endliche Menge aller Zeichen dieser Sprache, als auch ein Regelsystem, welches die Zusammensetzung dieser Zeichen beschreibt, die *Grammatik* der Sprache.[1] Beziehen sich die Regeln der richtigen Komposition auf die Zusammensetzung von Charakteren zu Wörtern, so spricht man von der *Morphologie*. Die Zusammensetzung von Wörtern zu Sätzen wird hingegen als *Syntax* oder Satzgrammatik bezeichnet. Diese Unterscheidung ist für formale Sprachen meist nicht sinnvoll, da in diesem Fall die Bedeutung eines Satzes auf der Ebene klar definierter Einzelzeichen geschieht und keine darunterliegenden Ebenen unterschieden werden. Im Bereich der natürlichen Sprachen ist sie jedoch unerlässlich.

[1] Die Verwendung des griechischen Majuskels Σ entspricht der gängigen Konvention in der Theorie formaler Sprachen (vgl. Erk und Priese 2008: 27). Diese Notation kann jedoch leicht zu Verwechselungen mit dem Summenzeichen ($\sum$) führen.

5.1 Maschinenlesbarer Text

Die Definition des Textes als einer Zeichenkette entspricht der Art und Weise in der „roher“ Text digital repräsentiert wird. Roher Text spielt dabei auf den Umstand an, dass eine solche Repräsentation meist die Ausgangslage für quantitative Textanalysen darstellt, allerdings in dieser Form nur sehr eingeschränkt verarbeitet werden kann. Eine Repräsentation von Text als Zeichenkette bezeichnet man als *String*. Strings stellen in vielen höheren Programmiersprachen einen der grundlegenden (primitiven) Datentypen dar, d.h. sie können nicht in grundlegendere Datentypen zerlegt werden. In Python[2] werden sie entweder mit einem einfachen ' oder einem doppelten Anführungszeichen " eingeleitet und ebenso beendet:

```
1 string1 = 'Das ist ein Text. Obwohl er nur zwei Sätze lang
      ist.'
2 string1
```

```
1 'Das ist ein Text. Obwohl er nur zwei S\xc3\xa4tze lang ist.'
```

```
1 string2 = "Dies      ist            ebenfalls        ein
      String"
2 string2
```

```
1 'Dies      ist            ebenfalls        ein  String'
```

An diesem kurzen Beispiel eines Strings lassen sich bereits eine Reihe von Besonderheiten und Problemen der Repräsentation von Text als String verdeutlichen. Zunächst fällt auf, dass ein String eine geschlossene Einheit ist, in der jedes Element ein Zeichen ist. Deshalb zählen auch die

[2] Diese und die folgenden Ausführungen beziehen sich auf Python 2.7+. In Python Version 3 und höher gibt es den Unterschied zwischen String und Unicode in dieser Form nicht mehr. Stattdessen nimmt Python 3 standardmäßig an, dass es sich um Unicode-codierten Text handelt. Dies führt zu einer radikalen Vereinfachung des Umgangs mit Texten. Gleichzeitig lassen sich die grundlegende Konzeption von Unicode und die damit verbundenen Probleme nicht mehr so einfach demonstrieren.

einzelnen Leerzeichen als Teile des Strings. Im Resultat sind beide Zeichenketten daher gleich lang:

```
1 print len(string1)
```

```
1 53
```

```
1 print len(string2)
```

```
1 53
```

Dies entspricht jedoch nicht der Art und Weise in der ein menschlicher Leser den Text auffassen würde. Leerzeichen und andere Zeichen, die für Menschen nicht direkt sichtbar sind (z.B.: Tabulatoren oder Absatzmarken) werden auch als „Whitespace“ bezeichnet, da sie für gewöhnlich nicht als Zeichen dargestellt werden. In der natürlichen Schriftsprache dienen diese Zeichen meist nur der Strukturierung des Textflusses und der Abtrennung von Wörtern. Daher haben diese Zeichen meist keinen eigenständigen, semantischen Wert. Wenngleich ihre Bedeutung für die Strukturierung des ganzen Textes enorm sein mag. In Maschinensprachen besitzen Leerzeichen hingegen eine spezifische Bedeutung und sind oft ein wichtiger Teil der Syntax. In Python werden zum Beispiel vier aufeinander folgende Leerzeichen, als ein Einschub für einen Codeblock interpretiert. Anders ausgedrückt, zählt in Maschinensprachen jedes einzelne Element eines Strings als potentielles „Wort“ (Token). Demgegenüber bedarf eine digitale Verarbeitung von natürlicher Sprache spezifischer Regeln um einzelne Token identifizieren zu können (siehe Abschnitt 5.2).

Die zweite Auffälligkeit besteht in der Verarbeitung der Sonderzeichen, welche im obigen Beispiel durch die Ersetzung des „ä“ in „Sätze“ mit dem Ausdruck `\xc3\xa4` geschieht. Dabei handelt es sich um einen sogenannten Unicode Codepoint. Der Grund dafür ist die Begrenzung des ursprünglichen Alphabets (ASCII: American Standard Code for Information Interchange) für die Eingabe von Zeichen in elektronische Datenverarbeitungssystem auf 128 Charaktere, welche der englischen Sprache entstammten. Folglich ist eine Reihe von Charakteren, wie beispielsweise die

Umlaute des Deutschen nicht im ASCII Code definiert. Um mit anderen Zeichensystemen umgehen zu können sind sogenannte Kodierschemata (Encodings) notwendig, die sich auf den Unicode Standard beziehen. Siehe hierzu auch Abschnitt 5.1.2, in dem der Unicode Standard behandelt wird.

Hier soll es darum gehen aufzuzeigen, dass Strings zunächst einmal nur die Ausgangslage quantitativer Textanalysen stellen. Die meisten „rohen" Daten liegen entweder als String oder Unicode String vor, weswegen ein grundlegendes Verständnis dieses Datentyps unerlässlich ist. Gleichzeitig zielt die Bearbeitung von Strings letztlich immer auf eine Transformation in leichter zu bearbeitende, sowie der Analyse zugänglichere Datenformate ab.

5.1.1 Bearbeitung von Strings

Standardmäßig bietet Python eine Reihe von Methoden zur Bearbeitung von Strings an. Ähnliche Verfahren finden sich in vielen anderen Skript- und Programmiersprachen. Sie lassen sich grob in Methoden der Prüfung von Strings und in Methoden zur Manipulation von Strings unterscheiden. Dabei muss beachtet werden, dass Strings als unveränderlich definiert sind. Jede Methode die einen String manipuliert gibt daher eine veränderte Kopie des ursprünglichen Strings zurück. Dies hat außerdem zur Folge, dass bestimmte Operationen nicht erlaubt sind. Beispielsweise das Einfügen von Elementen mittels des Indexes (slicing).

Da es sich bei Python Strings um Container Objekte handelt können sie über den *Index* angesprochen werden. Die Indexierung von Sequenzen ist eine der grundlegenden Techniken in Python, welche die Auswahl von einzelnen Elementen erlaubt. Um bestimmte Elemente auszuwählen werden diese in der Form `[i:j:k]` angegeben. Die drei Parameter können dabei wie folgt übersetzt werden: von der Stelle `i` (einschließlich) bis unter `j` in `k` Schritten. Werden die Parameter `i` oder `j` explizit ausgelassen, d.h. nur die Trennung mittels Doppelpunkten angegeben, so wird dies Selektion bis zum Ende/Anfang der Sequenz fortgesetzt.

Es ist dabei zu beachten, dass Pythons Indizes ihre Zählung bei `0` beginnen und auch negative Werte aufweisen können. Tabelle 5.1 zeigt die Stellen des Indexes sowie deren Korrespondenz mit den jeweiligen Elementen der Sequenz „String" an.

Mittels der Methode `.find()` kann die von links gelegene erste Stelle gefunden werden, an der ein bestimmter (Teil-)string auftritt. Diese Me-

thode des Durchsuchen eines Strings ist stark begrenzt, da sie nur den Index des Anfangs zurück gibt und nicht den des Endes.

```
string1.find('Text')
```

```
12
```

```
print string1[12]
print string1[12:16]
```

```
T
Text
```

Weitere wichtige Methoden zur Überprüfung und zum Durchsuchen von Strings umfassen:

- `.count(`*`sub`*`)`: Gibt die Anzahl des voneinander getrennten Auftretens des Teilstrings *`sub`* zurück.
- `.rfind(`*`sub`*`)`: Durchsucht den String von rechts nach links und gibt die Stelle des ersten Auftretens des Teilstrings *`sub`* zurück.
- `.isalnum()`: Prüft ob alle Charaktere des Strings alphabetische oder numerische Zeichen sind. Gibt booleschen Wert zurück (`True` oder `False`).
- `.isalpha()`: Prüft ob alle Charaktere des Strings alphabetische Zeichen sind. Gibt booleschen Wert zurück (`True` oder `False`).

S	t	r	i	n	g
0	1	2	3	4	5
-6	-5	-4	-3	-2	-1

Tabelle 5.1: Schematische Darstellung von Python-Indizes.

- `.isdigit()`: Prüft ob alle Charaktere des Strings numerische Zeichen sind. Gibt booleschen Wert zurück (`True` oder `False`).

Neben den Methoden der Überprüfung finden sich auch noch Methoden, die eine Transformation, oder genauer die Rückgabe einer veränderten Version des Ursprungsstrings, ermöglichen. Dadurch können wichtige Operationen wie zum Beispiel die Kürzung, Teilung und Veränderungen der Schreibweise durchgeführt werden. Die wichtigsten dieser Methoden sind:

- `.lower()` / `.upper()`: Gibt eine Kopie des ursprünglichen Strings zurück, bei dem sämtliche Zeichen klein- bzw. großgeschrieben wurden.
- `.split(`*`sep`*`)`: Gibt eine Liste von Strings (*list-of-strings*) zurück, indem der ursprüngliche string entlang jedes Auftretens des Teilstrings *`sep`* getrennt wird. Aufeinander folgende Instanzen von *`sep`* werden als ein Trennzeichen behandelt. Von rechts nach links: `.rsplit(`*`sep`*`)`.
- `'`*`str`*`'.join(`*`list`*`)`: Die inverse Operation zu `.split()`. Hier werden alle Elemente der Liste *`list`* durch *`str`* miteinander verknüpft. Das Ergebnis ist ein String.
- `.splitlines()`: Gibt eine Liste von Strings (*list-of-strings*) zurück, indem der ursprüngliche String entlang jedes Satztrennzeichens getrennt wird. Dabei kommt der *universal newlines* Ansatz zur Geltung, der folgende *end-of-line* Charaktere definiert: `'\n'` für Unix-Systeme, `'\r'` für ältere Macintosh-Systeme und `'\r\n'` entsprechend der MS-Windows Konvention.
- `.strip()`: Gibt eine Kopie des Strings zurück, bei der sämtlicher Whitespace am Anfang und Ende entfernt wurde. Was als Whitespace gilt wird durch die im Computer voreingestellten Gebietsschemaparameter (locale) definiert.
- `.replace(`*`old`*`, `*`new`*`)`: Gibt eine Kopie des Strings zurück, in der sämtliche Instanzen des Teilstrings *`old`* durch *`new`* ersetzt wurden.

Eine ausführlichere Beschreibung dieser und weiterer Methoden findet sich in der Python Dokumentation Kapitel 5.6.1.[3] Jenseits dieser spezifi-

[3] Die Dokumentation kann Online unter folgender Adresse abgerufen werden: https://docs.python.org/2/.

schen Methoden lassen sich Objekte vom Typ String in Python auch über eine Reihe von Operatoren bearbeiten. Diese Operatoren ähneln denjenigen, welche für andere Kontainer Datentypen (z.B. Listen, Tupel, etc.) in Python definiert sind und werden im Kapitel 5.6 der Python Dokumentation beschrieben.

5.1.2 Unicode

Bevor mit der Transformation von Texten und deren Analyse begonnen werden kann, muss man sich mit dem *encoding* des Textes beschäftigen. Wie oben schon angedeutet, können mittels des ASCII Schemas nur eine sehr begrenzte Anzahl von Zeichen dargestellt werden. Bei ASCII handelt es sich um eines der ursprünglichsten Kodierschemata, die alle die selbe Funktion erfüllen, nämlich die Repräsentation von Zeichen in einem Computer zu ermöglichen. Jedes Zeichen wurde dabei durch eine 8-Bit lange Zahl dargestellt, was zur technischen Begrenzung auf 256 Zeichen führte. In diesem Schema entspricht zum Beispiel der Zahlenwert 48 dem großgeschriebenen `H`.

Da die Zahl der in den menschlichen Sprachen verwendeten Schriftzeichen, inklusive anderer Symboliken, mathematischer Zeichen und dergleichen, weitaus größer als die zur Verfügung stehenden 256 Zeichen waren, wurden verschiedene alternative Kodierschemata entwickelt. Aus diesen Schemata entwickelte sich der heute weit verbreitete *Unicode Standard*, der vom Unicode Consortium verwaltet wird.[4] Zum gegenwärtigen Zeitpunkt definiert Unicode (8.0) 120.672 Zeichen mittels sogenannter *Code Points*. Jedes der so festgelegten Zeichen wird in der Form „U+“ gefolgt von einer hexadezimalen Zahl angegeben. Das großgeschriebene `H` wäre daher in Unicode Schreibweise „U+0048“. Die explizite Deklaration von Unicode geschieht in Python 2 durch das dem String vorangestellte `u`. Um einen Unicode Codepoint direkt anzusprechen muss dieser mit `\u` eingeleitet werden.

```
1 print u'\u0048'
```

```
1 H
```

[4] Eine weiterführende, technischere Beschreibung des Standards findet sich unter `http://www.unicode.org/standard/principles.html`.

Hier zeigt sich auch ein Vorteil, der für den enormen Erfolg von Unicode ausschlaggebend war, nämlich die Definition der ersten 256 Unicode Code Points entsprechend ihrer ASCII Vorbilder. Dadurch wurde die Umstellung vieler Systeme auf den Unicode Standard erleichtert, der sich im Zuge der Digitalisierung menschlicher Kommunikation zum dominanten globalen Standard entwickelte. Die Liste der definierten Unicode Zeichen umfasst heute nicht nur lebende Sprachen, sondern auch eine Reihe von historischen Zeichensystemen und sogar popkulturelle Symbole, wie Emoticons und Emoji.

Allerdings ist es wichtig zu verstehen, dass die Verwendung eines Unicode Schemas noch nicht bedeutet, dass der Text auf allen Maschinen in gleicher Art und Weise angezeigt wird oder überhaupt in lesbarer Form erscheint. Darüber wie ein Codepoint dargestellt wird entscheidet zum einen das *Encoding* eines Textes und zum anderen die vorhandenen Schriftarten (*fonts*). Letzteres ist vor allem wichtig, wenn es um das Setzen von Texten geht. Für die quantitative Analyse von Texten ist dies weniger entscheidend, da es hier vor allem um die richtige Zuordnung der Codepoints zu den Zeichen einer natürlichen Sprache ankommt, die durch das Encoding festgelegt ist. Das Problem besteht darin, dass Encodings oftmals mit Hinblick auf bestimmte Sprachregionen oder Systemarchitekturen definiert sind.[5] Zum Beispiel ist Latin-1 ein Encoding welches typischerweise für westeuropäische Sprachen Verwendung findet. Gleichzeitig gibt es noch eine Reihe weiterer Encodings, die diesen Sprachraum abdecken, wie die Windows Codepage cp1250 oder UTF-8. Diese Encodings unterscheiden sich zum Teil sehr tiefgreifend in ihren Definitionen des jeweiligen Codepoints.

Entsprechend des jeweils verwendeten Encodings, werden Unicode Zeichen unterschiedlich verarbeitet und gespeichert. So hängt beispielsweise die Schriftweise des €-Zeichens von der jeweils verwendeten Systemarchitektur ab. Da die folgenden Codebeispiele in einem Jupyter Notebook geschrieben wurde, entspricht die Ausgabe dem Standardencodings des Browsers, der für die Darstellung und Bearbeitung des Textes verwendet wurde. Daher wird das €-Zeichen mittels eines Latin-1 Encodings dargestellt.

```
1 '€'
```

[5] Ein Überblick über einige in Python verfügbare Standardencodings und die Sprachen für die sie am häufigsten verwendet werden findet sich in der Python Dokumentation in Kapitel 7.8.3.

```
'\xe2\x82\xac'
```

Gibt man das gleiche Zeichen jedoch als Unicode String ein, so erhält man einen String, dessen Codierung dem jeweiligen Encoding des Betriebssystem entspricht. In diesem Fall handelt es sich dabei um UTF-8.

```
u'€'
```

```
u'\u20ac'
```

Das Encoding eines Strings kann und muss in manchen Fällen explizit angegeben werden. In Python ist dies durch die Funktion `unicode()` bzw. die Methoden `.encode()` und `.decode()` möglich. Mittels der Methode `.decode()` kann ein String in ein bestimmtes Unicode Encoding überführt werden. Die folgende Zeile Code überführt das eingegebene €-Zeichen in einen cp1250 kodierten Unicode String. Das Argument (`encoding`) gibt hierbei das Zielencoding des resultierenden Unicode Strings an.

```
'€'.decode('cp1250')
```

```
u'\xe2\u201a\xac'
```

Das Gegenstück dazu ist die `.encode()` Methode, die einen Unicode String in einen normalen String überführt. Im Falle der hier verwendeten Encodings bedeutet dies aus einem UTF-8 kodierten String einen String zu machen, dessen nicht auf ASCII zurückführbaren Zeichen in die intern verwendete Kodierung der Textverarbeitung übertragen werden (hier: Latin-1). In diesem Fall bezieht sich das `encoding` Argument der Methode auf das ursprüngliche Encoding des Strings.

```
u'€'.encode('utf8')
```

```
1 '\xe2\x82\xac'
```

Diese Konzepte sind vor allem dann wichtig, wenn „rohe“ Strings, die mittels Text oder Data Mining Verfahren gewonnen wurden verarbeitet werden sollen. Bird et al. (2009) schlagen in ihrem Buch *Natural Language Processing with Python* ein decode/encode Schema vor, dass sicherstellen soll, dass Fehler bei der Bearbeitung von Strings reduziert werden und die Archivierung von Textdaten möglichst reibungslos funktioniert. Dabei werden die rohen Strings zunächst in Python Unicode transferiert (dekodiert), um dann als Unicode Strings im aktiven Speicher bearbeitet zu werden. Die Speicherung von Textdateien sollte anschließend ebenfalls mit einer expliziten Kodierung geschehen, damit sichergestellt werden kann, dass es bei einem Transfer zwischen verschiedenen Arbeitsumgebungen nicht zu einem Datenverlust kommt.

5.2 Token

Die Konzeption von Strings als unveränderlichen Zeichenketten, bei denen jedes Element eine eigene Bedeutung hat, ist vor allem dem grundlegenden Modell formaler Sprachen geschuldet. Deshalb eignet sich dieser Datentyp insbesondere um Anweisungen in Maschinensprachen zu verarbeiten. Natürliche Sprachen passen wesentlich schlechter in dieses Modell. Der wichtigste Unterschied liegt dabei in der Verarbeitung des Whitespace, also all jener Zeichen, die in menschlichen Sprachen meist zur Abgrenzung von Textbausteinen genutzt werden. Es sind jedoch genau diese symbolischen Einheiten, welche die Ausgangslage für die quantitative Analyse von Text darstellen. Daher müssen die einzelnen symbolischen Elemente, die als *Token* bezeichnet werden, identifiziert und in eine Datenstruktur überführt werden, die besser bearbeitbar ist als unveränderliche Strings. Diesen Vorgang bezeichnet man als Tokenisierung eines Textes.

Was genau einen Token kennzeichnet, d.h. einen Teil des Textes als eigenständige symbolische Einheit ausmacht, ist nicht immer direkt ersichtlich und hängt in nicht unerheblichem Maße von der verwendeten Sprache, vom Forschungsinteresse und der gewählten Perspektive ab. Token können daher auch nicht mit Wörtern gleichgesetzt werden. Vielmehr sind Token ein Platzhalter für Textausschnitte, die gemäß einer bestimmten *Tokenisierungsregel* erzeugt werden. Die damit verbundenen praktischen Probleme werden deutlich, wenn man sich vor Augen führt, dass

Wörter in verschiedenen Sprachen über einzelne Zeichen, über die Komposition von Zeichen und über die Mischung beider Konstruktionsregeln gebildet werden. Daher werden zum Teil sehr komplexe Tokenisierungsregeln notwendig.

Betrachten wir zunächst den relativ einfachen Fall eines in lateinischen Buchstaben verfassten, deutschsprachigen Textes. Zu diesem Zweck wird im folgenden Codebeispiel ein Objekt (mit dem Namen `text`) definiert, welches aus einem String besteht, der in Pythons internem Unicodeformat definiert wurde. Die dreifachen Apostrophen ermöglichen es einen Text zu schreiben der einfache Apostrophen sowie Anführungszeichen als Teile des Strings auffasst und umgebrochen werden kann.

```
1 text = u'''Hier sollte ein "anständiger" Text stehen. Am
      besten \
2 einer, der verschiedene Besonderheiten aufweist. Zum
      Beispiel, \
3 Zahlen (1999), Preise (19.99 €) und dergleichen. Außerdem
      sollten \
4 unterschiedliche Arten von Whitespace enthalten sein. Zum
      Beispiel \
5 ein Tabulator ("\t").'''
6
7 print text
```

```
1 Hier sollte ein "anständiger" Text stehen. Am besten einer,
      der verschiedene Besonderheiten aufweist. Zum Beispiel,
      Zahlen (1999), Preise (19.99 €) und dergleichen.
      Außerdem sollten unterschiedliche Arten von Whitespace
      enthalten sein. Zum Beispiel ein Tabulator ("   ").
```

Konzentrieren wir uns zuerst auf Wörter als Token und die Trennung des Textes entlang des sogenannten *Whitespace*. Gemeint ist damit jegliches Zeichen, dass in einem formatierten Text, für die menschliche Wahrnehmung, als eine Leerstelle erscheint. In diesem Fall enthält der Text zwei Arten von Whitespace, nämlich Leerzeichen und einen Tabulator (`\t`). Diese Tokenisierungsregel kann durch die Methode `.split()` angewendet werden.

```
1 print text.split()
```

```
1 [u'Hier', u'sollte', u'ein', u'"anst\xe4ndiger"', u'Text',
    u'stehen.', u'Am', u'besten', u'einer,', u'der',
    u'verschiedene', u'Besonderheiten', u'aufweist.',
    u'Zum', u'Beispiel,', u'Zahlen', u'(1999),', u'Preise',
    u'(19.99', u'\u20ac)', u'und', u'dergleichen.',
    u'Au\xdferdem', u'sollten', u'unterschiedliche',
    u'Arten', u'von', u'Whitespace', u'enthalten', u'sein.',
    u'Zum', u'Beispiel', u'ein', u'Tabulator', u'("', u'").']
```

Das Ergebnis dieser Operation ist ein Objekt vom Datentyp `list`. Listen eignen sich sehr viel besser zur Darstellung eines Textes in natürlicher Sprache. Da sie eine veränderliche Sequenz von beliebigen Objekten darstellen, können sie leichter transformiert und bearbeitet werden. Gleichzeitig demonstriert das Ergebnis, dass eine simple Tokenisierungsregel, wie die Auftrennung am Whitespace, nicht das gewünschte Ergebnis bringt. Zum einen werden die Satz- und Sonderzeichen nicht immer korrekt von den Worten getrennt, da in der Schriftsprache an vielen Stellen keine Leerzeichen verwendet werden, beispielsweise im Fall des Schlusspunktes eines Satzes. Zum anderen werden auch symbolische Einheiten getrennt, die je nach Kontext auch als ein Token aufgefasst werden können. Zum Beispiel das Eurozeichen, welches mit dem Preis zusammen eine symbolische Einheit bilden kann. Allerdings sind auch Texte oder Analysen vorstellbar in denen wir die numerische Angabe von Preisen und die Kennzeichnung der Währung getrennt behandeln würden.

Im Gegensatz zu formalen Sprachen ist die Bestimmung der relevanten, symbolischen Einheiten in den natürlichen Sprachen nicht von vorneherein möglich. Die geringere Eineindeutigkeit und das Fehlen einer formalen Grammatik können Tokenisierungsregeln notwendig machen, die sensibel für den Kontext und die Entstehungsbedingungen eines Textes sind. Noch wichtiger ist jedoch das leitende Interesse des Forschers. Konkrete Tokenisierungsregeln sollten je nach Forschungsfrage festgelegt werden und sollten die theoretischen Annahmen über den Gegenstandsbereich widerspiegeln. Eine solche *Kontextsensibilität* in der Tokenisierung ist jedoch mit grundlegenden Methoden wie `.split()` nicht zu erreichen. Eine Tokenisierung, die diesen Anforderungen gerecht werden kann, ist meist nur über reguläre Ausdrücke möglich. Da diesen jedoch eine eigene

Form der Grammatik zugrundeliegt, die für weit mehr als nur das Zerlegen eines Strings in Token eingesetzt werden kann, muss deren ausführlichere Diskussion an anderer Stelle (Abschnitt 5.4) erfolgen.

Um das allgemeine Vorgehen zur Tokenisierung von Texten beschreiben zu können wird im Folgenden auf die bestehenden Tokenisierungsfunktionen des *Natural Language Toolkit* (NLTK) zurückgegriffen. NLTK ist eine externe Programmbibliothek für Python (auch Modul oder Paket genannt), die eine Vielzahl an Funktionen und Ressourcen für die Bearbeitung von Texten umfasst. Der Fokus liegt dabei auf Routinen des *natural language processing*, also der Verarbeitung und Analyse von Texten nach computerlinguistischen Gesichtspunkten (Bird, Klein und Loper 2009: Kap. 0). Dazu gehören so wichtige Funktionalitäten, wie die Verwaltung großer annotierter Korpora, Stemming-Algorithmen und die Bestimmung von grammatikalischen Wortformen (*part-of-speech tagging*). Solange die Texte an der alltäglichen Sprache orientiert sind und nicht zu viele Formatierungsfehler aufweisen, eignen sich die NLTK-Verfahren in sehr hohem Maße für die Bearbeitung.

Im Folgenden wird das NLTK Paket importiert und die Tokenisierung des Beispieltextes mittels der `word_tokenize()` Funktion vorgenommen:

```
1 import nltk
2 tokens = nltk.word_tokenize(text)
3
4 print tokens
```

```
1 [u'Hier', u'sollte', u'ein', u'``', u'anst\xe4ndiger',
    u"''", u'Text', u'stehen', u'.', u'Am', u'besten',
    u'einer', u',', u'der', u'verschiedene',
    u'Besonderheiten', u'aufweist', u'.', u'Zum',
    u'Beispiel', u',', u'Zahlen', u'(', u'1999', u')', u',',
    u'Preise', u'(', u'19.99', u'\u20ac', u')', u'und',
    u'dergleichen', u'.', u'Au\xdferdem', u'sollten',
    u'unterschiedliche', u'Arten', u'von', u'Whitespace',
    u'enthalten', u'sein', u'.', u'Zum', u'Beispiel',
    u'ein', u'Tabulator', u'(', u'``', u"''", u')', u'.']
```

Das Ergebnis der Anwendung von `word_tokenize()` ist im Falle von Texten die wenige Sonderzeichen, bzw. keine Mischung verschiedener

Sprachen enthalten, relativ verlässlich. Texte die über eine ausgeprägte Fachsprache (z.B. wissenschaftliche Texte), uneinheitliche Schreibweisen (z.B. Online-Kommunikation) oder vom Fließtext abweichende Formatierungen (z.B. Tabellen in Geschäftsberichten) verfügen, sind für eine solche Tokenisierung oft ungeeignet.

Darüber hinaus bietet NLTK noch eine Reihe von weiteren, spezialisierteren Tokenisierungsfunktionen an.[6] Darunter findet sich unter anderem ein Tokenisierer für die Satzebene (`sent_tokenize()`) und für spezifische Textgattungen, wie zum Beispiel die `TweetTokenizer` Klasse. Grundsätzlich empfiehlt sich ein Blick in die Dokumentation der API, um den weitreichenden Inhalt besser abschätzen zu können.

Wegen der praktischen Bedeutung und auch um das Konzept der Tokenisierung noch eingängiger zu erläutern, wird hier noch kurz die Tokenisierung auf der Satzebene demonstriert.

```
1 sent_tokens = nltk.sent_tokenize(text)
2
3 print sent_tokens
```

```
1 [u'Hier sollte ein "anst\xe4ndiger" Text stehen.', u'Am
    besten einer, der verschiedene Besonderheiten
    aufweist.', u'Zum Beispiel, Zahlen (1999), Preise (19.99
    \u20ac) und dergleichen.', u'Au\xdferdem sollten
    unterschiedliche Arten von Whitespace enthalten sein.',
    u'Zum Beispiel ein Tabulator ("\t").']
```

Die `sent_tokenize()` Funktion operiert nach einem zweistufigen Verfahren. Zunächst wird versucht alle diejenigen Interpunktionen zu identifizieren, die keine Satzendungen darstellen (z.B. Abkürzungen). Dabei kommt der Algorithmus von Mikheev (2002) zum Einsatz. Zudem wird ein trainierter Klassifikationsalgorithmus (siehe auch Abschnitt 6.2.1) zur Identifikation von nicht-satztrennenden Zeichen verwendet. Das Training des Klassifikationsverfahrens erfolgte dabei an englischsprachigen Texten. Es eignet sich jedoch auch für andere indoeuropäische Sprachen, da der Satzbaus strukturell ähnlich ist. In manchen Fällen kann es aufgrund von problematischen Textformatierungen jedoch zu Problemen kommen.

[6] Eine genauere Beschreibung der grundlegenden Klassen und Objekte findet sich in der Dokumentation der NLTK API: http://www.nltk.org/api/nltk.tokenize.html.

Dies ist häufig der Fall, wenn Texte erst durch eine Umwandlung ihres Formats zu „rohem“ Text gemacht wurden. Beispielsweise durch die Konvertierung von PDF (Portable Document Format) zu einfachen Textdateien.

Im zweiten Schritt wird dann die Zerlegung des Textes anhand einer vordefinierten Liste von Unicode-Satzzeichen vorgenommen. Diese setzt sich wie folgt zusammen:

```
1 nltk.tokenize.punkt.PunktSentenceTokenizer.PUNCTUATION
```

```
1 (u';', u':', u',', u'.', u'!', u'?')
```

Der NLTK-Tokenisierer für die Wortebene baut auf dem gleichen, zweistufigen Verfahren auf um den Text in Token aufzuteilen. Erst nach einem Ausschluss von irrelevanten Trennzeichen, zum Beispiel einem Apostroph, wird der restliche Text entlang des im Unicode-Schema definierten Whitespaces aufgetrennt. Daher treffen die oben erwähnten Probleme auch auf die `word_tokenize()` Funktion zu.

Manche sprachlichen Phänomene können nur auf der Satzebene sinnvoll bestimmt werden, dies gilt insbesondere für die Grammatik. Gleichzeitig setzen die Verfahren zur Analyse grammatikalischer Eigenschaften in den meisten Fällen ebenfalls eine Tokenisierung auf Wordebene voraus. Dazu ist eine Kombination von Satz- und Wort-Tokenisierer notwendig. Das Resultat ist die etwas komplexere Datenstruktur einer Liste-von-Listen-von-Token:

```
1 print [nltk.word_tokenize(sent) for sent in sent_tokens]
```

```
1 [[u'Hier', u'sollte', u'ein', u'``', u'anst\xe4ndiger',
    u"''", u'Text', u'stehen', u'.'], [u'Am', u'besten',
    u'einer', u',', u'der', u'verschiedene',
    u'Besonderheiten', u'aufweist', u'.'], [u'Zum',
    u'Beispiel', u',', u'Zahlen', u'(', u'1999', u')', u',',
    u'Preise', u'(', u'19.99', u'\u20ac', u')', u'und',
    u'dergleichen', u'.'], [u'Au\xdferdem', u'sollten',
    u'unterschiedliche', u'Arten', u'von', u'Whitespace',
    u'enthalten', u'sein', u'.'], [u'Zum', u'Beispiel',
    u'ein', u'Tabulator', u'(', u'``', u"''", u')', u'.']]
```

Eine solche Datenstruktur kann auch genutzt werden um andere Ebenen von Texten und Diskursen abzubilden. Zum Beispiel als eine Liste der tokeniserten Absätze eines Textes oder als Liste von Dokumenten, die ebenfalls als Token repräsentiert werden. Allerdings ist es ratsam solche „nested lists“ nicht zu tief ineinander zu stapeln, da dies sehr schnell unübersichtlich und sehr schwer handhabbar werden kann. Ist der Korpus von einer Größe die es erlaubt ihn im Arbeitsspeicher zu halten, kann auch ein `pandas.DataFrame` zur Analyse und Datenverwaltung des Korpus genutzt werden. Das `pandas`[7] Paket bietet eine skalierbare und robuste Infrastruktur für die Verarbeitung von tabellarischen Daten in Python.

Mittels der `.apply()` Methode können Tokenisierungsfunktionen auf die Rohfassungen der Texte angewendet werden. Diese Vorgehensweise bietet zwei entscheidende Vorteile. Erstens erhält man so eine einheitliche Datenstruktur für den gesamten Korpus. Zweitens ist der `pandas.DataFrame` leicht in eine Reihe von statistischen Analyseverfahren einzubinden. Am Beispiel der Rohfassung des „Soziologie Abstracts“-Korpus (`SozAbstRaw.pkl`) soll dies kurz demonstriert werden. Eine ausführliche Beschreibung des Inhalts und Aufbaus dieses Korpus findet sich im Abschnitt 5.3.2.

Das folgende Codebeispiel importiert zunächst das `pandas` Modul und verwendet dessen `.read_pickle()` Methode um den Korpus von der Festplatte einzulesen. Anschließend wird die Tokenisierung durchgeführt, indem die bereits besprochene Tokenisierungsfunktion `nltk.word_tokenize()` per `.apply()` auf eine Sammlung von Abstracts ausgewählter Soziologiezeitschriften angewendet wird. Um die resultierende Datenstruktur aufzuzeigen, werden mittels der Methode `.head()` die ersten fünf tokenisierten Abstracts in verkürzter Form wiedergegeben. Dabei handelt es sich jeweils wieder um Listen die einzelne Token enthalten.

```
1 import pandas as pd
2
3 articles = pd.read_pickle('Daten/Soziologie/SozAbstRaw.pkl')
4
5 articles['Tokens'] = articles.Abstracts\
6                               .apply(nltk.word_tokenize)
```

[7] `http://pandas.pydata.org/`

```
7
8 articles.Tokens.head()
```

```
1 0    [This, paper, seeks, to, contribute, to, socia...
2 1    [It, is, often, suggested, that, the, politica...
3 2    [A, distinction, has, recently, been, proposed...
4 3    [Superficially, ,, Actor, Network, Theory, (, ...
5 4    [In, Bowling, Alone, Robert, Putnam, considers...
6 Name: Tokens, dtype: object
```

5.2.1 Methodologisches Caveat

Die Tokenisierung eines Textes stellt die Grundlage jeglicher quantitativen Textanalyse dar. Erst durch die Einteilung in vergleichbare und zählbare Sinneinheiten können die strukturellen Eigenschaften von Texten bestimmt und miteinander verglichen werden. In der Sprache der soziologischen Methodologie ausgedrückt, lässt sich der jeweilige Text als ein *Merkmalsträger* auffassen. Demzufolge stellen die Token die *Merkmale* bzw. *Variablen* dar, welche wir zur Beschreibung und Analyse des Textes heranziehen. Dies impliziert jedoch eine Gültigkeit dieser Variablen über die konkreten Messungen hinaus. Im Rückbezug auf die Diskussion der methodologischen Grundlagen einer soziologischen Symboltheorie, sehen wir hier die praktische Wendung des Problems des objektiven Charakters von Symbolen. Die Verwendung von Token als Analyseeinheiten basiert auf der Annahme, dass sie auch jenseits des Kontextes ihres Auftretens in einem objektiven Bezug zueinander stehen und sinnvoll voneinander abgrenzbar sind.

Ganz offensichtlich trifft die Annahme, dass Token mit sich selbst identisch sind, nicht immer zu. Zum Beispiel verkehren unterdrückte Minderheiten in ihren internen Diskursen oft die Bedeutung von Symbolen ins Gegenteil, die ursprünglich zu ihrer Herabwertung und Ausgrenzung gedacht waren (z.B. Croom 2013). Würde man Texte aus unterschiedlichen Diskursen heranziehen, in denen Token unterschiedliche Relationen zu einander aufweisen, so kann dies zu einer Verzerrung der Ergebnisse führen. Das Problem kann als die *Überlappung symbolischer Ordnungen* hinsichtlich ihrer Token verallgemeinert werden. Ein möglicher Weg damit umzugehen besteht darin, Diskurse, bei denen von einer sprachlichen Separation bei gleichbleibenden Token ausgegangen werden kann, zunächst

getrennt zu analysieren und nur dann zusammenzuführen, wenn man sich des Ausmaß des Problems bewusst ist.

Ein weiteres, methodisches Problem tritt auf, wenn Token nicht in allen Fällen mit sich selbst identisch sind. Bei natürlichen Sprachen kommt dies häufig in Form von Phrasen vor, also feststehenden Verkettungen von Wörtern. Diese *Phrasen* können als eigenständige, symbolische Einheiten aufgefasst werden, da sie nicht weiter zerlegt werden können ohne ihre Bedeutung zu verlieren. Beispielsweise wurde der Kommunismus in den siebziger Jahren oft als die „rote Bedrohung" bezeichnet. Eine Zerlegung dieser Phrase in „rote" und „Bedrohung" würde jedoch auch deren Bedeutungsgehalt als eigenständigen Begriff auflösen. Dieses Problem kann durch die Analyse der Kollokationen (bzw. N-Grame; siehe Abschnitt 5.5.1) eines Textes abgeschätzt und zumindest in Teilen behoben werden, indem man die identifizierbaren, festen Wendungen als eigenständige Token auffasst. Der Erfolg dieses Vorgehens ist allerdings davon abhängig wie gut solche Phrasen mittels automatischer Verfahren identifiziert werden können.

Methodische Probleme, wie die hier beschriebenen, die im Rahmen der Vorbereitung der Text auf die Analyse auftreten, werden im Bereich des Natural Language Processing oft mit der Formel *garbage in, garbage out* beschrieben. Damit wird auf die Unmöglichkeit hingewiesen eine fehlerhafte Textaufbereitung in der Analyse wieder ausgleichen, bzw. später „herausrechnen" zu können. Jegliche Modelle und Beschreibungen der in Texten zum Ausdruck kommenden symbolischen Ordnungen sind notwendigerweise durch die Qualität der Aufbereitung dieser Texte begrenzt. Daher kommt der Tokenisierung und vor allem den verwendeten Tokenisierungsregeln, sowie deren theoretischer Begründung, eine zentrale Bedeutung in der quantitativen Textanalyse zu.

5.2.2 Texte als Listen

Das Endergebnis jeglicher Tokenisierung ist eine Sequenz von Token. In Python werden hierfür meist `list` Objekte verwendet. Bei Pythons Listen handelt es sich um veränderbare Sequenzen von beliebigen Objekten, die über einen Index angesprochen werden können. Daneben könnten noch andere primitive Datentypen in Betracht gezogen werden, beispielsweise Tupel oder Arrays. Diese sind ebenso wie Listen nicht nur in Python implementiert, sondern finden sich prinzipiell in fast allen Programmierspra-

chen, wenngleich sich die Details der Implementation stark unterscheiden können.

Ein Vergleich von Tupeln und Arrays mit Listen kann helfen deutlich zu machen, warum gerade Listen eine intuitive Wahl zur Repräsentation von Texten darstellen. Tupel sind im Prinzip nichts weiter als Listen, die nach ihrer Erstellung nicht mehr verändert werden können. Aus diesem Grund sind sie tendenziell schwieriger zu transformieren. Arrays hingegen sind ähnlich transformierbar wie Listen, können jedoch nur Elemente ein und desselben Typs enthalten. Sie haben den großen Vorteil grundsätzlich effizienter hinsichtlich Rechenzeit und Speicherbedarf zu sein. Jedoch sind sie nicht standardmäßig in Python implementiert, was eine Installation des `numpy` Moduls notwendig macht. Zudem kann eine effiziente Transformation von Arrays sehr aufwendig werden. Da die Sequenzierung in Listen meist nur einen Zwischenschritt der Textanalyse darstellt und die reine Geschwindigkeit der Berechnung bei wissenschaftlichen Anwendungen kein Selbstzweck ist, stellt die Verwendung von Listen oft eine pragmatischere Lösung dar.

Eine Repräsentation von Texten als Listen von Token hat eine Reihe von Vorteilen gegenüber einer Darstellung als String. Dies ist hauptsächlich darauf zurückzuführen, dass es leichter ist über die Elemente einer Sequenz zu iterieren. Dadurch wird die Identifikation einzelner Elemente des Textes wesentlich vereinfacht. Zudem können die so identifizierten Token auch leichter verändert und gezählt werden. Diese drei Techniken werden im Folgenden kurz erläutert, da sie die Basis für weitere Standardoperationen der quantitativen Textanalyse liefern.

Die Iteration über eine Liste von Elementen wird in Python durch die Verwendung einer spezifischen Syntax, der *List Comprehension*, wesentlich vereinfacht. Dadurch können komplexe Schleifen und Bedingungen sehr kompakt ausgedrückt werden. Der Nachteil dabei ist die verringerte Lesbarkeit dieser Schreibweise für den menschlichen Betrachter. Der folgende Code verwendet zunächst die klassische Schreibweise einer *for*-Schleife um sämtliche großgeschriebenen Wörter in der Liste unserer Token zu identifizieren.

```
1 ## Definition einer leeren Liste
2 L = []
3
4 ## Schleife
5 for token in tokens:
```

```
6     if token[0].isupper():
7         L.append(token)
8
9 ## Liste ausgeben
10 print L
```

```
1 [u'Hier', u'Text', u'Am', u'Besonderheiten', u'Zum',
    u'Beispiel', u'Zahlen', u'Preise', u'Au\xdferdem',
    u'Arten', u'Whitespace', u'Zum', u'Beispiel',
    u'Tabulator']
```

Hier wird zunächst eine leere Liste definiert, um die identifizierten Token aufzunehmen. Danach wird über sämtliche Elemente der Sequenz `tokens` iteriert. Mittels der *if*-Bedingung wird geprüft ob sich an der nullten Indexstelle ein großgeschriebenes Schriftzeichen befindet. Da Groß- und Kleinschreibung für eine Reihe von natürlichen Sprachen definiert ist, sollte bei solchen Operationen immer auf Unicode-Strings zurückgegriffen werden. Würde hier ein Raw-String verwendet, so wäre die `.isupper()` Methode nicht in der Lage großgeschriebene Umlaute richtig zu identifizieren. Im Anschluss werden diejenigen Token an die Liste übergeben, für die diese Prüfung mit `True` evaluiert wurde.

Die folgende List Comprehension erzeugt dasselbe Ergebnis, ist jedoch sehr viel kompakter:

```
1 L = [token for token in tokens
2      if token[0].isupper()]
3
4 print L
```

```
1 [u'Hier', u'Text', u'Am', u'Besonderheiten', u'Zum',
    u'Beispiel', u'Zahlen', u'Preise', u'Au\xdferdem',
    u'Arten', u'Whitespace', u'Zum', u'Beispiel',
    u'Tabulator']
```

Gegenüber der ersten Schreibweise finden sich zwei Hauptunterschiede. Erstens erzeugt die List Comprehension, wie der Name schon andeutet, direkt eine Liste. Zweitens findet sich die Spezifikation des Elements,

welches die Elemente der neuen Liste ausmachen soll an der ersten Stelle, noch vor der Deklaration der Schleife. Die Schleife und die Bedingung werden jedoch weiterhin in derselben Reihenfolge geschrieben.

Mittels einer solchen Ausdrucksweise lassen sich die oben genannten Operationen (durchsuchen, transfomieren und zählen) an Texten sehr effizient ausführen. Eine sehr wichtige Standardoperation ist dabei die Veränderung von Strings um bestimmte, grammatikalische Eigenschaften zu entfernen (*stemming*) oder Vergleichbarkeit herzustellen. Da Stemming-Verfahren im nachfolgenden Unterkapitel eingängiger beschrieben werden, soll die Transformation von Texten zunächst am Beispiel der Groß- und Kleinschreibung demonstriert werden. Der folgende Code erzeugt eine Liste, bei der alle Worte kleingeschrieben sind. Dies ist insbesondere dann sinnvoll, wenn man nur an den spezifischen Wörtern eines Textes interessiert ist und die Informationen, die in den Unterschieden von Groß- und Kleinschreibung enthalten sind, für die Untersuchung nicht von Bedeutung sind.

```
1 tokens_lower = [token.lower()
2                 for token in tokens]
3
4 print tokens_lower
```

```
1 [u'hier', u'sollte', u'ein', u'``', u'anst\xe4ndiger',
    u"''", u'text', u'stehen', u'.', u'am', u'besten',
    u'einer', u',', u'der', u'verschiedene',
    u'besonderheiten', u'aufweist', u'.', u'zum',
    u'beispiel', u',', u'zahlen', u'(', u'1999', u')', u',',
    u'preise', u'(', u'19.99', u'\u20ac', u')', u'und',
    u'dergleichen', u'.', u'au\xdferdem', u'sollten',
    u'unterschiedliche', u'arten', u'von', u'whitespace',
    u'enthalten', u'sein', u'.', u'zum', u'beispiel',
    u'ein', u'tabulator', u'(', u'``', u"''", u')', u'.']
```

Die dritte Standardoperation on Texten ist das Zählen der enthaltenen Token. Dies kann zweierlei beinhalten, zum einen die Bestimmung der Anzahl der Token in einem Text und zum anderen die Häufigkeit des Auftretens eines bestimmten Tokens. Um alle enthaltenen Token zu zählen

kann die Länge der Liste mittels der Funktion `len()` bestimmt werden. In diesem Fall enthält der Text insgesamt 52 Token.

```
1 len(tokens)
```

```
1 52
```

Um die Häufigkeit eines bestimmten Tokens festzustellen wird die Methode `.count()` verwendet. Diese bestimmt, wie oft ein spezifisches Element in einer Liste vorkommt. Im folgenden Beispiel wird die Häufigkeit des Tokens `'ein'` abgefragt:

```
1 tokens.count('ein')
```

```
1 2
```

Die Bestimmung der numerischen Eigenschaften eines Textes stellt die Ausgangslage für die weitere Bearbeitung von Texten mit statistischen Verfahren dar. Deswegen bietet es sich an nicht nur die Häufigkeit eines spezifischen Tokens bestimmen zu können, sondern die Häufigkeitsverteilung aller Token, bzw. eines spezifischen Subsets (z.B. alle alphanumerischen Token), feststellen zu können. Um die Häufigkeitsverteilung bestimmen zu können bedarf es einer Eingrenzung auf einzigartige Werte, so dass jeder Token nur einmal in der Zählung vorkommt und eine Überführung in einen Datentyp, der eine Zuordnung der Häufigkeit zum jeweiligen Token möglich macht. Um alle einzigartigen Token zu erhalten, kann die `set()` Funktion verwendet werden. Für das weitere Arbeiten mit der durch Zählung gewonnen Häufigkeitsverteilung bietet sich der Datentyp des Python-Diktionärs an. Dabei werden einer Reihe von einzigartigen Schlüsseln spezifische Werte zugewiesen. Durch die Angabe des Schlüssels kann zu einem späteren Zeitpunkt der Wert zurückgeholt werden.

```
1 dictionary = {}
2
3 for unique in set(tokens):
4     dictionary[unique] = tokens.count(unique)
```

```
5
6 dictionary['ein']
```

```
1 2
```

Auf die Analyse von Texten ausgelegte Programmbibliotheken, wie zum Beispiel NLTK, verfügen oft über eigene Funktionen zur Erzeugung von Häufigkeitsverteilungen (NLTK: `nltk.FreqDist`). Diese bieten zudem noch eine Reihe von Methoden zur Visualisierung und weiteren Verarbeitung der Verteilung an. Dennoch basieren sie auf der selben Logik wie der obige Code. Für Sozialwissenschaftler bietet es sich zudem oft an die Häufigkeitsverteilung in eine Array oder eine Datentabelle zu überführen, um weiterführende statistische Analysen durchzuführen. Das konkrete Vorgehen wird in Abschnitt 5.6 genauer beschrieben.

5.2.3 Typen

Die oben beschriebene Vorgehensweise zum Zählen der Worthäufigkeiten verweist auf die Beziehung von Token und Typen, deren mengentheoretischer Hintergrund bereits im Abschnitt 2.3 erläutert wurde. Gemeint ist damit der Unterschied zwischen der konkreten Realisation eines Zeichens und der *Menge* der möglichen sprachlichen Zeichen, dem Alphabet Σ. Der einfachste Weg die Typen eines Textes festzustellen besteht in der Überführung der Liste der Token in eine Menge. Dies geschieht in Python mittels der Funktion `set()`, welche ein beliebiges, iterierbares Objekt in eine Menge von Typen umwandelt.

```
1 set('aaabbccccccccc')
```

```
1 {'a', 'b', 'c'}
```

Da es sich bei dem resultierenden Objekt um eine Menge handelt, sind sämtliche Standardoperationen der Mengenlehre möglich. Dies kann genutzt werden um Aussagen über die Gemeinsamkeiten und Unterschiede zwischen Texten hinsichtlich der Verwendung von Typen zu treffen.

Somit können, beispielsweise, Unterschiedlichkeiten im Jargon operationalisiert werden. Für eine Demonstration dieser Vorgehensweisen wird auf den Brown Korpus aus dem NLTK Paket zurückgegriffen.

Der Brown Corpus wurde 1961 von W. N. Francis und H. Kucera zusammengestellt und veröffentlicht.[8] Der Korpus umfasst eine Stichprobe von 374 Texten die im Jahre 1961 in den USA erschienen sind. Zudem ist der Text in eine Reihe von Kategorien unterteilt, die den Vergleich verschiedener Textsorten möglich macht.

```
1 from nltk.corpus import brown
2
3 # Kategorien ausgeben:
4 print(brown.categories())
```

```
1 [u'adventure', u'belles_lettres', u'editorial', u'fiction',
    u'government', u'hobbies', u'humor', u'learned',
    u'lore', u'mystery', u'news', u'religion', u'reviews',
    u'romance', u'science_fiction']
```

Zunächst werden aus den Texten der drei Kategorien: `'news'`, `'religion'` und `'romance'` Sets gebildet, die dann im nächsten Schritt durch die Verwendung von Standardoperationen der Mengenlehre genauer betrachtet werden können.

```
1 news = set(brown.words(categories='news'))
2
3 religion = set(brown.words(categories='religion'))
4
5 romance = set(brown.words(categories='romance'))
```

Schnittmenge: Die Menge U derjenigen Elemente, die in allen Mengen die zur Schnittmenge gehören enthalten sind.

$$U = A \cap B$$

[8] Genauere Informationen zu Geschichte und Aufbau des Korpus können entweder über die Readme Funktion (`nltk.corpus.brown.readme()`) oder unter `http://www.hit.uib.no/icame/brown/bcm.html` eingesehen werden

In Python lässt sich die Schnittmenge durch den &-Operator erzeugen. Die Betrachtung der Schnittmenge kann dazu genutzt werden Gemeinsamkeiten im Sprachgebrauch aufzuzeigen, sozusagen den kleinsten, gemeinamen Nenner verschiedener Texte und Textgattungen. Dabei handelt es sich allerdings noch nicht um eine eigenständige Analyse, sondern vielmehr um die Erzeugung eines deskriptiven Überblicks und eine Vorbereitung auf weiterführende Untersuchungen. Da die Ausgabe eine recht lange Liste von Typen ist, werden hier nur die ersten zehn Elemente angezeigt.

```
1 list(romance & religion & news)[:10]
```

```
1 [u'Night',
2  u'colleges',
3  u'yellow',
4  u'four',
5  u'facilities',
6  u'woods',
7  u'Communist',
8  u'oldest',
9  u'hate',
10 u'assembled']
```

Vereinigungsmenge: Die Menge U derjenigen Elemente, die mindestens in einer der Teilmengen vorkommen.

$$U = A \cup B$$

Dies lässt sich entweder durch die Verwendung eines |-Operators oder durch die set-Methode .union() erzeugen. Eine Vereinigungsmenge zu erzeugen ist meistens nicht sehr aussagekräftig, wenn sie nicht mit anderen Mengenoperationen verknüpft wird. Die resultierende Vereinigungsmenge kann zum Beispiel eingesetzt werden, um die Typen mehrerer Texte oder Korpora zusammenzufassen und die Restmenge im Verhältnis zu anderen Textgruppen zu bestimmen. Die Restmenge ist folgendermaßen definiert:

Differenz: Bestimmt hinsichtlich zweier Mengen (A, B) ist die Restmenge U die Menge aller Elemente die nur in A vorkommt.

$$A \setminus B$$

Sie wird in Python entweder durch den --Operator oder die Methode `.difference()` erzeugt.

Der folgende Code bestimmt alle Typen, die zum Set `religion` gehören, aber nicht zu `news` und `romance`:

```
list(religion - (romance | news))[:10]
```

```
[u"Newbiggin's",
 u'impersonalized',
 u'mid-week',
 u'dissolution',
 u'dynasty',
 u'hath',
 u'writings',
 u'fortiori',
 u'self-reliant',
 u'Liverpool']
```

Die symmetrische Differenz stellt eine Erweiterung der Differenz dar. Damit kann eine Menge derjenigen Elemente produziert werden, die ausschließlich in den jeweiligen Teilmengen existieren. In Python wird eine symmetrische Differenz mittels des `^`-Operators oder durch die Methode `.symmetric_difference()` gebildet. Dadurch lassen sich diejenigen Elemente identifizieren, die am besten zwischen den einzelnen Texten unterscheiden.

Symmetrische Differenz: Bestimmt hinsichtlich zweier Mengen (A, B), ist dies die Menge U aller Elemente, die nicht in der Schnittmenge $A \cap B$ enthalten sind.

$$A \triangle B$$

```
list(religion ^ (romance ^ news))[:10]
```

```
[u'stock',
 u'raining',
 u'belligerence',
 u'divinely',
 u'impersonalized',
 u'mid-week',
 u'sunbonnet',
 u'Elevated',
 u'narcotic',
 u'Pfc.']
```

5.2.4 Typen-Token Relation

Letztlich stellt die Dichotomie von Token und Typ auch erst die Möglichkeit bereit nach der Verknüpfung dieser beiden Elemente zu fragen. In gewisser Weise handelt es sich hier um das sprachtheoretische Gegenstück zur soziologischen Unterscheidung von Handlung und Struktur, wie sie im Vergleich von individuums-zentrierten und gesellschafts-orientierten Ansätzen beschrieben wurde. Dementsprechend wird auch hier das semiotische Dilemma deutlich. Die Typen einer Sprache sind unabdingbare Voraussetzung eines jeglichen Auftretens von Token, gleichzeitig sind nur die Token als empirische Ereignisse fassbar und wirkmächtig. Dieses Problem kann jedoch durch einen wechselseitigen Bezug dieser zwei Ebenen in Form eines zeitlich ausgedehnten Prozesses überwunden werden. Die vorhandenen Typen bilden die Grundlage jeglicher Verwendung eines Zeichensystems, d.h. sie Begrenzen den Raum der möglichen Äußerungen. Zugleich können Typen und ihre Beziehung zu anderen Typen im Rahmen ihrer konkreten Verwendung bestätigt oder verändert werden. Dieser wechselseitige Bezug macht Sprache zu einem Phänomen, dass sich trotz ständiger Veränderungen als so stabil erweist, dass es die Grundlage sozialer Koordination und Wissensverwaltung darstellen kann.

Somit erlaubt uns diese Unterscheidung auch eine Präzision der bisher verwendeten Begrifflichkeiten. Soziale Symbole sind demnach Typen, da es sich bei ihnen um sozial standardisierte Zeichen handelt, die relativ unabhängig von ihrer Verwendung in konkreten Handlungen existieren.

Grundsätzlich treten sie empirisch jedoch nur als Token in Erscheinung, d.h. als ein spezifischer, kontextabhängiger Gebrauch sozial definierter Symbole. Die symbolischen Ordnungen sind diejenigen Muster der Verwendung, welche sich im Prozess des Symbolgebrauchs herausbilden. Sie stellen somit ein Phänomen dar, welches sowohl eine Funktion der zur Verfügung stehenden Typen als auch der verwendeten Token ist. Betrachtet man symbolische Ordnungen unter dem Aspekt einer Auswahl aus der Menge möglicher Typen, so kann man die beobachteten Muster als Indikatoren für die *Bedeutung* der Symbole auffassen. Nimmt man hingegen vor allem die konkrete Verwendung als Token in den Blick, können Regelmäßigkeiten als *Wissen* interpretiert werden. Diese Auffassungen schließen sich nicht gegenseitig aus, da auch nicht davon auszugehen wäre, dass die Bedeutung von Worten und das damit zum Ausdruck gebrachte Wissen voneinander unabhängig wären. Hier wird vielmehr angenommen, dass es sich um ein Kontinuum handelt, welches je nach gewünschter Auflösung andere Verfahren und Daten erfordert.

Das Verhältnis von Typen und Token kann auch als ein Merkmal der Stilistik interpretiert werden. Betrachtet man innerhalb eines Textes das Verhältnis der möglichen Typen zu den realisierten Token, so lässt sich dies als ein Maß für die Variation des Wortschatzes und in gewisser Weise auch für die Komplexität des Textes nutzen. Dieses Maß wird als *Type-Token-Ratio* bezeichnet und wurde vor allem zur Abschätzung und Messung sprachlicher Ausdrucksfähigkeit eingesetzt (Templin 1957). Eine genauere Auseinandersetzung mit dieser Maßzahl und der Vergleich mit anderen Komplexitätsmaßen findet sich im Abschnitt 5.7.

5.3 Textdaten

Die Verbreitung neuer Medien und Kommunikationstechnologien hat auch zu einem enormen Zuwachs an Daten in Textform geführt. Bevor die Möglichkeiten der Verarbeitung dieser Daten näher erläutert werden können, bietet es sich an einen Blick auf mögliche Quellen und unterschiedliche Formen solcher Datensätze zu werfen.

Eine nach bestimmten Regeln geordnete und nach Auswahlkriterien zusammengestellte Sammlung von Texten wird in den Sprachwissenschaften als *Korpus* bezeichnet. Verfügt dieser Korpus zudem noch über eine Reihe von Metadaten, wie zum Beispiel die Klassifikation von Texten oder die grammatikalischen Formen der verwendeten Wörtern, so spricht man von einem *annotierten Korpus*. Demzufolge sind Texte die in Markup-

Sprachen wie HTML verfasst wurden immer als annotiert aufzufassen. Allerdings dient diese spezifische Annotation fast ausschließlich der Verbesserung der Lesbarkeit für menschliche Leser und ist daher für die computergestützte Analyse von Texten oft nicht relevant.

Korpora lassen sich entsprechend ihres sozialwissenschaftlichen Verwendungszwecks grob in zwei Kategorien unterteilen. Zum einen können Korpora Daten im herkömmlichen Sinne des Wortes sein, d.h. sie dienen der Analyse spezifischer Sachverhalte und werden als eine *Repräsentation der empirischen Wirklichkeit* aufgefasst. Andererseits kann ein bestimmter Korpus auch als *lexikalische Ressourcen* zur Analyse einer anderen Sammlung von Texten eingesetzt werden. Zum Beispiel kann ein Korpus der Informationen über die Phonetik der Wörter einer ausgewählten Sprache bereitstellt eingesetzt werden um in einem empirisch zu betrachtenden Korpus die Silbentrennung durchzuführen. Die Verwendung als lexikalische Ressource ist gängige Praxis in der Computerlinguistik, weswegen viele Standardprogramme (z.B. NLTK) für diese Zwecke eine Reihe von annotierten Korpora beinhalten. Für die Sozialwissenschaften ist dies jedoch eine ungewohnte Vorgehensweise, da Datensätze hier meistens isoliert betrachtet werden. Grundsätzlich ergibt sich hier die praktische Herausforderung mehrere Datenquellen – mit unterschiedlichen Formaten, Zielsetzungen, Aufbereitungen, etc. – zu verwalten und bei Bedarf in Einklang bringen zu können.

Diese größeren Einstiegshürden zu überwinden scheint jedoch grundsätzlich lohnenswert. Schließlich erlaubt uns die quantitative Analyse von Texten bis dato unbekannte Einblicke in die Natur sozialer Symbole und die daran anschließende Operationalisierung von so zentralen Begrifflichkeiten wie Kultur und Wissen. Mit der immer stärker werdenden Integration digitaler Medien in die alltägliche Kommunikation und dem immer einfacher werdenden Zugriff auf solche Daten, ergeben sich neue Forschungsfelder und letztlich die Möglichkeit die soziale Interaktion quasi in Echtzeit zu betrachten. Auch die praktische Bedeutung in technischen Anwendungsfeldern (von der Marktforschung bis hin zu Expertensystemen) sind nicht von der Hand zu weisen.

5.3.1 Verwaltung von Textdaten (Korpusmanagement)

Die Probleme im Umgang mit verschiedenen Datensätzen wird noch dadurch verstärkt, dass Textdaten im Vergleich zu Umfragedaten eine höhere Komplexität aufweisen. Zunächst einmal kommen Textdaten in den meis-

ten Fällen nicht in einer Form vor, die sie unmittelbar der quantitativen Analyse zugänglich macht. Transformation von Textdaten haben jedoch oft das Problem, dass sie ohne externe Informationen nicht mehr umkehrbar sind. Zugleich setzt ihre Umformung andere Techniken voraus, als diejenigen welche zur Transformation von Datentabellen eingesetzt werden können, da Wörter in Texten häufig sequentiell und hierarchisch geordnet sind. Auch in praktischer Hinsicht bringen textuelle Daten eine Reihe von neuen Herausforderungen mit sich. Dies gilt zum Beispiel für die bereits angesprochenen Kodierungsprobleme, aber auch für den generellen Speicherbedarf, der den Bedarf von rein numerischen Daten meist um einige Größenordnungen übertrifft.

Daten werden in den Sozialwissenschaften meist als Tabellen repräsentiert. In einer solchen Darstellungsform enthält die *Reihenfolge* in der die Variablen auftreten keine relevante Information. Die Reihenfolge von Zeichen in einer Zeichenkette wird hingegen durch die Grammatik bestimmt, die zu dieser spezifischen Sprache gehört. Werden die Token eines Textes als Variablen zur Beschreibung eines Textes verwendet, bezeichnet man die resultierende, tabellarische Darstellungsweise je nach ihrer Ausgestaltung als *Bag-of-Words*, bzw. *Term-Document* oder *Document-Term Matrix*. Wie bereits auf Seite 105 ausgeführt, ist eine solche Repräsentation die Grundlage für eine Reihe von quantitativen Analysen. Allerdings ist sie auch unumkehrbar, weswegen sich eine solche Darstellung nicht für die Verwaltung und die langfristige Arbeit mit einem Korpus eignet.

In einer wenig strikten Fassung können Tabellen jedoch durchaus genutzt werden um Korpora zu verwalten. In einer solchen Repräsentation beschreibt jede Zeile einen spezifischen Text, einschließlich dessen Token und Metainformationen. Schematisch würde ein solcher Korpus wie folgt aussehen:

	Roher Text	Liste der Token	Autorenschaft	Datum	etc.
Text 1					
...					
Text n					

Tabelle 5.2: Schematische Darstellung eines tabellarischen Text-Korpus.

Wird für die Verwaltung und Speicherung eines solchen Korpus ein Datenbanksystem (z.B.: SQL) verwendet, so können auch relativ große Textmengen verwaltet werden. Zudem ermöglicht ein solches Vorgehen auch die gezielte Auswahl von Texten entsprechend der vorhandenen Metadaten. Da der Bedarf an Arbeitsspeicher im Falle von Textdaten schnell problematisch werden kann, ist diese stückweise Bearbeitung von Texten oft eine praktische Notwendigkeit.

Für die praktische Umsetzung bietet sich hier ebenfalls das bereits erwähnte `pandas` Modul an. Es lässt sich zudem relativ einfach mit anderen Datenverwaltungssystemen (z.B. SQL, HDFS, etc.) koppeln und verfügt über eine Reihe von Tools zur Visualisierung von Daten, sowie der Kalkulation deskriptiver Statistiken. Der hauptsächliche Vorteil liegt jedoch in der direkten Unterstützung primitiver Python-Typen. Konkret ermöglicht uns dies tokeniserte Texte als Listen von Token abzuspeichern und auch als solche wieder zu laden. In anderen Systemen, wie SQL oder XML ist es oft aufwendiger eine Kompatibilität mit den Datentypen der verwendeten Analyseprogramme zu gewährleisten.

Neben der Sequentialität von Texten müssen Korpora unter Umständen auch den in Textdaten enthaltenen *Hierarchien* gerecht werden. In natürlichen Sprachen gliedern sich Texte oft in Absätze, die sich wieder in Sätze und diese wiederum in einzelne Worte unterteilen lassen. Zusätzlich erzeugt auch die Grammatik weitere hierarchische Unterteilungen in einem Text, indem zum Beispiel Haupt- und Nebensätze voneinander getrennt werden. Die Grammatik selbst ist ebenfalls hierarchisch aufgebaut, wie die Unterscheidung der Wortstämme und deren je nach Fall verschieden gearteten Ableitungen demonstriert. Diese Hierarchien sind in tabellarischer Form meistens nicht mehr sinnvoll darzustellen, zumindest nicht ohne die Erzeugung größerer Redundanzen.

Das grundsätzliche Problem kann am Beispiel des *NP-Chunkings* demonstriert werden. Ziel dieses Verfahren ist es sogenannte „noun phrases“, d.h. feste Verknüpfungen von Artikeln, Adjektiven und Substantiven, zu identifizieren. Diese sind hierarchisch auf verschiedenen Ebenen des Satzes angeordnet und können auch ineinander verschachtelt sein, deswegen werden sie oft als Baum-Graphen dargestellt. Aus dem Satz „Der kleine, gelbe Hund bellte die Katze an“ lassen sich beispielsweise zwei noun phrases extrahieren: „Der kleine, gelbe Hund“ und „die Katze“. Diese Technik und deren korrekte Anwendung werden im Rahmen der Netzwerk-Text-Analyse (siehe Abschnitt 6.4) eingängiger behandelt.

```
import nltk
sentence = [('Der', 'DT'), ('kleine', 'JJ'), ('gelbe', 'JJ'),
            ('Hund', 'NN'), ('bellte', 'VBD'), ('die', 'DT'),
            ('Katze', 'NN'), ('an', 'IN')]

grammar = 'NP: {<DT>?<JJ>*<NN>}'
cp = nltk.RegexpParser(grammar)
result = cp.parse(sentence)

result.pprint()
```

```
(S
  (NP Der/DT kleine/JJ gelbe/JJ Hund/NN)
  bellte/VBD
  (NP die/DT Katze/NN)
  an/IN)
```

Für die Darstellung hierarchischer Daten bieten sich vor allem XML (eXtensive Markup Language), sowie deren Derivate an. Grundsätzlich eignen sich alle Datenschemata, die Daten als Baumstrukturen auffassen. Als *Baum* (Tree) wird ein vollständig verbundener Graph bezeichnet, der von einem Knoten ausgeht (Wurzel oder root) und der keinen Zyklus beinhaltet, sprich die Kantenzüge des Graphen dürfen es nicht erlauben bei einer stetigen Bewegung vom Ausgangsknoten weg wieder bei einem bereits passierten Knoten zu landen. In XML wird ein solcher Baum durch die hierarchische Staffelung von Elementen erzeugt, die jeweils durch Tag-Paar begrenzt sind. Der tabellarische Beispielkorpus würde in dieser Darstellung wie folgt aussehen:

```
<root>
  <Text>
     <raw>Hier steht der rohe Text.</raw>
     <Tokens>
       <token>Hier</token>
       <token>steht</token>
       <token>...</token>
     </Tokens>
  </Text>
```

```
10 </root>
```

5.3.2 Textdaten als Forschungsobjekte

Zu den zentrale Annahme dieser Arbeit gehört der objektive Charakter von sozialen Symbolen, was zum Ausdruck bringen soll, dass sie bestimmten universellen Gesetzmäßigkeiten folgen. Die Erforschung dieser Gesetzmäßigkeiten ist jedoch nicht Gegenstand der Soziologie selbst, sondern fällt in die Domänen der Linguistik, Informationstheorie und ähnlicher Fachbereiche. Dennoch bildet der objektive Charakter von Zeichen die Vorbedingung für eine soziologische Analyse des Gebrauchs, der Strukturierung und der Relation von Symbolen und symbolischen Ordnungen. Bezogen auf Texte als Forschungsgegenstand bedeutet dies, dass prinzipiell alle in natürlichen Sprachen verfasste Texte allen Verfahren der quantitativen Textanalyse zugänglich sind.

In der Praxis sind Texte jedoch oft sehr stark durch soziale Regeln der richtigen Konstruktion und Argumentation gekennzeichnet. Ein Umstand, der sie überhaupt erst zu Gegenständen von soziologischem Interesse werden lässt. Gleichzeitig hat dies auch zur Folge, dass sich nicht jeder Text in gleichem Maße für die Erforschung eines bestimmten Phänomens und den Einsatz eines bestimmten Verfahrens eignet. Ein Extrembeispiel wäre die Analyse von Abstracts wissenschaftlicher Publikation hinsichtlich deren Sentiments, d.h. mit welcher Konnotation auf ein bestimmtes Thema Bezug genommen wird. Eine solche Einfärbung mag zwar grundsätzlich vorhanden sein, jedoch machen es die Regeln dieses spezifischen Diskurses relativ schwer ein sprachliches Muster zu isolieren. Dagegen spricht einerseits ein implizites Verbot direkter, wertender Aussagen, andererseits die relative Kürze und Zielsetzung eines wissenschaftlichen Abstracts. Grundsätzlich bedarf es daher eines gewisses Vorwissen bezüglich der sozialen Konstruktionsbedingungen und den sprachlichen Eigenheiten der zu untersuchenden Texte um abschätzen zu können welche Verfahren geeignet sind und ob eine bestimmte Fragestellung mit diesen Texten untersucht werden kann.

Im Folgenden werden der Aufbau und die Besonderheiten des hier hauptsächlich verwendeten Beispielskorpus vorgestellt. Besonderes Augenmerk liegt dabei auch auf den Fragestellungen, hinsichtlich derer die Zusammenstellung des jeweiligen Korpus erfolgte und wie sich diese gegebenenfalls erweitern ließe. Da das Hauptanliegen dieser Arbeit in der Entwicklung einer allgemeinen Theorie und Methodologie sozialer Sym-

bole liegt, ist die Analyse diese Korpus und der damit verbundenen symbolischen Ordnungen an einigen Stellen nur exemplarischer Natur. Die Interpretation vor dem Hintergrund anderer soziologischer Theorien oder die empirischen Implikationen für bestimmte Forschungsbereiche können daher ebenfalls nur angeschnitten werden.

Ein Großteil der folgenden Analysen bezieht sich auf einen Korpus von englischsprachigen Abstracts wissenschaftlicher Artikel: `SozAbst.pkl`. Der Korpus besteht aus den Abstracts von sozialwissenschaftlichen Artikeln aus den Jahren 1992-2010. Ziel der Zusammenstellung dieses Korpus war die Analyse der Wissens- und Diskursstrukturen der Soziologie, insbesondere hinsichtlich etwaiger Unterschiede, die sich durch eine Einbettung in nationalen Wissenschaftsfelder ergeben könnten. Für diesen Vergleich wurden Deutschland, Großbritannien und die USA gewählt. Die Auswahl soziologischer Texte, ebenso wie die der Länder hat dabei eine Reihe von Gründen. Zunächst einmal bieten sich die symbolischen Ordnungen der Soziologie für eine Demonstration und Kritik der hier diskutierten methodologischen Prinzipien in besonderem Maße an, da von den sozialwissenschaftlich interessierten Leser, die das Zielpublikum dieser Arbeit stellen, eine ausreichende Vertrautheit mit den Symbolen und Argumentationsregeln erwartet werden kann. Den Vergleich zum US-amerikanischen Diskurs herzustellen ist vor allem in wissenschaftssoziologischer Hinsicht von Interesse, da die USA in vielen Fällen den Dreh- und Angelpunkt internationaler Forschungsleistung stellen und die Struktur dieses Austausches maßgeblich beeinflussen (vgl. Das et al. 2009; Heiberger und Riebling 2015).

Die Entscheidung Abstracts anstelle des ganzen Textes zu verwenden ist dem Umstand geschuldet, dass diese immer auch in englischer Sprache vorliegen, was eine günstige Ausgangslage für einen Vergleich darstellt. Ein Vergleich unterschiedlicher Sprachen würde eine Vielzahl von methodischen Vorgehensweisen komplett ausschließen. Daraus kann sich jedoch ein weiteres Problem ergeben. Übersetzungen können potentiell zu einer Verzerrung der Daten führen. Nicht zuletzt deswegen, wurde auch eine Stichprobe von Abstracts aus der soziologischen Tradition Großbritanniens zusammengestellt. Damit wird sowohl die Untersuchung *soziologischer Wissenskulturen* gehaltvoller, als auch eine Vergleichsbasis geschaffen, die helfen kann rein sprachliche Unterschiede von kulturellen zu trennen. Ein weiterer Grund für die Wahl von Abstracts liegt in deren relativ geradlinigen Ausführung und klaren Zielsetzung. Gerade wissenschaftliche Texte weisen je nach Art des Kapitels (z.B.: Methoden, Stand

der Forschung, etc.) andere Bezugspunkte zum Gesamtdiskurs auf (vgl. Leydesdorff 1997). Ein Modell auf Basis des ganzen Textes würde daher tendenziell unterschiedliche wissenschaftliche Diskurse erfassen, die weit über den Bereich der Soziologie oder deren spezifischer Zeitschriften hinaus gehen würde. Die Konzentration auf Abstracts erscheint hier zielführender. Zumal diese explizit dafür gedacht sind die zentralen Aussagen des Textes einem weiteren Publikum zu signalisieren.

Die Zielsetzung des Korpus war dann auch ausschlaggebend für die konkrete Auswahl der Abstracts. Um einen Korpus zu erhalten der die grundlegende, symbolische Ordnung des jeweiligen Soziologiediskurses abdeckt, mussten Journals ausgewählt werden die hinsichtlich ihrer Position, Publikum und Ausrichtung vergleichbar waren. Voruntersuchungen hatten ergeben, dass nur hochrangige Zeitschriften, die auf den Fachbereich in seiner Gesamtheit ausgerichtet sind, dafür in Frage kommen konnten. Journale mit einem niedrigen Impact-Factor sind hingegen oft zu spezifisch auf bestimmte Themenbereiche ausgerichtet, als das eine größere Diskursstruktur damit beschrieben werden könnte. Hinzu kommt natürlich auch die Leuchtturm-Wirkung von hochrangigen Journals, die es erwarten lässt, dass prominente Theorien und Argumente früher oder später dort auftauchen werden. Basierend auf diesen Überlegungen wurden für jedes der drei Länder jeweils drei Zeitschriften ausgewählt, welche die Journals mit dem höchsten Impact-Factor im jeweiligen Land umfassen. Die folgende Liste gibt diese Journale wieder, sowie die Anzahl n der verwertbaren Abstracts pro Zeitschrift (Gesamt: $N = 5150$). In den Klammern wird die jeweilige ISO-4 Abkürzung des Journaltitels angegeben, die auch im Korpus genutzt wird um den Ursprung des jeweiligen Abstracts wiederzugeben.

- American Journal of Sociology (`Am. J. Sociol.`); $n = 650$.
- Annual Review of Sociology (`Annu. Rev. Sociol.`); $n = 411$.
- American Sociological Review (`Am. Sociol. Rev.`); $n = 842$.
- Berliner Journal für Soziologie (`Berliner J. Soz.`); $n = 410$.
- Kölner Zeitschrift für Soziologie und Sozialforschung (`Koln. Z. Soziol. Sozialpsych.`); $n = 574$.
- Zeitschrift für Soziologie (`Z. Soziol.`); $n = 463$.
- The British Journal of Sociology (`Br. J. Sociol.`); $n = 532$.

- Sociology (`Sociol.-J. Brit. Sociol. Assoc.`); $n = 777$.
- Work, Employment and Society (`Work Employ. Soc.`); $n = 491$.

Im weiteren Verlauf dieses Kapitels wird hauptsächlich die Rohfassung dieses Korpus herangezogen, um die grundlegenden Techniken der Textaufbereitung zu demonstrieren. Der daraus resultierende, aufbereitete Korpus wird im Kapitel 6 für die inhaltlichen Analysen und die Erläuterung des methodischen Vorgehens genutzt.

Weil sich dieser Korpus jedoch nicht für alle Formen der hier zu diskutierenden Analysen eignet, wird im Folgenden auch auf eine Reihe von bekannteren Korpora aus dem Bereich des Natural Language Processings zurückgegriffen. Diese Korpora haben sowohl den Vorteil, dass sie bereits für bestimmte Analysen aufbereitet sind, als auch den, dass ihre Eigenschaften sehr gut erforscht und dokumentiert sind. Eine kurze Beschreibung des jeweiligen Korpus findet sich an den entsprechenden Stellen im Text.

5.3.3 Textdaten als Forschungswerkzeuge

Neben dem Einsatz als Gegenstände der Forschung sind aufbereitete Korpora oft auch notwendig um andere Textkorpora erforschen zu können. Meistens geschieht dies in der Form eines Lexikons. Ein solcher Korpus weist bestimmten Token spezifische Eigenschaften zu. Eine Vielzahl von Eigenschaften ist dabei denkbar, zu den gängigsten gehören grammatikalische Wortformen, Silbentrennungsregeln, relative Worthäufigkeiten einer bestimmten Sprache und Sentimente. Solche Korpora sind nicht nur Voraussetzung quantitativer Textanalysen, sondern oft auch ein willkommenes Nebenprodukt oder im Falle der Corpuslinguistik sogar das Endprodukt systematischer Forschungen. Ihre Bedeutung für die weitere Forschung macht eine öffentliche Bereitstellung dieser Forschungsressourcen zu einem starken, forschungsethischen Imperativ. Im Bereich der Linguistik ist die Bereitstellung von Korpora zu Forschungszwecken daher die disziplinäre Norm. Allerdings ist eine bloße Publikation der Rohdaten oft nicht sonderlich hilfreich. Vielmehr bedarf es einer sinnvollen Strukturierung der Daten und einer ausreichenden Dokumentation des Korpusinhalts.

Solche Lexika und entsprechend aufbereitete Korpora stellen eine notwendige Vorausbedingung für eine sozialwissenschaftliche Erforschung symbolischer Ordnungen dar. Insbesondere im Bereich von Fachsprachen

und Soziolekten sind jedoch kaum ausreichende Ressourcen vorhanden. Das hat zur Folge, dass Texte, die in solchen Sprachen verfasst wurden oft nicht entsprechend aufbereitet werden können. Dieses Problem fand lange Zeit keine große Beachtung, da sich die Schriftsprache meistens an der dominanten Standardsprache eines Kulturkreises ausrichtete und geschriebener Text die primäre Datenquelle der sozialwissenschaftlichen Textanalyse war. Dieser Umstand war nicht zuletzt dem ungleichen Zugang zu Wissen und Schreibwerkzeugen geschuldet. Im Zuge der Digitalisierung hat sich dies jedoch ins Gegenteil verkehrt. Ein Großteil der menschlichen Online-Interaktionen wird durch Schriftverkehr vermittelt, der aufgrund sozialer Prozesse und technischer Rahmenbedingungen eine Vielzahl von sehr eigenen Kommunikationsformen und Soziolekten hervorgebracht hat.

Allerdings sind, auch schon allein aus urheberrechtlichen Gründen, nicht für alle Sprachen die gleichen Ressourcen vorhanden. Im Allgemeinen wird das Englische am weitreichendsten unterstützt. Gerade was die deutsche Sprache angeht muss jedoch oft auf externe Lösungen zurück gegriffen werden. Zum Beispiel gibt es für das automatische Durchführen eines NP-Chunkings bisher nur ein einziges nicht-kommerzielles Softwarepaket, welches auch die deutsche Sprache umfasst: den von Helmut Schmid an der Universität Stuttgart entwickelten TreeTagger.[9]

In gewisser Weise liegt hier eine der größten Hürden für die Adaption von Verfahren der quantitativen Textanalyse für sozialwissenschaftliche Zwecke. Ohne entsprechende Ressourcen, in Form von aufbereiteten Korpora, sprachspezifischen Algorithmen und Lexika, für die Erstellung neuer Verfahren und zum Einsatz in der Lehre, können eine Vielzahl soziologisch, relevanter Tatsachen nicht operationalisiert werden. Ebenso bleiben eine Reihe von Datenquellen dem Zugang verschlossen. Dies trifft in besonderem Maße auf eine Reihe von modernen Kommunikationsformen zu, wie z.B. den Einsatz vom Emoticons, die erst durch digitale Medien zugänglich wurden. Hier zeigt sich, dass das Problem der digital literacy nicht nur den einzelnen Forscher betrifft, sondern auch disziplinäre Auswirkungen hat. Zur Bereitstellung einer grundlegenden Infrastruktur an Daten und Forschungswerkzeugen sind vereinzelte Forscher schlicht nicht in der Lage. Erst ein relativ breit gestreutes Verständnis der Verfahren und technischen Details sowie der damit verbundenen Möglichkeiten

[9] Für dieses Programmpaket existieren eine Reihe von Python Wrappern, z.B.: `https://github.com/miotto/treetagger-python`.

der Analyse erlaubt die erfolgreiche Kooperation von Forschern, um den Datenschatz der Digitalisierung heben zu können.

5.4 Exkurs: Reguläre Ausdrücke

Reguläre Ausdrücke sind das wahrscheinlich mächtigste Werkzeug für die systematische Aufbereitung von maschinenlesbarem Text. Gleichzeitig können reguläre Ausdrücke äußerst komplex werden und sind berüchtigt dafür zu der Fehleinschätzung einzuladen, dass ein bestimmtes Problem einfach zu lösen wäre. In der Tech-Community wird dafür oft ein Zitat von Jamie Zawinski herangezogen, welches er wohl 1997 in der „alt.religion.emacs“ Usenetgruppe tätigte (vgl. Friedl 2006):

> Some people, when confronted with a problem, think “I know, I’ll use regular expression.” Now they have two problems.

Bei regulären Ausdrücken handelt es sich um eine Art Metasprache, in der jeder beliebige Zeichenkette über einem endlichen Alphabet spezifiziert werden kann. Entwickelt wurden sie von Stephen C. Kleene (1951), der sie als finite Automaten zur Beschreibung einer (potentiell unendlichen) Reihenfolge von Zuständen konzipierte. Als endliche Automaten bezeichnet man Algorithmen, die deterministisch verlaufen und dabei ohne „Gedächtnis“ auskommen, d.h. sie können nur eine endliche Anzahl von Zuständen annehmen.^[Eine genauere Betrachtung der Theorie endlicher Automaten würde an dieser Stelle zu weit führen. Eine hervoragende Einführung in dieses Thema findet sich bei Katrin Erk und Lutz Priese (2008). Diese Entdeckung war folgenreich für die theoretische Informatik und führte auch zu einer weitreichenden, praktischen Umsetzung in Form von Compilern für reguläre Ausdrücke, die sich heute in fast allen Betriebsystemen und Programmiersprachen finden lassen.

Unter der Bedingung das Σ das endliche Alphabet einer rechtslinearen Sprache ($\mathcal{L}_3$ in der Chomsky Hierarchie der formalen Sprachen) ist, lassen sich alle daraus konstruierbaren Sätze in regulären Ausdrücken spezifizieren. Mit rechtslinear ist gemeint, dass die Konstruktionsregeln für valide Sätze dieser Sprache nur ein Hinzufügen neuer Wörter ans Ende des bestehenden Satzes erlauben. Dies ist in der Grammatik natürlicher Sprachen sicher nicht der Fall, sehr wohl aber bei deren Repräsentation als Strings. Hierin liegt auch der Grund für die Eingangs erwähnte Einschränkung, die das Ersetzen von Zeichen in einem String (*slicing*) ver-

bietet. Gleichzeitig ist das Konkatenieren, d.h. das Zusammenfügen von Strings in der Reihenfolge von links nach rechts, eine erlaubte Operation.

In Python werden reguläre Ausdrücke über das Basismodul `re` bereitgestellt, dessen Dokumentation in der Python Standard Library Kapitel 7.2 eingesehen werden kann.[10] Darin sind Funktionen enthalten, die ähnliche Verfahren bereitstellen, wie die bereits vorgestellten Methoden zur Manipulation von Strings (siehe Abschnitt 5.1.1).

Ein regulärer Ausdruck wird in Python ebenfalls als String geschrieben. Allerdings empfiehlt es sich hier einen *raw string literal* zu verwenden. Ein solcher „wortwörtlicher" String wird definiert indem ihm ein kleines `r` vorangestellt wird. Die Verwendung eines solchen Strings ist zu empfehlen, da der Backslash Charakter \ in Python für das *Escaping* von Sonderzeichen verwendet wird. Escaping bedeutet, dass ein Backslash auch verwendet werden kann um die Bedeutung eines Sonderzeichens aufzuheben. Wollte man also eine Ausgabe des Ausdrucks `\n` haben, so müsste man den String folgendermaßen schreiben: `\\n`, da sonst ein Zeilenumbruch produziert werden würde. Dieses Problem hätte zur Folge, dass ein regulärer Ausdruck für einen Backslash – bei Verwendung eines normalen Python Strings – viermal wiederholt werden müsste (`\\\\`). Ein raw string literal umgeht dies, indem er alle Escape-Zeichen ignoriert, die nicht zur Beendigung einer String-Sequenz führen. Dies ist ausreichend, da in der Syntax regulärer Ausdrücke kein einzelner Backslash am Ende eines Strings vorkommen kann.

Betrachten wir der Einfachheit halber zunächst den Fall eines Alphabets von nur zwei Zeichen: $\Sigma = \{a, b\}$. In diesem Fall können wir eine Reihe von validen Sätzen bilden. Zum Beispiel $s_1 = \{aaaba\}$ oder $s_2 = \{bababa\}$. Gegeben Σ, sind $r_1 = \{a\}$ und $r_2 = \{ab\}$ Beispiele für zwei valide reguläre Ausdrücke. Bei der vorgeschriebenen Leserichtung von links nach rechts würde die Anwendung dieser regulären Ausdrücke auf diese Sätze bedeuten, dass r_1 das erste Zeichen in s_1 und das zweite in s_2 findet, r_2 würde hingegen die Stellen drei und vier in s_1 spezifizie-

[10] Grunsätzlich gibt es zwei Möglichkeiten in denen dieses Modul genutzt werden kann. Reguläre Ausdrücke können entweder in einer objekt-orientierten Schreibweise gebildet werden oder in einer funktionalen. Da sich diese Arbeit an ein Publikum in der wissenschaftlichen Praxis richtet, wird im Folgenden nur die funktionale Schreibweise verwendet. Diese entspricht eher den gängigen Praktiken im wissenschaftlichen Umgang mit Code und Programmiersprachenin, welche hauptsächlich durch die Verwendung prozeduraler Skripte sowie der Manipulation von Daten in einem Input-Output Schema gekennzeichnet ist.

ren sowie zwei und drei in s_2. Die Übersetzung dieses Beispiels in Python sieht folgendermaßen aus:

```
1 import re
2
3 s1 = 'aaaba'
4 s2 = 'bababa'
5
6 r1 = r'a'
7 r2 = r'ab'
```

Die einfachste Anwendung von regulären Ausdrücken ist das Auffinden des regulären Ausdrucks im String. Dafür stehen zwei Funktionen zur Verfügung, nämlich `re.match()` und `re.search()`. Bei `re.match()` wird vom Beginn des Strings nach rechts gehend geprüft ob der String dem regulären Ausdruck entspricht. In diesem Fall würde `r1` in `s1` gefunden werde. Alle anderen Kombinationen der regulären Ausdrücke mit den Strings würden jedoch keine Treffer erzeugen und dementsprechend nur den Wert `None` zurückgeben.

```
1 re.match(r1, s1)
```

```
1 <_sre.SRE_Match at 0x7f46276334a8>
```

```
1 type(re.match(r2, s1))
```

```
1 NoneType
```

Im Unterschied dazu sucht `re.search()` nach dem ersten Auftreten des regulären Ausdrucks im String, unabhängig von den Zeichen die diesem vorausgehen. Entsprechend produzieren im hier gewählten Beispiel alle regulären Ausdrücke einen Treffer in allen Strings.

Entspricht ein Teil oder der ganze String dem regulären Ausdruck, geben `re.match()` und `re.search()` jeweils ein `match` Objekt zurück. Dieses Objekt verfügt über eine Reihe von Methoden mit denen der durch den

regulären Ausdruck spezifzierte Teilstring zurückggeben werden kann (z.B.: `.group()`). Im formalen Beispiel wurde gesagt, dass der reguläre Ausdruck r_2 der zweiten und dritten Stelle des Satzes s_2 entspricht. Mit der Methode `.span()` können wir nach der Anfangs und Endposition des regulären Ausdruckes im String fragen und so die oben getroffene Aussage in Python überprüfen. Das Ergebnis ist auch nochmal eine Erinnerung daran, dass Indexierung in Python bei 0 beginnt und der Endwert nicht mit eingeschlossen wird. Somit entspricht die Angabe „von 1 bis unter 3“ den Positionen 2 und 3 im Beispiel.

```
1 match = re.search(r2, s2)
2 match.span()
```

```
1 (1, 3)
```

Neben diesen zwei Funktionen existieren noch weitere, die eine Reihe von wichtigen Aufgaben der Bearbeitung von Text abdecken. Die wichtigsten drei sollen an dieser Stelle kurz genannt werden, da sie auch in den weiteren Ausführungen eine Rolle spielen. Neben den aufgeführten Argumenten verfügen alle dieser Funktionen zusätzlich über ein optionales `flags=` Argument. Mit diesen kann die Evaluierung der regulären Ausdrücke gesteuert werden. Zum Beispiel sorgt die Flagge `re.IGNORECASE` dafür, dass der Unterschied zwischen Groß- und Kleinschreibung ignoriert wird. Auch hier muss für einen ausführlicheren Überblick auf die Python Standard Library (7.2.2) verwiesen werden.

- `re.findall(`*pattern*`,` *string*`)`: Findet alle nicht-überlappenden Vorkommnisse von des regulären Ausdrucks *pattern* in *string*. Gibt eine Liste der Treffer oder eine leere Liste zurück.
- `re.split(`*pattern*`,` *string*`,` `,` *maxsplit=0*`)`: Trennt *string* entsprechend dem regulären Ausdruck in *pattern* auf. Gibt eine Liste zurück.
- `re.sub(`*pattern*`,` *repl*`,` *string*`,` *count=0*`)`: Ersetzt jeden Teil von *string*, der *pattern* entspricht, mit *repl*. *repl* kann selbst reguläre Ausdrücke enthalten, insbesondere um Gruppen zu referenzieren, die in *pattern* spezifiziert wurden.

5.4.1 Reguläre Ausdrücke in der Praxis

Bisher wurden nur reguläre Ausdrücke angesprochen die einzelnen Buchstaben entsprechen. Darüber hinaus enthalten sie alle gültigen Sonderzeichen, wie zum Beispiel den Tabulator `\t` oder den Zeilenumbruch `\n`. Desweiteren gibt es eine Reihe von regulären Ausdrücken, die als Platzhalter für Kategorien von Ausdrücken dienen. Diese können auch explizit definiert werden, indem sie mit eckigen Klammer umschlossen werden. So lassen sich Ausdrücke zu Gruppen zusammenfassen. Allerdings spezifizieren diese Gruppen trotzdem immer nur eine Position in der Zeichenkette. D.h. der reguläre Ausdruck `[abc]` würde das erste Auftreten eines beliebigen Mitglieds dieser Gruppe spezifizieren. Eine Verneinung dieser Gruppen kann durch das Einsetzen eines Zirkumflex' `^` als dem erstem Zeichen der Gruppe erreicht werden. Ein einfacher Bindestrich `-` zwischen zwei Elementen wird innerhalb einer Gruppe als „von–bis" interpretiert. Demzufolge deckt der Ausdruck `[0-9]` alle natürlichen Zahlen von Null bis einschließlich Neun ab. Einige der wichtigeren Platzhalter finden sich in Tabelle 5.3.

Syntax	Gruppe	Bedeutung
`.`	`[^\n]`	Jegliches Zeichen außer dem Zeilenumbruch.
`\d`	`[0-9]`	Ganze Zahlen.
`\D`	`[^0-9]`	Alles was keine Zahl ist.
`\s`	`[ \t\n\r\f\v]`	Alles was ein Leerzeichen ist.
`\S`	`[^ \t\n\r\f\v]`	Alles was kein Leerzeichen ist.
`\w`	`[a-zA-Z0-9_]`	Alphanumerisches Zeichen und der Unterstrich.
`\W`	`[^a-zA-Z0-9_]`	Kein alphanumerisches Zeichen oder der Unterstrich.

Tabelle 5.3: Gruppen und Klassen von regulären Ausdrücken.

Eine Reihe anderer Sonderzeichen verdienen ebenfalls eine Erwähnung, nämlich `\b` bzw. `\B` oder `^` und `$`. Diese erfordern jedoch die Einführung eines weiteren Konzepts der Theorie formaler Sprachen. Der *leere String*, oft als Epsilon ϵ (für „empty") oder Lambda λ (für „leer") angegeben, bezeichnet einen String der Länge 0. Im Prinzip entspricht dies der leeren Menge der Mengenlehre und erfüllt eine ähnliche Funktion im Bereich formaler Sprachen. Es macht zum Beispiel die Abgrenzung von einzelnen Elementen eines Strings überhaupt erst möglich. In Python wird der leere String folgendermaßen angegeben: `''` oder `""`. Die regulären Ausdrücke `\b` und dessen Gegenstück (`\B`) sind definiert als der leere String am Beginn und Ende eines Wortes, d.h. die Grenze zwischen einem alpha-

numerischen Zeichen und einem Leerzeichen. Außerhalb einer Gruppe ([]) spezifiziert das Zirkumflex den leeren String am Anfang eines beliebig langen Strings. Dessen Inverse bildet das Dollarzeichen, welches das Ende eines Strings kennzeichnet.

Diese Klassen und Kategorien von Ausdrücken lassen sich als *Wildcards* verwenden. Gemeint ist damit die Suche und das Spezifizieren von Text, der unterschiedliche Schreibweisen aufweist. Dies ist vor allem in Bereichen des Text Mining von herausragender Bedeutung, da es die Feststellung von in Texten kodierten Daten erlaubt, die einem bestimmten, grundlegenden Schema folgen, aber unterschiedliche Werte aufweisen können. Beispiele dafür wären Datumsangaben, Postleitzahlen, unterschiedliche Schreibweisen eines Wortes und dergleichen. Der folgende Code kann helfen dies zu verdeutlichen:

```
1 text = 'Sie kam am 29.01.1956 zum ersten Mal zu Besuch.
      Nicht am 9. Mai oder am 04.08.82, aber am 28.11.1982.'
2
3 re.findall(r'(\d\d)\.(\d\d)\.(\d\d\d\d)', text)
```

```
1 [('29', '01', '1956'), ('28', '11', '1982')]
```

In diesem Beispiel wird zudem der Einsatz von Paranthesen demonstriert. Mit diesen lässt sich der Anwendungsbereich der regulären Ausdrücke sehr viel feiner steuern. Sie erlauben die Festlegung von sehr detailierten Regeln wie mit Teilbereichen des Ausdrucks umgegangen werden soll.[11] Auf diese Art lassen sich Texte *parsen*, d.h. ein Text kann aufgrund fester Regeln, die als reguläre Ausdrücke formuliert werden können, in eine Datenstruktur oder eine Reihe von Anweisungen überführt werden, die von einer Maschine verstanden werden können. Allgemeiner ausgedrückt: mittels einer Grammatik, also den Regeln der validen Satzkonstruktion, lässt sich die Semantik eines Textes erschließen. Eine genauere Betrachtung dieser Techniken und ihrer Implikationen würde jedoch ein eigenes Buch rechtfertigen. Es ist an dieser Stelle ausreichend darauf hinzuweisen, dass reguläre Ausdrücke prinzipiell dazu in der Lage sind die Zusammensetzung eines Strings abstrakt zu beschreiben. Genau darin

[11] siehe auch die Python Standard Library 7.2.1 für eine ausführlichere Beschreibung der Syntax

liegt ihr enormer Nutzen für die Aufbereitung von Texten und die Extraktion standardisierter Daten aus eben diesen.

Die im Beispiel beschriebene Vorgehensweise ist aber immer noch relativ unflexibel, da sie zwar mit unterschiedlichen Zahlenwerten umgehen kann, aber auf ein bestimmtes Schema der Datumsangabe festgelegt ist. Um diese Limitationen zu umgehen, können Operatoren genutzt werden, welche die Konstruktion komplexerer Ausdrücke ermöglichen.

Reguläre Ausdrücke können mittels dreier, grundlegender Operatoren miteinander verknüpft werden (vgl. Kleene 1951: 46ff). Der einfachste Operator, die *Verknüpfung* (oder Konkatenation), wurde implizit bereits angewendet. Der reguläre Ausdruck $r_2 = ab$ stellt demzufolge die Verknüpfung der Ausdrücke a und b dar. In Python wird dies einfach durch die Reihenfolge der regulären Ausdrück im rohen String erzeugt. Zu beachten ist hier lediglich, dass Gruppen immer nur einen regulären Ausdruck darstellen und deswegen nur eine Position im String einnehmen. Desweiteren ist die *Disjunktion* regulärer Ausdrücke erlaubt: $r_1 \vee r_2$ ist dann erfüllt, wenn an dieser Position im String entweder r_1 oder r_2 gegeben sind. Die Syntax des `re` Moduls sieht für eine Disjunktion den `|`-Operator vor.[12] Der dritte Operator ist die sogennante kleensche Hülle, bzw. *Kleenes Stern*. Der Ausdruck r^* besagt, dass der reguläre Ausdruck r beginnend an dieser Position gar nicht oder beliebig oft vorkommen kann.

> If E and F are events already constructed, then by E^*F we shall mean the event which consists of zero or more consecutive occurrences of E preceded by one of F. That is, E^*F can occur whenever
>
> $$\overbrace{E \cdots E}^{n \text{ times}} F$$
>
> occurs for some $n \geq 0$. (ebd.: 49)

Der Einschluss des Nicht-Auftretens ist formal nur deswegen möglich, weil die Definition regulärer Ausdrücke den leeren String beinhaltet. Pythons Syntax verwendet hierfür das Asterisk Zeichen (`*`). Aufbauend auf der Idee des Stern-Operators gibt es eine Reihe von syntaktischen Sonderformen, die eine genauere Spezifikation der Wiederholungen erlauben.

[12] Implementationen von regulären Ausdrücken sind im Großen und Ganzen relativ einheitlich. Jedoch kann es kleinere Unterschiede geben, weswegen sich die hier getroffenen Feststellungen explizit nur auf das `re` Modul zum Zeitpunkt der Fertigstellung dieses Textes beziehen.

Tabelle 5.4: Operatoren für reguläre Ausdrücke.

Syntax	Bedeutung
`*`	0 oder mehr Wiederholungen der vorangegangene RegEx.
`+`	1 oder mehr Wiederholungen der vorangegangene RegEx.
`?`	0 oder 1 Wiederholungen der vorangegangene RegEx.
`{n}`	Genau `n` Wiederholungen.
`{n,m}`	Von `n` bis einschließlich `m` Wiederholungen.
`{n,}`	Mindestens `n` Wiederholungen.
`{,m}`	Maximal `m` Wiederholungen.

Das Verwenden von Operatoren erlaubt einen sehr viel flexibleren Einsatz von Kategorien und Klassen von Ausdrücken und damit eine größere Abstraktion sowie kürzere Ausdrücke. Bezogen auf das vorherige Beispiel der Datumsangaben lässt sich somit eine allgemeinere Lösung finden:

```
1 re.findall(r'\d{2}\.\d{2}.\d{2,4}', text)
```

```
1 ['29.01.1956', '04.08.82', '28.11.1982']
```

Dieser reguläre Ausdruck lässt sich sogar auf geschrieben Datumsangaben erweitern. Unter der Annahme, dass Monatsnamen im Deutschen zwischen 3 und 9 Zeichen lang sind, kann dies als eine „oder“-Bedingung eingefügt werden: `\d{1,2}\. [A-z]{3,9}`.

```
1 re.findall(r'\d{1,2}\. [A-z]{3,9}|\d{2}\.\d{2}.\d{2,4}',
      text)
```

```
1 ['29.01.1956', '9. Mai', '04.08.82', '28.11.1982']
```

Grundsätzlich gilt, dass die Wiederholungsanweisungen `*` und `+` sowie die offenen Intervalle `{,m}` und `{n,}` gierige Algorithmen sind, d.h. sie versuchen soviele Zeichen wie möglich einem regulären Ausdruck zuzuordnen. Dieses Verhalten ist nicht immer gewünscht. Ein Beispiel dafür ist die Aufbereitung von Webseiten als Textdaten. Da es sich dabei meist um

HTML Dokumente handelt, enthalten diese für gewöhnlich eine Reihe von Formatierungsanweisungen in Form von sogenannten HTML-Tags. Beispielsweise zur Definition eines Absatzes:

```
<p>Hier steht ein Absatz.</p>
<p>Ein weiterer Absatz. Mit <em>inline</em> Formatierung.</p>
```

Für Textanalysen ist jedoch meistens nur der menschenlesbare Text von Interesse, weswegen diese Tags aus dem String entfernt werden müssen, bevor eine Tokenisierung durchgeführt werden kann. Eine mögliche Lösung ist die Entfernung der Tags mittels `re.sub()`.

```
html = '''<p>Hier steht ein Absatz.</p> \
<p>Ein weiterer Absatz. Mit <em>inline</em>
    Formatierung.</p>'''

re.sub(r'<.+>', '', html)
```

```
''
```

Die Anweisung lautete hier den Anteil des Strings mit einem leeren String zu ersetzen, der mit einem <-Zeichen beginnt und mit einem >-Zeichen beginnt. Da die Wiederholungsanweisung versucht möglichst viel von dem String zu verbrauchen, wird der gesamte String entfernt. Dieses Verhalten lässt sich abschalten, indem der Wiederholung ein ? hinzugefügt wird. In diesem Fall wird nur der minimalste Anteil des Strings genommen, für den die Bedingung zutrifft.

```
re.sub(r'<.+?>', '', html)
```

```
'Hier steht ein Absatz. Ein weiterer Absatz. Mit inline
    Formatierung.'
```

5.4.2 Spezifischere Tokenisierungsregeln

Neben der Extraktion und Aufbereitung von Texten ist das Erstellen von Tokenisierungsregeln eines der hauptsächlichen Anwendungsgebiete von regulären Ausdrücken in der quantitativen Textanalye. Wie bereits erwähnt, haben die Regeln mit denen Strings in Token zerlegt werden einen enorme Auswirkung auf die Qualität der daraus resultierenden Analysen der Texte. Sprachen und Texte weisen eine Vielzahl unterschiedlicher Muster und Entstehungskontexte auf, welche die Entwicklung einer generalisierten, universell einsetzbaren Tokenisierungsfunktion quasi unmöglich machen. Vielmehr bedarf es informierter, nachvollziehbarer Entscheidungen des Forschers, welche Tokenisierungsregeln sinnvoll sind. Reguläre Ausdrücke stellen die dafür notwendigen Werkzeuge bereit.

Im Folgenden soll ein (offensichtlich) fiktiver Chatbeitrag dazu dienen die Probleme der Tokenisierung von sehr spezifischen Texten zu verdeutlichen. Diese Wahl ist auch damit zu begründen, dass Texte die in Online-Interaktionen erzeugt werden von zunehmender Bedeutung für die die quantitative Analyse von Text und das Textmining sind.

```
1 chat = u'''haha.Wirklich??!?! Das ist ja intuitiv-lustig
      ...'''
```

Die spezifischen Eigenheiten von Chatbeiträgen, wie Schreibfehler, unkonventionelle Interpunktion und kreativer Umgang mit Sonderzeichen, können von standardisierten Tokenisierungsregeln oft nicht adäquat abgebildet werden:

```
1 import nltk
2 nltk.word_tokenize(chat)
```

```
1 [u'haha.Wirklich',
2  u'?',
3  u'?',
4  u'!',
5  u'?',
6  u'!',
7  u'Das',
8  u'ist',
```

```
 u'ja',
 u'intuitiv-lustig',
 u'\U0001f602\U0001f602',
 u'...']
```

Bevor wir uns den spezifischen Eigenheiten von Chatbeiträgen widmen, ist es sinnvoll die Tokenisierung von Texten noch einmal genauer zu betrachten und dies aus der Perspektive regulärer Ausdrücke heraus zu tun. Auf der grundlegendsten Ebene geht es bei der Tokenisierung von Strings um das Auffinden von Wörtern. In der Logik regulärer Ausdrücke bedeutet dies all diejenigen Teile des Strings zu lokalisieren, die nur aus alphanumerischen Zeichen bestehen: `\w+`. Alphanumerisch, weil die Kombination von alphabetischen und numerischen Zeichen oft als valide Ausdrücke in natürlichen Sprache vorkommen. Hätten wir Grund zu der Annahme, dass dies in einem spezifischen Sprachkontext nicht der Fall ist, so könnten wir stattdessen nur die alphabetischen Zeichen (`[A-z]`) verwenden oder eine spezifischere Definition versuchen.

```
re.findall(r'\w+', chat)
```

```
[u'haha', u'Wirklich', u'Das', u'ist', u'ja', u'intuitiv',
    u'lustig']
```

Hier wird allerdings ein erstes Problem deutlich. Die Verwendung von Verknüpfungszeichen, wie Bindestrich (`-`) oder Apostroph (`'`), wird nicht korrekt erkannt. Wir können dem Rechnung tragen, indem wir unsere Definition eines Wortes verfeinern.

```
re.findall(r'''\w+[\-\']?\w+''', chat)
```

```
[u'haha', u'Wirklich', u'Das', u'ist', u'ja',
    u'intuitiv-lustig']
```

Je nach Kontext können weitere Spezifikationen notwendig sein. Die einzige Möglichkeit besteht dann in einer sukzessiven Verfeinerung des

regulären Ausdrucks Anhand von Stichproben aus dem Korpus. zum Beispiel kann es in einem bestimmten Kontext gängig sein Worte mit Anführungszeichen zu versehen um Spott und Ironie auszudrücken, während dies in anderen Kontexten zur Kennzeichung von Zitationen verwendet werden würde. Je nach Fragestellung und Erkenntnisinteresse ist es notwendig solchen unterschiedlichen Verwendungen von Wörtern mit spezifischen Tokenisierungsregeln zu Rechnung zu tragen.

Im Falle von Chats ist es ebenfalls sinnvoll diesen Besonderheiten des Sprachgebrauchs Rechung zu tragen. Dies betrifft insbesondere die Verwendung von Interpunktion und Sonderzeichen wie Emoticons oder Emoji. Im Gegensatz zur Schriftsprache des Alltags werden multiple Interpunktionszeichen in Chats oft genutzt um eine spezifische Konnotation auszudrücken. Weil die Texte zudem recht kurz sind, haben diese Sonderzeichen eine dementsprechend größere, bedeutungstragende Funktion für den Text. Ähnliches gilt auch für die Sonderzeichen, die emotionale Zustände des Chattenden zu Ausdruck bringen sollen. Wir können beidem mit folgender Definition eines regulären Ausdrucks Rechung tragen:

```
1 re.findall(r'''[\.\?\!]+|\w+[\-\']?\w+|\S''', chat)
```

```
1 [u'haha',
2  u'.',
3  u'Wirklich',
4  u'??!?!',
5  u'Das',
6  u'ist',
7  u'ja',
8  u'intuitiv-lustig',
9  u'\U0001f602',
10 u'\U0001f602',
11 u'...']
```

Dabei ist die Reihenfolge der drei Definitionen von Bedeutung. Zuerst werden die multiplen Interpunktionszeichen durch den Ausdruck `[\.\?\!]+` erfasst. Im nächsten Schritt folgt die bereits bekannte Spezifikation von Wörtern. Daran anschließend werden alle bis jetzt nicht erfassten Sonderzeichen abgehandelt. Genau genommen besagt der letzte Aus-

druck (\S) alles zu extrahieren, was kein Leerzeichen ist. Entsprechend muss dieser Ausdruck als letztes gesetzt werden, da er sonst alle Nicht-Leerzeichen einzeln zurückgeben würde.

Abschließend muss hier noch einmal vor der Gefahr der unkritischen Übernahme von Tokenisierungsregeln gewarnt werden. Die hier beschriebene Annäherung an Chatbeiträge ist nicht ohne weiteres generalisierbar. Zum Beispiel werden weitere Besonderheiten spezifischer sozialer Medien, wie Hashtags oder Verlinkungen, nicht berücksichtigt. Wichtiger ist jedoch der Hinweis, dass eine Tokenisierung immer mit Hinblick auf eine konkrete Fragestellung stattfinden muss. Reguläre Ausdrücke stellen nur das adäquate Werkzeug dar um den jeweiligen Erfordernissen gerecht zu werden. Dies ist ein weiterer Hinweis auf die Bedeutung der „digital literacy“ im Bereich der quantitativen Textanalyse. Gerade weil natürliche Sprachen komplexe und kontextsensible Phänomene sind können ihnen standardisierte Lösungen nur bedingt gerecht werden.

5.5 Bestimmung textueller Eigenschaften

Sobald der Text in einer tokenisierten Form vorliegt, können seine spezifischen Merkmale mittels statistischer und computergestützter Verfahren analysiert werden. Dies kann sowohl für eigenständige Forschungsfragen genutzt werden, als auch zur weiteren Aufbereitung der Texte eines Korpus. Da diese Verfahren in den meisten Fällen relativ viel Rechenzeit erfordern ist es sinnvoll die Zwischenergebnisse zu speichern und später weiter zu verwenden.

NLTK bietet eine eigene Klasse für das Arbeiten mit tokenisierten Texten an. Die `Text`-Klasse wandelt eine Liste von Token in ein eigenständiges Objekt um, welches eine Reihe von grundlegenden Methoden der corpuslinguistischen Analyse von Texten bereitstellt. Hiermit lassen sich vor allem stilistische und strukturelle Merkmale einzelner Texte einfach vergleichen. Diese Verfahren können jedoch auch als eigenständige Funktionen gerufen werden. Ein solches Vorgehen hat den Vorteil, dass es ein flexibleres Vorgehen bei der Analyse erlaubt. Außerdem können Objekte der `Text`-Klasse nicht direkt von höherrangigen Datenstrukturen verarbeitet werden, die für eine statistische Analyse von Textmengen benötigt werden. Auch die Verarbeitung von Metadaten ist in spezialisierten Werkzeugen, wie dem `pandas.DataFrame`, sehr viel komfortabler. Daher wird die `Text`-

Klasse im Folgenden hauptsächlich für die Auswertung von Kollokationen herangezogen.[13]

5.5.1 N-Grame und Kollokationen

Eine der grundlegendsten Techniken zur Analyse und dem Vergleich von Texten ist die *Bestimmung der festen Wendungen* oder Phrasen, die diese enthalten. Dabei handelt es sich um konventionelle Verknüpfung von zwei oder mehr Worten zu einem zusammengesetzten Begriff, der eine eigenständige Bedeutung aufweist. Beispiele dafür wären feste Wendungen wie „soziale Gerechtigkeit“ oder „Verteilungsgerechtigkeit“. Die Studie dieser Begrifflichkeiten ist in den Sprachwissenschaften das Gebiet der Phraseologie (vgl. z.B.: Burger 2007). Sie sind jedoch auch im Bereich des Information Retrievals und der statistischen Analyse von Texten von Bedeutung. Hier werden sie vor allem als Prädiktoren der übergreifenden Bedeutung des Textes untersucht sowie als eine notwendige Vorstufe in der Aufbereitung von Textdaten.

Je nach der Morphologie der zugrundeliegenden Sprache kann die Zusammensetzung dieser Zeichen sehr unterschiedlich erfolgen. Aus diesem Grund werden im Englischen hierfür oft die Ausdrücke *n-grams* (N-Gramme) und *collocations* (Kollokationen) verwendet, die von einer Zusammensetzung mehrerer Worte in Form einer durch Leerstellen getrennten Sequenz von Worten ausgehen. Der Begriff der N-Gramme weist zudem darauf hin, dass es sich hier um längere Ketten von Wörtern handeln kann.[14] Formal kann gesagt werden, dass jedes Wort w eines endlichen Alphabets Σ der Länge n eine Teilmenge dieser Sprache darstellt, also $\{w_1, \cdots, w_n\} \in \Sigma^n$. Damit sind streng gennommen alle Worte einer natürlichen Sprache N-Gramme, da auch Unigramme – also Worte der Länge 1 – dazu gezählt werden müssen. Somit würde es sich bei „cultural studies“ um ein Bigramm handeln, während „cultural studies program“ ein Trigramm wäre.

In Sprachen, die Begrifflichkeiten mittels N-Grammen konstruieren die länger als ein Wort sind, stellt dies ein grundsätzliches Problem für die

[13] Weitere Methoden und eine ausführliche Erklärung der API finden sich in der NLTK Dokumentation: http://www.nltk.org/

[14] Zur Untersuchung ganzer Sprachen hinsichtlich der enthaltenen N-Gramme bietet die Firma Google ein öffentliches Werkzeug, den *Google N-Gram Viewer* an: https://books.google.com/ngrams. Hiermit lässt sich der digitale Bestand von Google Books im Zeitverlauf untersuchen. Zudem werden auch die Rohdaten (als N-Gram Listen) zugänglich gemacht.

Tokeniserung dar. Da die Zerlegung eines Textes in Token immer auf die Identifikation der Wortgrenzen abzielt, werden Token grundsätzlich als Unigramme konzipiert. Dies bedeutet, dass die Feststellung zusammengesetzter Begrifflichkeiten immer erst im Nachhinein geschehen kann und daher mit gewissen Unwägbarkeiten einhergeht, da Bigramme und dergleichen in diesem Fall erst mit statistischen Verfahren geschätzt werden müssen. In Sprachen wie dem Deutschen, in denen Kompositionen von Begriffen tendenziell als Unigramme konstruiert werden (z.B.: „Gesetzgebungsverfahren"), ist dieses Problem etwas geringer. Allerdings gibt es auch hier Phrasen, die aus mehreren Worten bestehen. Daher stellt eine Analyse der N-gramme eines Textes in vielen Fällen einen sinnvollen, ersten Schritt in der Datenaufbereitung dar.

Die Analyse von *Kollokationen* geschieht meist aus einer statistischen und informationstheoretischen Perspektive heraus. Demnach ist eine Kollokation die Wahrscheinlichkeit, dass auf ein Wort w in einer bestimmten Distanz ein bestimmtes anderes Wort folgt (vgl. Halliday 1961: 276).[15] Anders ausgedrückt, handelt es sich um die Wahrscheinlichkeit eine bestimmte Sequenz an Worten $P(w_{n+1} = x_{n+1} \mid w_1 = x_1, \cdots, w_n = x_n)$ als Ergebnis einer zufälligen Ziehung zu erhalten. Diese Überlegungen gehen auf die bereits beschriebene Vorstellung von Informationen als Zufallsverteilungen von Sequenzen zurück, wie sie von Hartley und Shannon als Grundlage der Informationstheorie beschrieben wurden. Genauer gesagt, als Markoff-Ketten, d.h. gedächtnislose Sequenzen, bei denen jeder neue Zustand unabhängig vom vorangegangenen ist (vgl. Shannon 1948: 387f). Dies ist natürlich eine grobe Vereinfachung, da die Auswahl einzelner Wörter in einem Satz nicht zufällig ist und sehr wohl von der Wahl der vorangegangenen Worte abhängt. Daher sind es genau diejenigen Abweichungen von einer zufälligen Verteilung, die eine systematische Analyse von Text überhaupt erst möglich machen. Daraus resultiert jedoch die Frage, welche Verteilungsannahmen hinsichtlich dieses Prozesses zu treffen wären und wie das Ausmaß der Assoziation abzuschätzen wäre.

Im Modul `nltk.collocations` sind eine Reihe von Assoziationsmaßen implementiert, denen unterschiedliche Annahmen über die Kon-

[15] Neben dieser wahrscheinlichkeitstheoretischen Auffassung existieren noch weitere Operationalisierungen des Konzepts der Kollokation, die an andere Definitionen, zum Beispiel aus dem Bereich der Semiotik oder der Diskurstheorie, anknüpfen. Da es hier hauptsächlich um die Verwendung von Kollokationen als einem Instrument der quantitativen Analyse von Text geht, werden diese hier nicht ausführlicher diskutiert. Für eine detailiertere Aufarbeitung des Begriffs der Kollokation sei Gledhill (2000) empfohlen.

struktionsweisen von Texte zugrundeliegen.[16] Davon wird hier nur das von Ted Denning (1993) entwickelte Assoziationsmaß auf Basis der *log-likelihood ratio* näher betrachtet, um die allgemeine Logik dieser Verfahren zu demonstrieren.

Die Wahl für dieses Verfahren ist jedoch nicht dessen Einfachheit geschuldet, sondern kann aus zeichen- und informationstheoretischen Überlegungen heraus begründet werden.[17] Die meisten anderen, gängigen Verfahren, die auch in NLTK implementiert sind, basieren auf der Annahme einer approximativen Normalverteilung von Worten. Dies kann bei sehr großen Textmengen eine brauchbare Annahme sein, berücksichtigt jedoch nicht die besonderen Verteilungeigenschaften von Worten in Texten und deren Aufbau als Sequenzen. Die Annahmen von normal- oder χ^2-verteilten Worten ist meistens nicht haltbar, da der Großteil der Worte relativ häufig in einem Text vorkommt, jedoch hinsichtlich des Vokabulars einer Sprache der überwältigende Teil der Worte sehr selten genutzt wird. Diese Besonderheit von Sprachen wird als Zipfs-Gesetz bezeichnet, welches im Abschnitt **??** noch ausführlicher behandelt werden wird. An dieser Stelle ist es ausreichend die Konsequenz zu benennen, die in einer Überschätzung der Bedeutung relativ häufiger Token liegt. Stattdessen sind es gerade die Worte mittlerer Häufigkeit, welche das beste Identifikationsmerkmal von Texten bereitstellen (vgl. ebd.).

Bei der Identifikation von Bigrammen mittels der log-likelihood ratio wird davon ausgegangen, dass die Aneinanderreihung von Wörtern in Texte als ein Bernoulli Experiment aufgefasst werden kann. Dabei handelt es sich um ein wiederholtes Zufallsexperiment, dessen jeweiliges Ergebnis von den vorangegangenen Ergebnissen unabhängig ist. Im Normalfall wird hierbei die *Binomialverteilung* zugrundegelegt, was die Frage nach Bigrammen in die Aussage transferiert, ob auf ein bestimmtes Wort ein anderes Wort folgt oder nicht. Ausgehend von der Wahrscheinlichkeitsfunktion der Binomialverteilung im diskreten Fall:

[16] Leider lässt die ansonsten hervorragende Dokumentation des NLTK-Moduls an dieser Stelle etwas zu wünschen übrig. Es gibt eine kurze Darstellung des Analyseablaufs in Form eines HOWTOs: http://www.nltk.org/howto/collocations.html. Um die Details der Implementation zu verstehen, kommt man um eine Auseinandersetzung mit dem Quellcode (http://www.nltk.org/api/nltk.html#module-nltk.collocations) nicht herum. Dies kann als ein weiterer Hinweis auf die Bedeutung der „Digital Literacy" in diesem Bereich gesehen werden.

[17] Allerdings ist die Einfachheit der Umsetzung dieses Verfahrens in Programmcode ebenfalls eine beachtliche Leistung. Eine äußerst elegante Umsetzung in zwei Zeilen R-Code findet sich auf Ted Dunnings Blog: `http://tdunning.blogspot.de/2008/03/surprise-and-coincidence.html`.

$$B(k;p;n) = \binom{n}{k} p^k (1-p)^{n-k},$$

lässt sich dies ausdrücken, als die Wahrscheinlichkeit des k-maligen Auftretens eines bestimmten Wortes in einer Sequenz der Länge n, bei einer allgemeinen Auftretenswahrscheinlichkeit dieses Wortes von p.

Unter diesen Grundannahmen können die Wahrscheinlichkeit für das Auftreten zweier Wörter in einer Sequenz ins Verhältnis zur Wahrscheinlichkeit ihres voneinander unabhängigen Auftretens gesetzt werden. Damit ergibt sich laut Dunning (1993: 67) folgendes Verhältnis der Likelihoodfunktionen (L):

$$\lambda = \frac{\max L(p;n_1;k_1)L(p;n_2;k_2)}{\max L(p_1;n_1;k_1)L(p_2;n_2;k_2)}$$

für die gilt

$$L(p;k;n) = p^k(1-p)^{n-k}.$$

Die jeweiligen Maximalwerte werden dann erreicht für $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$ und $p = \frac{k_1+k_2}{n_1+n_2}$. Daraus ergibt sich abschließend die log-likelihood ratio:

$$-2\log\lambda = -2[\log L(p;n_1;k_1) + \log L(p;n_2;k_2) - \log L(p_1;n_1;k_1) - \log L(p_2;n_2;k_2)].$$

Diese lässt sich äquivalent auch für eine multinomiale Verteilung konstruieren. In diesem Fall werden einfach Parameter für jedes Wort innerhalb des Korpus angenommen.

Die resultierenden log-likelihood ratios sind approximativ χ^2 verteilt, daher ist es möglich die Hypothese der Unabhägigkeit der Verteilung zweier Wörter auf Signifikanz zu prüfen. Allerdings wäre dies, aufgrund der hohen Anzahl an Freiheitsgraden, tendenziell aufwendig und im Ergebnis auch nicht sonderlich aussagekräftig, da anzunehmen ist, dass alle Worte in ihrer Auftretenswahrscheinlichkeit relativ stark abhängig von derjenigen der anderen Wörter wären. Für die Feststellung von Kollokationen ist dies jedoch auch nicht notwendig, da es ausreichend ist, die log-likelihood ratios für alle Wortpaare zu bestimmen und sie als eine Rangfolge zu betrachten. Damit lässt sich die Feststellung treffen, welche Wortpaare am stärksten in einer bestimmten Reihenfolge gemeinsam auftreten.

In NLTK stellt dieses Vorgehen das Standardverfahren zur Bestimmung von Kollokationen dar und ist als eine an Objekte der Klasse `Text` gebundene Methode implementiert. Das bedeutet, dass die entsprechende Liste von Token erst in ein `Text`-Objekt umgewandelt werden muss, bevor mittels dieser Methode Kollokationen identifiziert werden. Neben der Feststellung der Kollokationen als Merkmale von Texten und Korpora dient dieses Verfahren auch dazu den Text für weitere Analyseschritte aufzubereiten. Insbesondere wenn die Bedeutungen und Themen von Texten relativ zueinander bestimmt werden sollen, ist so eine Zusammenfassung angebracht. Der Grund dafür liegt hauptsächlich in der Behebung möglicher Verzerrung, die durch die Herstellung von Unigrammen im Verlauf der Tokenisierung zustande kommen können. Hat man die festen Wendungen eines Textes erst einmal ausfindig gemacht, können diese zu einzelnen Token zusammengefasst werden und somit als eigenständige Worte in die Analyse miteinfließen.

Davor muss der Text aber noch einen weiteren Vorbereitungsschritt durchlaufen, nämlich die Entfernung sogenannter Stopwörter. Dabei handelt es sich um die Wörter, welche am häufigsten in den Texten einer bestimmten Sprache vorkommen. Diese enthalten meist nicht viel eigene Information, im Sinne dienlicher Hinweise für die Unterscheidung verschiedener Texte. Da es sich dabei auch um die am häufigsten Token in individuellen Texten handelt, tragen sie nur sehr wenig zum Sinngehalt eines spezifischen Textes bei. Im Falle von Kollokationen kommt hinzu, dass eine Reihe von festen Wendungen gerade aus diesen häufigsten Wörtern gebildet werden. So würden im Deutschen die Kombination von Artikeln und Substantiven die Liste der Kollokationen dominieren, wenn man „der“, „die“, „das“, etc. nicht vorher entfernen würde. Weil das soziologische Forschungsinteresse auf die Erforschung von Kultur und Wissen, also den symbolischen Ordnungen hinter den Zeichen, ausgelegt ist, empfiehlt sich eine vorherige Entfernung von Stopwörtern auch bei fast allen anderen Analyseverfahren.

Da es sich hier um eine Standardoperation des Natural Language Processings handelt, sind Stopwortlisten in fast allen Programmpaketen aus diesem Bereich enthalten. Auch im NLTK findet sich ein Korpus für Stopwörter, der die häufigsten Stopwörter für eine Reihe von europäischen Sprachen enthält. Um einen Überblick über die enthaltenen Sprachen und die Argumente mit denen dieses abgerufen werden zu erhalten, kann die Methode `nltk.corpus.stopwords.fileids()` eingesetzt werden.

Für die Demonstration dieses Vorgehens und der anschließenden Analyse der Bigramme wird der bereits vorgestellt Korpus von Abstracts aus Soziologiezeitschriften verwendet. Bei der Entfernung der Stopwörter ist zweierlei zu beachten. Erstens kann es notwendig sein die Liste der Stopwörter um einige spezifische Begriffe zu erweitern. Im folgenden Codebeispiel wird die Liste der Stopwörter ergänzt, um eine Reihe von Begriffen aus dem Korpus zu filtern, die keinen zusätzlichen Informationsgehalt enthalten. Dabei handelt es sich in diesem Fall hauptsächlich um Informationen eines bestimmten Verlags. Zweitens muss beachtet werden, dass die Worte des NLTK Korpus für Stopwörter alle als kleingeschriebene Tokens enthalten sind. Da Python bei der Prüfung der Äquivalenz von Strings zwischen Groß- und Kleinschreibung unterscheidet, müssen die Tokens angepasst werden. Dies geschieht mittels der String-Methode `.lower()`. Dabei ist es empfehlenswert die Liste der Strings, aus der die Stopwörter entfernt wurden, in einer einheitlichen Schreibweise (Groß- oder Kleinschreibung) zu konstruieren. Ansonsten würde dies zu Verzerrungen bei späteren Analyseschritten führen, da unterschiedliche Schreibweisen als unterschiedliche Worte aufgefasst werden würden.

```
1  from nltk.corpus import stopwords
2
3  additional = ['elsevier',
4                'masson',
5                'sas',
6                'tous',
7                'droits',
8                'reserves',
9                'publie',
10               'par',
11               'medicales',
12               'scientifiques',
13               'inc']
14
15 stops = set(stopwords.words('english') + additional)
16
17 def stopped_tokens(tokens, stopwords,
18                     reg_ex=re.compile(r'\w+[\_\-]*\w*',
19                                       re.U)):
20     return [token.lower() for token in tokens
```

```
21              if token.lower() not in stopwords
22              and reg_ex.match(token)]
```

Der obige Code entfernt jedoch noch nicht die Stopwörter. Stattdessen wird die erweiterte Stopwortliste gebildet und eine generische Funktion zur Entfernung der Stopwörter definiert, die außerdem die Kleinschreibung aller Token herbeiführt und als weiteres Argument einen regulären Ausdruck akzeptiert. Der Standardwert dieses regulären Ausdrucks legt fest, dass nur alphanumerischen Zeichen, sowie deren Kombination mit den Trennungszeichen Unterstrich (_) und Minus (-) akzeptiert werden. Dadurch werden alle Sonder- und Satzzeichen aus dem Text entfernt. Eine Funktion zu schreiben erlaubt nicht nur deren Anpassung an verschiedene Datenlagen, sondern auch die einfache Durchführung der Entfernung von Stopwörter in einem `DataFrame` Objekt, indem wir die Funktion mittels `.apply()` vektorisieren. Dies hat den Vorteil, dass nicht mit unterschiedlichen Datenstrukturen gearbeitet werden muss und somit das Risiko von Fehlern bei der Datentransformation minmiert wird.

```
1 articles['Stopped'] = articles.Tokens.apply(stopped_tokens,
      stopwords=stops)
```

Es ist sinnvoll Kollokationen als ein Element der Sprache aufzufassen, da die überzufällig häufige Kombination von bestimmten Token in einzelnen Texten auch eine Folge des Stils eines bestimmmten Autors sein kann. Daher setzt eine Identifikation der Kollokationen eines Diskurses oder einer Sprache die Betrachtung der einzelnen Texte eines Korpus als einen Text voraus. Der folgende Code extrahiert die obersten 30 Kollokationen der geordneten Rangfolge, die durch Verwendung der oben beschriebenen log-likelihood ratio erzeugt wird.

```
1 all_tokens = sum(articles.Stopped, [])
2
3 all_text = nltk.Text(all_tokens)
4
5 all_text.collocations(30, 2)
```

```
1 united states; labour market; labor market; article
      examines; life
```

```
2 course; welfare state; young people; rational choice; men
      women;
3 results show; one hand; collective action; using data; labor
      force;
4 educational attainment; max weber; social capital; case
      study; human
5 capital; social movements; paper examines; socio-economic
      panel;
6 article argues; article explores; west germany; civil
      society; income
7 inequality; twentieth century; east germany; social sciences
```

Die resultierende Liste an Begriffen enthält bereits eine Reihe von Bigrammen, die sowohl der allgemeinen Sprache (z.B.: „united states“) als auch der spezifischen Fachsprache der Soziologie (z.B.: „rational choice“) zugeordnet werden können. Daneben finden sich jedoch auch noch eine Reihe fester Wendungen, die sich auf bestimmte Standardformulierungen beziehen (z.B.: „article examines“) und von denen in späteren Analyseschritten kein großer Erkenntnisgewinn zu erwarten ist. Gleichzeitig stellen sie auch kein allzu großes Problem für weitere Analysen dar, da sie Kollokationen umfassen, die spezifisch für diesen Diskurs sind.

Da die Berechnung der Kollokationen relativ viel Zeit in Anspruch nimmt und da sie primär der Vorbereitung auf weitere Analysen dient empfiehlt es sich die identifizierten Kollokationen in den Datensatz einzupflegen. Im `texttools` Modul (siehe Anhang A auf Seite 341) findet sich zu diesem Zweck die `combine_collocations()` Funktion. Diese nimmt drei Argumente: einen Text in der Form einer Liste von Token, eine Liste von Kollokationen in der Form von Tupeln und ein optionales Argument `using`, durch welches der String spezifiziert werden kann, der zum Zusammenfügen der Token verwendet werden soll. Um die Liste der Kollokationen als Tupel zu erhalten, kann man das Attribut `._collocations` der `text`-Klasse verwenden, nachdem man die Methode `.collocations()` ausgeführt hat.

5.5.2 Grammatik

Unsere bisherigen Betrachtungen von Texten haben sich hauptsächlich auf die Ebene des Vokabulars (Typen), bzw. deren Realisation (Token) bezogen. Neben dem Vokabular zeichnet sich jede Sprache durch eine Grammatik aus, d.h. den Regeln der Zusammensetzung von Token zu größe-

ren syntaktischen Einheiten (meistens Sätze). In der Linguistik und den Sprachwissenschaften sind die grammatikalischen Regeln natürlicher, erfundener und formaler Sprachen zentraler Forschungsgegenstand. Für die sozialwissenschaftliche Textanalyse ist vor allem die richtige Identifikation grammatikalischer Eigenschaften von Interesse. Es geht hier also nicht um die Grammatik selbst, sondern um deren Funktion als einer Ressource der Textanalyse.

Damit soll nicht gesagt sein, dass nicht auch die Grammatik selbst Gegenstand sozialer Auseinandersetzungen sein kann. Dies zeigt sich vor allem am Beispiel des Versuchs der Erzeugung einer „gendergerechten Sprache". Im Kern basiert diese Auseinandersetzung auf der Gleichsetzung des grammatikalischen Geschlechts von Wörtern und dem biologischen oder sozialen Geschlecht von Personen, worin sich auch noch einmal zeigt wie eng verbunden die menschlichen Selbstbeschreibungen und die sie umgebenden, symbolischen Ordnungen sind.[18] Dennoch erweist sich die Grammatik gemeinhin als ein relativ stabiles Phänomen. Verglichen mit dem Vokabular einer Sprache verändern sich grammatikalische Regeln selbst in historischen Zeiträumen kaum. Auch in der vorher erwähnten Debatte zeigt sich die Stabilität grammatikalischer Regeln. Zwar wird hier manchmal die Einführung geschlechtsneutraler Artikel und die Vermeidung von Pronomen angemahnt, jedoch handelt es sich hier bei genauerer Betrachtung auch nur um die Veränderung der verwendeten Worte und nicht der zugrundeliegenden, grammatikalischen Funktionen.

Für die sozialwissenschaftlich ausgerichtete Textanalyse ist die Grammatik insbesondere in zweierlei Hinsicht von Bedeutung. Zum einen ist es oft notwendig Wörter in deren grammatikalische Grundform zu überführen, um diese aus dem spezifischen Kontext ihrer Verwendung zu lösen und über eine Menge von Texten hinweg vergleichbar zu machen. Die Entfernung grammatikalischer Formen ist vor allem dann gerechtfertigt, wenn es um die Bestimmung der grundlegenden Themen und Wissensbestände einer Mehrzahl von Texten geht, da Grammatik in natürlichen Sprachen hauptsächlich auf die Konkretisierung von Sachverhalten ab-

[18] Ohne den Zusammenhang von Grammatik und Geschlechterbildern vertieft behandeln zu wollen, da mir diese Debatte zu stark von Ideologie durchsetzt zu sein scheint, möchte ich an dieser Stelle nur kurz die empirische Implikation aufzeigen. Nimmt man das Argument einer Beeinflussung von Geschlechterrollen durch grammatikalische Strukturen ernst, so würde dies bedeuten, dass Gesellschaften in denen die Grammatik der dominanten Sprache kein grammatikalisches Geschlecht kennt tendenziell egalitärere Gesellschaften sein müssten. In dieser Form liese sich somit auch eine empirische Prüfung durchführen.

zielt. Diese Kontextsensibilität und Spezifizität kann jedoch ebenso ein Werkzeug für die quantitative Textanalyse sein. Denn die Bestimmung der grammatikalischen Form von Wörtern in konkreten Sätzen (*part-of-speech tagging*) kann auch genutzt werden um Modelle zu erzeugen, die stärker auf die sprachlichen und kontextbezogenen Eigenschaften von Sprache eingehen. Die zweite Bedeutung der Grammatik in der Analyse von Texten, liegt demnach in der Ermöglichung von Modellen die sich stärker an den Eigenheiten sprachlicher Formen orientieren.

Stemming vs. Lemmatisierung

Zur Herstellung der grammatikalischen Grundform eines Wortes gibt es im Wesentlichen zwei Vorgehensweisen. Entweder man bedient sich eines Algorithmus der durch die iterative Anwendung bestimmter Regeln, z.B. im Falle indoeuropäischer Sprachen durch die Entfernung von spezifischen Wortendungen, eine Annäherung an die grammatikalische Grundform erzeugt (*Stemming*) oder man benutzt ein Lexikon um die korrekte Grundform zuzuordnen (*Lemmatisierung*). Da grammatikalische Regeln meist über eine Vielzahl von Ausnahmen verfügen, birgt das erstgenannte Verfahren das Risiko das unregelmäßige Wortformen nicht richtig reduziert werden oder das unterschiedliche Wörter auf die selben Grundformen reduziert werden. Lemmatisierung ist hingegen auf ein möglichst umfassendes Wörterbuch der jeweiligen Sprache angewiesen und dementsprechend nicht in der Lage mit falsch geschriebenen Wörtern umzugehen. Beide Methoden eignen sich nur für bestimmte Textgattungen und kommen insbesondere mit Fachsprachen und subkulturellen Dialekten nicht gut zurecht.

Einer der frühesten Stemming-Algorithmen wurde von Martin Porter (1980) entwickelt. Der sogenannte Porter-Stemmer-Algorithmus ist auch heute noch das weitverbreitetste Verfahren für Texte in englischer Sprache (vgl. Willett 2006). 2001 veröffentlichte Porter zudem eine Erweiterung in Form einer eigenständigen Sprache für Stemming-Algorithmen: Snowball. Darauf basierende Snowball-Stemmer sind heute standardmäßig in einer Vielzahl von Anwendungen enthalten und bieten Unterstützung für die meisten gängigen europäischen Sprachen. Auch das NLTK Modul enthält sowohl den Porter-Stemmer als auch dessen Verallgemeinerung. Daneben sind auch noch eine Reihe weiterer Stemming-Algorithmen für nicht-europäische Sprachen Bestandteil des `nltk.stem` Submoduls.

```
## Importieren des Porter-Stemming Algorithmus
from nltk.stem import PorterStemmer

## Importieren des Snowball-Stemming Algorithmus
from nltk.stem import SnowballStemmer
```

Um die Ergebnisse der beiden Stemmer miteinander und mit einem Lemmatisierer vergleichen zu können, muss ein englischer Text genutzt werden. Dazu wird im Folgenden ein Zitat von Norbert Elias (1991: 96) herangezogen, welcher inmitten der etwas technischeren Ausführungen dieses Kapitels helfen kann noch einmal die Gründe für eine soziologische Beschäftigung mit Symbolen zu verdeutlichen.

```
elias = """Awareness of the social character of \
languages, of their function as means of communication \
between a plurality of human beings is essential for \
the understanding of their symbolic function and thus \
for the term 'meaning'"""

elias_tokens = nltk.word_tokenize(elias)
```

Zum stemmen muss der jeweilige Token der `.stem()` Methode des instantierten Stemmers übergeben werden. Die Instanz einer beliebigen Klasse wird in Python durch den Aufruf und die Zuweisung dieser Klasse erzeugt.

```
## Instanz des Stemmers
stemmer = PorterStemmer()

## Stemmen
stemd = [stemmer.stem(token) for token in elias_tokens]
```

Da Porter und Snowball im Falle des Englischen weitestgehend äquivalent sind, ist ein direkter Vergleich dieser beiden Stemmer hier nicht sinnvoll. Damit kurz das Erstellen einer Snowball-Instanz für eine andere Sprache demonstriert werden kann, wurden eine Reihe von Wörtern der deutschen Sprache als Beispiel herangezogen. Gleichzeitig dient dieses Beispiel auch als eine Illustration der allgemeinen Fallstricke beim Einsatz von Stemming-Algorithmen.

```
stemmer = SnowballStemmer(language='german')

tokens = ['ging', 'gehen', 'gegangen',
          'soziologisch', 'Soziologe', 'Soziologie',
          'Soziolog_innen']

[stemmer.stem(token) for token in tokens]
```

```
[u'ging',
 u'geh',
 u'gegang',
 u'soziolog',
 u'soziolog',
 u'soziologi',
 u'soziolog_inn']
```

Wie aufgrund der algorithmischen Vorgehensweise zu erwarten, zeigt sich, dass irreguläre Wortformen nicht korrekt erkannt werden können. Die Entfernung bestimmter Wortendungen ist außerdem nicht auf Verben begrenzt. Auf den ersten Blick mag es nützlich erscheinen, dass „soziologisch“ und „Soziologe“ zu einem Ausdruck zusammengefasst werden, da man hier durchaus von einem gemeinsamen Wortstamm sprechen kann. Allerdings werden auch hier nur bestimmte Endungen entfernt, was zur Folge hat das „Soziologie“ als eigenes Wort bestehen bleibt. Ein weiteres Problem stellen Token wie „Soziolog_innen“ dar, die inhaltlich gut zu stemmen wären, aber nur sehr begrenzt den grammatikalischen Regeln der Standardsprache folgen und daher von den relativ rigiden Regeln eines Stemming-Algorithmus nicht richtig erkannt werden.

Für die Lemmatisierung stellt das NLTK Modul das Objekt `WordNetLemmatizer` bereit. Dieser Lemmatisierer verwendet das WordNet Diktionär um die Grundform von Wörtern zu finden. Auf der Homepage des Projektes findet sich eine gute Zusammenfassung WordNet des Aufbaus dieser Datenbank, die ein wichtige Ressource für eine Vielzahl von quantitativen Textanalysen darstellt:

> WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets

> are interlinked by means of conceptual-semantic and lexical relations.[19]

Dieser Lemmatisierer wird, ähnlich wie die Stemming-Algorithmen, erst initialisiert und dann auf die einzelnen Worte angewandt. Der folgende Code wendet sowohl Stemming als auch Lemmatisierung auf den ausgewählten Text (das Elias-Zitat) an. Die resultierenden Token werden miteinander verglichen und nur diejenigen in Tabelle 5.5 dargestellt, bei denen beide Resultate zu unterschiedlichen Ergebnissen kommen.

Es wird schnell ersichtlich, dass die Lemmatisierung von beiden Verfahren das präzisere ist. Allerdings sind auch diesem Verfahren zwei sehr eindeutige Grenzen gesetzt. Kommt das Wort nicht im entsprechenden Diktionär vor, so kann auch keine Lemmatisierung erfolgen . Dies ist gerade bei Fachsprachen und Subkulturen fast durchgehend der Fall. Des Weiteren kann es oft schwierig sein eine offen zugängliche Datenbank zur Lemmatisierung von Token einer spezifischen Sprache zu finden.[20] Außerdem muss man auch in Betracht ziehen, dass je nach Implementation und Textlänge eine Lemmatisierung mit einem sehr viel höheren Rechenaufwand verbunden sein kann. Gerade bei großen Textkorpora können solche Transformationen daher schnell unpraktikabel werden.

Gleichzeitig wird in Tabelle 5.5 auch das Problem einer Überspezifizierung von Lemmas deutlich. „Symbolic“ wird als ein eigenes Wort gezählt. Da es allerdings im gesamten Buch, aus dem dieser Beispielsatz entnommen wurde, hauptsächlich um das Konzept von Symbolen geht, ist zu erwarten, dass dieser Begriff in sehr vielen Permutationen zu finden sein wird. Möchte man nur dieses Buch analysieren, so wäre es in den meisten Fällen wünschenswert die unterschiedlichen Wortformen in denen der Begriff „Symbol“ vorkommt zu erhalten, um zum Beispiel die Schwerpunktsetzung von Kapiteln unterscheiden zu können. Stellt das Buch jedoch nur einen Datenpunkt in einer Vielzahl unterschiedlicher Texte dar, so kann es zweckmäßig sein die gröbere Bestimmung der Wortformen zu verwenden, die durch das Stemming erzeugt wird. Wenn das Ziel der Untersuchung die übergreifenden Themen und symbolischen Ordnungen sind, kann es sinnvoll sein die Variation verschiedener Wortformen möglichst stark zu begrenzen, um die dominanten Unterschiede zwischen den Themen besser erkennen zu können.

[19] `http://wordnet.princeton.edu/`

[20] Selbst für die deutsche Standardsprache sind bisher nur wenige Ressourcen öffentlich verfügbar. Ausnahmen stellen die Erweiterungsbibliothek GermaNLTK und die eigenständige Anwendung TreeTagger dar.

	Lemmas	Original	Stems
0	The	The	The
1	question	question	question
2	be	is	is
3	be	being	be
4	scrutinize	scrutinized	scrutin
5	how	how	how
6	to	to	to
7	integrate	integrate	integr
8	into	into	into
9	a	a	a
10	general	general	gener
11	theory	theory	theori
12	of	of	of
13	action	action	action
14	the	the	the
15	two	two	two
16	type	types	type
17	of	of	of
18	rational	rational	ration
19	action	action	action
20	distinguish	distinguished	distinguish
21	by	by	by
22	Max	Max	Max
23	Weber	Weber	Weber
24	base	based	base
25	on	on	on
26	value	value	valu
27	rationality	rationality	ration
28	and	and	and
29	instrumental	instrumental	instrument
30	rationality	rationality	ration

Tabelle 5.5: Vergleich von Stemming und Lemmatisation.

Letztlich ist die Entscheidung für oder gegen Stemming-Algorithmen bzw. Lemmatisierung im konkreten Forschungshandeln sowieso nicht einfach durch „Standards“ oder „gängige Praktiken“ zu treffen. Stattdessen müssen die Eigenarten der jeweiligen Verfahren in der Kombination mit den zu untersuchenden Texten berücksichtigt werden. Es liegt im Ermessensspielraum des Forschers einzuschätzen, welche Art der Verzerrung damit einhergeht oder ob eine Herstellung der Wortstämme überhaupt einen Unterschied macht, wenn die bedeutungstragenden Worte nicht richtig erkannt werden. Insofern wird auch hier noch einmal deutlich, dass es in der Bearbeitung komplexer Daten keinen Königsweg gibt.

POS-Tagging

Im Gegensatz zu Stemming und Lemmatisierung geht es beim Part-of-Speech-Tagging und dem damit eng verwandten Chunking um die korrekte *Identifikation der grammatikalischen Form* eines Wortes im Kontext des jeweiligen Satzes. Die dabei eingesetzten Verfahren ähneln den bereits beschriebenen, setzen jedoch fast immer eine Kombination von Wörterbüchern und Algorithmen ein. Dies ist notwendig, da die grammatikalische Funktion eines Wortes in einem Satz relativ zu den Positionen der anderen Wörter ist. Daher handelt es sich um ein sehr viel schwereres Problem als die Rückführung auf die grammatikalische Grundform, die tendenziell unabhängig von der konkreten Position des Wortes versucht werden kann.

Die unterschiedlichen grammatikalischen Regeln von natürlichen Sprachen bedeuten auch, dass die Feststellung der Wortformen immer auf eine bestimmte Sprache begrenzt ist. Daher sind sozialwissenschaftliche Forscher, für die die Feststellung der grammatikalischen Form nur ein Zwischenschritt zu weiteren Analysen ist, hier auf lexikalische Ressourcen aus dem Bereich der Corpuslinguistik angewiesen.[21] NLTK bietet hierzu einen eigenen „off-the-shelf“ POS-Tagger an. Die entsprechende Funktion `pos_tag()` verarbeitet Listen von Token und gibt eine Liste von Tupeln zurück, bei denen der jeweilige Token an der ersten Stelle steht, gefolgt vom entsprechenden „Tag“, der die jeweilige grammatikalische Wortform angibt.

[21] Das `nltk`-Modul beinhaltet grundsätzlich alle Funktionalitäten um einen eigenen Algorithmus zur Bestimmung von Wortformen zu entwickeln. Die dafür notwendigen Werkzeuge finden sich in `nltk.tag`. Allerdings dürfte der Aufwand oft den Nutzen für die angewandte Forschung übersteigen. Eine gute Einführung in die Entwicklung und Logik von Tagging und Chunking findet sich hier: http://www.nltk.org/book/ch05.html.

Die Bezeichnung der grammatikalischen Wortformen entspricht dabei dem sogenannten Penn Treebank Tagset (vgl. Taylor, Marcus und Santorini 2003). Um die Bedeutung des jeweiligen Tags zu verstehen gibt es eine entsprechende Hilfsfunktion: `nltk.help.upenn_tagset()`. Wird diese Funktion ohne Argumente gerufen, so erhält man eine Liste des gesamten Tagsets mit entsprechenden Beispielen. Zusätzlich nimmt die Funktion auch einen regulären Ausdruck an, der einen oder mehrere Tags spezifiziert. Der folgende Code demonstriert dies, indem er sämtliche Arten von Substantiven der englischen Sprache (*nouns*) ausgibt, die im Penn Treebank Tagset definiert sind.

```
1 nltk.help.upenn_tagset(r'NN*')
```

```
1 NN: noun, common, singular or mass
2     common-carrier cabbage knuckle-duster Casino afghan shed
          thermostat
3     investment slide humour falloff slick wind hyena
          override subhumanity
4     machinist ...
5 NNP: noun, proper, singular
6     Motown Venneboerger Czestochwa Ranzer Conchita Trumplane
          Christos
7     Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin
          ODI Darryl CTCA
8     Shannon A.K.C. Meltex Liverpool ...
9 NNPS: noun, proper, plural
10    Americans Americas Amharas Amityvilles Amusements
          Anarcho-Syndicalists
11    Andalusians Andes Andruses Angels Animals Anthony
          Antilles Antiques
12    Apache Apaches Apocrypha ...
13 NNS: noun, common, plural
14    undergraduates scotches bric-a-brac products bodyguards
          facets coasts
15    divestitures storehouses designs clubs fragrances
          averages
16    subjectivists apprehensions muses factory-jobs ...
```

Neben der `pos_tag()` Funktion bietet NLTK auch eine Funktion für das POS-Tagging auf der Ebene von Sätzen an (`pos_tag_sents()`). Als Eingabe wird dementsprechend eine Liste bestehend aus Listen von Token erwartet. Im Folgenden werden beide Verfahren kontrastiert um die Funktionsweise und mögliche Probleme besser verstehen zu können. Zur Feststellung der grammatikalischen Wortformen wird ein Abstract aus dem Korpus `SozAbstRaw.pkl` herangezogen. Die Auswahl dieses spezifischen Textes hat keine inhaltlichen Gründe, er eignet sich schlichtweg um einige Eigenheiten des POS-Taggings zu demonstrieren.[22]

Um die notwendige Datenstruktur für `pos_tag_sents()` zu erzeugen, wird auf die Funktion `texttools.sent_word_tokenize()` zurückgegriffen. Diese führt eine zweistufige Tokenisierung eines Textes durch, bei der der rohe Text mittels NLTK Funktionen zuerst in Sätze zerlegt wird, um dann anschließend auf der Ebene von Wörtern tokenisiert zu werden. Das Resultat ist eine Liste von Listen, die Wörter enthalten.

```
1 from texttools import sent_word_tokenize
2
3 text = articles.Abstracts[5124]
4
5 sentences = sent_word_tokenize(text)
6
7 sent_tagger = sum(nltk.pos_tag_sents(sentences), [])
8 word_tagger = nltk.pos_tag(nltk.word_tokenize(text))
```

In Tabelle 5.6 finden sich die Unterschiede in der Zuordnung der grammatikalischen Grundformen für unterschiedliche Ebenen des Taggings. In allen drei Fällen beziehen sich die unterschiedlichen Zuordnungen der grammatikalischen Formen auf denselben Satz:

> Women fare relatively well in the area of access, less well in terms of the college experience, and are particularly disadvantaged with respect to the outcomes of schooling.

Das Wort „Women“ wird vom Satz basierten Tagger als Mehrzahl eines *common noun* (Substantiv) klassifiziert, wohingegen der wortbasierte Al-

[22] Da die Implementation dieses spezifischen und anderer Tagger ständig verfeinert wird, kann es sein, dass sich bei zukünftigen Versionen von NLTK unterschiedliche Ergebnisse zeigen. Die hier verwendete Version war: 3.2.1. Die grundsätzlichen Unterschied im Vorgehen von POS-Tagging auf der Ebene von Sätzen und auf der Ebene einzelner Wörter sollten von zukünftigen Änderungen aber nicht berührt werden.

gorithmus hier den Tag *proper noun, singular* vergibt, der die Einzahlform eines Eigennamen beschreibt. Ebenso wird „fare“ im ersten Fall als Gegenwartsform eines Verbs erkannt, während das Tagging auf der Wortebene zu dem Schluss gelangt, dass es sich hier um ein Substantiv handelt, also eine „Fahrkarte“ oder einen „Tarif“. Zuletzt wird „schooling“ auf der Satzebene als einfaches Substantiv eingeordnet, während es auf der Wortebene als Verlaufsform eines Verbes erkannt wird.

	Sentences	Words
Women	NNS	NNP
fare	VBP	NN
schooling	NN	VBG

Tabelle 5.6: Vergleich von POS-Tags auf Satz- und Wordebene

Hier wird deutlich, dass Algorithmen zur Feststellung der grammatikalischen Wortform sowohl auf die Eigenschaften des Wortes selbst zurückgreifen, als auch auf dessen relative Position zu anderen Worten. Da die `pos_tag()` Funktion den Anfang und das Ende eines Satzes nicht erkennen kann, deutet sie die Großschreibung von „Women“ als ein Zeichen für einen Eigennamen. Aus dem selben Grund wird auch das zusammengesetzte Substantiv „outcomes of schooling“ nicht richtig erkannt, da der Algorithmus in diesem Fall den Satz nicht als eine geschlossene Einheit betrachtet und so nicht erkennen kann, dass es sich hier um eine feste Wendung und das Objekt des Nebensatzes handelt. Auch die Sequenzialität des Vorgehens wird in diesem Beispiel deutlich. Statt „fare“ als das zentrale Verb des Satzes zu identifizieren, nimmt der Algorithmus eine feste Wendung im Sinne eines „Frauentarifs“ an. Dies ist dem Umstand geschuldet, dass die einstufung eines Wortes als Noun die Standardeinstellung fast aller POS-Tagger ist, da davon ausgegangen wird, dass ein unbekanntes Wort am ehesten eine Eigenname ist.

Diese Besonderheiten müssen in der praktischen Anwendung berücksichtigt werden, da die Brauchbarkeit der gewonnen Resultate und deren Weiterverwendung in darauf aufbauender Forschung stark davon abhängig sind. Es empfiehlt sich daher POS-Tagging wenn möglich auf der Ebene von Sätzen durchzuführen, da dies eine korrektere Zuordnung der Tags sicherstellt. Natürlich gilt diese Empfehlung auch für Algorithmen aus anderen Programmmpaketen oder Modulen.

Das Chunking stellt eine Sonderform des POS-Taggings dar. Zunächst werden die grammatikalischen Formen der Wörter bestimmt, um dann in einem nächsten Schritt zu sogenannten *Chunks* zusammengefasst zu werden. Diese Zusammenfassung geschieht jedoch nicht automatisch, sondern setzt die Spezifizierung eines Parsers voraus, der den klassifizierten Satz in eine Baumstruktur überführt. Ein solches Vorgehen wurde bereits in den Ausführungen bezüglich der hierarchischen Struktur von Grammatiken und lexikalischen Daten näher erläutert (siehe Abschnitt 5.3.1). Zum jetzigen Zeitpunkt werden diese Verfahren hauptsächlich im linguistischen Bereich zur Analyse von Satzstrukturen oder in der Informatik zur Analyse von Code eingesetzt. In der sozialwissenschaftlichen Forschung spielt dieses Verfahren bisher nur eine sehr untergeordnete Rolle. Einzig bestimmte Bereiche der Netzwerk-Text-Analyse scheinen davon inspiriert worden zu sein. Daher findet eine eingehendere Betrachtung dieses Verfahrens in Abschnitt 6.4 statt.

5.6 Bestimmung numerischer Eigenschaften

Die Überführung von Textdaten, d.h. Strings oder Listen von Strings, in *numerische Indikatoren* ist die notwendige Voraussetzung für jegliche quantitative Analyse. Im Gegensatz zu Datenbeständen die sich auf Personen beziehen, müssen Textdaten erst mit dem Hinblick auf die zu stellenden Forschungsfragen aufbereitet werden bevor sie in numerische Indikatoren überführt werden können. Entscheidungen für oder gegen die Identifikation von Kollokationen, Stemming oder Lemmatization sowie die Entfernung von Stopwörtern basieren in vielerlei Hinsicht auf den besonderen Merkmalen von Texten. Nach einer Umwandlung in numerische Indikatoren sind viele dieser Verfahren nicht mehr oder nur unter erheblichem zusätzlichen Aufwand durchführbar. Zudem können je nach Art der numerischen Repräsentation Informationen verloren gehen, z.B. die Groß- und Kleinschreibung oder die Reihenfolge der Wörter. Daher ist es grundsätzlich empfehlenswert die numerischen Repräsentationen getrennt vom ursprünglichen Textkorpus zu erstellen.

Im Prinzip lassen sich eine Reihe von Eigenschaften von Texten in numerische Indikatoren überführen. Beispielsweise die Anzahl der Wörter, die Häufigkeit bestimmter Phrasen, die Länge des Textes oder von Teilen davon. Gerade die Vielfalt der Eigenschaften, die extrahiert werden können, sorgt dafür, dass in den meisten Programmpaketen keine vorgefertigten Lösungen enthalten sind. Daraus resultiert die Notwendigkeit

eigene Methoden und Funktionen zu programmieren, um der eigenen Forschungsfrage gerecht werden zu können. Gleichzeitig hat dies auch zur Folge, dass eine ausführliche Behandlung aller erdenklichen Formen von numerischen Repräsentationen hier weder sinnvoll noch möglich ist. Daher konzentriert sich der folgende Abschnitt auf die Feststellung der Häufigkeit von Token und den damit verbundenen Konzepten.

Das heute meist verwendete Modell für die numerische Repräsentation von Textdaten ist das *vector space model* (VSM), bei dem die textuellen Eigenschaften als Vektoren in einem Raum dargestellt werden. Alternative Bezeichnungen dafür sind term-document matrix (bzw. document-term matrix) sowie „bag-of-words". Insbesonders letztere weist daraufhin, dass in dieser Repräsentation die Worte für sich genommen werden, d.h. sowohl die Reihenfolge als auch die Relationen zwischen den Tokens nicht repräsentiert werden. Das Vektorraum-Modell wird auf Gerard Salton (1979) zurückgeführt, der es in seinem Artikel *Mathematics and Information Retrieval* als Erster im Kontext der Verarbeitung von Textdaten beschrieb. Für Sozialwissenschaftler ist dieses Modell vor allem deswegen interessant und anschlussfähig, da es der üblichen Darstellung von Daten in tabellarischer Form entspricht, d.h. die Merkmalsträger (Zeilen) stellen die einzelnen Texte dar, während die Spalten die Häufigkeiten der im Text enthaltenen Wörter (Typen) als Variablen abbilden. Dies hat zur Folge, dass die Standardprogramme und -verfahren der Sozialwissenschaften problemlos auf diese Darstellungsweise angewendet werden können.

Für die Zählung der Wörter wird im Folgenden die Funktion `nltk.FreqDist()` verwendet, die einen tokenisierten Text nimmt und ein geordnetes Diktionär mit den Häufigkeiten zurückgibt. Die List Comprehension produziert eine Liste dieser Diktionäre, die dann in ein Pandas `DataFrame` Objekt umgewandelt wird.

```
1 dtm_counts = (nltk.FreqDist(tokens) for tokens in
      articles.Stopped)
2
3 dtm = pd.DataFrame(dtm_counts)
```

Die Spalten der resultierende Datentabelle entsprechend der alphabetischen Ordnung der Typen. Eine Besonderheit, ist die enorme *Sparsity* der resultierenden Datentabelle. Wie bereits angesprochen ist ein Großteil der resultierenden Zellen nicht besetzt, da die meisten Worte einer Sprache nur sehr selten als Token in spezifischen Texten vorkommen. Um ei-

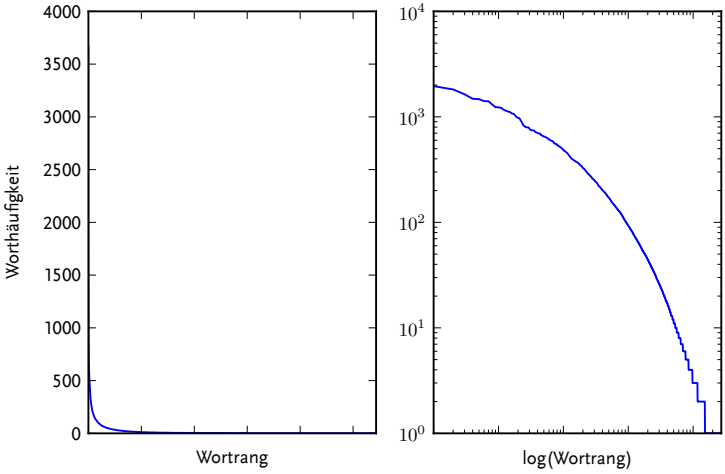

Abbildung 5.1: Beziehung von Wortrang und Worthäufigkeit auf normalen Achsen sowie logarithmierten x und y-Achsen.

nen Überblick über die Verteilung der Worthäufigkeiten im gesamten Korpus zu erhalten, können die Spaltensummen der Datentabelle mittels der `.cum()` Methode betrachtet werden. In der grafischen Darstellung (siehe Abbildung 5.1) wird deutlich, dass es sich um eine sehr extreme Verteilung handelt, bei der eine kleine Mehrheit der Worte sehr häufig vorkommt, während der Großteil der Typen sehr selten auftaucht.

Diese besondere Art der Verteilung, die mit Pareto's Gesetz und der Powerlaw-Verteilung eng verwandt ist, wird als Zipf's Gesetz bezeichnet und scheint ein generelles Merkmal von Texten zu sein, die in natürlichen Sprachen verfasst wurden (vgl. Zipf 1949). Zipf's Gesetz besagt, dass sich die Rangordnung r der Typen eines Textes invers proportional zur Häufigkeit der Typen n verhält, d.h. formal ausgedrückt: $n \propto \frac{1}{r^\alpha}$ (vgl. Moreno-Sánchez, Font-Clos und Corral 2016: 1). Der Parameter α gibt hierbei die Form der Verteilung an. Diese Formulierung zeigt auch, dass die Zipfsche Verteilung der Powerlaw-Verteilung entspricht, die meist in der Form $p(x) \propto x^{-\alpha}$ angegeben wird (vgl. Clauset, Shalizi und Newman 2009: 662). Ob es sich im konkreten Fall um eine Zipfsche Verteilung handelt, kann am einfachsten durch eine Logarithmierung beider Seiten (Häufig-

keiten und Ränge) festgestellt werden. Daraus ergibt sich $\ln n \propto \alpha \ln r$ und damit ein linearer Zusammenhang mit der Steigung α.

Allerdings produziert dieses Vorgehen in den meisten empirischen Fällen keine belastbaren Ergebnisse, da insbesondere der Anfang der Verteilung nur relativ wenige Werte aufweist und sich daher nicht gut zur Schätzung eignet. Dies zeigt sich auch im obigen loglog-Plot, bei dem eine einfache Betrachtung ausreicht, um einer Powerlaw-Verteilung gegenüber skeptisch zu sein. Des Weiteren führt die vorangegangene Entfernung der Stopwörter und die Zusammenfassung der Kollokationen, die in diesem spezifischen Fall zum Einsatz kamen, zu einer weiteren Abweichung von einem linearen Zusammenhang. Grundsätzlich sind für eine Überprüfung von Zipf's Gesetz numerische Verfahren sehr viel besser geeignet (vgl. Clauset, Shalizi und Newman 2009).

Die Passgenauigkeit von Zipf's Gesetz in natürlichen und auch in künstlichen Sprachen, wie zum Beispiel Esperanto, konnte in einer Reihe von Untersuchungen gezeigt werden (vgl. Moreno-Sánchez, Font-Clos und Corral 2016; Manaris et al. 2006). Obwohl Zipf's Gesetz für Texte und eine Reihe von anderen Phänomenen eine gute Beschreibung zu sein scheint, ist die Ursache für diese spezifische Art der Verteilung weitgehend ungeklärt. Zipf (1949) selbst sah die Ursache im psychologischen Prinzip des *least efforts*, welches seiner Ansicht nach einen möglichst komprimierten Wortschatz für die alltägliche Kommunikation begünstigte. Im Zuge der allgemeineren Diskussion der letzten Jahre um Powerlaw-Verteilungen wurden eine Reihe von alternativen Erklärungsvorschlägen vorgebracht. Besonders viel Beachtung wurde dabei Mechanismen wie dem *preferential attachement*, dem *yule process* und der *self-criticality* geschenkt (vgl. Newman 2005). Bis jetzt scheint es jedoch keine einheitliche oder auch nur prinzipiell verallgemeinerbare Antwort auf die Frage nach dem „Warum" des Zipfschen Gesetzes zu geben.

Für die quantitative Analyse von Texten hat Zipf's Gesetz jedoch eine Reihe von forschungspraktischen Implikationen. Zum einen macht es deutlich, dass die oben erwähnte, geringe Besetzung der Zellen der Datentabelle der besonderen Natur von Textdaten geschuldet ist. Zudem lassen sich daraus auch Implikationen für die Analyse von Texten gewinnen. Da der Großteil der Token aus denen Texte bestehen recht häufig in allen Texten vorkommt, eignen sich diese kaum um zwischen Texten zu unterscheiden oder Rückschlüsse auf den Inhalt zu ziehen. Somit liefert Zipf's Gesetz die theoretische Begründung für die Entfernung von Stopwörtern, da von diesen keine hohe Aussagekraft bezüglich der jeweiligen Texte zu

erwarten ist. Zugleich weist die Zipfsche Verteilung aber auch auf Typen hin, von denen wir einen enorm hohen Beitrag zur Unterscheidbarkeit von Texten erwarten können. Dies sind die sogenannten *Hapaxen*, d.h. Typen die in einem bestimmten Korpus nur ein einziges Mal vorkommen. Es ist einfach ersichtlich, dass diese als perfekte Identifikatoren des jeweiligen Textes fungieren. Gleichzeitig weisen sie keinerlei Bezüge zu anderen Texten auf. Daher ist ihre Bedeutung für die *inhaltliche* Analyse von Texten ebenfalls sehr gering.

5.7 Symbolische Komplexität

Die Komplexität von symbolischen Ordnungen spielt in einer Vielzahl soziologischer Theorien eine zentrale Rolle. Auch in diesem Fall findet sich die Unterscheidung von individualistischer und gesellschaftlicher Perspektive wieder. Für eine auf das Individuum und dessen Handlungen ausgerichtete Soziologie ist symbolische Komplexität das Produkt spezifischer Kompetenzen. Eine ähnliche Auffassung vertritt vor allem die psychologische Forschung, in der Textkomplexität zur Bestimmung von sprachlicher Kompetenz genutzt wird. Hierbei wird die Komplexität von Texten als die externe Manifestation von Sprachverständnis, Beherrschung des Wortschatzes und allgemeiner lexikalischer Kompetenz aufgefasst. In der Literatur werden die dazu eingesetzten Verfahren oft unter dem Sammelbegriff der *Lexical Diversity Measures* diskutiert (vgl. McCarthy und Jarvis 2010). Forschungsfragen in diesem Bereich beziehen sich unter anderem auf den Einfluss von Lebensumständen, Bildung, sozialer Herkunft und Mehrsprachlichkeit (vgl. Malvern 2008).

Demgegenüber wird die Komplexität von Texten in der Linguistik und den Sprachwissenschaften hauptsächlich als eine Eigenschaft von Texten, Textgattungen und spezifischen Sprachen aufgefasst. In dieser Perspektive wird die Komplexität des Textes als ein Resultat der Komplexität des Inhalts, bzw. als ein Resultat der Entstehungsbedingungen, aufgefasst. Die entspricht der gesellschaftlichen Sichtweise in der die symbolische Komplexität eine vom Individuum relativ unabhängige Eigenschaft von Kulturen und Artefakten ist. Einsatzbereiche finden sich vor allem in der Evaluierung von Lehrbüchern (vgl. Flory, Phillips Jr. und Tassin 1992), der Verständlichkeit von medizinischen Fachausdrücken für Laien (vgl. Fitzsimmons et al. 2010) und der politischen Bildung (vgl. Mihm 1973). Diese Verfahren werden gemeinhin als *Readability Scores* bezeichnet (vgl. Meade und Smith 1991).

Unabhängig von der weiterführenden Interpretation basieren die Maßzahlen zur Bestimmung textueller Komplexität weitestgehend auf den selben Überlegungen. Grundsätzlich wird versucht die Komplexität eines Textes allein durch die Betrachtung struktureller Eigenschaften zu bestimmen. Dazu zählen in den meisten Fällen die durchschnittliche Länge von Wörtern, Sätzen und Texten, die Anzahl mehrsilbiger Wörter sowie die Relation von Typen und Token.

5.7.1 Messung lexikalischer Diversität

Wie der Name schon nahe legt, bauen die Lexical Diversity Measures auf dem Konzept der lexikalischen Diversität auf, welches in seiner einfachsten Ausprägung der Type-Token-Relation entspricht. Oft wird dieser Begriff auch mit der lexikalischen Diversität gleichgesetzt und mit *LD* abgekürzt. Für einen tokenisierten Text T ergibt sich die Typen-Token-Relation TTR_T als Differenz der Mächtigkeit der Menge der vorkommenden Typen (Wortschatz) W_T und der Anzahl der Token n_T:

$$\text{TTR}_T = \frac{W_T}{n_T}$$

Die resultierende Maßzahl schwankt zwischen 0 und 1. Sie wird dann maximal, wenn jede Type genau einmal als Token im Text vorkommt. Umso öfter Token wiederholt werden, desto kleiner wird die lexikalische Diversität. Damit handelt es sich um ein Maß für die Redundanz der Token eines Textes.

Betrachten wir die Anwendung zunächst am Beispiel einiger klassischer Texte aus dem Gutenberg Korpus (siehe Tabelle 5.7). Hierbei weisen die gesammelten Gedichte von William Blake, mit einem Wert von $0{,}22$, den höchsten Grad an lexikalischer Diversität auf. Anders ausgedrückt, 21% des Textes entfallen auf Token die nur einmal verwendet werden. Die King James Bible bildet hingegen das Schlusslicht mit ca. 1% einzigartiger Token. Es ist wenig überraschend, dass die Spitze der Verteilung von lyrischen Texten dominiert wird. Schließlich liegt der Reiz und die Kunstferigkeit der Poesie ja gerade im Spiel mit multiplen Bedeutungen und eleganten Wortwendungen. Darüber hinaus tragen auch die spezifischen Regeln der Rhythmik und des Reimschema zu einer weniger eineindeutigen Sprache mit einem geringeren Grad an Redundanz bei.

Allerdings wird hier eine problematische Eigenheit von endlichen Zeichensystemen (z.B.: natürlichen Sprachen) deutlich. Mit der Zunahme der

	Type-Token-Relation	Textlänge
Poems by William Blake 1789	0.217860	8354
The Tragedie of Macbeth by William Shakespeare 1603	0.173596	23140
The Tragedie of Hamlet by William Shakespeare 1599	0.145798	37360
The Tragedie of Julius Caesar by William Shakespeare 1599	0.137808	25833
Paradise Lost by John Milton 1667	0.111035	96825
The Man Who Was Thursday by G. K. Chesterton 1908	0.098349	69213
The Wisdom of Father Brown by G. K. Chesterton 1914	0.096429	86063
The Adventures of Buster Bear by Thornton W. Burgess 1920	0.093023	18963
Leaves of Grass by Walt Whitman 1855	0.092515	154883
The Ball and The Cross by G. K. Chesterton 1909	0.092241	96996
Alice's Adventures in Wonderland by Lewis Carroll 1865	0.088420	34110
Stories to Tell to Children by Sara Cone Bryant 1918	0.079549	55563
Moby Dick by Herman Melville 1851	0.074063	260819
Persuasion by Jane Austen 1818	0.062462	98171
Sense and Sensibility by Jane Austen 1811	0.048264	141576
The Parent's Assistant, by Maria Edgeworth	0.045537	210663
Emma by Jane Austen 1816	0.040592	192427
The King James Bible	0.013624	1010654

Tabelle 5.7: Typen-Token-Relation ausgewählter Texte aus dem Gutenberg Korpus.

Token sinkt immer auch das Verhältnis von Typen zu Token. Unter der Bedingung eines endlichen Wortschatzes weisen längere Texte daher immer eine niedrigere Typen-Token-Relation auf. Dieser Umstand ergibt sich aus dem Bezug von Textlänge und Größe des verwendeten Vokabulars, welcher in der Linguistik als *Herdans Gesetz* und im Bereich des Information Retrievals als *Heaps Gesetz* bekannt ist (vgl. Herdan 1960; Heaps 1978). Leo Egghe (2007) zeigte, dass beide Formulierungen mathematisch äquivalent sind und durch eine Formel beschrieben werden können. Angepasst an die hier verwendete Notation ergibt sich folgender Zusammenhang:

$$W_T = k n_T^{\theta}$$

für $k, \theta > 0$ und $\theta < 1$. Demzufolge kann die Anzahl der distinkten Typen W_T eines Textes als eine Funktion der Textlänge n_T beschrieben werden. Bei k und θ handelt es sich um Konstanten, die der Anpassung an den jeweiligen empirischen Fall einer spezifischen Sprache dienen. Unabhängig davon ergibt sich ein assymptotischer Zusammenhang zwischen der Länge des Textes und der Diversität des verwendeten Wortschatzes. Daraus folgt, dass die lexikalische Diversität mit zunehmender Textlänge ebenfalls abnehmen wird. Insofern ist es auch nicht überraschend, dass die King James Bible ein geringere Typen-Token-Ratio aufweist, enthält sie doch ca. 120 mal so viele Token wie die gesammelten Gedichte von Edward Blake.

Gerade bei sehr kurzen Texten, wie zum Beispiel den wissenschaftlichen Abstracts wird dieses Problem sehr deutlich. Der höchste Wert ist bereits sehr nahe am theoretischen Höchstwert von 1. Im Durchschnitt sind somit ca. 62% der Token in den einem Abstract aus dem Soziologie-Korpus einzigartige Wörter.

```
1 articles['TTR'] = articles.Tokens.apply(ttr)
2
3 articles.TTR.describe()
```

```
1 count    5150.000000
2 mean        0.612603
3 std         0.061939
4 min         0.353175
5 25%         0.571429
6 50%         0.609929
```

```
75%         0.652105
max         0.944444
Name: TTR, dtype: float64
```

Eine Betrachtung des Abstracts mit der höchsten Typen-Token-Relation (94,4%) zeigt, dass dies auf den Umstand zurückzuführen ist, dass es sich dabei nur um einen Satz handelt. Der negative Zusammenhang von Type-Token-Ratio und Länge des Textes (Anzahl der Token) wird jedoch auch bei der Betrachtung des gesamten Korpus deutlich. Zeichnet man eine Regressionsgerade in die grafische Darstellung des Verhältnisses von Textlänge und TTR, so weist diese ein klar ersichtliche negative Steigung auf.

```
articles.Abstracts[articles.TTR.idxmax()]
```

```
'Taking contemporary Poland as a case in point, the emerging
    processes of class formation attendant on privatization
    and democratization are examined using traditional
    concepts of class analysis allied to more recent social
    capital theory.'
```

Um dieses Problem in den Griff zu bekommen, wurden eine Reihe von erweiterten Messverfahren für die lexikalische Diversität entwickelt. In den letzten Jahren setzte sich dabei das von Malvern und Richards (1997) entwickelte Maß mit dem Namen *voc-D* zunehmend durch. Bei diesem Verfahren werden wiederholt Stichproben aus dem Text gezogen. Die in den Stichproben enthaltene lexikalische Diversität wird dann mehrfach gewichtet und auf den gesamten Text umgerechnet. Es konnte jedoch gezeigt werden, dass diese Art der Berechnung nicht nur umständlich ist, sondern ab einer bestimmten Textlänge (ca. 250 Wörtern) nicht mehr invariant gegenüber dem Wortschatz ist (vgl. McCarthy und Jarvis 2007).

Der Vollständigkeit halber und um die Logik hinter diesen Verfahren deutlich zu machen, wird hier die von McCarthy und Jarvis (ebd.) vorgeschlagene Alternative *HD-D* kurz erläutert. Bei diesem Vorgehen wird die lexikalische Diversität durch eine wiederholte Stichprobenziehung geschätzt. Die resultierende Maßzahl ist die Summe der hypergeometrisch verteilten Wahrscheinlichkeiten, dass das jeweilige Wort t in der Stichprobe r, welche aus dem Text T gezogen wurde, enthalten ist.

$$\text{HDD} = \sum_{i=1}^{N} \frac{1}{r}(1 - P(t_i|x = 0)), \quad P(t) \sim \text{Hypergeom}(x, r, n, N)$$

Wobei x die Anzahl des Vorkommens von t_i in T_r angibt, hinsichtlich dessen die Wahrscheinlichkeit bestimmt werden soll. Desweiteren sei r die Textlänge der Stichprobe, n die Häufigkeit von t_i in T und N die Größe der Stichprobe r. Unter der Annahme einer hypergeometrischen Verteilung können nun die Wahrscheinlichkeiten für das Auftreten von einer bestimmten Häufigkeit des Ereignisses t_i bestimmt werden. Da jedoch nur das generelle Auftreten eines spezifischen Token von Interesse ist, kann die Wahrscheinlichkeit für mindestens ein t_i zweckmäßiger als die Gegenwahrscheinlichkeit zu $x = 0$ bestimmt werden. Für die Beerchnung von HD-D wird normalerweise eine Stichprobengröße N von 42 Token verwendet, um den Bezug zur ursprünglichen Formulierung voc-D zu erhalten (vgl. McCarthy und Jarvis 2007: 472).

Neben dem Problem der Textlänge wurden weitere Herausforderungen bezüglich der Messung lexikalischer Diversität festgestellt. Dazu zählen vor allem die *Homogenität* des Textes und dessen *sequentielle Ordnung* (vgl. McCarthy und Jarvis 2010: 382). Homogenität nimmt auf den Umstand Bezug, dass Texte oft nicht einem einzigen Genre angehören und auch nicht unbedingt nur einen Autor aufweisen. Dies kann zu sehr unterschiedlichen Stilen und damit einer erhöhten textuellen Diversität führen, die dann jedoch nicht auf die im Text kodierten Bedeutungen oder die Sprachkompetenz des jeweiligen Autors zurückzuführen werden. Demgegenüber deutet Sequentialität auf den Umstand hin, dass ein sich fortwährend entwickelndes Narrativ eine höhere lexikalische Diversität zur Folge hätte, während dies bei Texten ohne Narrativ, zum Beispiel bei Sachtexten oder Bedienungsanleitungen, nicht zwangsläufig der Fall wäre. Die Beeinflussung lexikalischer Diversität durch diese Faktoren wird vor allem deshalb kritisch gesehen, weil sie der Erforschung individueller Merkmale, wie zum Beispiel dem Stil eines Autors oder Sprachkompetenzen, dienen soll. Da diese Merkmale als relativ konstante Eigenschaften des Individuums konzipiert werden, muss ihre Unabhängigkeit von der Textlänge und damit auch vom Genre und den sozialen Regeln des richtigen Schreibens sicher gestellt werden.

Aus diesen Gründen heraus schlagen McCarthy und Jarvis (ebd.: 384f) den *MTLD* Index als ein valideres Maß zur Messung von lexikalischer Diversität vor. Zur Berechnung dieses Index wird die Liste der Token in soge-

nannte Faktoren zerlegt und deren durchschnittliche Länge ermittelt. Ein Faktor wird gebildet indem sukzessive über die einzelnen Elemente t_i für $i = \{1, 2, ...n\}$ des Textes iterieriert wird, bis die Type-Token-Ratio einen vorher festgelegten Schwellenwert fs (als Standardwert wird hier $0{,}72$ vorgeschlagen (ebd.: 384)) erreicht hat. Für die so ermittelte Teilmenge der Token gilt: $\mathrm{TTR}_{[t_i,t_j]} \leq \mathrm{fs}$. Dies wird so lange wiederholt bis sich kein Faktor mehr konstruieren lässt. Für die verbleibenden Wörter, die keinen ganzen Faktor mehr ergeben wird ein *partial factor* ps kalkuliert, der sich aus dem prozentualen Anteil von $\mathrm{TTR}_{\mathrm{ps}}$ an der Spannbreite zwischen dem gewählten Schwellenwert fs und dem theoretischen Maximum von 1 ergibt:

$$\mathrm{ps} = \frac{1 - \mathrm{TTR}_{\mathrm{ps}}}{1 - \mathrm{fs}}.$$

Der MTLD_T Wert eines Textes T ergibt sich dann als die Textlänge n_T geteilt durch die Summe der gebildeten Faktoren $|\mathrm{fs}_T|$ und des verbleibenden Rests ps_T. Um der unterschiedlichen Sequentialisierung von Texten gerecht zu werden, wird der MTLD Wert sowohl für den Text in seiner ursprünglichen als auch in umgekehrter Reihenfolge berechnet und der Durchschnitt der beiden Werte ermittelt:

$$\mathrm{MTLD}_{\overleftrightarrow{T}} = \frac{\mathrm{MTLD}_{\overrightarrow{T}} + \mathrm{MTLD}_{\overleftarrow{T}}}{2} = \frac{\frac{n_T}{|\mathrm{fs}_{\overrightarrow{T}}|+\mathrm{ps}_{\overrightarrow{T}}} + \frac{n_T}{|\mathrm{fs}_{\overleftarrow{T}}|+\mathrm{ps}_{\overleftarrow{T}}}}{2}.$$

Der Python-Code für die Implementierung dieses Indizes, sowie für *HD-D*, würde an dieser Stelle zuviel Platz in Anspruch nehmen. Für den interessierten Leser findet sich der entsprechende Code im Anhang A auf Seite 339.

5.7.2 Messung der Lesbarkeit

Während die Verfahren zur Messung der lexikalischen Diversität hauptsächlich auf die den Texten zugrundeliegenden, sprachlichen Strukturen abzielen, gehen die Verfahren zur Messung der Lesbarkeit von der Schwierigkeit sprachlicher Konstruktionen aus. Dementsprechend sind diese Verfahren viel stärker an die Besonderheiten spezifischer Sprachen und Themen gebunden. Im Gegensatz zur lexikalischen Diversität beruhen sie zudem auf empirischen Beobachtungen und sind daher hauptsächlich für die Textgattungen geeignet an denen sie geeicht wurden. Hier

wird also die Schwierigkeit des Textes als variabel betrachtet, während das Textverständnis oder die sprachliche Kompetenz als konstant angenommen werden.

Trotz der Vielzahl an unterschiedlichen Vorgehensweisen im Bereich der algorithmischen Ermittlung der Lesbarkeit, basieren die einzelnen Verfahren mehr oder minder auf dem selben Grundgedanken. Ausgangslage ist stets eine Menge von Texten deren Lesbarkeit bekannt ist, d.h. entweder gibt es eine *quantifizierbare Einschätzung von Lesern* oder es handelt sich um Texte die ein *bestimmtes Niveau an Lesefähigkeit* vermitteln sollen (z.B.: Schulbücher). Von diesem Korpus werden eine Reihe von sprachlichen Merkmalen erfasst, wie z.B. die durchschnittliche Satzlänge, die dann mittels feststehender Parameter auf die bestehende Lesbarkeitsskala der Texte übertragen werden. Der so erstellte Index weist somit eine nahezu perfekte Übereinstimmung mit dem Korpus auf, aus dem er hervorgegangen ist. Das hat zur Folge, dass man die resultierenden Werte nicht einfach auf andere Sprachen übertragen kann. Die Übertragung auf andere Textgattungen wird zwar in der Literatur nicht kritisch hinterfragt, die Funktionsweise dieser Verfahren lässt hier jedoch auch eine gewisse Skepsis angebracht erscheinen.

Die im Folgenden dargestellten Lesbarkeitsindizes sollen helfen diesen Umstand zu verdeutlichen. Der Name in Klammern gibt die Bezeichnung an unter der die Maßzahl im Modul `texttools.measures` zu finden ist. Die Details der Implementation können ebenfalls im Anhang A auf Seite 339 eingesehen werden. Folgende Notation wird für die Textparameter aller ausgewählter Komplexitätindizes verwendet:

- n_T: Anzahl der Token im Text.
- c_T: Anzahl der Zeichen im Text.
- s_T: Anzahl der Sätze im Text.
- $\bar{s}_T$: Durchschnittliche Satzlänge.
- syl_T: Anzahl der Silben im Text.
- $\mathrm{polysyl}_T$: Anzahl der Worte mit drei oder mehr Silben (Mehrsilber).

Liste der Maßzahlen:

1. *automated readability index* (`ari`):

Entwickelt von Senter und Smith (1967). Bezieht sich auf die Notenstufen des US-Schulsystems und wird ähnlich wie der Coleman-Liau Test nicht über Mehrsilber, sondern über Länge der Zeichenkette berechnet. Daher ist dieser Index verhältnismäßig einfach zu berechnen und weist auch die konsistentesten Werte auf, da seine Definition nicht auf den Regeln der Feststellung von Silben beruht. Der ARI Index kann als eine Schätzung der Jahrgangsstufe des US-Schulsystems aufgefasst werden. Ein Wert von 1 bezieht sich dabei auf Kindergarten oder Vorschule (5.-6. Lebensjahr), während das theoretische Maximum bei 14 (College) liegt. Dezimalstellen müssen dabei stets zur ganzen Zahl aufgerundet werden (10,1 wird zu 11):

$$4{,}71\left(\frac{c_T}{n_T}\right)+0{,}5\left(\frac{n_T}{s_T}\right)-21{,}43$$

2. *Flesch reading ease* (`fre`):

   Entwickelt von Rudolf Flesch (1948) im Kontext militärischer Bedienungsanleitungen und Dienstanweisungen. Ein höherer Flesch Index bedeutet ein leichter zu lesendes Dokument. Von 0-30 gilt ein Dokument als sehr schwer zu lesen, während Dokumente im Bereich von 90-100 als äußerst leicht verständlich eingestuft werden, d.h. als äquivalent zu Textbüchern der 5. Klasse des US-amerikanischen Schulsystems.

$$206{,}835-1{,}015\left(\frac{n_T}{s_T}\right)-84{,}6\left(\frac{\mathrm{syl}_T}{n_T}\right)$$

3. *FRE für Deutsch* (`fre_d`):

   Von Toni Amstad (1978) vorgenommene Anpassung des flesch reading ease Index an die größere durchschnittliche Wortlänge in der deutschen Sprache.

$$180-\left(\frac{n_T}{s_T}\right)-58{,}5\left(\frac{\mathrm{syl}_T}{n_T}\right)$$

4. *Flesch-Kincaid grade level* (`fkg`):

   Weiterentwicklung des FRE Indexes durch Kincaid et al. (1975). Der wesentliche Unterschied besteht darin, dass die resultierenden

Maßzahlen als Notenstufen gemäß des US-Schulsystems interpretiert werden können. Die Lesart entspricht damit der des ARI Index.

$$0{,}39\left(\frac{n_T}{s_T}\right) + 11{,}8\left(\frac{\mathrm{syl}_T}{n_T}\right) - 15{,}59$$

5. *Simple Measure of Gobbledygook* (`smog`):

   Enwickelt von Harry McLaughlin (1969). Wird vor allem in der Analyse der Lesbarkeit von medizinischen Dokumenten eingesetzt und hat sich dort als Goldstandard durchgesetzt (vgl. Davis et al. 1990; Meade und Smith 1991; Fitzsimmons et al. 2010). Der SMOG Index gibt die Lesbarkeit in Jahren des Schulbesuchs an, die für ein Verständnis des Textes notwendig sind.

$$1{,}043\sqrt{\mathrm{polysyl}_T \frac{30}{s_T}} + 3{,}1291$$

6. *Wiener Sachtextformel* (`wst4`):

   Von Bamberger und Vanecek (1984) vorgestellte Lesbarkeitsformel, die spezifisch auf deutschsprachige Sachtexte abgestimmt ist. Daneben existieren eine Reihe von weiteren Formeln, die entweder auf andere Textarten bezogen sind, oder feste Wortlisten verwenden. Da die von Bamberger und Vanecek erhobenen Wortlisten heute nicht mehr ohne weiteres verfügbar sind, wird hier nur die Version verwendet, die ohne externe Informationen auskommt.

$$0{,}2656\bar{s}_T + 0{,}2744\left(\frac{\mathrm{polysyl}_T}{n_T}100\right) - 1{,}693$$

Aufgrund der Funktionsweise dieser Maßzahlen ergeben sich eine Reihe von methodischen Problemen. Wie schon angesprochen, führt die „Maßschneiderung“ auf bestimmte Textsorten zu einer sehr starken Einengung der praktischen Anwendbarkeit. Bestimmte Arten von Textbausteinen, wie zum Beispiel mathematische und chemische Formeln können zudem zu sehr seltsamen Ergebnissen führen. Zum Beispiel wird in einigen Anekdoten von Anwendern davon berichtet, dass chemische Formeln zu hohen negativen Werten bei Indizes führen, die auf Silbenzählung basieren. Dieses Problem kann am Beispiel des FRE index leicht nachvollzogen werden. Das Verhältnis von Anzahl der Wörter (chemische

Elemente) und Silben (Bindungen) in einer chemischen Formel würde den zweiten Term der Gleichung enorm große werden lassen. Auch die Interpretation der jeweiligen Werte kann entsprechend schwierig sein, wenn man nicht mit dem Schulsystem vertraut ist auf das sich Lesbarkeit-Indizes beziehen.

Ein weiteres Problem ist die Definition der zugrundeliegenden Textparameter, die bei genauerer Betrachtung sehr schwer umzusetzen sind. In den meisten Fällen ist einfach von Wörtern, Zeichen oder Silben, die Rede. Wie jedoch bereits im Kapitel über die Tokenisierung dargestellt, erzeugen unterschiedliche Tokenisierungsregeln sehr unterschiedliche Token. Was ein Satz ist, ist ebenfalls eine täuschend simple Frage. Zählen nur Hauptsätze? Was ist mit Satzkonstruktionen die ein Semikolon verwenden? Die Verwendung von unscharf definierten Begriffen ist sicherlich auch dem Umstand zu verdanken, dass die meisten Maßzahlen vor der Verwendung maschineller Textverarbeitung entstanden. Daher sind sie oft nicht angepasst an die exakten Notationen, die für die computergestützte Verarbeitung von Texten notwendig ist.

Eine Übersicht ausgewählter Komplexitäts- und Lesbarkeitsmaße findet sich in Tabelle 5.8. Daran lassen sich die bereits beschriebenen Herausforderungen dieser Maßzahlen noch einmal verdeutlichen. Innerhalb der jeweiligen Gruppe (Komplexität: TTR, HDD, MTLD; Lesbarkeit: ARI, FKG, SMOG) scheint es eine Übereinstimmung in der allgemeinen Tendenz zu geben. Für die Komplexitätsmaße gilt dies insbesondere dann, wenn das TTR-Maß außen vor gelassen wird, da sowohl HDD als auch MTLD in Abgrenzung dazu entwickelt wurden. Die Ähnlichkeiten zwischen den Lesbarkeitsindizes sind hingegen noch prägnanter, weisen aber auch auf ein grundsätzliches Problem hin. Die drei Werke von William Shakespeare, die im Korpus enthalten sind, erhalten sehr niedrige Lesbarkeitswerte. Im Falle von *Julius Caesar* ergibt sich sogar ein negativer ARI Wert. Der Grund dafür liegt in der Definition eines Satzes. Zur Bestimmmung der Anzahl der Sätze werden für gewöhnlich die Satzzeichen gezählt. Dazu gehört auch der Doppelpunkt, der im Drama zur Kennzeichnung des Sprechers verwendet wird. Dies erhöht die Anzahl der Sätze beträchtlich und reduziert damit auch solche Textparameter, wie die durchschnittliche Anzahl von Worten pro Satz. Daher wird hier noch einmal deutlich, dass diese Maßzahlen nur im Kontext der jeweiligen Textgattung Sinn ergeben.

	TTR	HDD	MTLD	ARI	FKG	SMOG
Emma by Jane Austen 1816	0.040592	0.852534	83.407714	2.241120	3.140423	7.442995
Persuasion by Jane Austen 1818	0.062462	0.844237	83.028028	2.729410	3.293418	7.554174
Sense and Sensibility by Jane Austen 1811	0.048264	0.849555	85.250894	3.003209	3.666423	7.810890
The King James Bible	0.013624	0.802139	43.489551	0.317108	0.813708	5.496798
Poems by William Blake 1789	0.217860	0.853047	72.184383	0.391842	0.734942	4.793556
Stories to Tell to Children by Sara Cone Bryant 1918	0.079549	0.834875	59.863013	0.939266	1.370564	5.532821
The Adventures of Buster Bear by Thornton W. Burgess 1920	0.093023	0.845315	69.040256	1.605429	1.942912	5.708533
Alice's Adventures in Wonderland by Lewis Carroll 1865	0.088420	0.828723	64.701360	1.206616	2.034218	6.137913
The Ball and The Cross by G. K. Chesterton 1909	0.092241	0.855142	76.846277	3.237452	3.541300	7.589518
The Wisdom of Father Brown by G. K. Chesterton 1914	0.096429	0.854192	86.657838	3.218667	3.484913	7.502382
The Man Who Was Thursday by G. K. Chesterton 1908	0.098349	0.848863	78.010688	2.907637	3.333609	7.649576
The Parent's Assistant, by Maria Edgeworth	0.045537	0.841840	67.614937	0.957264	1.856505	6.404432
Moby Dick by Herman Melville 1851	0.074063	0.845850	83.755920	2.545585	2.687860	6.986918
Paradise Lost by John Milton 1667	0.111035	0.849484	110.787379	2.292366	1.735177	6.242244
The Tragedie of Julius Caesar by William Shakespeare 1599	0.137808	0.838800	72.042427	-0.072610	0.597847	5.212268
The Tragedie of Hamlet by William Shakespeare 1599	0.145798	0.844249	75.911864	0.076894	0.598222	5.481250
The Tragedie of Macbeth by William Shakespeare 1603	0.173596	0.838165	80.080991	0.284673	0.478772	5.170106
Leaves of Grass by Walt Whitman 1855	0.092515	0.809163	42.469812	1.698947	2.082555	6.581154

Tabelle 5.8: Vergleich von Komplexitäts- und Lesbarkeitsmaßen.

5.8 Texte als methodisch-praktische Herausforderung

Der Vergleich von unterschiedlichen Maßzahlen zur Erfassung der Komplexität symbolischer Ordnungen hilft zwei Besonderheiten der quantitativen Analyse von Texten hervorzuheben, die von allgemeiner Bedeutung für ein Verständnis der besonderen Eigenschaften von Texten sind. Zum einen wird hier noch einmal der janusköpfige Charakter symbolischer Ordnungen deutlich, die sowohl als Repräsentationen kognitiver, wie auch genuin überindividueller Phänomene aufgefasst werden können. Andererseits zeigt sich auch, dass eine Methodologie sozialer Symbole nicht umhinkommt die besonderen Gesetzmäßigkeiten von Texten zu berücksichtigen.

Es muss allerdings auch festgestellt werden, dass diese Doppeldeutigkeit sozialer Symbole eine Funktion spezifischer Perspektiven und nicht der Sprache selbst ist. Man kann lexikalische Diversität als einen Ausdruck von Sprachtalent und Stilistik sehen, kommt jedoch nicht um die Tatsache herum, dass diese trotzdem den allgemeinen Gesetzen der Textproduktion gehorcht und deswegen eben auch als eine Funktion der Schwierigkeit des Textes aufgefasst werden kann. Gleiches gilt für die grammatikalischen Eigenschaften eines Textes, die sowohl als Grammatik einer Sprache, wie auch als Stilistik aufgefasst werden können. Auch hinsichtlich der numerischen Eigenschaften kann dies gezeigt werden. Genau wie die Sprache als Ganzes gehorchen auch die einzelnen Texte Zipfs Gesetz und weisen damit auf die Eigengesetzlichkeit der Sprache hin, die sie aber auch erst zu einem brauchbaren Instrument der Kommunikation und des Wissensaustausches macht.

Die besonderen Verteilungseigenschaften von Texten, die in den Formulierungen von Herdan (1960) und Heap (1978) ebenso zum Ausdruck kommen wie in Zipfs (1949) Gesetz und den informationstheoretischen Ansätzen von Hartley (1928) und Shannon (1948), stellen eine Herausforderung für die Tradition der sozialwissenschaftlichen Datenanalyse dar. Gängige Vorannahmen, wie die Normalverteilung von Eigenschaften und deren grundsätzliche Unabhängigkeit, greifen im Falle von Texten und damit auch für die Erforschung symbolischer Ordnungen oft zu kurz. Stattdessen müssen die Besonderheiten sprachlicher Konstruktionen bei der Anwendung und Entwicklung von Methoden berücksichtigt werden. Auch hinsichtlich der Aufbereitung der Daten und deren Überführung in numerische Indikatoren, kommt man um die Auseinandersetzung mit gram-

matikalischen und sprachlichen Eigenheiten und deren Übersetzung in formale Prinzipien nicht herum.

Eine Aufnahme der quantitativen Textanalyse in den methodischen Werkzeugkoffer der Soziologie kann deswegen nicht einseitig und partiell erfolgen. Ein Übertragung der gängigen Denk- und Arbeitsweisen der empirischen Sozialforschung gelangt hier an ihre Grenzen. Vielmehr bedarf es einer stärkeren Formalisierung der Vorgehensweisen und einem flexiblen Umgang mit den Herausforderungen komplexer Datentransformationen. Auch unsere grundlegendste Perspektive der Trennung von Sozietät und Individuum, von Datenpunkt und Population wird durch die Funktionsweise symbolischer Ordnungen konterkariert. Allerdings, so scheint es, beinhalten diese Herausforderungen aber auch die grundsätzliche Möglichkeit für eine Weiterentwicklung der Soziologie und zur Überwindung überkommener methodischer und theoretischer Paradigmen.

6 Symbolische Strukturen

Soziale Symbole, so die hier vertretene Auffassung, sind sozial standardisierte Zeichen, d.h. physikalische Objekte, die Informationen kodieren können. Die bisherigen Ausführungen bezogen sich jedoch hauptsächlich auf Symbole in der Form von Schriftzeichen. Prinzipiell können die vorangegangenen Betrachtungen zu Texten, Token, Typen und der Messung von Symbolen auf alle sozial standardisierten Zeichen angewandt werden. Man könnte also auch Musik, Bilder, Videoaufzeichungen und eine Reihe von anderen symbolischen Phänomenen hinzuziehen, solange für diese sinnvollerweise angenommen werden kann, dass ihre Elemente in einer vom Zufall unterscheidbaren Weise geordnet sind. Allerdings ist unser gegenwärtiger Stand des Wissens bezüglich nicht-sprachlicher, symbolischer Systeme sehr gering. So etwas wie die Token oder Typen eines Bildes können zum jetzigen Zeitpunkt nicht bestimmt werden, da wir weder das Alphabet noch die Grammatik der *Bildsprache* kennen. Abstrahiert man von diesen konkreten methodischen Problemen, die im Folgenden noch eingehender betrachtet werden, so kann man dennoch sagen, dass die bisherigen Äußerungen ebenso wie deren weitere Ausarbeitung für alle sozialen Symbole Gültigkeit beanspruchen.

In den theoretischen Grundannahmen wurde bereits festgestellt, dass die Symbole selbst nur Indikatoren für diejenigen sozialen Phänomene sind, denen das eigentliche Interesse der sozialwissenschaftlichen Forschung gilt. Gemeint sind damit die *symbolischen Ordnungen*, d.h. eine relativ beständige Verknüpfung von Symbolen oder symbolischen Ordnungen, die in einem konkreten Verhältnis zueinander stehen. Sind diese symbolischen Ordnungen von Menschen internalisiert worden, so ist die Rede von individuellem Wissen. Wenn symbolische Ordnungen auf einer Gruppenebene abgebildet werden, zum Beispiel als eine Sammlung von Texten, die symptomatisch für einen bestimmten Diskurs stehen, so können symbolische Ordnungen als überindividuelle Kultur aufgefasst werden. Beide Perspektiven ändern jedoch zunächst einmal nichts an der grundlegenden Auffassung von symbolischen Ordnungen als einem eigenständigen und objektiven, sozialen Phänomen.

Zwei Merkmale symbolischer Ordnungen wurden bisher nur in theoretischer Hinsicht verhandelt. Dies sind zum einen die *Bedeutungen*, von denen gesagt wurde, dass sie in symbolischen Ordnungen kodiert wären und zwar in den konkreten Relationen aus denen diese symbolischen Ordnungen bestehen. Das zweite Merkmal ist dann auch die Art und Weise, wie diese Relationen zu begreifen wären oder anders ausgedrückt, wie die *Struktur symbolischer Ordnungen* adäquat beschrieben werden kann.

6.1 Relationale Bedeutungen

Bedeutungen spielen sowohl in einer gesellschaftstheoretischen, als auch in einer individualistischen Auffassung sozialer Phänomene eine zentrale Rolle. Beide Perspektiven unterscheiden sich jedoch im jeweiligen Verhältnis zwischen Symbolen und Bedeutungen. In der individualistischen Sichtweise sind die konkret geäußerten Symbole eine Folge individueller Bedeutungskonstruktionen, während sie in der gesellschaftstheoretischen Fassung Repräsentanten übergreifender, sozialer Phänomene und kulturell geteilter Bedeutungen sind. Demgegenüber geht eine prozesstheoretische Sichtweise, ähnlich wie die monadische Interpretation der Semiotik, davon aus, dass symbolische Ordnungen sowohl den Überbau als auch den Unterbau symbolischer Prozesse darstellen. Anders ausgedrückt, das individuelle Denken ist ebenso symbolischer Art wie das überindividuelle Wissen und die Kultur, die jenes Denken überhaupt erst intersubjektiv nachvollziehbar werden lassen. Symbolische Ordnungen entstehen und wandeln sich demzufolge im Wechselspiel mit individuellen und überindividuellen Prozessen. Wie bereits dargelegt, ähnelt dieser Prozess der systemtheoretischen Auffassung, allerdings mit dem entscheidenden Unterschied, dass Symbole hier ein eigengesetzliches, aber nicht ein autopoietisch geschlossenes System darstellen, vielmehr sind sie als eine Form der Interpenetration von Individuen und Gesellschaft aufzufassen.

Zu sagen, dass Symbole Bedeutungen in ihren Relationen kodieren, ist äquivalent zu der Aussage, dass der symbolische Prozess eine *spezifische Struktur* aufweist, die sich pfadabhängig wandelt. Dies führt allerdings zu der zweiten, oben aufgeworfenen Frage, nämlich wie diese Strukturen und damit die darin kodierten Bedeutungen zu modellieren wären. Dass es einen spezifischen und berechenbaren Zusammenhang zwischen Symbolen und dem was sie ausdrücken sollen gibt ist offensichtlich, da jegliche symbolisch vermittelte Kommunikation sonst prinzipiell unmöglich wäre. Allerdings ist auch anzunehmen, dass ein solcher Zusammenhang

in unterschiedlichen Symbolsystemen sehr verschiedene Formen annehmen kann. Dies kann anhand zweier Dimensionen verdeutlicht werden (siehe Tabelle 6.1). Zum einen die Eineindeutigkeit, mit der einem Symbol oder einer symbolischen Ordnung eine spezifische Position in einem Zeichensystem zugewiesen werden kann und zum anderen, wie komplex der symbolische Raum ist, d.h. wie viele unterschiedliche Symbole und Konstruktionsregeln potentiell zur Bildung von Bedeutungen herangezogen werden können.

Eineindeutigkeit kann in diesem Zusammenhang als ein Maß für die Striktheit der Zuordnung von Zeichen und Bedeutungen aufgefasst werden. In einem absolut eineindeutigen Zeichensystem kann jede Bedeutung genau einer symbolischen Ordnung zugewiesen werden und perfekt umkehrbar sein. In der Sprache der Semiotik lässt sich dies als der idealtypische Fall fassen in dem Zeichen, Objekt und Representamen vollkommen deckungsgleich sind. In soziologischer Hinsicht wäre dies gegeben, wenn die in der jeweiligen symbolischen Ordnung kodierte Information bei einem Austausch zwischen sozialen und psychischen Systemen ausgelesen und wieder in eine identische Ordnung rückübersetzt werden kann. Allerdings wird die Eineindeutigkeit von symbolischen Ordnungen hier als ein graduelles Phänomen aufgefasst, demzufolge können konkrete Zeichensysteme mehr oder weniger eineindeutig sein. Natürliche Sprachen weisen im Durchschnitt eine relativ geringe Eineindeutigkeit auf, da unterschiedliche Sachverhalte prinzipiell auf sehr viele verschiedene Arten und Weisen ausgedrückt werden können. Es gibt in ihnen jedoch bestimmte symbolische Ordnungen, wie zum Beispiel Eigennamen oder bestimmte hochsynthetische Konzepte (z.B. Freiheit, Gerechtigkeit, Wahrheit, etc.), die in den meisten Fällen nicht durch andere symbolische Ordnungen ersetzt werden können.

Eine hohe Eineindeutigkeit findet sich vor allem im Falle von formalen Sprachen, Programmiersprachen oder nicht-sozialen Zeichensystemen, wie beispielsweise der DNA. Gerade letzteres Beispiel kann helfen die Konzepte von Eineindeutigkeit und Bedeutung eingehender zu erläutern. DNA kann demnach als ein aus vier Zeichen (den Basen) bestehendes Zeichensystem aufgefasst werden, dass nach bestimmten grammatischen Regeln ausgelesen werden kann um 22 grundlegende Aminosäuren zu erzeugen (vgl. Ambrogelly, Palioura und Söll 2007). Dabei bilden Triplets von Basen (die sogenannten *Codons*) die Wörter, die sowohl die einzelnen Aminosäuren kodieren, als auch die Anweisungen des richtigen Kodierens und Dekodierens enthalten. Rein rechnerisch können mit einem aus

	Niedrige Komplexität	Hohe Komplexität
Niedrige Eineindeutigkeit	Konventionelle Sprachen (z.B.: Alltagssprache)	Abstrakte Sprachen (z.B.: Lyrik)
Hohe Eineindeutigkeit	Konkrete Sprachen (z.B.: DNA)	Formale Sprachen (z.B.: Mathematik)

Tabelle 6.1: Analytische Klassifizierung von Sprachen entsprechend ihrer Eineindeutigkeit und Komplexität.

vier Zeichen bestehenden Alphabet genau 64 Wörter mit drei Zeichen gebildet werden. Dementsprechend ist es nicht verwunderlich, dass es Synonyme im genetischen Code gibt, dass also ein und dieselbe Aminosäure auf verschiedene Arten codiert sein kann. Allerdings sind nicht alle Kodierungen gleich wahrscheinlich, ein Phänomen das als *codon bias* bezeichnet wird und eine Reihe von praktischen Implikationen hat (vgl. Plotkin und Kudla 2011). Für die Eineindeutigkeit des Zeichensystems DNA bedeutet dies jedoch, dass es sich hier nicht um perfekte Eineindeutigkeit handelt, da zwar von einem bestimmten Codon auf eine spezifische Aminosäure geschlossen werden kann, jedoch nicht umgekehrt. Verglichen mit natürlichen Sprachen ist dies jedoch immer noch ein sehr hohes Ausmaß an Eineindeutigkeit.

Die zweite wichtige Dimension zur analytischen Betrachtung symbolischer Ordnungen ist die *Komplexität der Konstruktionsregeln*. Gemeint ist damit die Anzahl der potentiell möglichen Zusammensetzungen, von einzelnen Symbolen oder Zeichen, zu validen Ausdrücken. In diesem Sinne kann zum Beispiel die Komplexität des Zeichensystems DNA als relativ niedrig im Vergleich zu natürlichen Sprache angesehen werden, da die Konstruktionsregeln valide Ausdrücke auf vier Zeichen in linearen Sequenzen begrenzen. Demgegenüber verfügen natürliche Sprachen über einen sehr viel größeres Alphabet und eine Vielzahl von Konstruktionsregeln.

Das Ausmaß der Möglichkeiten zur Formierung gültiger Aussagen ist dabei jedoch von Sprache zu Sprache recht verschieden und unterscheidet sich auch hinsichtlich des Anwendungsbereichs. So weisen technische und fachspezifischer Zeichensysteme oft eine geringere Anzahl von Regeln der richtigen Komposition auf, die dafür jedoch im allgemeinen präziser sind. Texte der Lyrik und Prosa erlangen ihre Wirkmächtigkeit hingegen oft durch eine geschickte Erweiterung des Rahmens der bestehenden Sprachkonvention.

Die Dimensionen der Eineindeutigkeit und Komplexität sind ein hilfreiches Werkzeug um Passgenauigkeit verschiedener Methodologien bezüglich unterschiedliche Texte und Sprachen besser einschätzen zu können. Gleichzeitig dient sie aber auch als Erinnerung daran, dass spezifische Modellierungen von Bedeutungen nicht zwangsläufig auf alle Texte und Textarten anwendbar sind.

6.2 Schlagworte und Information Retrieval

Der grundlegendste Ansatz um Bedeutungen zu modellieren besteht in der Annahme einer Hierarchie von Ausdrücken, bei denen einzelnen Symbolen oder symbolischen Ordnungen eine einzigartige Bedeutung zugewiesen werden kann. Diese Bedeutung wäre dabei wiederum in einem spezifischen Symbol kodierbar. Eine solche Auffassung kommt zum Beispiel in der Benutzung von Schlagworten zur Sortierung und Zuordnung von Texten in Datenbanken zum Ausdruck. Aber auch in alltäglichen Diskursen wird oft davon ausgegangen, dass Worte, Sätze und ganze Texte eindeutige Bedeutungen haben. In gewisser Weise entspricht diese Sichtweise der bereits diskutierten Vorstellung, dass Symbole als Repräsentation von nicht-symbolischen Objekten aufgefasst werden können, wie sie in bestimmten Richtungen der Semiotik und Soziologie anzutreffen ist. Auch das Codieren von qualitativen Interviews geht von einer eben solchen Vorstellung aus.

In der *hierarchischen Beziehung* von Texten und Schlagworten kommt zum Ausdruck, dass der Beitrag den Symbole zum Verständnis der kodierten Bedeutungen leisten können sehr unterschiedlich sein kann. Manche Worte umfassen ganze Themenbereiche und können vielleicht sogar den ganzen Text hinreichend beschreiben, während andere so generischer Art sind, dass sie zwar unverzichtbare Elemente einer Sprache sind, zur Kategorisierung von spezifischen Texten jedoch nicht brauchbar sind. Wie schon in den Ausführungen zur Entfernung von Stopwörtern diskutiert, liegen die Ursachen für die unterschiedliche Bedeutungsstärke von Symbolen zumindest teilweise in den spezifischen Verteilungseigenschaften (z.B.: Zipfs-Gesetz) von Sprachen begründet.

Neben ihrer Orientierungsfunktion für Personen, die nach bestimmten Informationen suchen, bieten Schlüssel- und Stichwörter auch eine Reihe von Ansatzpunkte für eine soziologische Erforschung. Sie beinhalten Informationen über Abwägungs- und Bewertungsprozesse die unter bestimmten Bedingungen dadurch rekonstruiert werden können. Die Ver-

teilung von Schlagwörtern über eine Menge von Texten hinweg kann, gerade im Fall von verschlagworteten Wissensbeständen, als eine Art kognitiver Landkarte aufgefasst werden. Zudem kann in der Hierarchie von Symbolen eine Ursache für die Fokusierung von Aufmerksamkeiten in einem spezifischen Diskurs gesehen werden. Auch in praktisch-methodischer Hinsicht ist eine Beschäftigung mit der Zuweisung und Konstruktion von Schlüsselwörtern von soziologischem Interesse, da sie zu einem besseren Verständnis und einer potentiellen Verbesserung qualitativer Codierstrategien beitragen kann.

Ob sich die Untersuchung von Stichwörtern für eines der oben genannten Ziele eignet, hängt in einem nicht unerheblichem Ausmaß von der Art und Weise ab in der die Schlagworte erzeugt und zugewiesen wurden. Im Prinzip lassen sich hierbei drei Fälle unterscheiden: eine Zuweisung mittels einer *Heuristik*, durch einen *Algorithmus* (im strengen Sinne des Wortes) oder durch eine *Kombination* von beidem. Unter einer Heuristik wird dabei eine auf Erfahrungswissen basierende Regel verstanden. Hierunter fallen also sowohl menschliche Coder, die nach einem persönlichen Erfahrungsschatz operieren, als auch Computerprogramme die in der Lage sind zu „lernen". Bei einem Algorithmus handelt es sich hingegen um eine klar angegebbare, allgemeine Regel, die vom Kontext ihrer Anwendung unabhängig ist. Diese Unterscheidung ist jedoch etwas unscharf, da das *Lernen* selbst, also die Anpassung von zukünftigen Entscheidungen an die gemachten Erfahrungen, als ein deterministischer Algorithmus aufgefasst werden kann. Demnach bezieht sich der Begriff Heuristik nicht auf die Art des Lernens, sondern nur auf den Umstand, dass die Entscheidungen mit Hinblick auf ein gelerntes Wissen erfolgen, d.h. an einem bestimmten Erfahrungshorizont ausgerichtet sind. Die Unterscheidung von Heuristik und Algorithmus soll auch vor Augen führen, dass die konkrete „Maschine" die diese Regel ausführt letztlich nicht ins Gewicht fällt. Jeder Algorithmus kann ebenso von einem Menschen wie von einem Computer ausgeführt werden.[1]

Die hier getroffene Unterscheidung zwischen Heuristik und Algorithmus ähnelt dem Begriffspaar der *Supervised* und *Unsupervised Classification* aus dem Bereich des *Machine Learning*. Regeln zur Klassifikation werden dabei als Supervised Classification bezeichnet, wenn sie an einem Korpus trainiert werden, der bereits eine korrekte Zuordnung von Texten zu Katgorien und Schlüsselwörtern vorgibt (Bird, Klein und Loper 2009: Kap.

[1] Geschichtlich betrachtet ist jedoch die Ausführung von Rechenregeln (Algorithmen) durch Menschen der Normalfall.

6.1). Für den „Lernenden“ muss also eine klare Unterscheidung von Erfolg und Irrtum möglich sein. Im Falle einer Unsupervised Classification existiert diese Rückmeldung jedoch nicht. Stattdessen wird die Zuordnung aufgrund von statistischen und mathematischen Eigenschaften der Texte ermittelt. Auch die Klassifikationen werden in diesem Falle aus dem Text selbst gewonnen, bzw. sind nicht als Attribute des Textes vorgegeben.

6.2.1 Lernende Maschinen

Das überwachte, maschinelle Lernen (Supervised Classification) basiert auf Verfahren, welche die Eigenschaften von Objekten nutzen um diese bestimmten Kategorien zuzuordnen. Als Grundlage dieser Zuordnung dient dabei ein Sammlung von Objekten, deren korrekte Klassifikation bereits bekannt ist. Im Falle einer Supervised Classification von Texten sprich man hier auch von einem Trainings-Korpus. Unabhängig vom konkreten Verfahren, folgt das Vorgehen bei dieser Art der Klassifikation stets einem allgemeinen Schema:

1. *Korpora erstellen*. Um den Trainings-Korpus zu erstellen bedarf es einer Sammlung von Texten, deren Zuordnung zu den zu untersuchenden Kategorien und Schlüsselwörtern bekannt ist. Diese Zuordnung kann zum Beispiel durch eine händische Kodierung gemäß qualitativer Kodierungstechniken erzeugt werden. In diesem Fall würde es sich um die oben erwähnte Verknüpfung von menschlichen und maschinellen Heuristiken handeln. Alternativ kann die Kodierung auch durch Metadaten oder durch den die Texte hervorbringenden Prozess zustande kommen. Dies wäre zum Beispiel der Fall, wenn die politische Ausrichtung des Autors als eine Kategorisierung des Textes aufgefasst würde (vgl. z.B.: Klemmensen, Hobolt und Hansen 2007). Aus dem Trainings-Korpus wird stets eine zusätzliche Stichprobe entnommen (der Test-Korpus), die zur späteren Überprüfung der Klassifikation verwendet wird.

2. *Eigenschaften definieren*. Die Zugehörigkeit eines Textes zu einer Kategorie wird über die Eigenschaften des jeweiligen Textes bestimmt. Welche davon für die Schätzung herangezogen werden muss je nach Fragestellung entschieden werden. Für die Analyse von Texten bieten sich dabei zunächst die einzelnen Token an. Allerdings können prinzipiell alle Eigenschaften, die sinnvoll bestimmt werden können herangezogen werden. Beispiel wären die Eingrenzung auf Typen

einer bestimmten grammatikalischen Form, die Textlänge, geeignete Komplexitätsmaße, etc.

3. *Modell trainieren*. In diesem Schritt wird ein spezifisches Verfahren festgelegt, welches die korrekte Zuordnung von Kategorien zu Texten aus den spezifizierten Eigenschaften lernt. Dazu werden die Eigenschaften des Trainings-Korpus als unabhängige Variablen und die zu wählende Klassifikation als abhängige Variable aufgefasst. Daher muss die Klassifikation auch stets eindimensional sein.

4. *Klassifizieren*. Der durch das trainieren gewonnene *Klassifizierer* kann dann Texte aufgrund der gewählten Eigenschaften den Kategorien zuordnen. Im Prinzip ist dies für jedes Dokument möglich, aus dem die dafür notwendigen Eigenschaften extrahiert werden können. Damit das Modell im nächsten Schritt evaluiert werden kann wird die Klassifizierung zunächst auf den Test-Korpus angewandt.

5. *Modell überprüfen*. Durch die Klassifizierung des Test-Korpus lassen sich die durch den Klassifizierer geschätzten Zuordnungen mit den empirischen Zuordnungen vergleichen. Aus der Abweichung der Schätzung von den tatsächlichen Kategorien können verschiedene Maßzahlen berechnet werden, die eine Einschätzung der Modellgüte erlauben. Aufbauend auf diesen kann der Klassifizierer durch eine Anpassung der Eigenschaftsdefinitionen verbessert werden.

Klassifizierer erstellen

Zur Demonstration des konkreten Vorgehens wird auf den aufbereiteten Korpus der Abstracts soziologischer Fachzeitschriften zurückgegriffen (`SozAbst.pkl`; siehe auch 5.3.2). Dabei stellen die Abstracts den zu klassifizierenden Text und die von den Autoren vergebenen Schlagworte die bestehende Klassifikation dieser Abstracts. Zwei Einschränkungen müssen hierbei vorgenommen werden. Zum einen sind die Schlagworte nicht eindimensional, da sie sich stets auf den ganzen Text beziehen kommt es zu Überschneidungen zwischen den vergebenen Schlüsselwörtern. Zum anderen sind die Schlagworte von sehr unterschiedlichem Synthesegrad. So umfasst zum Beispiel das Schlagwort „Gender“ rein konzeptionel eine größere Anzahl an soziologischen Forschungsvorhaben als „Transgender“, welches als eine Spezifizierung des „Gender“-Themas aufgefasst werden kann.

Dies zeigt sich empirisch auch in einer Betrachtung der Häufigkeiten. Das Schlagwort „Gender“ wurde im gesammten Korpus 181 mal vergeben, während „Transgender“ nur ein einziges Mal vorkommt. Dies ist auch dadurch zu erklären, dass Autoren für spezifischere Konzepte oft unterschiedliche Begrifflichkeiten verwenden, bzw. spezifischere Konzepte schon per Definition auch eine Mehrzahl von Konzepten auf einer vergleichbaren Stufe der Hierarchie der Begrifflichkeiten implizieren. So findet man neben „Transgender“ auch Schlagworte wie „Gender Transition“ oder „Transsexualism“, die zwar unterschiedliche Bedeutungen aufweisen, aber sich in etwa auf einer ähnlichen Synthesestufe befinden. Das es sich hierbei um Unterarten von „Gender“ handelt, sieht man auch daran, dass diese Begriffe tendenziell gemeinsam vorkommen.

Um das Kriterium der Eindimensionalität zu erfüllen bietet es sich an den Klassifizierer nur hinsichtlich des Schlagwortes „Gender“ zu trainieren und alle anderen Abstracts als „nicht-Gender“ zu klassifizieren. „Gender“ bietet sich in diesem Fall aufgrund der relativen Dominanz im Datensatz an, die auch auf eine zentrale Stellung dieses Themas in der Sozialwissenschaften hindeutet. Prinzipiell ließen sich noch andere Schlüsselwörter in die Klassifikation miteinbeziehen, allerdings müsste dazu sichergestellt werden, dass keine Überschneidungen vorhanden sind und sich die Konzepte auf einer ähnlichen Synthesestufe befinden.

```
1 import pandas as pd
2
3 articles = pd.read_pickle('Daten/Soziologie/SozAbst.pkl')
4
5 keywords = articles['Author Keywords']
6
7 keywords = keywords.str.split(r'; *', expand=True)\
8            .unstack().str.lower()
9
10 gendered = keywords[keywords=='gender']\
11                    .index.get_level_values(1)
```

Aus der Klassifikation und den vorhandenen Abstracts kann nun ein Korpus von klassifizierten Texten erstellt werden. Dazu werden zunächst alle Abstracts ausgewählt, an die das Schlagwort „Gender“ vergeben wurde. Darauf aufbauend wird eine neue Variable (`KW_Gender`) konstruiert. Diese erhält den Wert 1 wenn die Autoren dieses Artikel „Gender“ als

Schlagwort ausgewählt haben. Im gegeteiligen Fall wird die 0 vergeben. Dabei muss dafür Sorge getragen werden, dass nur diejenigen Abstracts in den Korpus aufgenommen werden die tatsächlich Schlüsselworte aufweisen.

```
articles['KW_Gender'] = 0
articles.loc[gendered, 'KW_Gender'] = 1

missings = articles['Author Keywords'].isnull()
corpus = articles[['Abstracts', 'KW_Gender']][~missings]
```

Im darauffolgenden Schritt wird diese Textsammlung in einen *Trainings-* und einen *Test-Korpus* zerlegt. Dies muss mit einer Stichprobenziehung geschehen, die eine gleichmäßige Verteilung der Klassifikation über beide Korpora ermöglicht. Pandas `DataFrame` Objekte stellen hierfür die Methode `sample()` bereit, die mittels des Schlüsselarguments `frac=1`[2] gerufen werden kann, um eine komplette Durchmischung des Datensatzes zu erzeugen. Dabei bleibt die Zuordnung von Text und bestehender Klassifikation erhalten.

Nach der Durchmischung muss der Korpus zunächst in eine numerische Repräsentation überführt werden. In diesem Fall wurde dafür die Klasse `CountVectorizer` aus dem `scikit-learn` Modul verwendet. Diese Programmbibliothek ist auf Verfahren des maschinellen Lernens ausgelegt. Der `CountVectorizer` erlaubt die direkte Übersetzung eines Textes der als String repräsentiert ist in eine Dokument-Wort-Matrix. Dabei werden weiterführende Aufgaben wie Tokenisierung und die Entfernung von Stopwörtern ebenfalls von diesem Objekt übernommen. Das instantierte Objekt verfügt über eine `.fit_transform()` Methode, welche die numerische Repräsentation einer Liste von Texten übernimmt. Das Resultat ist eine $n \times m$ Matrix, die in einen Trainings-Korpus ($n = 1500$) und einen Test-Korpus ($n = 806$) aufgeteilt wird.

```
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np

```

[2] `frac` gibt den Anteil der Stichprobe am Gesamtdatensatz an. Ein Wert von eins gibt daher den kompletten Datensatz zurück.

```
pattern = r'(?u)\b[^0-9\W]\w+\b'
stopws = stopwords.words('english'))

vectorizer = CountVectorizer(token_pattern=pattern,
                             stop_words=stopws)

corpus = corpus.sample(frac=1)

X = vectorizer.fit_transform(corpus.Abstracts)
feature_names = np.array(vectorizer.get_feature_names())

X_train, X_test = X[:1500], X[1500:]
```

Analog zu dieser Vorgehensweise wird das zu klassifizierende Merkmal, also die abhängige Variable (y), in der gleichen Art und Weise aufgeteilt:

```
y = corpus['KW_Gender']

y_train = y[:1500]
y_test = y[1500:]
```

Um ein Modell zu bilden wird der Trainingskorpus (`X_train`) zur Schätzung der bereits bestehenden Klassifikation (`y_train`) der Dokumente herangezogen. Dies entspricht im Prinzip dem generellen, sozialwissenschaftlichen Vorgehen bei statistischen Schätzungen. Die einzelnen Textmerkmale, in diesem Fall die Häufigkeiten des jeweiligen Wortes, werden dabei im Sinne von unabhängigen Variablen genutzt und fungieren als Prädiktoren für die Ausprägung einer abhängigen Variablen, der Klassifikation des Textes. Dementsprechend ist dies die Stelle an der theoretische Überlegungen in die Modellbildung einfließen können und auch sollten. Dies gilt insbesondere für die Auswahl der Merkmale und der Verfahren, welche für die Klassifikation herangezogen werden.

Im Folgenden werden zwei Verfahren näher betrachtet, die im Bereich des maschinellen Lernens vor allem für die Analyse von Text eingesetzt werden (vgl. Bird, Klein und Loper 2009: Kap. 6). Diese Verfahren sind dementsprechend in den meisten Programmbibliotheken enthalten (z.B.: NLTK und scikit-learn). Es handelt sich dabei einerseits um die *Naive-Bayessche Klassifikation* unter der Annahme einer multinomialen Verteilung und andererseits um die *logistische Regression*, die im Bereich des

maschinellen Lernens auch als Maximum-Entropie Schätzung bezeichnet wird.

Die Naive-Bayesschen Verfahren gehören zu den ersten automatisierten Klassifikationsverfahren die für Texte entwickelt wurden (z.B.; Lewis und Gale 1994; Friedman 1997). Die Grundannahmen ist, dass die Klassifikation eines Textes als eine bedingte Wahrscheinlichkeit $P(y \mid x_1, \ldots x_n)$ aufgefasst werden kann. Dabei bezieht sich y auf die spezifische Ausprägung der Klassifikationsvariablen, während $x_1, \ldots x_n$ einen Vektor von Textmerkmalen bezeichnet. Mittels des *Satz von Bayes* lässt sich dann die Wahrscheinlichkeit für die Klassifikation eines Textes unter der Bedingung der Merkmale wie folgt bestimmen:

$$P(y \mid x_1, \ldots x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots x_n)}.$$

Die „Naivität“ besteht hierbei in der Annahme der Unabhängigkeit der einzelnen Merkmale voneinander. Dies bedeutet, dass die Wahrscheinlichkeit $P(x_1, \ldots x_n \mid y)$ als das Produkt der Einzelwahrscheinlichkeiten $P(x_i \mid y)$ aufgefasst werden kann. Bezogen auf Texte würde dies bedeuten, dass die Auftretenswahrscheinlichkeit eines bestimmten Wortes vollkommen unabhängig von den anderen Typen ist. Diese Annahme kann zu Recht als naiv bezeichnet werden, sagt aber zunächst einmal nichts über die Passgenauigkeit des Modells aus, welches sich trotzdem in einer Vielzahl praktischer Anwendungen (z.B.: Spam-Filter) bewährt hat.

Unter der Annahme der Unabhängigkeit lässt sich die gesuchte, bedingte Wahrscheinlichkeit wie folgt angeben:

$$P(y \mid x_1, \ldots x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y).$$

Mittels dieses Zusammenhangs lässt sich die Zugehörigkeit eines bestimmten Textes zu einer bestimmten Klasse schätzen. Die spezifischen Verfahren variieren je nachdem welche Verteilungannahme bezüglich der bedingten Wahrscheinlichkeiten der Merkmale $P(x_i \mid y)$ gewählt wird. Wie der Name schon vermuten lässt geht die multinomiale Naive-Bayes Klassifikation von einer multinomialen Verteilung aus.

```
1 from sklearn.naive_bayes import MultinomialNB
2
3 mb_clf = MultinomialNB()
```

```
4 mb_clf.fit(X_train, y_train)
```

```
1 MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

Bei der logistischen Regression wird versucht die Zugehörigkeit zu einer Klasse mittels einer logistischen Verteilung zu schätzen. Aus den möglichen Zufallsverteilungen, welche die empirische Verteilung modellieren können, wird diejenige ausgewählt, die die Entropie maximiert. Konkret bedeutet dies, dass die Wahrscheinlichkeit der Zugehörigkeit eines Merkmalsträgers zur Klasse y folgendermaßen angegeben werden kann:

$$P(y \mid x_1, \ldots x_n) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 \ldots \beta_n x_n}}.$$

Die β-Koeffizienten im Exponenten der eulerschen Zahl e geben das eigentliche Modell wieder. Dies kann als eine verallgemeinerte Form der linearen Regressionsgleichung aufgefasst werden, bei der β_0 den *Achsenabschnitt* (intercept) der Funktion darstellt und $\beta_1, \ldots \beta_n$ die *Koeffizienten*, d.h. den Beitrag des jeweiligen Token x_i zur Steigung der logistischen Kurve. Für sich genommen sind diese Koeffizienten relativ schwer zu interpretieren, da sie gleichermaßen skaliert sein müssten, um vergleichbar sein zu können. Da es jedoch keine „natürliche“ Obergrenze für das Vorkommen eines Wortes in einem Text gibt, bzw. diese aufgrund Zipf's Gesetz nicht für jedes Wort gleich wäre, können die Koeffizienten zunächst nicht direkt interpretiert werden. Dennoch erlauben sie eine ungefähre Einschätzung der relativen Bedeutung eines Merkmals im Gesamtmodell.

```
1 from sklearn.linear_model import LogisticRegression
2
3 lr_clf = LogisticRegression()
4 lr_clf.fit(X_train, y_train)
```

```
1 LogisticRegression(C=1.0, class_weight=None, dual=False,
      fit_intercept=True,
2           intercept_scaling=1, max_iter=100,
                multi_class='ovr', n_jobs=1,
3           penalty='l2', random_state=None,
                solver='liblinear', tol=0.0001,
```

```
4           verbose=0, warm_start=False)
```

Klassifikation evaluieren

Die beiden Klassifikationsobjekte (`mb_clf` und `lr_clf`) enthalten nach der Anwendung der `.fit()` Methode die statistischen Informationen der jeweiligen Modelle. Daher liegt es nahe die β-Koeffizienten der logistischen Regression (`.coef_`) und die logarithmierten Wahrscheinlichkeiten der multinomialen Naive-Bayes Klassifikation (`.feature_log_prob_`) für eine erste Analyse und ein besseres Verständnis des Modells heranzuziehen. Dies ist jedoch mit einer Reihe von Problemen behaftet.

Der folgende Code gibt die zehn höchsten Wahrscheinlichkeitswerte für die Klassifikation eines Textes mit dem Schlagwort „Gender“ an. Aus Gründen der besseren Lesbarkeit wurden die logarithmierten Wahrscheinlichkeiten in ihre Dezimalschreibweise überführt. Dies illustriert sowohl den relativ geringen Beitrag eines spezifischen Merkmals zur Klasssifikation eines Textes, als auch das Problem einer unabhängigen Betrachtung dieser Merkmale. Für sich genommen erscheinen diese Merkmale wenig aussagekräftig, da es sich hier um sehr allgemeine, soziologische Begrifflichkeiten handelt. Die Distinktion zwischen verschiedenen Dokumenten ergibt sich erst durch die Kombination dieser Merkmale im konkreten Fall.

```
1 mb_feats = pd.Series(np.exp(mb_clf.feature_log_prob_[1]),
2                       index=feature_names)
3
4 mb_feats.sort_values(ascending=False).head(10)
```

```
1 women      0.007368
2 gender     0.005315
3 social     0.004832
4 work       0.004671
5 men        0.003382
6 article    0.003060
7 class      0.002738
8 time       0.002617
9 labour     0.002255
```

```
10 family     0.002053
11 dtype: float64
```

Die Koeffizienten der logistischen Regression erscheinen hingegen schon etwas distinkter, da die Worte mit den höchsten β-Koeffizienten einen relativ klaren Zusammenhang mit der Kategorie „Gender“ aufweisen. Allerdings ist auch hier vorstellbar, dass spezifische Begriffe, wie zum Beispiel „relations“, ebenso in anderen Kontexten Verwendung finden. Es wäre also auch hier eine Betrachtung sämtlicher Merkmale nötig, die jedoch Aufgrund der Menge der Eigenschaften ($m = 14760$) nicht möglich ist.

```
1 lr_feats = pd.Series(lr_clf.coef_[0],
2                       index=feature_names)
3
4 lr_feats.sort_values(ascending=False).head(10)
```

```
1 gender          1.057258
2 gendered        0.845434
3 women           0.626044
4 de              0.609770
5 workforce       0.561663
6 female          0.526468
7 feminist        0.509226
8 relations       0.493862
9 reproductive    0.484968
10 girls          0.437823
11 dtype: float64
```

Normalerweise ist die Aussagekraft des Beitrags eines spezifischen Textmerkmals zum Gesamtmodell relativ gering. Die Güte des Gesamtmodells ist davon jedoch nicht zwangsläufig beeinträchtigt. Gerade im Bezug auf die Naive-Bayes Schätzung ist die Güte der Klassifikation sowohl in theoretischer Hinsicht (vgl. Friedman 1997), als auch in einer Vielzahl praktischer Anwendungen bestätigt worden. Dabei ist insbesondere hervorgehoben worden, dass die Probleme der korrekten Schätzung von $P(x_i \mid y)$ sich nicht eins zu eins auf die Modellgüte übertragen, da diese sich hauptsächlich aus der Summe der Merkmale ergibt (vgl. Zhang 2004). Wegen der Menge der zur Schätzung verwendeten Merkmale ist

der Beitrag eines einzelnen Merkmals zum Gesamtmodell relativ gering. Ähnliches gilt auch für die Schätzung mittels einer logistischen Regression.

Die Menge der unabhängigen Variablen verweist jedoch auch auf ein grundlegenderes Problem in der Modellbildung, welches auch als „curse of dimensionality" bezeichnet wird (vgl. Bellman 1957). Im Kontext des Machine Learning wird darunter die zunehmende Schwierigkeit einer treffenden Klassifikation bei *steigender Dimensionalität des Raums der Merkmale* verstanden. Die Unterscheidung zwischen einzelnen Klassen wird schwieriger, da die einzelnen Datenpunkte mit steigender Dimensionalität geringere Distanzen zu einander aufweisen. Dies ist leicht einsichtig, wenn man sich vor Augen führt, dass die meisten Texte nur einen Bruchteil der Worte des gesamten Raums enthalten und an allen anderen Stellen eine Null aufweisen. Mit der Zunahme der Nullstellen, d.h. mit einer Zunahme der Dimensionalität der Beobachtungen, sinkt demzufolge auch die Distanz zwischen den Vektoren, da sie denselben Wert auf den neu hinzugefügten Dimensionen aufweisen, wodurch es schwerer wird sie voneinander zu unterscheiden.

Dieser Effekt wirkt sich jedoch stärker auf die Modellierung des jeweiligen Beitrags der unabhängigen Variablen aus, als auf die klassifikatorische Leistung des Gesamtmodells (vgl. Friedman 1997). Zudem ist es durchaus plausibel anzunehmen, dass die *Interdependenz zwischen Merkmalen* einen mäßigenden Effekt auf den Fluch der Dimensionalität hat. Da die Auswahl der Worte in einem Text eigentlich nie unabhängig von anderen Worten stattfindet, ist es anzunehmen, dass sich gewisse Gemeinsamkeiten in der Wortwahl ergeben, die wiederum die Distanz zwischen unterschiedlichen Texten erhöhen und so die Klassifikation erleichtern. Dennoch stellt der Fluch der Dimensionalität ein zentrales Problem in der Modellierung von Texten dar. Im Kontext von überwachten Klassifikationsalgorithmen kann damit auf verschiedene Art und Weise umgegangen werden (vgl. Zimek, Schubert und Kriegel 2012). Grundlegendstes Mittel zur Verbesserung der Klassifikationsleistung ist dabei die Auswahl der Merkmale die zur Modellbildung herangezogen werden.

Ob eine Veränderung des Modells überhaupt notwendig ist entscheidet sich erst mit der Betrachtung der Klassifikationsleistung. Dazu wird der zu Beginn beiseite gelegte Test-Korpus herangezogen. Durch die Anwendung des Klassifikationsmodell auf diesen Korpus können die geschätzten Zuordnungen mit der bestehenden verglichen werden. Als einfachstes Gütekriterium kann die Schätzung der *Akkuratheit* (accuracy) des Modells

	Multinomial Naive-Bayes	Logistische Regression
Akkuratheit	0.904467	0.913151
Genauigkeit	0.181818	0.500000
Trefferquote	0.028571	0.157143
F1	0.049383	0.239130

Tabelle 6.2: Metriken für Schätzgenauigkeit bei überwachtem Lernen. Vergleich von Multinomial Naive-Bayes und logistischer Regression.

gelten. Die Akkuratheit gibt die relative Anzahl der richtig klassifizierten Merkmale an. Da am Anfang der Modellbildung eine zufällige Zuteilung der Texte in Trainings- und Test-Korpus stattfand, fällt diese Akkuratheit mit jeder neuen Berechnung anders aus.

Die Akkuratheit alleine sagt jedoch noch nichts über die Fehler einer Klassifikation aus. Um einen besseren Überblick über das Modell zu erhalten, ist es ebenfalls notwendig die Menge der *falsch-positiv* und der *falsch-negativ* klassifizierten Fälle hinzu zu ziehen. Daraus ergibt sich die *Genauigkeit* (*Precision*), als das Verhältnis aller Fälle einer Klasse, zur Summe aus der Menge aller Fälle zuzüglich der falsch-positiv Klassifikationen. Mit der Zunahme falsch-positiver Klassifikationen nähert sich die Genauigkeit dementsprechend 0 an. Bei der *Trefferquote* (*Recall*) wird hingegen das Ausmaß der falsch-negativ Fehler herangezogen und ins Verhältnis zur Anzahl der *richtig-positiv* klassifizierten Fälle gesetzt. Ein Wert nahe Null deutet daher auf einen Klassifizierer hin, der kaum in der Lage ist die positiven Fälle richtig zuzuordnen. Im hier betrachteten Fall würde dies bedeuten, dass das Schlagwort „Gender“ vom Algorithmus nicht zugewiesen wurde, obwohl es auf dem ursprünglichen Artikel zu finden war. Kombiniert man Genauigkeit und Trefferquote mittels des harmonischen Mittels, so erhält man den *F1-score*.[3]

Eine Betrachtung der Metriken für die Modellgüte (siehe Tabelle 6.2) macht zweierlei klar. Zum einen ist die Akkuratheit des Gesamtmodells in diesem Fall anscheinend relativ unaussagekräftig. Genauigkeit, Trefferquote und F1-Score deuten draufhin, dass die positiven Werte, also das Schlagwort „Gender“, oft auf die falschen Fälle angewendet wurden. Zum anderen ist das Klassifikationsmodell der logistischen Regression auf den

[3] All diese Metriken und einige weitere sind in `sklearn.metrics` implementiert.

ersten Blick zutreffender, da in diesem Fall zumindest die Quote der falschen Zuweisung des Schlagwortes geringer ist. Insgesamt scheint die logistische Regression in diesem Fall ein besseres Modell zu liefern. In ihrem Vergleich von multinomialer Naive-Bayes Klassifikation und logistischer Regression kommen Ng und Jordan (2002) zu dem Schluss, dass logistische Regressionen vor allem bei großen Datensätzen bessere Prädiktoren für die Klassenzugehörigkeit liefern, während Naive-Bayes Klassifikation bei kleineren Datensätzen tendenziell besser klassifizieren.

Die allgemeine Fehleranfälligkeit beider Modelle bei gleichzeitig hoher Akkuratheit kann auch auf den Umstand zurückgeführt werden, dass die Klassifikationsmerkmale sehr ungleich verteilt sind. Da es sehr viel mehr nicht-„Gender" Artikel gibt, werden diese eher richtig erkannt. Für die vergleichsweise wenigen „Gender"-Artikel, deren korrekte Klassifikation schließlich das Ziel der Modellbildung war, gilt dies jedoch nicht. Eine mögliche Strategie die Modellgüte zu verbessern besteht im künstlichen Herstellen einer Gleichverteilung der Klassen sowie in einer gezielteren Auswahl der Merkmale die für die Klassifikation herangezogen werden. Da es sich bei dem hier verwendeten Schlagwort jedoch nur um ein arbiträr gewähltes Beispiel handelt, welches, wie schon erwähnt, keine richtige Eindimensionalität im Falle der abhängigen Variablen aufweist, sind der Verfeinerung dieses Modells natürliche Grenzen gesetzt.

6.2.2 Möglichkeiten und Grenzen der Klassifikation

Die Verfahren des überwachten Lernens erlauben die Klassifikation von Texten anhand bereits feststehender Bedeutungszuschreibungen. Zu deren Schätzung werden dabei multidimensionaler Merkmal herangezogen. Als Resultat dieser Methodologie erscheinen Bedeutungen stets als eindimensionale Phänomene. Konkret bedeutet dies, dass sich spezifische, von einander abgrenzbare Bedeutungen mit diesen Verfahren sehr gut modellieren lassen.

Ein typisches Beispiel für einen solchen Anwendungsbereich ist die Sentimentanalyse (*sentiment analysis*), welche die Bestimmung der intendierten Richtung von Aussagen zum Ziel hat, d.h. ob Personen sich positiv oder negativ äußern (vgl. Pang, Lee und Vaithyanathan 2002; Nasukawa und Yi 2003). Auch dazu bedarf es eines Korpus, der sowohl Texte zur Extraktion der Merkmale, als auch eine eindimensionale Bewertung des Gegenstandes, auf den sich der Text bezieht, beinhaltet. Aufgrund der modernen Informationstechnologie ist dies bei einer Reihe von prozess-

generierten Daten des World Wide Webs der Fall. Beipiele hierfür sind „like"-Buttons in diversen sozialen Netzwerken oder Online-Reviews von Produkten, wie sie in praktisch allen Online-Kaufhäusern vorhanden sind. Insbesondere Twitter hat sich in den letzten Jahren als primäre Datenquelle und Untersuchungsgegenstand für Sentimentanalysen herauskristallisiert (z.B.: Pak und Paroubek 2010; Agarwal et al. 2011). Besonders hervorzuheben sind die praktischen Anwendungen auf der Basis von Klassifikationstechniken, die in diesem Bereich entstanden sind. So zum Beispiel Systeme zur Unterstützung der Diagnose von suizidalen Tendenzen durch Analyse des Nutzerverhaltens in sozialen Netzwerken (De Choudhury et al. 2016).

Im Falle von Daten die nicht bereits mit einer Klassifikation versehen sind, muss eine Kodierung durch trainierte Personen erfolgen. Hierin liegt ein großes und bisher nicht ausreichend genutztes Potential der hier diskutierten Klassifikationsverfahren zu einer Integration in die Sozialwissenschaften. Wie bereits angesprochen (siehe Abschnitt 4.2.5), ist die Anwendung von quantitativen Verfahren der Textanalyse nicht als ein Ersatz oder eine Bedrohung qualitativer Verfahren zu verstehen. Gerade die klassifikatorischen Ansätze weisen diesbezüglich ein großes integratives Potential auf. Einerseits kann die Zuweisung der Texte zu Klassen von den *Kodiertechniken* der qualitativen Methodenlehre profitieren. Andererseits stellt die Kombination von menschlicher Kodierung und überwachtem Lernen die Möglichkeit einer *expliziteren Modellbildung* bereit und erlaubt eine Überprüfung dieser Modelle anhand verschiedener Testverfahren. Letztlich wird es damit auch möglich, ein durch intensive Lektüre gewonnenes Modell auf größere Textkorpora zu übertragen und es damit in größeren Zusammenhängen testbar zu machen. Eine konsequente Kombination dieser Vorgehensweisen könnte es ermöglichen die oft problematisierten Limitationen der qualitativen Sozialforschung, wie mangelnde Generalisierbarkeit und fehlende Überprüfbarkeit, aufzuheben.

Vor dem Hintergrund der unterschiedlichen Sichtweisen auf soziale Symbole, die in den Sozialwissenschaften dominieren, kann hier festgestellt werden, dass sich die Klassifikationsverfahren auf den ersten Blick insbesondere für eine individualistische Sichtweise auf soziale Symbole eignen. Bei genauerer Betrachtung funktionieren sie jedoch nur unter der Prämisse eines geteilten kulturellen Rahmens, in dem soziale Symbole ihre Bedeutung durch den Gebrauch erhalten. Dies sieht man sowohl an der Modellierung der Merkmale als einem übergreifenden multidimensionalen Raum, als auch in der Messung der abhängigen Variablen, ent-

lang einer Dimension, auf der sich die spezifischen Äußerungen einordnen lassen. Die Unterscheidung von Mikro- und Makroebene hält, zumindest für diese Art der Modellierung sozialer Symbole, keinen zusätzlichen Erkenntnisgewinn bereit. Stattdessen erweist es sich in diesem Fall als zweckmäßig Symbole als einen eigenen Gegenstandsbereich aufzufassen, der sowohl Aussagen über die Ausprägung von Bedeutungen an bestimmten Punkten als auch bezüglich deren Verteilung im Raum zulässt.

Die Verfahren des überwachten Lernens modellieren symbolische Ordnungen und die in ihnen kodierten Bedeutungen als eine Hierarchie. Demzufolge können die Symbole eines Textes zur Bestimmung der spezifischen Ausprägungen einer Bedeutung genutzt werden. Bedeutungen werden dabei als *eindimensionale Variablen* aufgefasst, die zunächst einmal von der Existenz anderer Bedeutungen unabhängig ist. Dies ist durchaus eine plausible Annahme, wenn es sich um einen kurzen Text handelt der auf eine zentrale Aussage hin ausgerichtet ist. In dem Maße, in dem Texte länger werden und die darin kodierten Bedeutungen komplexere, vielschichtigere Strukturen aufweisen, wird eine Modellierung mit Klassifikationsverfahren problematischer. Dies wurde auch in der mangelnden Eindimensionalität des Schlagwortes „Gender" deutlich. Die vergebenen Stichworte weisen eine eigene Struktur auf, bei denen bestimmte Worte in anderen enthalten sind, wie zum Beispiel das Schlagwort „Transgender" welches inhaltlich in der „Gender"-forschung verordnet werden kann. Mit den hier diskutierten Verfahren können diese strukturellen Eigenheiten nicht modelliert werden.

Eine zweite Limitation der Klassifikationsverfahren besteht in dem Umstand, dass die abhängige Variable bereits vorhanden sein muss. Das bedeutet, dass auf bereits feststehende Bedeutungszuweisungen zurückgegriffen werden muss. Entweder in Form von prozessgenerierten Variablen die im Laufe der Erzeugung des Textes festgeschrieben wurden (z.B.: die Vergabe von Sternchen bei Kundenrezensionen) oder durch eine manuelle Kodierungen von Texten. In beiden Fällen hat man es mit einer *vorausgehenden Interpretation* zu tun, die man weder ignorieren noch explizit im Modell berücksichtigen kann. Daraus ergibt sich, dass die Qualität der Klassifikation in entscheidendem Maße von der korrekten Messung dieser Bedeutungszuweisungen abhängig ist. Alle dabei auftretenden methodischen Probleme übertragen sich letztlich auch auf die Erstellung einer überwachten Klassifikation.

6.3 Symbolische Ähnlichkeiten

Der wissenschaftliche Forschungsbereichs des *Information Rerievals* entwickelte sich unter dem Eindruck ähnlicher Fragestellung, wie sie Bereich des maschinellen Lernens vorherrschen. Grob gesagt geht es in beiden Fällen um die Verwaltung von Informationen auf der Grundlage von Bedeutungen. Im Gegensatz zu einer direkten Datenbankabfrage, bei der ein Nutzer spezifische Datenpunkte auswählt, steht dabei das Problem im Vordergrund, wie die Informationen so geordnet werden können, dass eine Suchanfrage aufgrund inhaltlicher Kriterien geschehen kann.

> In the most general case, a retrieval system might be designed to handle any kind of query, and the system might furnish direct replies to such queries. In such question answering, or fact retrieval systems a wide variety of different types of information identifiers may be needed, and the answers may have to be based not only on a deep analysis of each individual information item, but also on general world knowledge and other extraneous factors. (Salton 1979: 1)

In den hier verwendeten Begrifflichkeiten lässt sich dies als ein Problem der Zuordnung von spezifischen symbolischen Ordnungen (z.B.: dem Wissen einer Person) zu einem übergreifenden System symbolischer Ordnungen (z.B.: dem Wissensbestand einer wissenschaftlichen Disziplin) auffassen. Wie schon erwähnt gibt es zunächst keinen Grund anzunehmen, dass diesen beiden Bereiche fundamental inkompatibel wären.

Das grundlegende Vorgehen jeglichen Information Retrievals geht vom Abgleich einer Suchanfrage mit einem Index aus, der die Dokumente hinsichtlich ihrere Relevanz bezüglich einer Suchanfrage ordnet. Dies ermöglicht es diejenigen Dokumente auszuwählen, welche der Suchanfrage am ehesten entspechen. Konkret bedeutet dies auch, dass die Relevanz eines Dokumentes für eine spezifische Suchanfrage eine Funktion der Relation dieses Dokumentes zu allen anderen Texten des Korpus sein muss.

Es gibt eine Vielzahl verschiedener Verfahren, welche der obigen Beschreibung des Information Retrievals gerecht werden. An dieser Stelle konzentrieren wir uns jedoch auf das Vektorraum-Modell (vgl. ebd.: 6ff). Dabei handelt es sich um das wahrscheinlich prominenteste und am weitesten eingesetzte Verfahren des Information Retrievals. In der Tat wird dieser Ausdruck mittlerweile als ein Synonym für den Forschungsbereich selbst verwendet (vgl. Dubin 2004). Zudem passen die Formalisierungen

dieses Ansatzes im weitesten Sinne zur hier diskutierten Methodologie, deren Ziel ja nicht die Beschäftigung mit unterschiedlichen Formen von Suchanfragen, sondern die Modellierung symbolischer Ordnungen als einem sozialen Phänomen ist.

6.3.1 Das Vektorraum-Modell

Wie bereits ausführlich behandelt, werden Texte hier als interdependente Sequenzen von Symbolen aufgefasst. Für die Modellierung der darin enthaltenen symbolischen Ordnungen ist es aber notwendig diese Texte in ihre Attribute und Merkmale zu zerlegen. Dieser Schritt enthält stets einen Verlust von Informationen, da eine Selektion hinsichtlich der Merkmale des Textes stattfinden muss. Um eine Menge von Texten und damit das *System symbolischer Ordnungen* eines bestimmten Diskurses oder Wissensbestandes zu beschreiben, müssen die ausgewählten Textmerkmale als Repräsentationen eines gemeinsamen Vektorraums aufgefasst werden. Das heißt im Prinzip nichts anderes als das angenommen wird, dass alle Texte über die selben Merkmale verfügen können. Formell ausgedrückt gilt daher für ein Dokument d, das einem Korpus $D = \{d_1, d_2, \ldots, d_n\}$ angehört, dass es als ein Vektor seiner Merkmale interpretiert werden kann: $d_i = (a_{i1}, a_{i2}, \ldots, a_{it})$ (vgl. Salton 1979: 7). Verwendet man die Häufigkeit der Token eines Korpus als Merkmale der Texte, so ist das Resultat nichts anderes als die Dokument-Wort-Matrix, deren Grundlage die Transformation von Tokenlisten in Häufigkeiten von Typen ist.

Die Repräsentation einer Sammlung von Texten als einen Raum von Vektoren hat zwei wichtige Konsequenzen. Zum einen hat dies die *Unabhängigkeit* der einzelnen Vektoren voneinander zur Folge. Diese Unabhängigkeit ist dabei eine notwendige Konsequenz der Auffassung von Wissen als einem Vektorraum (vgl. Dubin 2004). Eine solche Feststellung schließt jedoch nicht aus, dass Beziehungsmuster zwischen einzelnen Wörtern und Texten dadurch aufgedeckt werden können. Die zweite Folge ist die Möglichkeit einer Verortung der Texte relativ zueinander, da sie als einzelne Elemente eines *gemeinsamen Vektorraums* aufgefasst werden können.

Das Modell des Vektorraums hat eine Reihe wünschenswerter Qualitäten für Modelle des Information Retrievals:

> Vector space models have attractive qualities: processing vector spaces is a manageable implementational framework, they are mathematically welldefined and understood, and they are intuitively appealing, conforming to everyday metaphors such

> as "near in meaning". In this way, vector spaces can be interpreted as a model of meaning, as semantic spaces. (Karlgren, Holst und Sahlgren 2008: 531)

Auch die hier vetretenen Auffassung symbolischer Ordnungen als Relationen von Symbolen und die in einen übergreifenden Raum (Kultur) eingeordnet sind, ist mit der grundlegende Vorstellung eines Vektorraums in hohem Maße kompatibel. Dementsprechend wären Bedeutungen als Muster in den relativen Positionen der Symbole denkbar.

Es muss jedoch auch betont werden, dass der Fokus des Information Retrievals zunächst nicht auf der Modellierung absoluter Bedeutungen liegt. Vielmehr geht es hier darum eine Anfrage, die als ein Vektor $Q = (q_1, q_2, \ldots, q_t)$ modelliert wird, so in diesen Vektorraum abzubilden, dass eine Rangfolge der Dokumentvektoren erzeugt werden kann, die deren Entfernung zu den Dokumentvektoren entsprechend einer gewählten Metrik wiedergibt. Damit gibt r_i die Ähnlichkeit eines Dokumentvektors d_i mit der Suchanfrage Q wieder: $r_i(Q, d_i)$ (vgl. Salton 1979: 7ff). Je nach gewählter Metrik ergeben sich unterschiedliche Interpretationen der Ähnlichkeit, dennoch können all diese Maße als verschiedene Formen von relativer Distanz zwischen allen Vektoren verstanden werden.

In Python enthält das Modul `scipy.spatial.distance` eine Vielzahl von Distanzmetriken und entsprechende Hilfsfunktionen.[4] Im Folgenden werden die Distanzmetriken der *Kosinus-Ähnlichkeit* (auch *Salton's cosine* gennant) (vgl. ebd.: 8) und der *Jaccard-Index* (im Original: *coefficient de communauté*) (vgl. Jaccard 1912) näher betrachtet.

Der Jaccard-Index misst die Überschneidung zweier Mengen und kann mengentheoretisch als das Verhältnis der Mächtigkeiten von Vereinigungs- und Schnittmenge aufgefasst werden. Daraus ergibt sich auch, dass der Jaccard-Index nur bei absoluten Unterschieden Sinn ergibt und deshalb tendenziell nur auf binäre Vektoren angewendet werden kann.[5] Die Jaccard Ähnlichkeit zwischen einem beliebigen Dokument und der Suchanfrage kann dann angegeben werden, als die Summe des Minimums an der jeweiligen Stelle des Vektors, geteilt durch das Maximum aller Elemente des Vektors, für die gilt $q_t, a_j \geq 0$:

[4] Seinem Namen entsprechend berechnet dieses Modul stets die Entfernung bzw. die Unähnlichkeit von Vektoren.

[5] Tendeziell deshalb, da es stets möglich ist eine diskrete Variable in $k-1$ Dummy-Variablen zu zerlegen und damit in einen binären Vektor zu überführen.

$$J_i(Q, d_i) = \frac{\sum_j \min(q_j, a_{ij})}{\sum_j \max(q_j, a_{ij})}.$$

Ähnlichkeit wird in diesem Sinne als eine Symmetrie der Elemente der Vektoren aufgefasst. Die Jaccard Ähnlichkeit nimmt einen Wert von 1 an, wenn die binären Vektoren vollkommen gleich ausgeprägt sind und wird 0 sobald keine Übereinstimmungen vorliegen.

Die Kosinus-Ähnlichkeit ist im Prinzip nichts weiter als die Verallgemeinerung des Kosinus zu einem Winkelmaß für mehrdimensionale Vektoren. Im zweidimensionalen Raum gibt $\cos(\theta)$ das Verhältnis der am Winkel θ anliegenden Kantenseite mit der Hypothenuse eines rechtwinkligen Dreieicks an. Bei einem Winkel von $0°$ oder $180°$ ergibt sich ein Kosinus von 1 bzw. -1, diese Übereinstimmung in der Orientierung wird als das Ausmaß der Ähnlichkeit interpretiert. Ein rechter Winkel ($90°$ oder $270°$) ergibt hingegen 0 und steht somit für die maximale Unähnlichkeit. Da das *Skalarprodukt* (*inneres Produkt*) zweier Vektoren beliebiger Länge (aber größer als null) folgendermaßen definiert ist: $\vec{a} \cdot \vec{b} = |\vec{a}|\,|\vec{b}|\,\cos(\theta)$, kann der Kosinus zweier Vektoren als das Verhältnis des Skalarprodukts zum Produkt der normierten Länge beider Vektoren aufgefasst werden. Daraus ergibt sich die verallgemeinerte Version als:

$$\cos_i(Q, d_i) = \frac{Q \cdot d_i}{|Q|\,|d_i|} = \frac{\sum_{j=1}^{t} a_{ij} q_j}{\sqrt{\sum_{j=1}^{t} a_{ij}^2}\sqrt{\sum_{j=1}^{t} q_j^2}}.$$

In dieser Form entspricht die Kosinus-Ähnlichkeit dem generalisierten Rangkorrelations-koeffizienten (vgl. Kendall 1944).

Obwohl die Verallgemeinerung des Kosinus mathematisch korrekt ist, ist die Interpretation dieses Maßes in hochdimensionalen Räumen alles andere als intuitiv. Wie schon beim Fluch der Dimensionalität angesprochen, verhalten sich Objekte in hochdimensioneln Räumen oft nicht mehr so wie es deren Gegenstücke in den uns vertrauten drei Dimensionen erwarten lassen. Mit steigender Dimensionalität steigt auch die relative Distanz der Objekte, da schlichtweg „mehr“ Raum vorhanden ist. In Abhängigkeit der Dimensionalität des Raumes kann zum Beispiel eine Kosinus-Ähnlichkeit von 0.707 bzw. ein Winkel von $45°$ ein sehr hohes Ausmaß von Ähnlichkeit bedeuten. Hinzu kommt, dass die konkreten Zusammenhänge zwischen diesen Vektoren, also die Strukturen des semantischen Raums, nicht zwangsläufig auf Clusterbildung oder andere intuitiv erfassbare Strukturen hinauslaufen. Stattdessen deutet einiges auf filamentar-

tige Ketten von Typen hin (vgl. Karlgren, Holst und Sahlgren 2008). Dies kann jedoch als ein empirischer Hinweis auf die spezifischen Interdependenz von Bedeutungen aufgefasst werden.

6.3.2 Berechnung und Gewichtung

Im Information Retrieval wird der Vektor der Anfrage Q für gewöhnlich durch die Eingabe eines Nutzers erzeugt und im Anschluss mit den vorhandenen Dokumentvektoren D verglichen. Aus einer sozialwissenschaftlichen Perspektive sind jedoch die konkreten Beziehungen im empirischen Material von vorrangigem Interesse. Der einfachste Anwendungsfall ist daher die Erzeugung einer symmetrische Distanz- bzw. Ähnlichkeitsmatrix, indem die Distanzmetriken auf alle Vektoren des Vektorraums angewendet werden. Dies erlaubt Aussagen darüber, welche Texte sich ähneln und welche lokalen Muster diese Ähnlichkeiten aufweisen. Eine weitere Möglichkeit der Anpassung dieser Techniken auf sozialwissenschaftliche Fragestellungen besteht in der Konstruktion der Vektoren im Hinblick auf spezifische Fragestellungen. Aber auch das grundlegende Vorgehen des Information Retrievals selbst kann bis zu einem gewissen Grad nutzbar gemacht werden, zum Beispiel durch die theoriegeleitete Konstruktion von Texten, die dann mit empirischen Realisationen verglichen werden können.

Da der Jaccard-Index auf die Betrachtung absoluter Unterschiede, d.h. boolscher Vektoren, ausgelegt ist, eignet er sich zum Beispiel für einen Vergleich der Schlagworte von Texten oder anderweitiger Kodierungen, die als ja-nein Unterscheidungen vorliegen. Dazu wird ein entsprechender Vektorraum konstruiert und in eine Matrix paarweiser Ähnlichkeiten überführt. Dies geschieht mit der SciPy-Funktion `pdist()`, welche unter der Angabe des Vektorraums (als n-dimensionaler Arry) und der jeweiligen Methode (hier `'jaccard'`) die obere Diagonale einer Distanzmatrix produziert. Mittels `squareform()` lässt sich dies in eine vollständige, symmetrische Matrix überführen, die in den meisten Fällen einfacher handzuhaben ist. Diese Matrix wird an jeder Stelle von 1 abgezogen, um daraus eine symmetrische Ähnlichkeitsmatrix zu generieren, welche die Ähnlichkeit aller Texte mit allen Texten angibt.

```
1 from scipy.spatial.distance import pdist, squareform
2 import nltk
3
```

```
keyws = articles['Author Keywords'].dropna()
data = [nltk.FreqDist(keys.lower().split(r'; '))
        for keys in keyws]

kw_dtm = pd.DataFrame(data, index=keyws.index)

sim = 1 - pdist(kw_dtm.fillna(0), 'jaccard')
```

Dabei fällt auf, dass der Großteil (ca. 98%) der möglichen Textpaare keinerlei Ähnlichkeit ($J_i = 0$) aufweisen.

```
np.count_nonzero(sim) / float(len(sim))
```

```
0.02335132531752497
```

Dies ist hier vor allem auf die geringe Anzahl der Schlagworte pro Text in Relation zur Menge aller Schlagworte des Korpus zurückzuführen. Allerdings ist auch bei regulären Textdaten mit einer geringen Dichte des numerischen Vektors zu rechnen, die nicht minder problematisch wäre.

Desweiteren gibt es eine kleinere Gruppe von Textpaaren, die eine perfekte Übereinstimmung ($J_i = 1$) ihrer Schlagworte aufweisen. Eine genauere Betrachtung der entsprechenden Texte zeigte, dass es nur zwei Fälle gibt, in denen eine solche Synchronizität der Schlagworte zustande kommt. Erstens, bei der Wiederveröffentlichung von Artikeln anlässlich eines Jubiläums oder zu Todestagen berühmter Forscher. Zweitens, im Falle von Repliken und Gegenrepliken zu einem bestimmten Artikel, welche dann mit den gleichen Schlagworten gekennzeichnet werden.

Transponiert man den ursprünglichen Vektorraum, so erhält man Vektoren, welche die Verwendung der einzelnen Typen über die Texte hinweg abbildet. Treten zwei unterschiedliche Typen im selben Text als Token auf, so haben sie an derselben Stelle den Wert 1. Dieser Vektorraum wird auch als *co-occurence matrix* bezeichnet. Ermittelt man deren Distanzen, so erhält man einen Überblick darüber, welche Worte ähnliche Funktionen in Texten haben. Da `gender` das meistgebrauchte Schlagwort im Korpus ist, lohnt es sich hier mit einer eingehenderen Betrachtung zu beginnen.

```
sim = 1 - squareform(pdist(kw_dtm.fillna(0).T, 'jaccard'))

```

```
kw_sim_df = pd.DataFrame(sim,
                         columns=kw_dtm.columns,
                         index=kw_dtm.columns)

kw_sim_df['gender'].sort_values(ascending=False)[:10]
```

```
gender         1.000000
class          0.096916
family         0.075829
employment     0.063063
work           0.056338
race           0.051887
ethnicity      0.051402
inequality     0.048673
identity       0.048193
feminism       0.046392
Name: gender, dtype: float64
```

Die obersten zehn Ähnlichkeiten des Schlagwortes „gender“ dürften selbst für Personen die mit dem sozialwissenschaftlichen Diskurs nur oberflächlich vertraut sind wenig überraschend sein. Im Prinzip finden wir hier einen Niederschlag des Forschungsprogramms der „Intersektionalität“. Es zeigt sich also, dass der Begriff „Gender“ in dem hier betrachtenen Diskurs hauptsächlich mit Ungleichheitskonzeptionen (z.B.: „class“, „feminism“ und „inequality“) und den Merkmalen von Individuen (z.B.: „race“, „ethnicity“ und „identity“) in Verbindung steht.

Gleichzeitig wird hier ein Unterschied in der *Synthesestufe von Begrifflichkeiten* deutlich, der wiederum auf die *hierarchische Ausrichtung von Schlagworten* verweist. In abnehmender Reihenfolge umfassen diese Begriffe immer spezifischere Phänomene. Dieser Effekt ist höchstwahrscheinlich auch darauf zurückzuführen, dass bei der Wahl der Schlagwörter zwei Hauptziele miteinander konkurrieren. Einerseits soll ein möglichst großes Publikum über den Inhalt informiert werden und andererseits möglichst präzise auf das spezifische Thema hingewieden werden.

Der Vergleich mit anderen Schlagworten, wie zum Beispiel „class“ weist zudem auf die *Kontextabhängigkeit der Wortwahl* hin. Eine hohe Ähnlichkeit in der Verwendung zwischen zwei Worten („gender“ und „class“) sagt demnach noch nichts über die Verwendung in anderen Kontexten aus.

Stellt man den Ähnlichkeiten des Schlagwortes „gender“ die von „class“ gegenüber, so zeigt sich hier eine stärkere Annäherung an Begrifflichkeiten die zur Beschreibung von gesellschaftlichen Phänomenen verwendet werden (z.B.: „stratification“, „status“ und „individualization“). Somit kann davon ausgegangen werden, dass eine Kombination von „gender“ und „class“ in einer Liste von Schlagworten eine andere Bedeutung kodiert, als zum Beispiel „class“ und „individualization“.

```
kw_sim_df['class'].sort_values(ascending=False)[:10]
```

```
class                1.000000
gender               0.096916
inequality           0.059829
status               0.058824
stratification       0.050505
race                 0.047619
ethnicity            0.046729
habitus              0.045977
individualization    0.044444
employment           0.042373
Name: class, dtype: float64
```

Die Kosinus-Ähnlichkeit betrachtet nicht nur die An- oder Abwesenheit einer Eigenschaft, sondern auch die Häufigkeit des Auftretens. Somit lassen sich Ähnlichkeiten zwischen Texten als eine Funktion des Gebrauchs der Worte abschätzen. Da Worthäufigkeiten in Texten sehr ungleich verteilt sind (Zipfs-Law) und sich die Verwendung der absoluten Häufigkeiten in der Praxis nicht bewährt hat, wird in den meisten Fällen auf eine Gewichtung zurückgegriffen. Das dominante Gewichtungsmaß ist dabei das sogenannte *TFiDF* Maß, dessen Buchstaben für *term frequency inverse document frequency* stehen. Dieses von Karen Spärck Jones (1972) vorgeschlagene Maß gibt die Bedeutung eines Tokens t in einem Text d_i als das Produkt von dessen Häufigkeit $\text{tf}(t, d_i)$ gewichtet mit dem logarithmierten Verhältnis der Anzahl aller Texte $N = |D|$ zur Anzahl der Texte in denen das spezifische Wort vorkommt $n_t = |\{d_i \in D : t \in d_i\}|$. Damit ergibt sich folgende Formel für die Gewichtung:

$$\text{tfidf}(t, d_i, D) = f(t, d_i) \log \left(\frac{N}{n_t} \right).$$

Daraus folgt, dass die TFiDF-Gewichtung mit der Anzahl die ein Wort in allen Texten vorkommt sinkt, während sie in dem Maße ansteigt mit dem dieses Wort auf eine geringe Anzahl von Texten begrenzt ist.

Aufgrund seiner hohen praktischen Relevanz finden sich eine Reihe von Derivaten und Anpassungen dieser Gewichtung (vgl. Paltoglou und Thelwall 2010) sowie eine Implementation in fast allen Programmpaketen, die sich mit der quantitativen Analyse von Texten beschäftigen. Trotz dieser weiten Verbreitung und der erwiesenen Nützlichkeit, fehlt eine solide, theoretische Begründung für die Angemessenheit dieser Gewichtung. Es finden sich eine Reihe von Lösungsvorschlägen, die unter anderem von einer Anlehnung an Zipf's Gesetz, informationstheoretischen Überlegungen oder probabilistischen Modellen ausgehen, jedoch allesamt mit bestimmten Problemen behaftet sind (vgl. Robertson 2004). Die zum Teil sehr unterschiedlichen Vorstellungen bezüglich natürlicher Sprachen machen eine schnelle Lösung dieses Problems sehr unwahrscheinlich.

Da die Berechnung der Kosinus Ähnlichkeit über alle Texte sehr unübersichtlich wäre und um eine Grundlage für spätere Vergleiche mit anderen Verfahren zu erhalten, wird im Folgenden die Kosinus-Ähnlichkeit der gebündelten Abstracts der Soziologiezeitschriften betrachtet. Diese Zusammenfassung der einzelnen Texte kann in zweierlei Hinsicht plausibilisiert werden. Zum einen ist die Entscheidung ein Manuskript einzureichen immer auch von den Themen, die einer bestimmten Zeitschrift zugeschrieben werden gerahmt. Zweitens ist die Veröffentlichung der Texte das Resultat der gemeinsamen Entscheidung der Reviewern und Herausgebern. Diese orientieren sich neben inhaltlichen Qualitätsmerkmalen immer auch an der thematischen Ausrichtung der Zeitschrift. Daher kann davon ausgegangen werden, dass sich diese Rahmenbedingungen der Textproduktion in den veröffentlichten Abstracts niederschlagen und es daher legitim ist Aussagen über die thematische Ausrichtung einer Zeitschrift auf der Basis des zusammengefassten Textmaterials zu treffen.

Die resultierende Dokument-Wort-Matrix entlang der Zeitschriften wurde im Anschluß mit einer TFiDF-Gewichtung versehen. Zur Darstellung der Ähnlichkeiten wurde eine Heatmap (siehe Abbildung 6.1) verwendet, deren unteres Dreieck – d.h. unterhalb der Selbstähnlichkeits-Diagonalen – den Kosinus für die TFiDF-gewichteten Vektoren abbildet. Das obere Dreieck enthält dementsprechend die Werte ohne Gewichtung.

Hier zeigt sich nur ein geringer Unterschied zwischen den gewichteten und den ungewichteten Werten. Die auf TFiDF basierenden Kosinus-Maße sind im allgemeinen etwas höher. Dies ist sehr wahrscheinlich auf

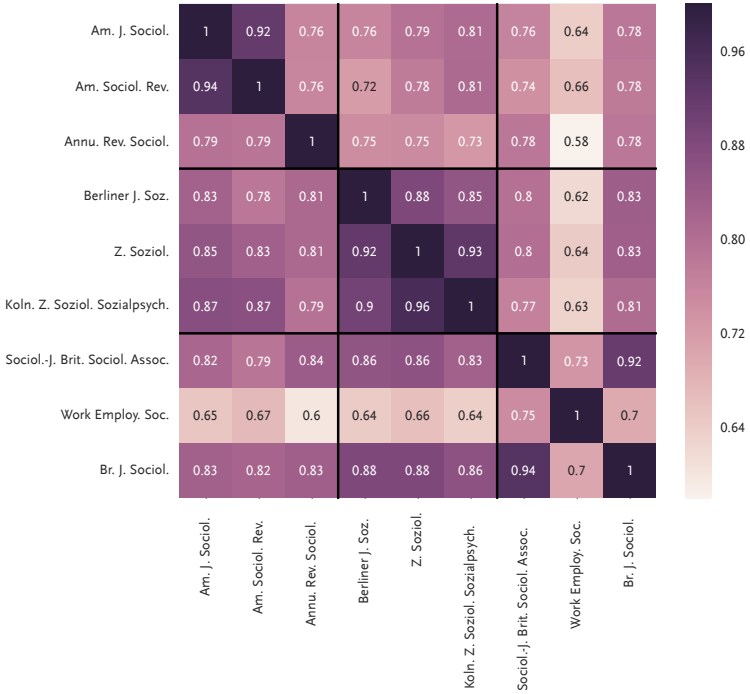

Abbildung 6.1: Kosinus-Ähnlichkeiten von der Abstracts von Soziologie-Zeitschriften als Heatmap. Werte unterhalb der Selbstähnlichkeitsdiagonalen mit TFiDF-Gewichtung.

die Verwendung einer spezifischen Fachsprache zurückzuführen. Fachausdrücke die sich auf bestimmte Forschungsbereiche beziehen haben grundsätzlich eine Tendenz zu mittleren Worthäufigkeiten, weshalb sie ein höheres Gewicht erhalten. Da auch die Soziologie eine international vernetzte Wissenschaft ist, ist es zu erwarten, dass Ähnlichkeiten dadurch prägnanter in Erscheinung treten.

Bei der Betrachtung der Kosinus-Ähnlichkeiten zeigen sich die stärksten Ähnlichkeiten (0,9 und höher) innerhalb der drei Ländergruppen. Allerdings bedeutet dies nicht, dass die Ländergruppen homogen wären, da mit *Annual Review of Sociology* und *Work, Employment & Society* zwei Zeitschriften vorhanden sind, die sich jeweils sehr stark von ihrer Ländergruppe unterscheiden und höhere Ähnlichkeiten zu externen Journals aufweisen. Dabei ist die letztgenannte Zeitschrift auch dadurch gekennzeichnet, dass sie im allgemeinen die geringste Ähnlichkeit zu den anderen Zeitschriften aufweist. Dies kann durch den sehr spezifischen Fokus dieser Zeitschrift auf Probleme der Ungleichheit am Arbeitsmarkt erklärt werden. Des Weiteren zeigt sich, dass jeder der drei Gruppen über einen „Ausreißer" verfügt, der relativ betrachtet weniger Gemeinsamkeiten mit den anderen Gruppenmitgliedern aufweist, als diese untereinander. Bezogen auf die in Deutschland angesiedelte Gruppe von Zeitschriften, kann zudem festgestellt werden, dass hier die größte Homogenität innerhalb einer Gruppe erreicht wird.

6.3.3 Symbolische Vektorräume

In den vorangegangenen Betrachtungen der Techniken und Verfahren des Information Retrievals wurde zunächst einmal deutlich, dass hier im Vergleich zur sozialwissenschaftlichen Analyse von Texten eine große Differenz in der Zielsetzung besteht. Zwar spielt in beiden Bereichen die Frage nach der Ähnlichkeit von symbolischen Ordnungen eine große Rolle, im Information Retrieval wird diese jedoch hinsichtlich extern bereitgestellter Suchkriterien entschieden. Demgegenüber sind sozialwissenschaftliche Analysen stärker an den Mustern im empirischen Material interessiert.

Der enorme Beitrag des Information Retrievals zu einer Methodologie sozialer Symbole liegt in der theoretischen Ausarbeitung des Vektorraummodells und im praktischen Nachweis von dessen Relevanz. Damit wurde die grundsätzliche Möglichkeit geschaffen Bedeutungen und symbolische Ordnungen als Phänomene in einem Vektorraum aufzufassen. Dies

ermöglicht es Konzepte wie *Distanz*, *Synchronizität* und *Orientierung* auf symbolische Phänomene anzuwenden und diese entsprechend zu analysieren. Wie schon in der Aufarbeitung der theoretischen Konzeptionen symbolischer Ordnungen gezeigt wurde, ist die Metapher des Raums zentraler Bestandteil des Großteils dieser Auffassungen. Der große Verdienst des Vektorraummodells liegt in der Entwicklung dieser Metapher hin zu einem eigenständigen Modell.

Aus der Perspektive einer sozialwissenschaftlichen Methodologie zur Untersuchung symbolischer Phänomene gibt es allerdings auch zwei wesentliche Limitationen des ursprünglichen Vektorraummodells. Zum einen die Begrenzung auf den direkten Vergleich von Texten bzw. von Worten. Dementsprechend können Bedeutungen und symbolische Ordnungen nur indirekt als Relationen von Texten bzw. Worten bestimmt werden. Die konkrete Struktur der symbolischen Ordnungen und die damit einhergehende Operationalisierung von Themen, Bedeutungen und dergleichen, kann hiermit zunächst nicht adäquat erfasst werden. Zweitens, bringt die Modellierung von Texten als hochdimensionaler Räume eine Reihe von methodischen Problemen mit sich, dessen zentralstes der Fluch der Dimensionalität ist. Dadurch wird die Bestimmung spezifischer Strukturen symbolischer Ordnungen in einem Vektorraum enorm erschwert und in manchen Fällen auch gänzlich verhindert.

6.4 Netzwerk-Text-Analysen

Die Analyse von symbolischen Ordnungen und Bedeutungen als Netzwerk-Graphen ist eine Erweiterung des vorangegangenen hierarchischen Modells. Wie bei diesem auch, werden Bedeutungen als eine Funktion der Beziehung zwischen konkreten Symbolen aufgefasst. Der wesentliche Unterschied besteht in der Annahme, dass Bedeutungen eine Funktion der konkreten *Beziehungsstrukturen* sind und nicht des Verhältnisses der einzelnen Symbole. Anders ausgedrückt, Bedeutungen werden im Beziehungsgeflecht der Symbole konstruiert und formieren so ein System von Bedeutungen, welches weit über den konkreten Text hinausweist:

> A semantic network system includes not only the explicitly stored net structure but also methods for automatically deriving from that a much larger structure or body of implied knowledge. For example, the assertion in Figure 1 that Toby is hungry implies that he is a conscious animal, and everything

> true of conscious animals is automatically true of Toby. Almost all systems have structured concept-hierarchies or taxonomies used for this kind of derivation [...], and these hierarchies themselves are also 'semantic networks.' (Lehmann 1992: 2)

Im Vergleich zur Auffassung des Textes als einer Hierarchie von Schlagworten, kommt hier eine Annahme größerer Komplexität der Konstruktionsregeln einer Sprache zum Ausdruck. Bedeutung kann demzufolge nicht nur durch spezifische Signalwörter bestimmt werden, sondern ergibt sich im Kontext und in der konkreten Relation zu anderen Symbolen.

Formale Grundlage der Analyse von Texten als Netzwerke ist die Graphentheorie. Ein Netzwerk ist demnach eine Menge von Einheiten, sowie deren Relationen. Formal betrachtet handelt es sich bei einem Netzwerk um einen Graph $\mathcal{G}$ der als ein geordnetes Paar $(\mathscr{N}, \mathscr{L})$ dargestellt werden kann.[6] $\mathscr{N}$ bezeichnet dabei die Menge von Objekten oder Knoten (engl. *Vertices* oder *Nodes*) aus denen der Graph besteht, während $\mathscr{L}$ die Menge der Verbindungen zwischen diesen Objekten (engl. *Edges*) darstellt. Die Menge aller Knoten eines Netzwerkes kann somit als $\mathscr{N} = \{n_1, n_2, \ldots, n_g\}$ angegeben werden, wobei N die Anzahl der Knoten angibt, welche auch als Größenordnung oder Knotenzahl bezeichnet wird. Die Definition für die Menge der Kanten unterscheidet sich je nach Art des Graphen. In einem *ungerichteten* Graphen ist der Kantenzug $\{n_i, n_j\}$ identisch mit der Kante $\{n_j, n_i\}$, wir sprechen dabei von einem ungeordneten Paar. Demgegenüber ist bei einem *gerichteten* Graphen die Reihenfolge der Knoten in einem Kantenzug entscheidend. In diesem Fall werden die Kantenzüge als 2er-Tupel aufgefasst und dementsprechend die formale Schreibweise (n_i, n_j) verwendet. Daneben existieren eine Reihe weiterer Arten von Graphen, die jedoch hier nicht von hervorgehobener Bedeutung sind.

Zur Bearbeitung von Netzwerkdaten in Python finden sich eine Reihe frei verfügbarer Programmbibliotheken, unter denen NetworkX die wahrscheinlich grundlegendste und flexibelste Implementation darstellt.[7] Da NetworkX ausschließlich in Python geschrieben wurde ist die Bedienung für Personen die mit Python vertraut sind relativ einfach und intuitiv. Allerdings geht dies an manchen Stellen auch auf Kosten der Geschwindig-

[6] Für einen Überblick über die Graphentheorie und die soziale Netzwerkanalyse sei Wasserman & Faust (1994) empfohlen.

[7] Die Dokumentation findet sich unter: `https://networkx.readthedocs.io/en/stable/#`

keit. Für die Ansprüche der Netzwerk-Text-Analyse scheint es jedoch vollkommen ausreichend. Zudem spielt die Kompatibilität mit den basalen Python-Typen, wie Strings und Listen, eine wichtigere Rolle für die Transformation, Repräsentation und Analyse von maschinenlesbaren Texten als Netzwerke.

```
1 import networkx as nx
```

Die Betrachtung von Texten als Netzwerk-Graphen beginnt in den 60er Jahren mit dem Versuch Wissensbestände als „semantic networks“ aufzufassen (vgl. Quillian 1967; Brachmann 1979; Lehmann 1992). Bei dieser Vorgehensweise werden semantische Informationen in der Form von Graphen gespeichert, die aus Konzepten und deren jeweiligen Beziehungen bestehen. Aussagen der Form *Subjekt-Prädikat-Objekt* werden dabei als gerichtete Kantenzüge (n_i, n_j, α) geschrieben, bei denen α das jeweilige Prädikat als ein Attribut des Kantenzuges darstellt. Der aus solchen Kanten bestehende Graph wird dann als Repräsentation der Bedeutungen einer bestimmten Wissensordnung aufgefasst. Theoretische wäre es damit möglich semantische Anfragen an eine solche Datenbank zu stellen und eine entsprechende Antwort zu erhalten.

Während die computerwissenschaftliche Beschäftigung mit semantischen Netzwerken hauptsächlich auf deren Einsatz im Bereich formaler Sprachen, bzw. auf die Übersetzung von Wissensbeständen in formale Sprachen abzielt, will die sozialwissenschaftliche Netzwerkanalyse die grundlegenden, *semantischen Strukturen* beschreiben, die in konkreten Texten vorkommen. Dieses Forschungsgebiet entstand Mitte der 90er Jahren in Anlehnung an die bereits erwähnten Arbeiten zu semantischen Netzwerken und in Auseinandersetzung mit der Co-Word Analysis (vgl. Callon, Law und Rip 1986; Leydesdorff 1989). Die frühen Arbeiten der Netzwerk-Text-Analyse zielten auf die Erfassung und die Repräsentation der Relationen von Konzepten und Token, die mittels qualitativer Codierung aus Texten und Transkripten gewonnen wurden (vgl. Carley und Kaufer 1993; Carley 1997). Mit der Ausweitung der Rechenleistung von Computern sowie der breiteren Verfügbarkeit von Verfahren des Natural Language Processing wurde eine Analyse größerer Netzwerke und eine direkte Extraktion der relationalen Daten möglich (vgl. Diesner und Carley 2010).

Grundlage der Netzwerk-Text-Analyse ist die Bestimmung von Objekten und deren konkreten Relationen, die dann in Form eines Netzwerk

Graphens analysiert werden. Je nach Fragestellung, Datenlage und Aufbereitung der Texte sind unterschiedliche Verfahren und Vorgehensweisen notwendig (vgl. ebd.). Trotzdem lässt sich eine Abfolge von einzelnen Schritten feststellen, die allen Verfahren aus diesem Bereich gemeinsam zu sein scheinen. Corman et al. (Corman et al. 2002) diskutieren die Analyse von diskursiven Systemen mittels des Verfahrens der *Centering Resonance Analysis* (CRA), wobei sie dieses Vorgehen in vier Einzelschritte unterteilen: *Selection, Linking, Indexing* und *Application*. Diese Abfolge kann als eine grundlegenden Blaupause für eine netzwerkanalytische Untersuchung von Texten übernommen werden.

Selection (Auswahl) und Linking (Verknüpfung) umfassen die *Konstruktion* von Knoten und Kantenzügen, entsprechend des Forschungsinteresses und der gewählten Perspektive (vgl. ebd.: 174f). Demgegenüber konzentrieren sich die Schritte Indexing (Vermessung) und Application (Anwendung) auf die *Auswertung* des Netzwerkes. Beim Indexing handelt es sich allgemein gesagt um die Berechnung netzwerkanalytischer Maßzahlen über ein Netzwerk bestehend aus Symbolen. Dieser Arbeitsschritt ist der methodologisch aufwendigste, da hierbei die theoretische Vorstellung von Sprache, Symbolen und Bedeutungen in die Maßzahlen und Formalismen der Netzwerktheorie übersetzt werden müssen (vgl. ebd.: 176ff). Darauf folgt mit dem Anwendungsschritt die abschließende Interpretation der resultierenden Netzwerkstruktur (vgl. ebd.: 181).

6.4.1 Konstruktion

Der formale Charakter der Graphentheorie erlaubt es eine Vielzahl von Phänomenen als Netzwerke zu modellieren. Solange sich Objekte identifizieren und direkte Beziehungen zwischen diesen konstruieren lassen, kann die resultierende Struktur als ein Netzwerk aufgefasst und entsprechend analysiert werden. Dies hat zur Anwendung von Netzwerkmodellen auf eine Vielzahl unterschiedlicher Strukturen und Prozesse geführt, wie zum Beispiel Protein-Interaktionen in Hefemolekülen (Jeong et al. 2001), Ausbreitung von Krankheiten (Newman 2002) oder die Kollaborationsbeziehungen zwischen Filmschauspielern (Watts und Strogatz 1998). Ein solches Ausmaß von Interdisziplinarität und Wissenstransfer wurde nur möglich, da Netzwerke ein relativ einfaches, *formales Modell von Struktur* darstellen. Solange sich Knoten und deren Beziehungen feststellen lassen, kann man die Verfahren der Netzwerkanalyse anwenden. Die Abwesenheit von konkreten Festlegungen und ontologischen Vorannahmen hat

aber auch zur Folge, dass die Auswahl von Knoten und Kanten stets vor dem Hintergrund theoretischer Überlegungen stattfinden muss. Die bloße Tatsache, dass ein Netzwerk konstruiert werden kann heißt noch lange nicht, dass es sich zur Modellierung des zu untersuchenden Phänomens eignen würde.

Je nach Fragestellung und vor dem Hintergrund theoretischer Festlegungen, ziehen unterschiedliche Autoren unterschiedliche *Textelemente als Knoten* heran. Carley und Kaufer (1993) sprechen diesbezüglich von „focal concepts", welche sie als übergreifende Konzepte von hoher symbolischer und sozialer Bedeutung beschreiben. In diesem Fall kann eine Identifikation solcher Konzepte zum Beispiel durch ausreichende Übereinstimmung von menschlichen Codern hinsichtlich der Bedeutung dieser Zeichen stattfinden (vgl. ebd.: 197). Die Autoren der Centering Resonance Analysis schlagen einen ähnlichen Weg ein, gehen jedoch von linguistischen Überlegungen aus und definieren „Center" als die entscheidenden Knoten ihres semantischen Netzwerks:

> In CRA, we unitize communication in terms of words contained in the noun phrases that make up utterances. Utterances are sentences or the conversational equivalent thereof [...] and they represent finite groups of centers constructed by communicators to fit into a coherent stream of other utterances. Noun phrases identify the centers, and the words making them up are the codable (linkable) units. (Corman et al. 2002: 173)

Konkret bedeutet dies, dass *Noun Phrases* als die zentralen und bedeutungsstiftenden Elemente von Sätzen angenommen werden. In der anschließenden Konstruktion des Netzwerkes dienen sie als Knoten. Noun Phrases bezeichnet dabei Substantive und Eigennamen zuzüglich der sie beschreibenden Adjektive, falls diese vorhanden sind (vgl. ebd.: 174). Token die eine andere grammatikalische Funktionen erfüllen werden in diesem Verfahren nicht beachtet.

Somit setzt die CRA eine Feststellung der grammatikalischen Formen mittels *POS-Tagging* oder ähnlichen Verfahren voraus (siehe Abschnitt 5.5.2). Allerdings ist dies alleine nicht ausreichend, da Noun Phrases Kombinationen verschiedener Wortformen darstellen und auch mehrfach in einem Satz vorkommen können. Das generelle Verfahren zur Feststellung von festen Wendungen wird im Allgemeinen als *Chunking* bezeichnet und wurde bereits kurz in Abschnitt 5.3.1 angesprochen. Aus Platzgründen kann hier nur auf diejenigen Aspekte des Verfahrens eingegangen wer-

den, die für eine Identifikation von Noun Phrases unerlässlich sind. Für eine ausführlichere Einführung in dieses komplexe und weitreichende Teilgebiet der Computerlinguitik wird Bird et al. (2009: Kap. 7) empfohlen.

Mittels einer Kombination der Funktionen `pos_tag_sents()` und `sent_word_tokenize()` wird im Folgenden das erste Abstract aus dem Soziologiezeitschriften-Korpus in Tokens zerlegt sowie deren grammatikalische Form bestimmt. Das Resultat ist eine Hierarchie von drei Datentypen. Jeder Text eines Abstracts ist eine Liste von Sätzen, die wiederum eine Liste von Tupel beinhaltet. An der ersten Stelle des Tupels findet sich der entsprechende Token, während der jeweilige POS-Tag die zweite Stelle einnimmt.

```
1 from texttools import sent_word_tokenize
2
3 text = articles.Abstracts[0]
4 sents = sent_word_tokenize(text)
5 tagged_sents = nltk.pos_tag_sents(sents)
```

Um die Sätze in *NP-Chunks* zu zerlegen, muss eine reguläre Grammatik konstruiert werden, die eine Regel für die Identifikation der gewünschten Satzteile angibt. NLTK bietet hierfür eine leicht modifizierte Version der Python-Syntax für reguläre Ausdrücke an. Zur Spezifizierung der Wortformen werden dabei die Tags des UPenn Tagsets verwendet. Für Noun Phrases muss eine reguläre Grammatik erstellt werden, die Substantive (`NN`) und Adjektive (`JJ`) umfasst. Mit dieser Grammatik wird daraufhin ein `nltk.RegexpParser` Objekt instantiert, welches bereits markierte Sätze in einzelne Chunks überführt. Die Grammatik des `nltk.RegexpParser` besteht aus einem Namen für den Chunk, in diesem Fall `NP`, gefolgt von einer Beschreibung des Chunks in geschwungenen Klammern. Der Ausdruck `<JJ>*<NN>+` gibt dabei folgende Regel wieder: Ein Chunk umfasst 0 oder mehr Adjektive/Adverben gefolgt von mindestens einem Substantiv. Das Resultat des Parsings ist ein Baumgraph, dessen „Wurzel“ der Satz ist. Die darunterliegenden Ebenen geben die einzelnen „Blätter“ wieder, die durch die reguläre Grammatik beschrieben wurden.

```
1 grammar = 'NP: {<JJ>*<NN>+}'
2
3 parser = nltk.RegexpParser(grammar)
4 tree = parser.parse(tagged_sents[0])
```

Da die weitere Struktur des Satzes für die CRA hier nicht von Bedeutung ist, wird der Baum im Folgenden auf die Noun Phrases eingegrenzt.

```
1 for subtree in tree.subtrees(filter=lambda t:
      t.label()=='NP'):
2     print subtree.leaves()
```

```
1 [('paper', 'NN')]
2 [('social', 'JJ'), ('capital', 'NN'), ('research', 'NN')]
3 [('social', 'JJ'), ('capital', 'NN')]
4 [('social', 'JJ'), ('mobility', 'NN')]
5 [('impact', 'NN')]
6 [('social', 'JJ'), ('trust', 'NN')]
```

Die so identifizierten Noun Phrases können dann als Knoten zur Konstruktion eines semantischen Netzwerks herangezogen werden (*Selection*). Um die Kanten zwischen diesen Knoten zu definieren, muss festgestellt werden, was in diesem Fall eine Beziehung konstituiert. Wie beim Auswählen der Knoten auch, bieten sich hierbei je nach Fragestellung sehr unterschiedliche Vorgehensweisen an. Die wahrscheinlich meistgenutzte Variante des *Linkings* ist das Auftreten im selben Kontext als eine Beziehung zu definieren. Im Falle der CRA wird dafür das gemeinsame Vorkommen innerhalb einer Noun Phrase verwendet. Ausgehend von der Annahme das „Autoren" Äußerungen aus einzelnen Symbolen zusammensetzen, werden alle Kombinationen (ohne Wiederholungen und ohne zurücklegen) zwischen den Token der jeweiligen Phrase als Kantenzüge aufgefasst (vgl. Corman et al. 2002: 175f). Konkret bedeutet dies, dass aus einer Phrase wie „social capital research" drei Kantenzügen gebildet werden: {social, capital}, {social, research} und {capital, research}.

Dieses Konstruktionsprinzip wird auf alle in einem Text vorkommenden Noun Phrases angewendet, um daraus das finale Netzwerk zu konstruieren. Treten Kantenzüge wiederholt auf, so werden diese mit der Anzahl ihres Auftretens gewichtet. Da die Reihenfolge keine Rolle spielt, resultiert dieses Konstruktionsprinzip in einem ungerichteten Graph mit gewichteten Kanten. Für die Konstruktion eines Graphen gemäß dieser Prinzipien wird die Funktion `cra_graph()` in `texttools.nta` bereitgestellt. Sie fasst alle der oben dargestellten Einzelschritte zusammen und

produziert aus einer Liste von Sätzen, die wiederum aus Token mit POS-Tags bestehen, einen NetworkX-Graphen.

```
1 from texttools.nta import cra_graph
2
3 G = cra_graph(tagged_sents)
```

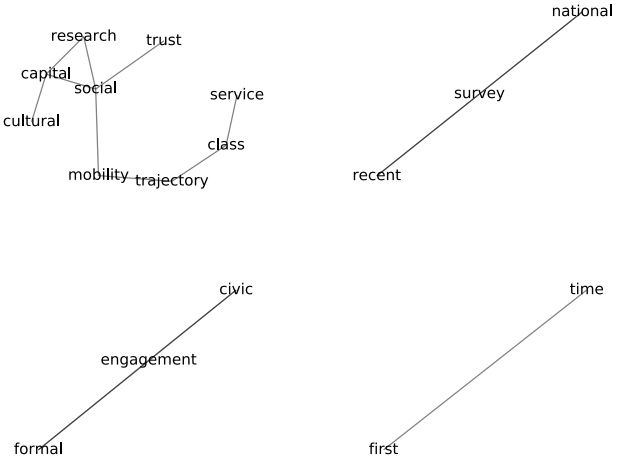

Abbildung 6.2: Darstellung eines Abstracts als Centering Resonance Analysis Graphen.

Der resultierende Graph zerfällt in vier Subgraphen, die in Abbildung 6.2 dargestellt sind. Da es sich hier um einen relativ kurzen Text handelt, ist es nicht verwunderlich, dass der Graph nicht vollständig verbunden ist. Von den vier Subgraphen erscheint vor allem der größte interessant, da es sich hierbei um die zentralen Begrifflichkeiten des Textes handelt, durch die ein bestimmtes Thema abgesteckt wird. Hier deutet das semantische Netzwerk auf Forschungen im Bereich sozialer Mobilität und soziologischer Kapitaltheorien hin. Die anderen Graphen enthalten Worte die

eher spezifische Besonderheiten des Textes zum Ausdruck bringen oder diesen in einem externen Diskurs verorten. Zum Beispiel die Datenbasis, die durch die Knoten „recent“, „national“ und „survey“ abgebildet wird. Das Teilnetzwerk aus „first“ und „time“ markiert hingegen scheinbar das Innovationspotential des Artikels, während „formal“, „civic“ und „engagement“ auf den Gegenstandsbereich oder aber ein Anwendungsfeld hindeuten.

Ein Vergleich mit dem gesamten, zugrundeliegenden Text zeigt, dass die Repräsentation des Textes als Netzwerk zentraler Begrifflichkeiten zwar eine Verkürzung darstellt, aber dennoch die Aussagen des Textes treffend zusammenfasst:

```
1 print text
```

```
1 This paper seeks to contribute to social capital research by
    linking measures of formal and informal forms of social
    capital to social mobility trajectories and assessing
    their impact on social trust. Drawing on data from a
    recent national survey - Cultural Capital and Social
    Exclusion (2003/2004) - we analyse formal civic
    engagement and informal social connections. The latter
    data are obtained using, for the first time in a study
    in Britain, Lin's (2001) 'Position Generator' approach
    as a means to identify the volume, range and position of
    individuals' informal social contacts. The pattern of
    contacts suggests that access to social ties is strongly
    conditioned by mobility trajectory. We also show that
    civic engagement in formal associations is especially
    high among second-generation members of the service
    class. It is also shown that both class trajectory and
    possession of two types of social capital have
    significant impacts on trust. Among the social groups
    disadvantaged in terms of bridging social ties are not
    only those in lower classes but also women and members
    of minority ethnic groups.
```

Da die CRA den internen Zusammenhang von Noun Phrases zur Verlinkung heranzieht, ist sie nur für Sprachen geeignet, die feststehende

Ausdrücke als eine *Kombination von Token* darstellen. Sprachen die zusammengesetzte Begriffe verwenden, wie zum Beispiel das Deutsche können hingegen nicht einfach auf die hier beschriebene Art und Weise in eine Netzwerkstruktur überführt werden. Stattdessen ist es in diesem Fall nötig den Kontext des gemeinsamen Auftretens anzupassen, der zur Konstruktion einer Beziehung im Netzwerk verwendet wird. Denkbare Verwendungskontexte sind hier vor allem textuelle Bedeutungseinheiten wie Satz oder Absatz. Damit wird jedoch auch die zugrundeliegende, theoretische Konzeption maßgeblich verändert. Die Konstruktion eines Netzwerk-Graphen hängt somit ab von der verwendeten Sprache, der Textart sowie dem Wissen um die jeweiligen Praktiken der Textproduktion.

6.4.2 Analyse

Mit der Konstruktion des Netzwerks unter der Anleitung theoretischer Festlegungen ist die Voraussetzung für eine netzwerkanalytische Untersuchung von Texten geschaffen. Die entscheidende Frage ist dabei, wie die formalen Maßzahlen der Netzwerkanalyse auf Symbole und symbolische Ordnungen übertragen werden können. Um dieser Frage nachgehen zu können ist es notwendig sich zunächst vor Augen zu führen, dass alle Netzwerkmaße unabhängig von ihrer Komplexität auf drei grundlegenden Eigenschaften von Netzwerken aufbauen:

- Art und Anzahl der Knoten,
- Art und Anzahl der Kanten,
- Wege durch das Netzwerk, entlang der Kantenzüge.

Daraus ergeben sich alle weiterführenden Überlegungen zur Struktur von Netzwerken. Das Ziel des *Indexings* ist es diese grundlegenden Eigenschaften sowie die daraus folgenden Netzwerkmaße festzustellen.

In ihrer Analyse von semantische Netzwerke aus focal concepts, beziehen sich Carley und Kaufer (Carley und Kaufer 1993: 187ff) auf diese Merkmale und leiten daraus drei analytische Kategorien ab: *Density* (bezogen auf Knoten), *Conductivity* (bezogen auf Pfade) und *Consensus* (bezogen auf Kanten). Unter Density wird dabei die Anzahl der Beziehungen verstanden, die auf ein bestimmtes Konzept verweisen. In Begrifflichkeiten der Netzwerkanalyse entspricht dies dem Knotengrad (*Degree*) eines Knoten. Im Falle von gerichteten Netzwerken handelt es sich dabei um den *Indegree*, zu dessen Berechnung nur die eingehenden Beziehungen

herangezogen werden. Demgegenüber bezeichnet Conductivity die Weiterleitungsfunktion eines Konzeptes oder Symbols, welches als die Anzahl der kürzesten Pfade, die durch ein Netzwerk verlaufen, interpretiert wird. Dies entspricht in graphtheoretischer hinsicht der *Betweenness Centrality*, welche definiert ist als die relative Anzahl der kürzesten Pfade durch einen Knoten (vgl. Freeman, Roeder und Mulholland 1979: 221f). Das Konzept des Consensus weicht in gewisser Weise von den Vorangegangenen ab, da es stärker an der spezifischen Art und Weise ausgerichtet ist, mit der Carley und Kaufer ihre Netzwerke konstruierten. Konsens bezieht sich auf das Ausmaß der Übereinstimmung zwischen Personen bei der Interpretation der Beziehungsstruktur des semantischen Netzwerks (vgl. Carley und Kaufer 1993: 189ff). Da diese Netzwerke durch von Menschen getroffene Entscheidungen zustande kamen, kann die Übereinstimmung unter den Codern hinsichtlich der Menge der Beziehungen als ein Ausmaß des Konsens gelten.

Die CRA nähert sich dem Problem der Übertragung von Netzwerkeigenschaften auf Symbole vor allem über die Idee der Zentralität von Konzepten an:

> To the extent that a CRA network is structured, some words are more influential than others in channeling flows of meaning. They are literally more meaning-full than other words in the network. Thus, identifying the structural influence of the words allows one to measure this property. We operationalize this idea of influence as the centrality of a given word in the CRA network. Although a variety of measures could be used, centering theory points us most clearly toward betweenness centrality. (Corman et al. 2002: 176)

Diese Auffassung entspricht weitestgehend dem Konzept der Conductivity, als einer Weiterleitung semantischer Gehalte durch Worte. Ein solcher „Flow of Meaning" schließt an ein Verständnis von Netzwerken und Zentralitäten als Repräsentation von dynamischen Prozessen an, dem sogenannten „Network Flow" (vgl. Borgatti 2005).

Der hier vertretenen Auffassung von symbolischen Ordnungen als Relationen von Symbolen kommt die Feststellung von Pfaden und Distanzen ebenfalls sehr nahe. Damit ließe sich fassen, warum manche Konzepte stärker mit anderen assoziert sind und inwiefern sich Wissensordnungen in ihrer Struktur ähneln. Allerdings ist eine solche Interpretation auch stark abhängig von der jeweiligen Konstruktionsmethode des Netzwerkes.

Problematisch erscheint hier die Festlegung auf eine Interpretation der Beziehungsnetzwerke als eine Repräsentation des „Flow of Meaning“ bzw. der Konduktivität, da sich beide Konzepte letztlich auf Dynamiken und nicht auf statische Netzwerkstrukturen beziehen. Es ist nicht ohne weiteres ersichtlich, was es ist, dass zwischen den Konzepten fließt. Stattdessen wird im Folgenden die Ansicht vertreten, dass semantische Netzwerke eine Repräsentation der statischen Struktur einer symbolischen Ordnung sind. Die semantische Entfernung würde demzufolge etwas darüber aussagen, wie groß die (Netzwerk-) Distanz zwischen Symbole in bestimmten Verwendungskontexten ist.

In der Graphentheorie bezieht sich Distanz stets auf die geodätische Distanz zwischen Knoten in einem vollständig verbundenen Netzwerkgraphen. Ein Pfad umfasst dabei stets eine Menge von Kantenzügen, die auf dem Weg von n_i zu n_j überquert werden müßen. Der Pfad darf jeden Knoten und jeden Kantenzug dabei nur ein einziges Mal passieren. Dies gilt jedoch nicht für den Start- und den Endknoten, die identisch sein dürfen. Die Anzahl der Kantenzüge ergibt dabei die Länge des Pfades. Als kürzester Pfad (*shortest path*) oder auch geodätische Distanz (*geodesic distance*)[8], wird die minimale Länge aller Pfade d_{ij} zwischen n_i und n_j bezeichnet. Besteht keine durchgehende Verbindung zwischen dem Anfangs- und Endknoten, so geht man von einer unendlich langen Distanz aus.

Somit wären Pfade als *Konstruktionsdistanz* zwischen spezifischen Konzepten aufzufassen. Konkret bedeutet dies, dass die Zusammensetzung von Worten gemäß der Regeln der Sprachpraxis etwas darüber aussagt, wie die spezifischen Worte in einem konkreten Diskurskontext kombiniert werden um Bedeutungen auszudrücken. In dem hier betrachteten Netzwerk (Graph oben links in Abbildung 6.2) ist die direkte Verknüpfung von „social“ und „capital“ relativ entscheidend für die Bedeutung des Textes ($d = 1$). Demgegenüber sind „social“ und „service“ mit vier Kantenzügen relativ weit voneinander entfernt. Obwohl es sich im allgemeinen Sprachgebrauch um eine valide Kombination handelt, scheint sie in diesem spezifschen Text keine Rolle zu spielen.

Die *Betweenness-Zentralität* eines Knoten misst das Ausmaß in dem dieser Knoten ein Teil der kürzesten Pfade zwischen allen anderen Knoten ist, ohne dabei Anfangs- oder Endknoten zu sein (vgl. Freeman, Roeder

[8] Im deutschsprachigen Raum wird oft fälschlicherweise die geodätische Distanz als Bezeichung für die „mittlere Pfadlänge“ zwischen zwei Knoten angegeben, zum Beispiel im gleichnamigen Wikipediaartikel. Dabei handelt es sich anscheinend um eine fehlerhafte Übersetzung eines Textes von Mark Newman (2003), in dem es um die „average geodesic path lenght“ geht.

und Mulholland 1979: 223; Wasserman und Faust 1994: 190).[9] Es sei $\hat{d}_{ij}$ die Anzahl aller kürzesten Pfade zwischen n_i und n_j durch ein Netzwerk. Somit kann die Anzahl der kürzesten Pfade, die den Knoten n_k enthalten, beschrieben werden als: $\hat{d}_{ij}(n_k)$. Unter der Annahmen, dass alle kürzesten Pfade gleich wahrscheinlich sind, ergibt sich die Wahrscheinlichkeit für einen Pfad der n_k enthält in der Relation zu allen möglichen, kürzesten Pfaden:

$$C_B(n_i) = \sum_{i<j}^{g} \hat{d}_{ij}(n_k)/\hat{d}_{ij}.$$

Dieses Maß für die *man-in-the-middle* Postion eines Knotens wird für gewöhnlich auf die theoretische Höchstmenge von kürzesten Pfaden standardisiert, diese wiederum ist abhängig von der Anzahl der Knoten N. Daher ergibt sich die abschließende Formel für die Betweenness-Zentralität als:

$$C'_B(n_i) = \frac{\sum_{i<j}^{g} \hat{d}_{ij}(n_k)/\hat{d}_{ij}}{(N-1)(N-2)/2}.$$

In seiner standardisierten Form schwankt dieses Maß zwischen 0 und 1, was einen Vergleich von verschiedenen Netzwerken und Zentralitätsmaßen ermöglicht.

Je nach Konstruktion des Netzwerkes muss die Betweenness jedoch anders interpretiert werden. Im hier vorliegenden Netzwerk aus Noun Phrases, deutet eine hohe Betweenness-Zentralität auf ein Wort hin, dass bei der Konstruktion von Centern eine vermittelnde Position einnimmt (siehe Tabelle 6.3). Dabei bedeutet eine hohe Betweenness jedoch nicht zwangsläufig, dass es sich bei dem jeweiligen Wort (Type) um eine zentrale Aussage des Textes handelt. Vielmehr muss man sich darunter jene Symbole vorstellen, die den Text zusammenhalten und ihn zu einer geschlossenen Einheit integrieren. Deswegen ist es nicht verwunderlich, dass das Wort „social" den höchsten Wert an Betweenness aufweist, da dies die grundlegende Ausrichtung des Textes wiedergibt. Sozusagen den kleinsten gemeinsamen Nenner an Bedeutung. Gleichzeitig verweist es auf den weiteren Entstehungskontext des Textes, nämlich das es sich um ein sozialwissenschaftliche Arbeit handelt. Worte wie „class" und „capital" nehmen

[9] Die Notationen sind bezüglich dieser Maßzahl etwas inkonsistent, weswegen eine Notation gewählt wurde, deren Hauptanliegen es ist, Konsistenz innerhalb dieser Arbeit zu bewahren. Gleichzeitig sollte der Bezug zu den erwähnten Arbeiten ersichtlich sein.

	Betweenness (C'_B)	Closeness (C'_C)	Eigenvector (C_e)
social	0.678571	0.533333	0.648516
mobility	0.535714	0.500000	0.178506
trajectory	0.428571	0.421053	0.049116
capital	0.250000	0.421053	0.621680
class	0.250000	0.333333	0.013448
cultural	0.000000	0.307692	0.159073
research	0.000000	0.400000	0.325011
service	0.000000	0.258065	0.003441
trust	0.000000	0.363636	0.165938

Tabelle 6.3: Vergleich unterschiedlicher Zentralitätsmaße eines NP-Netzwerks.

hingegen relativ geringe Werte an. Gleichzeitig sind diese aber von größerer Bedeutung, wenn es um die Bestimmung der inhaltlichen Bedeutung des Textes selbst geht. Betweenness-Zentralität scheint somit die *potentielle Anschlussfähigkeit* von Symbolen zu messen und stärker auf eine Einbettung in einen bestimmten Entstehungskontext sowie die spezifische Fachsprache zu verweisen.

Betrachtet man zudem andere Zentralitätsmaße, wie die Closeness und die Eigenvector Centrality, so erhärtet sich der Eindruck, dass sehr unterschiedliche Eigenschaften symbolischer Ordnungen durch Netzwerke repräsentiert werden können. Gleichzeitig wird deutlich, wie notwendig eine Einbettung in theoretische Konzeptionen symbolischer Ordnungen für die Interpretation der Netzwerkeigenschaften ist.

Closeness-Zentralität bezeichnet die relative Distanz eines Knoten n_i zu allen anderen Knoten (vgl. Wasserman und Faust 1994: 184f). Ihre normalisierte Form kann folgendermaßen angegeben werden:

$$C'_C(n_i) = (N-1)\sum_{j=1}^{g} d_{ij}.$$

Dieses Maß ist nur für vollständig verbundene Netzwerke geeignet, da die geodätische Distanz zu einem nicht-verbundenen Knoten unendlich groß ist.

Wenn die Annahme korrekt ist, dass Noun Phrases als eine Kombination einzelner Symbole gebildet werden und diese Inklusion als ein Form

der Bezugnahme auf andere Symbole geschieht, würden Worte mit hoher Closeness Überleitungen in der Struktur des Textes darstellen. Dies ließe erwarten, dass sie sehr viel stärker auf die Eigenheiten der Wortwahl des Textes reagieren, als dies bei der Betweenness der Fall war. Und in der Tat weisen einige Wörter eine vergleichsweise höhere Closeness-Zentralität auf. Der Unterschied ist insbesondere bei Begrifflichkeiten wie „capital" und „trust" aussagekräftig, da diese sich stärker auf den spezifischen Inhalt des Textes stützen und eher an den Rändern des Netzwerks positioniert sind. Daher kann die Closeness eines Knotens in diesem Fall als ein Maß für dessen Bedeutung bei der *Verknüpfung lokaler Bedeutungen* eines semantischen Netzwerkes angesehen werden.

Im Unterschied zu den vorrangegangenen Zentralitätsmaßen bezieht sich die *Eigenvektor-Zentralität* C_e nicht nur auf die Distanzen, sondern wird relativ zur Zentralität aller anderen Knoten bestimmt. Die Zentralität eines Knoten n_i ergibt sich dabei als eine Funktion des C_e aller Knoten, mit denen er eine Verbindung aufweist. Dieses Maß bezieht somit auch die Transitivität von Zentralität mit ein. Bezogen auf Machtdynamiken lässt sich hiermit der Einfluss einer Einzelperson in Abhängigkeit vom Einfluss aller anderen modelliern. Anders ausgedrückt, „in a power hierarchy, one's power is a positive function of the powers of those one has power over" (Bonacich 1987: 1171).

Die Bestimmung der Eigenvektor-Zentralität kann als das Eigenwert-Problem der linearen Algebra aufgefasst werden. Allgemein ausgedrückt handelt es sich bei einem Eigenvektor $\vec{e}$ um einen vom Nullvektor verschiedenen Vektor, der multipliziert mit einer Matrix A eine Streckung seiner selbst um einen konstanten Wert λ, den Eigenwert, ergibt. Formal ausgedrückt, entspricht dies der folgenden Gleichung:

$$A\vec{e} = \lambda\vec{e}.$$

Für eine gegebene Matrix gibt es immer eine Reihe von Eigenvektoren, die diese Gleichung erfüllen. Von diesen wird der größte Eigenwert und der damit korrespondierende Eigenvektor gewählt. Im Falle der Netzwerkanalyse ist es die Adjazenzmatrix der Knoten, die in ihre Eigenvektoren zerlegt wird.

Die *Adjazenzmatrix*, die auch als Soziomatrix bezeichnet wird, ist eine symmetrische Matrix, die in jeder Zelle die Beziehungsstärke zum jeweiligen Knoten angibt. Damit ergibt sich die Eigenvektor-Zentralität eines Knoten n_i als eine rekursive Funktion der Eigenvektor-Zentralitäten aller anderen Knoten zu denen er eine Verbindung aufweist:

$$C_e(n_i) = \frac{1}{\lambda} \sum_i A_{ij} C_e(n_j).$$

Im Falle ungewichteter Netzwerke ist die Adjazenzmatrix nur mit Nullen und Einsen besetzt. Aufgrund des Konstruktionsprinzips von CRA-Graphen, handelt es sich hierbei jedoch um gewichtete Graphen, welche in ihrer Repräsentation als Adjazenzmatrix das jeweilige Gewicht des Kantenzuges zur Darstellung der Verbindung zweier Knoten verwenden. Die Funktion `cra_graph()` hinterlegt zu diesem Zweck die „Beziehungsstärke" des jeweiligen Kantenzuges automatisch in dessen `weight`-Attribut.

Bezogen auf ein Netzwerk auf Noun Phrases kann die Eigenvektor-Zentralität als das Ausmaß der *Bedeutung eines Symbols in lokalen Hierarchien* aufgefasst werden. Daher kann man davon ausgehen, dass es sich bei den Typen mit hoher Zentralität um Symbole handelt, die im jeweiligen Kontext einen hohen Signalcharakter haben und die für die Bildung von lokalen „Bedeutungs"-Clustern ausschlaggebend sind. Wörter mit einem hohen C_e Wert sagen somit am meisten über die zentralen Themen eines spezifischen Textes aus. Im Vergleich dazu geben die anderen Maße einen eingehenderen Überblick über die Struktur einer spezifischen Fachsprache (Closeness) sowie diejenigen Begrifflichkeiten die einen Text intern zusammenhalten und externe Anschlussfähigkeit gewährleisten (Betweenness).

Mit dem Indexing wird die Grundlage geschaffen die Struktur eines Netzwerkes weiter zu analysieren und zu visualisieren. Dieser Analyseschritt wird hier als *Application* bezeichnet und umfasst insbesondere auch den Vergleich verschiedener Netzwerke, sowie die Interpretation der Strukturen und Maßzahlen hinsichtlich einer konkreten Fragestellung. Um diesen Analyseschritt besser beurteilen zu können, wird er im Folgenden auf die Frage bezogen, welche Themen jeweils die drei primären Soziologiezeitschriften der USA, Großbritanniens und Deutschlands strukturieren und wie sich diese zueinander verhalten. Konkret bedeutet dies, die zentralen Begrifflichkeiten zu identifizieren, welche die Bedeutungskonstruktion der semantischen Netzwerke dominieren, und deren (Un-)Ähnlichkeiten zu modellieren.

Als Indexing Maß wird auf die Eigenvektor-Zentralität zurückgegriffen, da diese die interne Strukturierung eines Diskurses am besten darstellt. Die Betweenness-Zentralität würde hingegen tendenziell diejenigen Symbole identifizieren, welche die höchste Anschlussfähigkeit im jeweiligen Diskurs aufweisen. Demzufolge würde eine klassische CRA eher Aussa-

gen über die Konstruktion von soziologischen Texten mittels einer Fachsprache und unter der Bedingung der Anpassung an die Binnenkultur der jeweiligen Fachzeitschrift treffen.

Da sich das Modell sehr stark an einzelnen Texten orientiert, wird es für größere Textmengen sowohl unübersichtlicher als auch sehr viel aufwendiger zu berechnen. Daher beziehen sich die folgenden Analysen auf die Ebene der Zeitschriften, soll heißen deren gebündelte Abstracts. Dieses Vorgehen entspricht demjenigen, dass bereits im Falle des Vektorraummodels angewendet wurde (siehe Abschnitt 6.3.1). Desweiteren wird dadurch auch der inhaltliche und methodologische Abgleich mit den Kosinus-Ähnlichkeiten der Textvektoren ermöglicht.

Für die so konstruierten NP-Graphen wird im nächsten Schritt die Eigenvektor-Zentralität bestimmt. Die Betrachtung der zehn Typen mit der höchsten Zentralität (siehe Tabelle 6.4) liefert einen ersten Überblick über die zentralen Wörter in der semantischen Struktur des jeweiligen Journals. Darin zeigen sich Unterschiede innerhalb der Ländergruppen ebenso wie übergreifende Gemeinsamkeiten. So sind zum Beispiel die *Zeitschrift für Sozialforschung*, *Work Employment & Society* und die *American Sociological Review* wohl relativ ähnlich ausgerichtet. In allen drei Fällen nehmen Worte Spitzenpositionen ein, die mit Arbeitsmärkten und einer Lebensverlaufsperspektive in Verbindung zu stehen scheinen. Gleichzeitig werden hier auch Unterschiede zu anderen Gruppen und innerhalb der Gruppe deutlich.

Insgesamt entsteht der Eindruck einer Zweiteilung des Diskurses dieser spezifischen Zeitschriften. Diejenigen Zeitschriften in denen Worte wie „social", „class" und „theory" eine dominante Stellung aufweisen beziehen sich wohl anscheinend mehr auf allgemeine Theoriepositionen der Soziologie, wohingegen die drei vorher erwähnten Journals eine Konzentration auf Ungleichheitsthemen aufweisen. Allerdings handelt es sich hier zunächst nur um einen groben Überblick, der weder die konkreten Verteilungen der Eigenvektor-Zentralität berücksichtigt, noch einen Maßstab für die Unterschiedlichkeit der NP-Netzwerke aufweist.

Um einen sytematischeren Vergleich zu erhalten, wird im Folgenden wieder auf die Berechnung der Kosinus-Ähnlichkeit zurückgegriffen. Wie in den Ausführungen bezüglich des Information Retrievals bereits dargelegt, handelt es sich dabei um ein Verfahren, welches die Ähnlichkeit von Texten als Orientierung im Raum modelliert. Die Voraussetzung dafür ist, dass die Vektoren die gleiche Dimensionalität aufweisen. Prinzipiell ist dies hier gegeben, da es sich nur um Texte der englischen Sprache

AJS	ASR	ARS	BerJS	ZS	KZSS	SJBSA	WES	BJS
social	labor	research	social	labor	social	social	labour	social
capital	market	future	theory	market	market	class	market	class
movement	force	social	structure	german	theory	capital	participation	theory
theory	social	recent	inequality	force	inequality	theory	process	mobility
structure	participation	empirical	integration	participation	origin	analysis	force	capital
network	capital	sociological	order	social	action	research	position	structure
organization	movement	science	change	theory	structure	mobility	emotional	science
change	structure	work	action	system	labor	science	local	change
control	female	literature	policy	other	rational	sociological	flexibility	control
science	economic	life	research	training	research	change	experience	life

Tabelle 6.4: Die obersten zehn Worte entsprechend ihrer Eigenvektor-Zentralität je Soziologie- Zeitschrift.

handelt. Problematisch sind jedoch Unterschiede in den Amerikanischen (z.B.: „labor“) und Britischen (z.B.: „labour“) Schreibweisen. Dies dürfte tendenziell zu einer Überschätzung der Distanz im Bezug auf englische Texte resultuieren. Da diese Unterschiede jedoch nur bei wenigen Worten vorkommen, ist die generelle Aussagekraft des Vergleichs davon nicht berührt.

Die resultierende Matrix der Distanzen in Abbildung 6.3 repräsentiert das Ausmaß der Kosinus-Ähnlichkeit der jeweiligen Strukturen der semantischen Netzwerke, bezogen auf deren Eigenvektor-Zentralitäten. Dadurch ergibt sich ein wesentlich differenzierteres Bild der Lage. So zeigt sich im Falle der Britischen Soziologiezeitschriften der größte interne Gruppenunterschied. Während *British Journal of Sociology* und *Sociology Journal – The British Sociological Association* eine ähnliche semantische Struktur aufweisen, unterscheiden sich beide sehr stark von *Work, Employment & Society*. Letztgenanntes weist mehr Ähnlichkeit mit der *Zeitschrift für Sozialforschung*, der *American Sociological Review*, sowie der *Kölner Zeitschrift für Sozialforschung und Sozialpsychologie* auf. In den beiden anderen Ländergruppen sind diese Unterschiede nicht so stark ausgeprägt. Es findet sich stattdessen eher eine Dreiteilung. Ein Themenbereich ähnelt dem WES hinsichtlich der Konzentration auf Themen des Arbeitsmarkts und der Ungleichheit, während Zeitschriften wie ASR und ZS zwar ebenfalls einen Bezug zum Arbeitsmarkt herstellen, zugleich aber einen stärkeren Bezug zur Sozialstrukturanalyse aufweisen, was vor allem in ihrer Distanz zum dritten Themenbereich, der eher theoretischen, allgemeinen Soziologie, deutlich wird.

Die Ähnlichkeiten der Eigenvektor-Zentralitäten entsprechen in ihrem grundlegenden Muster dejenigen, die aus dem Vektorraummodel abgeleitet werden können. Das zwei verschiedene Verfahren zu ähnlichen Schlüs-

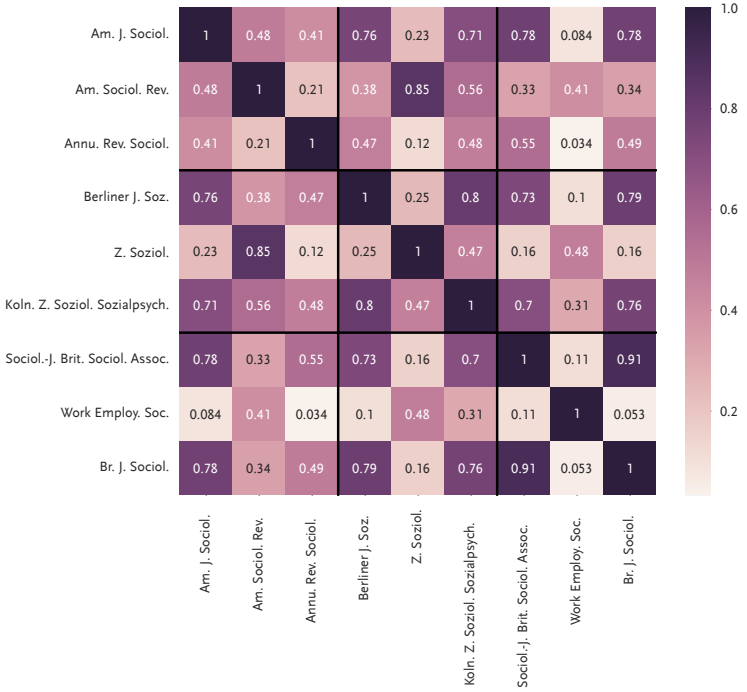

Abbildung 6.3: Kosinus-Ähnlichkeit der Eigenvektor-Zentralitäten der CRA Graphen als Heatmap.

sen kommen, unterstreicht die Belastbarkeit dieser Ergebnisse. Zugleich wird hier ein Vorteil der semantischen Netzwerkanalyse deutlich. Im Vergleich mit den Kosinus-Ähnlichkeiten des Vektorraummodels, haben wir hier ein Modell, dass eine sehr viel explizitere Konstruktion seiner semantischen Objekte erlaubt. Daher ist es in diesem Fall einfacher zu klären, welche konkreten Symbole für die beobachteten Unterschiede verantwortlich sind.

6.4.3 Texte als semantische Netzwerke

Im Gegensatz zum Machine Learning, bedarf es bei der NTA keines vorgegebenen Trainingskorpus, d.h. das Modell wird nicht an einer bestehenden Klassifikation ausgerichtet. Stattdessen handelt es sich hier um ein *exploratives Vorgehen*, welches darauf abzielt den Bedeutungsgehalt aus den strukturellen Gegebenheiten eines Textes abzuleiten. Deshalb besteht auch keine direkte Möglichkeit die Repräsentation eines Textes als einer Netzwerkstruktur auf ihre Passgenauigkeit mit menschlicher Interpretation hin zu untersuchen. Zwar ist es generell möglich dafür Experimente heranzuziehen oder die Ergebnisse der Netzwerkanalyse mit anderen Informationen über die Texte abzugleichen, eine Bestimmung der Präzision und des Recalls wie es beim Machine Learning der Fall ist, ist aber nicht möglich. Der Hauptgrund hierfür ist ein methodologischer. Die NTA geht nicht von einem spezifischen Bedeutungsgehalt eines Textes aus, der als abhängige Variable modelliert werden könnte. Vielmehr ist der Bedeutungsgehalt eines Netzwerkes dessen spezifische Struktur und kann somit auch nur relativ zu anderen Strukturen bestimmt und interpretiert werden. Das bedeutet aber auch, dass die enorme Vielfalt der Verfahren der Netzwerkanalyse auf semantische Netzwerke angewendet werden kann.

Zu den großen Vorteilen der NTA zählt gerade die Möglichkeit die Struktur von Texten explizit modellieren zu können. Mit der Anwendung unterschiedlicher Konstruktionsprinzipien, d.h. je nach dem wie Selection und Linking gestaltet werden, ergeben sich Repräsentation unterschiedlicher, struktureller Parameter. Statt der hier verwendeten NP-Netze wären auch eine Reihe alternativer Modellierungen möglich gewesen, zum Beispiel die Darstellung von Subjekt-Objekt Beziehungen, bei denen das jeweilige Verb als ein Kantenattribut fungiert.

Gleichzeitig liegt hier auch die größte Herausforderung für die Weiterentwicklung und Anwendung der Netzwerk-Text-Analyse. Was als ein Objekt und was als eine Beziehung angenommen wird, muss durch Theo-

rien der Zeichen- und Symbolkonstruktion flankiert werden. Ansonsten ergeben sich schlichtweg zuviele Möglichkeiten für die Konstruktion eines Netzwerkes. Zudem ist eine Interpretation der Netzwerkmaße und -strukturen nicht möglich, ohne Überlegungen bezüglich der sozialen Regeln und Praktiken von Sprache anzustellen.

Die hier vertretene Auffassung einer Kodierung von Bedeutungen in den Relationen von Symbolen passt intuitiv zur Repräsentation dieser Relationen in Form eines Netzwerkes. Dies findet sich so auch in den Überlegungen einer Reihe von Autoren zur symbolischen Konstruktion von Bedeutungen (vgl. Lehmann 1992: 3ff). Allerdings enthält diese Auffassung ein weitreichendes Problem, welches in der Diskussion der vorangegangenen Analysetechniken bereits angedeutet wurde. In semantischen Netzwerken werden Knoten als *eineindeutige Symbole* aufgefasst, die zwar unterschiedlich kombiniert werden können, dabei aber immer mit sich selbst identisch bleiben. Am Beispiel eines NP-Graphen kann dies demonstriert werden. Nimmt man die beiden Phrasen „social science" und „social work", so würde daraus ein Graph entstehen, der alle drei Worte als Knoten miteinander verbindet. Obwohl „social" in beiden Fällen gleich geschrieben wird, verschiebt sich die Bedeutung des Wortes durch seine Kombination mit „science" und „work". Im jeweiligen Kontext kann „social" alles meinen was im weitesten Sinne der Interaktion von Menschen geschuldet ist oder sich auf die spezifische Motivation des Handelns als etwas „für die Gemeinschaft" beziehen.

Es ist sicherlich korrekt anzunehmen, dass Bedeutungen in symbolischen Beziehungen kodiert sind. Allerdings erscheint es fraglich, ob die Repräsentation durch konkrete Beziehungen in Form von Kantenzügen für alle symbolischen Ordnungen geeignet ist. In spezifischen Sprachen, in denen Symbole in ihren Bedeutungen fixiert und eineindeutig sind, ist eine Repräsentation als Netzwerk sinnvoll. Für andersgeartete Fälle mag sie sich besser oder schlechter eignen, je nach den konkreten Konstruktionsprinzipien des zugrundeliegenden Netzwerks. Dennoch bleiben generelle Limitationen erhalten. Mehrdeutige Worte (Polysemie und Homonymie) und Bedeutungsverschiebungen durch die Kombination von Worten (Kollokationen) stellen ein unlösbares Problem für die Netzwerk-Text-Analyse dar.

6.5 Latente Ordnungen

Die vorangegangenen Modelle weisen eine gemeinsame Limitation auf. Sie können Bedeutungen nur indirekt erfassen. Im Falle der überwachten Klassifikationsverfahren muss auf eine extern festgestellte Bedeutung zurückgegriffen werden. Das Modell des Vektorraums kann die Ähnlichkeit zwischen Texten darstellen, erlaubt aber kaum Aussagen darüber, wodurch diese zustande kommen. Ein ähnliches Problem existiert auch im Bereich der Analyse semantischer Netzwerke, da diese ebenfalls eine Repräsentation der Worte und Strukturen des Textes liefern, aber nur indirekt auf Bedeutungen hinweisen können, bzw. auf externe Bedeutungszuschreibungen durch menschliche Coder angewiesen sind. Dies hat zur Folge, dass Bedeutungen hier nicht direkt gemessen oder modelliert werden können, sondern nur als interpretative Zuschreibungen möglich sind.

Im Bereich der qualitativen Beschäftigung wird diese interpretative Leistung als eine exklusive Domäne des Menschen angesehen. In einigen Fällen wird zudem davon ausgegangen, dass eine sinnvolle Interpretation nur durch besonders geschulte Interpreten möglich ist. Es ist jedoch angebracht zu fragen, worin diese besondere Qualität des Menschen denn besteht und warum diese einem Algorithmus grundsätzlich verwehrt bleiben sollte. Vor dem Hintergrund der hier vertretenen Auffassung des objektiven Charakters symbolischer Ordnungen, ist die Annahme einer unberechenbaren Interpretationsleistung grundsätzlich eine wenig befriedigende Perspektive.

Gehen wir von der grundlegenden Annahme aus, dass Bedeutungen in symbolischen Ordnungen kodiert werden und diese in spezifischen Relationen zueinander stehen, so bedeutet dies, dass wir symbolische Ordnungen als relativ stabile Verknüpfungen von Tokens auffassen können. Statt anzunehmen, dass der Sinn einer Äußerung in direkter Weise an die verwendeten Worte gebunden ist, suchen wir nach Regelmäßigkeiten im Gebrauch von Worten. Symbolische Ordnungen können demzufolge als Funktion von Token und nicht als eine Funktion von Typen angesehen werden. Im Prinzip bedeutet dies nichts anderes, als anzunehmen, dass sich, bis zu einem gewissen Grad, mit unterschiedlichen Worten dieselben Sachverhalte ausdrücken lassen. Gleichzeitig impliziert dies auch, dass so verstandene, symbolische Ordnungen nur als Regelmäßigkeiten im Gebrauch, über eine Vielzahl von konkreten Texten hinweg, identifiziert werden können. Es geht also darum die latenten Ordnungsprinzipien zu finden, die das Auftreten bestimmter Token erklären können.

Grundsätzlich finden sich zwei Vorgehensweisen, die beide dem Bereich des Information Retrievals entstammen und für eine solche Modellierung von symbolischen Ordnungen zweckmäßig erscheinen. Zum einen kann versucht werden symbolische Ordnungen als eine Art der Dimensionsreduktion zu fassen, wie es bei der sogenannten Latent Semantic Analysis (vgl. Deerwester et al. 1990) geschieht. Andererseits wurde in den letzten Jahren eine Reihe neuer Verfahren entwickelt die gemeinhin als *Topic Models* oder auch *Generative Models* bezeichnet werden (vgl. Steyvers und Griffiths 2007). Diese versuchen die Konstruktion eines Textes mittels mehrstufiger, probabilistischer Verfahren zu modellieren.

6.5.1 Latent Semantic Analysis

Die Analyse latenter Semantiken wurde von Deerwester et al. (1990) entwickelt, um die hohen Dimensionalitäten des Vektorraummodels im Rahmen des Information Retrievals handhabbar zu machen. Zudem lieferte ihr Verfahren eine Antwort auf die Probleme, die aus der Existenz von Synonymen und Polysemie hervorgehen. Dabei gehen sie von der Überlegung aus, dass Worte nur Hinweise auf die in Texten kodierten Semantiken geben, nicht aber mit diesen gleichzusetzen sind. Vielmehr wird angenommen, dass die verwendeten Worte als *Konkretisierungen zugrundeliegender Bedeutungen* aufgefasst werden. Daher bedarf es eines Verfahrens, dass den Korpus auf seine latente Bedeutungsstrukturen reduziert.

Da das Vektorraummodell die Grundlage für die Beschreibung des Korpus liefert, ist es naheliegend auf Techniken der linearen Algebra zurückzugreifen, die eine Zerlegung dieses Raumes in eine Repräsentation niederer Dimensionalität erlauben. Im Prinzip ist dies nichts anderes als die Verallgemeinerung des bereits besprochenen Eigenwertproblems. Eine symmetrische Matrix lässt sich in eine Reihe von Eigenvektoren zerlegen, die zusammen mit den jeweiligen Eigenwerten eine Matrix in Form eines Vektors repräsentieren können. Die Verallgemeinerung dieser Vorgehensweise auf arbiträre, rechteckige Matrizen ist die *Singulärwertzerlegung* (SVD), die eine Standardanwendung in einer Vielzahl wissenschaftlicher Disziplinen darstellt (vgl. Martin und Porter 2012). Jede Matrix M, die aus Elementen der realen oder komplexen Zahlen besteht kann als ein Produkt dreier Matrizen aufgefasst werden:

$$M_{m\times n} = U_{m\times m} \cdot \Sigma \cdot V^{H}_{n\times n}.$$

Die Matrix $U_{m\times m}$ ist die unitäre Matrix die aus den linken Singulärvektoren besteht, während $V^H_{n\times n}$ eine adjungierte Matrix ist, welche die rechten Singulärvektoren abbildet. Demgegenüber enthält die Diagonalmatrix Σ die sogenannten Singulärwerte σ_i in geordneter Reihenfolge.

Die Singulärwertzerlegung hat eine Reihe bedeutsamer mathematischer Eigenschaften. Für die Modellierung latenter Semantiken wird der Umstand genutzt, dass es dadurch möglich ist eine niedrigdimensionale Repräsentation des zugrundeliegenden Raumes zu erhalten. Da die maximale Anzahl von Singulärwerten dem Minimum der Dimensionen der ursprünglichen Matrix entsprechen muss ($k = \min(m, n)$), können wir σ_i so wählen, dass $\sigma_0 < \sigma_i < \sigma_k$ gilt. Reduzieren wir die Matrizen der Singulärvektoren ebenso und multiplizieren wir alle drei Matrizen miteinander, so erhalten wir $\hat{M}$ als eine Approximation von M, welche wir als eine Reduktion auf die wesentlichsten Merkmale der ursprünglichen Matrix auffassen können. Es gilt also:

$$\hat{M} = U_{m\times i} \cdot \Sigma \cdot V^H_{n\times i}$$

für eine Reduktion von M auf den Rang i.

Je nach der Art der zugrundeliegenden Matrix M beschreibt dieses Vorgehen ein anderes, etabliertes, sozialwissenschaftliches Verfahren. Wird eine Korrelations-Matrix auf diese Art und Weise zerlegt, so spricht man von einer *Faktoranalyse* oder genauer, einer *Principal Component Analysis*. Durch die Zerlegung einer Matrix aus χ^2-Residuen ergibt sich eine *Multiple Korrespondenzanalyse*. Legt man die Kovarianzmatrix zugrunde, so handelt es sich um die *Canonical Correlation*. Im Falle der Analyse latenter Semantiken wird von einer TFiDF-gewichteten Dokument-Wort-Matrix ausgegangen, die als eine Repräsentation des Bedeutungsgehaltes der einzelnen Worte relativ zu ihrem Auftreten im gesamten Korpus aufgefasst werden kann.

Reduktion semantischer Räume

Die Durchführung einer latenten Semantikanalyse ist mit fast allen gängigen Programmpaketen möglich, die Funktionen der linearen Algebra enthalten. So ist die Singulärwertzerlegung beispielsweise in der SciPy-Bibliothek als Funktion `scipy.linalg.svd()` enthalten. Obwohl die Berechnung mittels dieser Algorithmen relativ effizient ist, kann eine Weiterverabeitung und Anpassung der Ergebnisse in vielen Fällen langwierig sein. Daher gibt es spezifischere Programmbibliotheken, welche die

Modellierung semantischer Räume unterstützen und wesentlich komfortabler machen. Im Folgenden wird daher auf das `gensim`-Modul (vgl. Řehůřek und Sojka 2010) zurückgegriffen, welches eine Vielzahl von unterschiedlichen Modellen für semantische Räume beinhaltet.[10] Gensim greift intern auf die erwähnten `scipy`-Funktionen und damit letztlich auf deren zugrundeliegenden C und FORTRAN Funktionen zurück, was eine relativ hohe Performanz der Berechnung garantiert.

Ausgangslage für die Modellierung ist auch in diesem Fall ein Korpus, der im Format einer *Liste von Listen mit Token* gehalten ist. Für die Konstruktion der Dokument-Wort-Matrix besitzt Gensim eigene Funktionen, welche den Korpus intern in eine Sparse Matrix transformieren.[11] Bei diesem Format werden nur diejenigen Zellen (in Koordinatenschreibweise) angegeben, die von einem Standardwert (für gewöhnlich Null bzw. leere Zellen) verschieden sind. Diese Konzeption ist gerade bei relativ großen Korpora wichtig, da so die Menge des notwendigen Arbeitsspeichers signifikant reduziert werden kann. Desweiteren sind die meisten Funktionen von Gensim auf eine sogenannte *out-of-core* Schätzung ausgelegt, d.h. die einzelnen Berechnungen können nacheinander ausgeführt werden, was die Speicherlast enorm reduziert. Im Gegenzug wird dadurch jedoch die Rechenzeit verlängert. Die nativen Gensim-Funktionen zu verwenden ist auch deshalb angeraten, weil die Sparse Matrix mit einem dazugehörigen Diktionär generiert wird. Dadurch wird es möglich die einzelnen Token wieder ihrer numerischen Identifikationsnummer zuzuordnen.

Um einen Korpus in eine Dokument-Wort-Matrix zu überführen, erfordert Gensim zunächst die Erstellung eines Diktionärs, welches die enthaltenen Token als ganze Zahlen anstelle von Strings repräsentiert. Dies ist sinnvoll, da so der Speicherbedarf für den transformierten Korpus weiter sinkt. Die `corpora.Dictionary`-Instanz verfügt zudem über eine eigene Speichermethode, so dass das Diktionär nur einmal erstellt werden muss.

```
1 from gensim import corpora, models
2
3 docs = articles.StopCollTokens.tolist()
```

[10] Ein Überblick über die in Gensim implementierten Verfahren findet sich im dazugehörigen Online-Tutorial: `https://radimrehurek.com/gensim/tutorial.html`.

[11] An dieser Stelle muss darauf hingewiesen werden, dass die klassische Formulierung von Deerwester et al. (1990) von einer Wort-Dokument-Matrix ausgeht. Daher erscheinen im Vergleich die linken und rechten Singulärvektoren vertauscht. Der alleinige Grund dafür ist die sozialwissenschaftliche Konvention, welche die Beobachtungen als Zeilen und die Variablen als Spalten auffasst.

```
4
5 dictionary = corpora.Dictionary(docs)
6
7 dictionary.save('Daten/SozAbst.dict')
```

Im nächsten Schritt wird dieses Diktionär auf die Liste der Tokenlisten angewendet um eine Sparse Matrix (*Bag-of-Words*) zu erhalten. Dies geschieht über die `doc2bow()` Methode. Auch hier kann der resultierende Korpus in einer Reihe von Formaten gespeichert werden.[12] In diesem Fall wird das *Matrix Market Exchange Format* verwendet.

```
1 corpus = [dictionary.doc2bow(doc) for doc in docs]
2
3 corpora.MmCorpus.serialize('Daten/SozAbst.mm', corpus)
```

Danach erfolgt die Transformation des Korpus mittels einer TFiDF-Gewichtung und dessen anschließende Zerlegung in seine zugrundeliegende Dimensionalitäten. Ein weiterer Vorteil von Gensim ist, dass an dieser Stelle ein Schlüsselwortargument (`num_topics`) für den maximalen Rang i bis einschließlich dessen die Singulärwerte ermittelt werden sollen vergeben werden kann. Der von Radim Řehůřek (2011) entwickelte Algorithmus erlaubt eine sehr effiziente Singulärwertzerlegung, die mit einer Zerlegung von Teilräumen auskommt und daher sowohl Rechenzeit spart sowie den Speicherbedarf reduziert.

```
1 tfidf = models.TfidfModel(corpus)
2
3 tfidf_corpus = tfidf[corpus]
4
5 lsa = models.LsiModel(tfidf_corpus,
6                       id2word=dictionary,
7                       num_topics=300)
```

Mittels der Methode `.show_topic()` können die einzelnen Dimensionen als eine Liste von Tupeln zurückgegeben werden. Das erste Argument dieser Methode gibt die jeweilige Dimension an, während das zweite Argument die Länge des anzuzeigenden Singulärvektors angibt bis zu dem

[12] Details finden sich in der Gensim-Dokumentation: `https://radimrehurek.com/gensim/tut1.html#corpus-formats`

die Skalarwerte in absteigender Reihenfolge angezeigt werden sollen. Die resultiertenden Tupel enthalten an erster Stelle das jeweilige Merkmal, in diesem Fall handelt es sich dabei um die im Korpus vorkommenden Wörter, sowie an zweiter Stelle den jeweilige Wert des Singulärvektors auf dieser Dimension. Jede dieser Dimensionen kann als eine Art durchgehendes Thema, beziehungsweise als eine zugrundeliegende Ordnung der verwendeten Symbole aufgefasst werden. Die absoluten Werte des Singulärvektors geben die relative Bedeutung dieses Wortes für die jeweilige Dimension wieder.

```
1 lsa.show_topic(0, 10)
```

```
1 [(u'women', 0.12656175627791472),
2  (u'social', 0.12066168357075997),
3  (u'work', 0.11468763709363815),
4  (u'political', 0.099034949033273867),
5  (u'research', 0.098626393006953361),
6  (u'class', 0.09655170686067141),
7  (u'employment', 0.094657606244561973),
8  (u'theory', 0.092660327358501673),
9  (u'sociology', 0.087036030498619882),
10  (u'economic', 0.082036612466961686)]
```

Dabei ist zu beachten, dass diese Werte auch negativ werden können. Da es sich um eine lineare Kombination der zugrundeliegenden Basisvektoren handelt, sind die Vorzeichen der Werte der Singulärvektoren grundsätzlich nicht bestimmt. Ihre Vergabe hängt vom konkreten Lösungsverfahren (Algorithmus) ab, mit dem die Faktorisierung durchgeführt wird.[13] Aus mathematischer Sicht ist dabei nur wichtig, dass die Einzigartigkeit der Lösung gegeben ist, also das unterschiedliche Lösungsansätze zu – von Rundungsfehlern einmal abgesehen – gleichen Werten gelangen. Ausschlaggebend für die Interpretation ist daher nur der absolute Wert, als Beitrag des Wortes zur jeweiligen Dimension. Zwar gibt es erste Lösungsansätze für dieses Problem (vgl. Bro, Acar und Kolda 2007), da diese jedoch

[13] In den meisten Fällen wird zur Durchführung einer SVD auf die LAPACK Bibliothek zurückgegriffen. So auch in Gensim, welches den LAPACK „divide-and-conquer“ Algorithmus (xGESDD) verwendet.

voraussetzen, dass die ursprünglichen Werte negativ oder positiv werden können, werden sie hier nicht weiter diskutiert.

Neben einer Betrachtung und Interpretation der einzelnen Dimensionen, erlaubt die Zerlegung von Texten in eine niederdimensionale Repräsentation auch die Analyse des Verhältnisses der Texte zu den Dimensionen, d.h. deren thematischen Bezug. Für eine Vielzahl von Fragen ist es notwendig arbiträre Textvektoren im Raum der latenten Semantiken zu verorten. Zum Beispiel, wenn man wissen möchte, wie sich nicht im ursprünglichen Korpus enthaltene Texte zu den bestehenden verhalten oder Kombinationen der vorhandenen Texte im Raum abgebildet werden sollen.

Mittels einer Prozedur, die als *Folding-In* bezeichnet wird (vgl. Deerwester et al. 1990: 394-399), ist es möglich Dokumentvektoren im Raum der latenten Dimensionen abzubilden. Dabei wird für einen Dokumentvektor, oder eine entsprechende Matrix von Dokumenten M_q, der geometrische Schwerpunkt in Abhängigkeit von der relativen Bedeutung der enthaltenen Worte entlang der latenten Dimensionen berechnet.

Im Prinzip heißt das nichts anderes, als die Dokumente auf Basis der darin enthaltenen Worte in diesem Raum zu verorten und diese Vektoren mit den entsprechenden Singulärvektoren für die einzelnen Worte zu gewichten. Dementsprechend kann die Repräsentation D_q eines Dokumentes oder eines Korpus als das Produkt des Dokumentvektors M_q, welcher die Singulärvektoren der Worte reduziert auf den entsprechenden Rang i enthält, und der invertierten Σ-Matrix, als einem Gewichtungsfaktor, ausgedrückt werden:

$$D_q = M_q \cdot V_{n \times i} \cdot \Sigma^{-1}.$$

Solange wir die Annahme vertreten können, dass die grundlegenden Dimensionen des Vektorraums durch die Reduktion nicht zu stark verzerrt wurden und wenn der zu projezierende Text ein Mindestmaß an Wörtern mit dem ursprünglichen Korpus gemein hat, ist eine solche Abbildung möglich und sinnvoll.

Interpretation latenter Semantiken

Was aber heißt es von latenten Dimensionen der Semantik zu sprechen und wie lässt sich dieses Modell für eine soziologische Betrachtung symbolischer Ordnungen nutzen? Zur Beantwortung dieser Frage erscheint es passend, die Soziologie selbst zum Gegenstand einer solchen Betrachtung

zu machen und das im vorangegangenen Kapitel konstruierte LSA-Modell eingehender zu interpretieren. Die Wahl eines vertrauten Subjekts macht es zudem einfacher die resultierenden Dimensionen auf ihre Plausibilität hin zu prüfen.

Zum Verständnis der Dimensionen einer latenten Semantikanalyse ist es zunächst einmal wichtig sich vor Augen zu führen, dass die Dimensionen in gewisser Weise aufeinander aufbauen. Alle Dimensionen werden *orthogonal* zueinander konstruiert, d.h. es wird jeweils versucht einen rechten Winkel zu allen anderen Dimensionen zu finden und die einzelnen Objekte (Worte und Dokumente) darin zu verorten. Im Resultat bedeutet dies auch, dass der Anteil der in den Dimensionen enthaltenen Informationen bezüglich der ursprünglichen Matrix sukzessive abnimmt. Ist die grundlegende Matrix eine Korrelationsmatrix, so können die Singulärwerte σ_i als das Ausmaß der erklärten Varianz aufgefasst werden. Bei einer TFiDF-Gewichtung und einer asymmetrischen Matrix ist eine so präzise Interpretation zwar nicht möglich, aber es zeigt sich trotzdem dieselbe, abnehmende Tendenz der in den Singulärvektoren enthaltenen Information. In diesem Fall können die Singulärwerte als ein Indikator für den relativen Beitrag der jeweiligen Dimension an den Strukturen des gesamten Raumes interpretiert werden.

Daraus ergeben sich zwei bedeutsame Konsequenzen für die Ergebnisse der LSA. Zum einen enthält die erste Dimension den größten Anteil der zugrundelegenden Strukturen des semantischen Raumes und kann daher als eine Art „allgemeiner" Dimension aufgefasst werden. Es ist daher zu erwarten, dass sich hier diejenigen Symbole wiederfinden, welche die dominanten Begrifflichkeiten für den zu untersuchenden Korpus enthalten, weswegen sie für den Diskurs insgesamt kennzeichnend sind. Zweitens hat dies auch zur Folge, dass die Themen unabhängig von ihrer gewählten Anzahl sind. Das hier verwendete Modell wurde mit dreihundert Dimensionen initialisiert, dies geschah jedoch nur aus Gründen einer Schonung der Ressourcen. Prinzipiell wäre auch eine Zerlegung der gesamten Matrix möglich gewesen, wegen der zunehmend geringeren Information wäre das Resultat jedoch kaum informativer gewesen.

Daraus ergibt sich die zentrale Frage, wie die Anzahl der Dimensionen sinnvoll festgelegt werden kann. Man könnte auf das aus der Faktoranalyse bekannte Elbogenkriterium zurückgreifen, d.h. man sucht nach dem Punkt an dem der Plot der Singulärwerte sichtbar abflacht. Alternativ sind jedoch auch andere Entscheidungskriterien denkbar. Ihnen allen ist jedoch gemein, dass sie von einer „richtigen" Anzahl von Dimensionen und

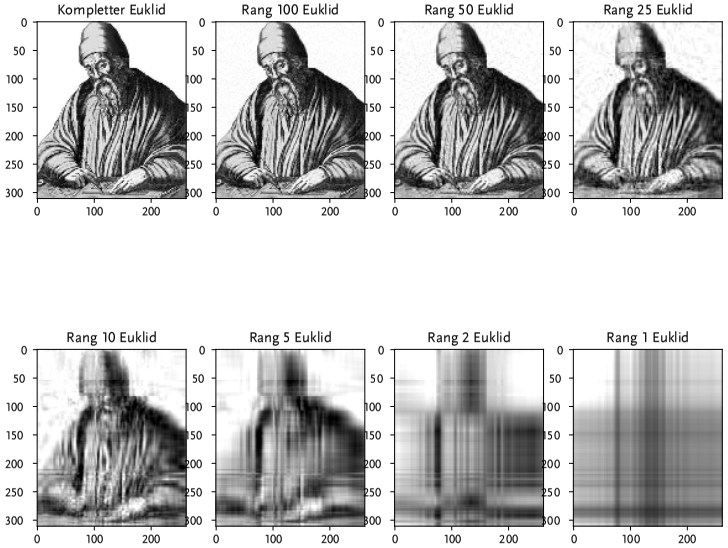

Abbildung 6.4: Sukzessive Singulärwertzerlegung von Bilddaten auf unterschiedlichen Ränge.

damit auch von einer „richtigen“ Anzahl an Themen oder symbolischen Ordnungen ausgehen. Allerdings legen die hier gemachten Überlegungen nahe, dass es sinnvoll ist symbolische Ordnungen als ein *hierarchisches Mehrebenenphänomen* zu betrachten. Es ist intuitiv ersichtlich, dass die in einem Schriftstück angesprochenen Themen im Verhältnis zueinander einen sehr unterschiedlichen Auflösungsgrad haben können.

Zudem legt auch die grundlegende Funktionsweise der Singulärwertzerlegung eine solche Modellierung nahe. Die Wahl unterschiedlicher Ränge zur Approximation und Verdichtung der zugrundeliegenden Matrix muss eher als eine Art Auflösunggrad verstanden werden. Man kann mit diesem Verfahren eine präzisere oder gröbere Auflösung wählen, je nach grundlegender Fragestellung und im Einklang mit praktischen Überlegungen.

Das bisher Gesagte kann durch eine Anwendung auf Bilddaten verdeutlicht werden (siehe Abbildung 6.4). Dies macht es einfacher die teilweise recht abstrakte Mathematik und die vorher diskutierten Besonderheiten besser zu verstehen. Dazu wird auf eine fiktive Illustration des Mathematikers Euklid von Alexandria zurückgegriffen. Dieses Bild wird zunächst in eine Matrix überführt, die anschließend in ihre Singulärwerte zerlegt und für bestimmte Ränge wieder zusammengesetzt wird. Man sieht hier sowohl den in niedrigeren Rängen überproportional zunehmenden Verlust an Genauigkeit der Approximation, als auch die Erhaltung der grundlegenden, strukturellen Eigenschaften des Bildes. Selbst bei einem Rang von $i = 10$ ist das Bild zumindest noch in seinen Umrissen erkennbar und für eine Person, die mit dem Original vertraut ist, auch identifizierbar.

Im Folgenden werden nur die obersten zehn Dimensionen für den Korpus der Soziologie-Abstracts (siehe Tabelle 6.5) näher betrachtet. Diese Begrenzung ist darauf zurückzuführen, dass die darin enthaltenen symbolischen Ordnungen sich alle auf einem relativ ähnlichen Abstraktionsniveau zu befinden scheinen und zudem als eine grundlegende Repräsentation des soziologischen Diskurses angesehen werden können. Diese Interpretation wird auch durch eine Betrachtung der Σ-Werte (Singulärwerte) gestärkt (siehe Abbildung 6.5 auf Seite 274). Hier zeigt sich sowohl die Dominanz der ersten Dimension als auch ein weiterer Knick im Bereich der zehnten Dimension, was darauf hindeutet, dass es sich bei den vorangegangenen Dimensionen um einen vergleichbaren Synthesegrad handelt. Wie schon dargelegt, hat eine Darstellung weiterer Dimensionen keine Veränderungen der vorangegangenen zur Folge. Die Charakterisierung dieser Dimensionen erfolgt jeweils durch die obersten zwanzig Wörter.

Die darin enthaltenen Dimensionen geben auf den ersten Blick eine relativ akkurate Beschreibung der soziologischen Themen- und Arbeitsbereiche wieder. Wie erwartet enthält die erste Dimension („Allgemein") vor allem recht allgemeine Begriffe, die auf gewisse Art den soziologischen Mainstream wiedergeben. Hier sind nicht nur die grundlegenden Spezialgebiete der Soziologie enthalten, wie zum Beispiel „women" und „employment", sondern auch die dominanten, fachspezifischen Bezüge und Demarkationslinien: „sociology", „political", „economical". Darauf folgt eine Dimension, die vor allem Problemfelder und Diskussionsansätze der soziologischen Theorie mit einem starken Bezug zum Thema Gender abbildet („Theorie/Gender"). Zusammen mit der dritten Dimension, die Begriffe der Arbeit und des Berufslebens erfasst, scheinen diese drei Dimensionen die grundlegenden Probleme und abstrakten Begrifflichkeiten der Soziologie zu charakterisieren.

Die daran anschließenden Dimensionen erscheinen im Vergleich dazu auf konkretere Problemfelder bezogen zu sein. Sie umfassen Dimensionen, die sich auf die Analyse sozialer Klassen sowie auf Fragen der Religion, Demographie und Diskriminierung beziehen. Hinzu kommen „Rasse/Ethnizität", thematische Bezüge zu Wohlfahrtsstaaten und eine Dimension, die Netzwerke, Austausch und Macht beinhaltet. Gerade an der letztgenannteren Dimension wird deutlich, dass die Dimensionalitäten dieses Raumes eher semantische Gemeinsamkeiten in einem breiten Rahmen rekonstruieren und mögliche theoriepolitische Auseinandersetzungen um die Deutung von Begriffen nicht wiederspiegeln.

Die soziologische Fassung der sozialen Netzwerkanalyse (SNA) und die damit verbundenen Begrifflichkeiten folgen der allgemeinen Tendenz der Soziologie thematisch geschlossene Gruppierungen hervorzubringen (vgl. z.B.: Abbott 2001) und neuen Begrifflichkeiten mit Inkorporation zu begegnen. Deutlich wird dies vor allem in der Herausgabe spezifischer Fachpublikationen, sowie in der Existenz von spezifischen Vertretungen im Rahmen disziplinärer Dachorganisationen, wie zum Beispiel der Sektion *Soziologische Netzwerkforschung* in der Deutschen Gesellschaft für Soziologie.[14] In den letzten Jahren wurde spezifisch im deutschsprachigen Raum der paradigmatische Charakter der Netzwerkforschung im Sinne einer „relationalen Soziologie" verhandelt (vgl. Fuhse und Mützel 2010).

[14] An dieser Stelle ist weder eine Wertung der Sektion, noch der allgemeinen Segregierungstendenzen der Soziologie intendiert. Vielmehr geht es hier alleine um die Berücksichtigung dieser disziplinären Besonderheiten im Hinblick auf die Interpretation des semantischen Raums.

Damit ging auch eine Diskussion um die Frage einher, wie dieses Paradigma in die dominanten Fraktionen der Soziologie zu integrieren wäre, zum Beispiel entlang des Mikro-Makro Bruchs (vgl. z.B.: Bernhard 2012) oder im Hinblick auf die Quali-Quanti Debatte (Diaz-Bone 2008). Keine dieser Brüche werden jedoch in der zehnten Dimension unseres Modells deutlich. Es finden sich sowohl die klassischen Begrifflichkeiten der Netzwerkforschung wieder, insbesondere „network“ und „ties“, als auch allgemeinere Begriffe, wie „exchange“, „power“ und „inequality“. Damit scheint die zehnte Dimension so etwas wie die soziologische Semantik zwischenmenschlicher Beziehungen und deren Merkmale abzubilden.

Dies verdeutlicht noch einmal die Besonderheit der LSA als Zerlegung eines Diskurses von „oben“ herab. Die Konstruktion der latenten Dimen-

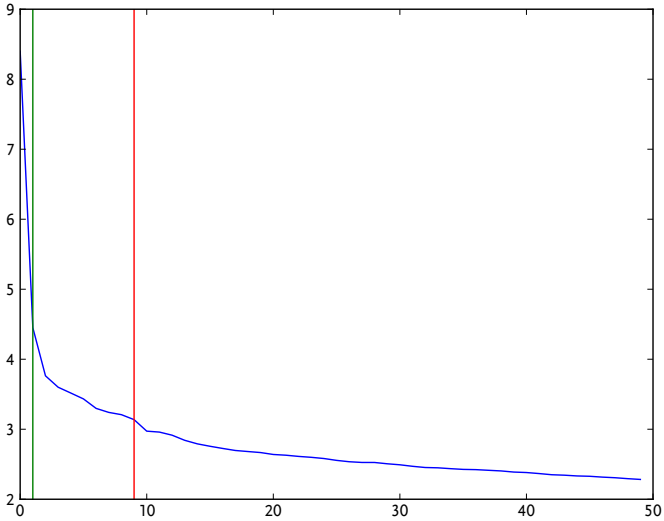

Abbildung 6.5: Σ-Werte für die ersten 50 Dimensionen der LSA der Soziologie Abstracts. Die grüne Linie gibt den Übergang von Dimension 0 zu Dimension 1 an, während die rote Line ein zweites (geringeres) Abflachen der Σ-Werte am Übergang von der neunten zur zehnten Dimension beschreibt.

sionen geschieht dabei ausgehend vom gesamten Korpus und geht mit aufsteigender Dimensionalität in Richtung einer immer kleinteiligeren Repräsentation. Das hat zur Folge, dass die unterschiedlichen Praktiken und Zielsetzungen von Autoren bei weitem nicht so ausschlaggebend sind wie die (Un)-Ähnlichkeiten in der Intention der Autoren. Anders und formaler ausgedrückt, die LSA erzeugt unterschiedliche Dimensionen die orthogonal zu einander stehen, was zur Folge hat, dass diese nicht wechselseitig ausschließend sind. Erst wenn wir die Muster der Verteilung über bestimmte Autoren oder andere Einheiten betrachten ist die Analyse von Unterschieden und die Konstruktion von Typologien möglich.

Wie die Elemente des semantischen Raums über die Dimensionen verteilt sind, kann mittels Folding-In modelliert werden, als eine Abbildung entsprechender Textvektoren auf diesen Raum. Damit lässt sich zum Beispiel die Frage beantworten, die beim Vergleich der Ähnlichkeiten der einzelnen Soziologiezeitschriften aufgeworfen wurde: Welche Inhalte sind bezeichnend für deren Relationen im semantischen Raum?

Voraussetzung einer Abbildung der Dokumentvektoren ist die Kombination der Texte und eine Konvertierung in das von Gensim verwendete Datenformat. Hierbei ist zu beachten, dass die LSA nur Worte als Merkmale von Texten bearbeiten kann, die bereits zur Initialisierung des ursprünglichen Modells verwendet wurden. Daher ist es notwendig das bereits erstellte Diktionär zur Zuordnung der numerischen Repräsentation der Token zu verwenden.

Nach der Erstellung des Korpus kann die bereits kalkulierte *TFiDF*-Gewichtung ebenso auf die neuen Dokumentvektoren angewandt werden. Grundsätzlich stellt sich hier aber die Frage, ob eine solche Gewichtung sinnvoll ist, da sie von den Häufigkeiten des ursprünglichen Korpus ausgeht. Im konkreten Fall ist dies jedoch kein Problem, da die entsprechenden Texte auch zur Modellbildung verwendet wurden. In anderen Fällen und insbesondere bei einer drastischen Veränderung der Worthäufigkeiten kann dieser Schritt jedoch unangebracht sein. Daher empfiehlt es sich in diesem Fall das Modell von vorneherein neu aufzubauen, ein Vorgehen, dass in praktischen Anwendungen (z.B.: Suchmaschinen) oft nur schwer realisierbar, in der sozialwissenschaftlichen Forschung jedoch ohne weiteres möglich ist. Als letzter Schritt wird der Korpus mit den neu generierten und gewichteten Dokumentvektoren in den Raum der latenten Semantiken abgebildet.

```
1 ## Gruppieren der Texte:
```

Allgemein	Theorie/Gender	Arbeit	Soziale Klasse	Demographie	Religion	Diskriminierung	Rasse/Ethnizität	Wohlfahrtsstaat	Macht
women	women	class	class	children	religious	women	racial	children	exchange
social	employment	employment	religious	class	women	religious	black	political	network
work	sociology	work	women	family	religion	political	race	theory	sociology
political	children	women	class_analysis	political	political	work	employment	women	inequality
research	theory	religious	gender	sociology	children	workers	violence	parents	women
class	men	children	classes	women	religiosity	religion	theory	citizenship	networks
employment	family	workers	organizations	parents	secularization	citizenship	white	family	power
theory	political	educational	mobility	mobility	school	gender	german	state	organizations
sociology	action	labour	religion	organizations	educational	men	action	model	racial
economic	occupational	education	work	firms	gender	job	system	families	class
new	gender	labour_market	identity	organizational	network	class	identity	religious	actors
family	education	management	occupational	workers	identity	state	educational	action	ties
analysis	labour_market	school	social_class	state	inequality	employees	health	mothers	research
model	social	parents	organizational	mothers	exchange	management	education	welfare	firms
gender	concept	employees	children	work	citizenship	mobility	neighborhood	earnings	organizational
within	earnings	job	network	families	europe	labour_market	crime	status	children
different	marriage	effects	consciousness	countries	politics	labour	neighborhoods	men	segregation
effects	mothers	religion	population	inequality	participation	black	germany	network	marriage
article	sociological	inequality	effects	market	pluralism	feminist	countries	exchange	ethnic
paper	job	ethnic	rates	gender	racial	education	research	sociology	educational

Tabelle 6.5: Top-10 LSA-Dimensionen der Soziologie-Abstracts.

```
grouped = articles.groupby('ISO-4 Journal Abbreviation')
docs = grouped.StopCollTokens.sum()

## Korpus erstellen:
corpus = [dictionary.doc2bow(doc) for doc in docs]
corpora.MmCorpus.serialize('Daten/SozISO4.mm', corpus)

## TFiDF Gewichtung:
iso4_tfidf = tfidf[corpus]

## Folding-In und Repräsentation als Matrix:
data = [dict(journal) for journal in lsa[iso4_tfidf]]

## Als DataFrame
journal_df = abs(pd.DataFrame(data,
                              index=docs.index))
journal_df = journal_df.iloc[:,:10]
journal_df.columns = columns
```

Die Ergebnisse einer Betrachtung des Verhältnisses der Soziologiezeitschriften zu den ersten zehn semantischen Dimensionen findet sich in Tabelle 6.6. Darin wird zunächst einmal deutlich, dass die bereits beobachteten, relativen Ähnlichkeit auch auf der Ebene der Dimensionen zu finden sind. Zum Beispiel im Falle der relativ hohen Ähnlichkeit der deutschen Soziologiezeitschriften, die auch auf der Ebene der semantischen Dimensionen deutlich wird. Hinzu kommen hier aber auch noch die Möglichkeit Unterschiedlichkeiten inhaltlich zu interpretieren. Die Hauptunterschiede finden sich in diesem Falle hauptsächlich auf der Ebene der theoretischen Begrifflichkeiten mit einem gendertheoretischen Einschlag, einem Bereich also, der von den verwendeten Worten her eine gewisse Schnittmenge mit der kritischen Soziologie aufzuweisen scheint. Insbesondere das *Berliner Journal für Soziologie* scheint innerhalb dieser Gruppe eine hohe Affinität dazu aufzuweisen. Bedeutsam erscheint hier auch der Unterschied auf den Dimensionen „Demographie“ und „Wohlfartsstaat“, zwei Dimensionen die häufig zusammen betrachtet werden. Während *Berliner Journal* und *Zeitschrift für Soziologie* höhere Werte auf der „Demographie“-Dimension aufweisen, sind es beim Thema „Wohlfahrtsstaat“ vor allem die *Kölner Zeitschrift* und die *Zeitschrift für Soziologie*. Dies hängt wohl einerseits mit einer stärkeren Orientierung der *KZfSS* auf Makrosoziologien und Gesellschaftstheorien zusammen und andererseits mit einer gewis-

sen Ferne zur empirischen Sozialforschung, wie sie insbesondere in der Demographie zum Ausdruck kommt.

Auch die bereits in anderen Modellen gefundene Sonderstellung von *Work, Employment & Society* kann durch die Betrachtung der latenten Dimensionen inhaltlich besser erklärt werden. Hier fällt zunächst auf, dass die *WES* relativ betrachtet am weitesten vom allgemeinen Sprachduktus der Soziologie entfernt zu sein scheint (Dimension: „Allgemein"). Desweiteren weist sie relativ hohe Werte in den Bereichen „Theorie/Gender", „Arbeit" und „Diskriminierung" auf. Somit scheint dies Zeitschrift auch von ihren thematischen Bezügen her eher auf Fragen der Genderforschung, der Diskriminierung und des Arbeitslebens abzuzielen. Eine Richtung die in der deutschen Soziologie wohl am ehesten mit dem Begriff „kritische Soziologie" umschrieben werden könnte.

Diese Beobachtungen und Vergleiche geschehen jedoch unter der Prämisse eines geteilten, semantischen Raumes. Mit der Kombination der Dokumente zu einem gemeinsamen Vektorraum wurde diese Feststellung axiomatisch getroffen, die im Rahmen der darauf aufbauenden Verfahren nicht mehr überprüft werden kann. Für thematische Unterschiede ist dies normalerweise kein Problem, da diese sich in den Dimensionen niederschlagen würden. Systematische Unterschiede im Sprachgebrauch könnten hingegen zu Verzerrungen führen, müssten dazu jedoch auf fast alle Worte des Korpus zutreffen.

6.5.2 Generative Modelle

Das Verfahren der Analyse latenter Semantiken mittels einer Singulärwertzerlegung, wurde aufgrund enormer Erfolge in der Praxis zu einer wichtigen Grundlage weiterer Forschungsbemühungen im Bereich des Information Rettrievals (vgl. Landauer et al. 2013). Eine Reihe von Erweiterungen und neuer Verfahren, die ebenfalls von der Idee einer Modellierung der Themen und Bedeutungen eines Textes als latenten Dimensionen ausgingen, schlossen daran an. Im Unterschied zur LSA bauen viele dieser Verfahren auf statistischen Annahmen auf und versuchen Texte als Zufallsverteilungen zu beschreiben, weswegen man diesbezüglich auch von *probabilistischen Modellen* spricht (vgl. Hofmann 2001). Da die manifesten Verteilungen der Worte als eine Funktion latenter Zufallsverteilungen aufgefasst werden, wird hierfür in der Literatur auch der Begriff der „generativen Modelle" verwendet (vgl. Steyvers und Griffiths 2007). Die

	Allgemein	Theorie/Gender	Arbeit	Soziale Klasse	Demographie	Religion	Diskriminierung	Rasse/Ethnizität	Wohlfahrtsstaat	Beziehungen
Am. J. Sociol.	0.852785	0.013531	0.124362	0.129528	0.181193	0.072979	0.066942	0.155186	0.059106	0.126887
Am. Sociol. Rev.	0.868659	0.176518	0.175200	0.170147	0.123687	0.055036	0.065736	0.164676	0.034256	0.089164
Annu. Rev. Sociol.	0.807804	0.138138	0.065565	0.044074	0.060900	0.051034	0.068156	0.186692	0.008252	0.118775
Berliner J. Soz.	0.808181	0.259036	0.034275	0.050978	0.019091	0.069503	0.087408	0.236448	0.059800	0.082032
Z. Soziol.	0.846021	0.123979	0.040545	0.078918	0.056241	0.077804	0.007446	0.250766	0.098067	0.013587
Koln. Z. Soziol. Sozialpsych.	0.844453	0.013691	0.127403	0.095368	0.007910	0.078320	0.028164	0.268406	0.101094	0.011853
Sociol.-J. Brit. Sociol. Assoc.	0.854061	0.139713	0.061203	0.192441	0.229945	0.072647	0.041311	0.137620	0.076753	0.006368
Work Employ. Soc.	0.757616	0.238939	0.430087	0.063210	0.047994	0.016441	0.240635	0.036488	0.073874	0.028173
Br. J. Sociol.	0.874098	0.138622	0.037509	0.139526	0.084691	0.112062	0.023441	0.013479	0.075023	0.048544

Tabelle 6.6: Beziehungen zwischen den ausgewählten Soziologie-Zeitschriften und den ersten zehn Dimensionen des semantischen Raumes.

grundlegende Idee ist demzufolge eine Nachbildung des Prozesses mit dem Text generiert wird.

Die latenten Bedeutungen werden in diesen Verfahren als *Themen* (Topics) bezeichnet und als Zufallsverteilungen über alle Worte einer spezifischen Sprache angesehen. Entsprechend der Notation von Steyvers und Griffith (2007) kann die Auftretenswahrscheinlichkeit $P(w_i)$ eines bestimmten Wortes w_i in einem Text als eine Funktion der Themen $z = \{z_1, \ldots, z_T\}$ eines Korpus modelliert werden. Somit ergibt sich die Wahrscheinlichkeit eines spezifischen Wortes aus der Wahrscheinlichkeit für dieses Wort unter der Bedingung eines spezifischen Themas in diesem Dokument: $P(w_i \mid z_i = j)$ mal der allgemeinen Wahrscheinlichkeit für dieses Thema in diesem Dokument: $P(z_i = j)$. Daraus folgt

$$P\left(w_i\right) = \sum_{j=1}^{T} P(w_i \mid z_i = j)\, P(z_i = j)$$

für die Bestimmung dieser Wahrscheinlichkeiten in Abhängigkeit von allen Themen. Allgemeiner ausgedrückt, gibt dieses Modell die Wahrscheinlichkeit an, mit der ein „Autor" ein bestimmtes Thema auswählt und dementsprechend bestimmte Wörter wählt um dieses Thema zu „formulieren".

Um dieses allgemeine Modell nutzbar zu machen ist eine Schätzung der einzelnen Parameter notwendig. Die zu modellierenden Parameter sind zum einen die Wahrscheinlichkeit für ein Wort gegeben ein Thema $P(w_i \mid z_i = j)$, die als Modellparameter ϕ wiedergegeben wird, und zum anderen die Wahrscheinlichkeit für die Existenz eines Themas in einem Dokument $P(z_i = j)$, die durch den Parameter θ beschrieben wird. Weiterhin wird angenommen, dass beide Modellparameter einer multinomialen Verteilung folgen. Von einem solchen Modell der Textgenese ausgehend, entwickelten Blei et al. (2003) das Verfahren der *Latent Dirichlet Allocation* (LDA). Dabei wird angenommen, dass es eine Anzahl von k latenten Themen gibt, die einer multinomialen Verteilung folgen, welche annähernd Dirichlet verteilt ist: $\beta_k \sim Dir(\eta)$. Für eine im vorneherein festgelegte Anzahl von Themen lässt sich dann die Verteilung dieser Themen in einem Dokument θ_d durch die Ziehung aus Dir (α) schätzen. Der α-Parameter ist ein Vektor, der die zu erwartende Zufallsverteilung über alle Themen beschreibt: $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$. Darauf aufbauend können die restlichen Parameter des Modells bestimmt werden: „for each word i in the document, draw a topic index $z_{\text{di}} \in \{1, \ldots, K\}$ from the topic weights $z_{\text{di}} \sim \theta_d$ and draw the observed word w_{di} from the selected topic,

$z_{\text{di}} \sim \beta_{z_{\text{di}}}$“ (Hoffman, Bach und Blei 2010: 2). Zur Schätzung der Parameter dieses Modells wird ein *expectation-maximation* (EM) Algorithmus verwendet (vgl. Blei, Ng und Jordan 2003: 1005).

Anwendung

Aus der Konvergenz dieses Modells ergeben sich die ex post Wahrscheinlichkeiten für die beiden Modellparameter ϕ und θ. Da es sich um eine Mehrzahl von Themen, Typen und Dokumenten handelt, können die Modellparameter für den ganzen Korpus in Matrixform angegeben werden. Dabei beinhaltet Φ die posteriori Wahrscheinlichkeit mit der ein Wort ϕ_j auf ein bestimmtes Thema $z_i = j$ hinweist, während die Θ-Matrix die Wahrscheinlichkeiten θ_d für die Anwesenheit eines bestimmten Themas in einem Dokument d wiedergibt. Diese Aufteilung in zwei Matrizen, die zum einen die *Relation zwischen Typen und Dokumenten* sowie zum anderen die *Beziehung von latenten Themen und Dokumenten* wiedergibt, entspricht konzeptionell den Matrizen U und V^H der Latent Semantic Analysis. Allerdings handelt es sich im Fall der LDA um Wahrscheinlichkeitsvektoren und nicht um Singulärvektoren.

Ebenso wie im Falle der LSA, gibt es eine Reihe von Paketen, welche die LDA implementieren. Aus Gründen der Robustheit und Effizienz der Implementation wird an dieser Stelle erneut auf das Gensim-Paket zurückgegriffen. Im Prinzip ändert sich so nur die Wahl des zur Modellierung verwendeten Objektes (`models.LdaModel()`).

Ein weiterer Unterschied besteht jedoch auch in der Gewichtung. Es finden sich eine Reihe von Tutorials und Fachmeinungen, die auch im Falle der Latent Dirichlet Allocation eine Verwendung der TFiDF-Gewichtung anraten. Dies ist jedoch mit Vorsicht zu genießen, da die Modellspezifikation mittels einer Dirichlet-Verteilung eigentlich diskrete Werte voraussetzt. Da bisher noch keine überzeugenden Argumente für eine Gewichtung vorgebracht oder systematische Forschungen zu möglichen Verzerrungen durchgeführt wurden, wird im Folgenden auf den Einsatz von TFiDF verzichtet.

Ausgangslage ist auch hier das von Gensim verwendete Format einer Sparse Matrix. Daher kann das bereits zugewiesene `corpus`-Objekt ebenso wie das Diktionär weiterverwendet werden. Die praktische Berechnung des Modells durch die `LdaModel` Funktion unterscheidet sich dementsprechend kaum von der Durchführung der LSA. Allerdings handelt es sich bei den resultierenden Themen, die ebenfalls mit `.show_topic()` angezeigt

werden können, um die Wahrscheinlichkeit mit der das jeweilige Wort ein ausgewähltes Thema beschreibt.

```
lda10 = models.LdaModel(corpus,
                        id2word=dictionary,
                        num_topics=10)

lda30 = models.LdaModel(corpus,
                        id2word=dictionary,
                        num_topics=30)

lda10.show_topic(0, 10)
```

```
[(u'political', 0.0098920340557711803),
 (u'among', 0.005747660176139811),
 (u'class', 0.0050643531848046999),
 (u'state', 0.0042045745182763697),
 (u'data', 0.004163722932610866),
 (u'social', 0.0040058956325786273),
 (u'women', 0.003994820867623196),
 (u'analysis', 0.0039386197832171049),
 (u'economic', 0.0031726729595395191),
 (u'identity', 0.00317230953308313)]
```

Trotz des ähnlichen Ablaufs ergeben sich eine Reihe von Unterschieden in der Verwendung und Interpretation dieses Modells. Da der EM-Algorithmus die Parameter des Modells schätzt, indem er iterativ *Zufallsschätzungen* vornimmt, die auf den Parametern der vorangegangenen Iteration beruhen, ergibt sich je nach gewählter Anzahl der Themen k ein anderes Modell. Zudem muss bedacht werden, dass der EM-Algorithmus auf Zufallsziehungen aufbaut und daher jede neue Berechnung des Modells Themen produziert, die sowohl in ihrer Reihenfolge als auch in den sie beschreibenden Worten voneinander abweichen.

Zur weiteren Bearbeitung des Modells stehen dieselben Techniken wie im Falle der LSA zu Verfügung. Zum Beispiel können Folding-In sowie Ähnlichkeitsmaße auf die gleiche Art und Weise berechnet werden. Da das Resultat der LDA Wahscheinlichkeitsvektoren sind können diese auch

entsprechend in weiterführende, statistische Analysen eingebunden werden.

Besonderheiten der Interpretation

Im Folgenden werden zwei LDA-Modelle mit $k = 10$ (siehe Tabelle 6.7 auf Seite 284) und $k = 30$ (siehe Tabelle 6.8 auf Seite 286) miteinander verglichen, um ein besseres Verständnis für die Besonderheiten eines LDA Modells zu erhalten. Ein erster Vergleich der beiden Tabellen deutet darauf hin, dass das Verfahren anscheinend dazu neigt, bei vergleichsweise niedrigen Werten von k, Worte mit relativ hohen Häufigkeiten zur Beschreibung der Themen zu verwenden. Bei einer Betrachtung der zehn ausgewählten Themen und dem Vergleich mit dem vorangegangenen Modell – welches mit $k = 10$ initialisiert wurde – fällt zunächt einmal die größere Differenziertheit der Tehmen für $k = 30$ ins Auge. So findet sich zum Beispiel das Wort „social" im ersten Modell neun mal unter den ersten drei Wörtern, die ein Thema beschreiben. Demgegenüber ist dies im zweiten Modell nur ca. in der Hälfte der Themen der Fall. Generell lässt sich erkennen, dass das zweite Modell bereits sehr viel kleinteiligere und deutlich abgegrenztere Themen identifiziert. Dieser Unterschied kann auch dadurch erklärt werden, dass eine geringen Anzahl von wählbaren Themen den Algorithmus dazu zwingt kleinere Unterschiede in den Häufigkeiten für eine Entscheidung heranzuziehen.

Im Vergleich dazu erzeugt das Modell mit $k = 30$ eine Reihe von Themen die sich besser inhaltlich von einander abgrenzen lassen. Da sich, wie bereits erwähnt, die Zuordnung der Themen zu den Worten mit jedem Durchlauf verändert und die Darstellung von 30 verschiedenen Themen hier nicht zweckmäßig erscheint, wird in Tabelle 6.8 nur eine kurze Auswahl an relativ stabilen und interpretierbaren Themen wiedergegeben. Relativ stabil bedeutet in diesem Fall, dass sich die entsprechenden Worte, die diesen Themen zugeordnet wurden, sich durch eine erneute Berechnung des Modells nicht so sehr verschieben, dass eine Interpretation nicht mehr möglich wäre.

Es lassen sich in den ausgewählten Themen drei unterschiedliche Arten von Semantiken ausmachen. Zum einen finden sich Themen, die bestimmte *Forschungsgegenstände* der Soziologie beschreiben. In dieser Auswahl zählen dazu die Themen: 7 (Organisationen), 8 (Gender), 9 (Handlungstheorie), 17 (ethnische Segregation) und 19 (Identitäten/Rollen). Im Unterschied dazu scheinen die Dimensionen 11, 26 und 29 eher diejenigen Begrifflichkeiten abzudecken, die mit unterschiedlichen *Perspekti-*

Thema 0	Thema 1	Thema 11	Thema 17	Thema 19	Thema 26	Thema 29	Thema 7	Thema 8	Thema 9
power	children	political	racial	identity	model	social	organizational	women	social
social	blacks	cultural	black	ethnic	effects	political	organization	research	theory
reform	foreign	social	neighborhood	religious	size	chapter	markets	data	collective_action
american	social	historical	neighborhoods	segregation	authors_find	party	social	social	action
theory	investment	review	whites	gender	educational	structure	network	countries	immigrant
officials	law	politics	state	identities	rates	support	organizations	effects	collective
two	justice	movements	ties	social	social	sociology	sociology	differences	theories
theories	influence	states	conflict	asian	protest	state	work	men	trust
also	same-sex	war	work	also	labor_markets	inequality	networks	models	model
family	parental	economic	residential_segregation	within	performance	institutional	exchange	new	article
individual	capital	institutional	race	women	data	analysis	forms	employment	scores
kin	new	process	also	work	using	welfare	new	using	interactions
sex_segregation	volunteering	society	police	change	theory	class	within	also	system
injustice	wealth	theory	institutional	use	family	social_movements	groups	political	actors
analysis	courts	programs	african_americans	different	education	elections	analysis	female	process

Tabelle 6.7: Latent Dirichlet Allocation der Soziologie-Abstracts ausgewählter Zeitschriften für $k = 10$ Themen.

ven und Zielsetzungen der Forschung einhergehen. Thema 11 zeichnet sich vor allem durch Worte aus die typisch für makrosoziologische Forschungen sind, wie zum Beispiel „political“, „historical“ und „institutional“. Bei Thema 26 sind hauptsächlich Wörter und Phrasen zu finden, die auf empirisch ausgerichtete Forschung hindeuten, während Thema 29 eine politische Ausrichtung andeutet. Die Themen 0 und 1 weichen davon insofern ab, als es sich hier anscheinend um *Formen der Konnotation* handelt. So wird Thema 0 durch Worte beschrieben, in denen eine politische Kritik mitschwingt, wie „reform“, „injustice“ oder „power“. Ähnliches trifft wohl auch auf Thema 1 zu, welches Minderheiten thematisiert und mit Begrifflichkeiten wie „law“ und „justice“ in Verbindung bringt, was den Schluss nahelegt, dass es hier um den Schutz von Minderheiten vor Diskriminierung geht.

Wenn aber davon auszugehen ist, dass sich die Zuordnungen der Themen zu den Wörtern bei gleichbleibendem k und wiederholter Berechnung des Modells verändern, wie ist dann eine sinnvolle Interpretation möglich? Als eine erste Lösungsstrategie sollten stets multiple Berechnungen des gleichen Modells vorgenommen werden, damit ein Gefühl für die enthaltenen Themen gewonnen werden kann. Lässt sich dabei feststellen, dass ein Thema relativ invariant in seiner Bedeutung bleibt, also wiederholt das selbe Muster in den Wörtern zum Ausdruck kommt, so kann es plausibel als ein Thema interpretiert werden.

Dies ist allerdings nur eine notwendige Vorausbedingung und kein hinreichender Lösungansatz. Ob ein bestimmtes theoretisch relvantes oder explorativ gefundenes Thema existiert, sollte auch deduktiv geprüft werden. Dazu kann man sich der Techniken Verfahren des Information Retrievals bedienen. Indem ein „künstlicher“, theoretisch hergeleiteter Dokumentvektor erzeugt wird, kann geprüft werden ob die Interpretation eines spezifischen Themas stimmig ist. Die aus der Diskussion der LSA bereits bekannte Technik des Folding-In erlaubt es über eine rein explorative Nutzung semantischer Modelle hinauszugehen und Themen gezielt und unter Rückgriff auf theoretische Überlegungen zu modellieren. Im Folgenden wird dafür ein Vektor (`gendervec`) erzeugt, der zentrale Schlagworte des Themas „Gender“ ausdrücken soll.

```
1 genderdoc = ['women', 'men', 'gender', 'sex']
2
3 gendervec = dictionary.doc2bow(genderdoc)
4
```

Thema 0	Thema 1	Thema 11	Thema 17	Thema 19	Thema 26	Thema 29	Thema 7	Thema 8	Thema 9
power	children	political	racial	identity	model	social	organizational	women	social
social	blacks	cultural	black	ethnic	effects	political	organization	research	theory
reform	foreign	social	neighborhood	religious	size	chapter	markets	data	collective_action
american	social	historical	neighborhoods	segregation	authors_find	party	social	social	action
theory	investment	review	whites	gender	educational	structure	network	countries	immigrant
officials	law	politics	state	identities	rates	support	organizations	effects	collective
two	justice	movements	ties	social	social	sociology	sociology	differences	theories
theories	influence	states	conflict	asian	protest	state	work	men	trust
also	same-sex	war	work	also	labor_markets	inequality	networks	models	model
family	parental	economic	residential_segregation	within	performance	institutional	exchange	new	article
individual	capital	institutional	race	women	data	analysis	forms	employment	scores
kin	new	process	also	work	using	welfare	new	using	interactions
sex_segregation	volunteering	society	police	change	theory	class	within	also	system
injustice	wealth	theory	institutional	use	family	social_movements	groups	political	actors
analysis	courts	programs	african_americans	different	education	elections	analysis	female	process

Tabelle 6.8: Latent Dirichlet Allocation von zehn ausgewählten Themen der Soziologie-Abstracts für ein Gesamtmodell mit $k = 30$ Themen.

```
5 lda30[gendervec]
```

```
1 [(8, 0.80666666666618536)]
```

Es zeigt sich, dass der künstliche Dokumentvektor mit einer Wahrscheinlichkeit von $\theta_{\text{gender}} = 0{,}81\%$ das Thema 8 enthält, welches bereits in der Betrachtung der Wahrscheinlichkeitsvektoren der Worte ϕ_8 als das Thema „Gender“ interpretiert wurde. Des Weiteren bedeutet die Rückgabe nur eines relevanten Themas, dass die Wahrscheinlichkeit für ein anderes Thema im Rahmen dieses Modells verschwindend gering ist. Dies unterstützt die bestehende Interpretation dieses Themas, da die gewählten Worte auch in einer Reihe anderer Themen eine Rolle spielen.

Eine weitere Kontrastierung der bestehenden Interpretation mit „künstlichen“ Vektoren scheint deren Angemessenheit zu bestätigen. Ersetzt man das Wort „men“ durch „identities“, so wird erwartungsgemäß Thema 19 (Identitäten) als wahrscheinlichstes Thema ausgegeben. Dies illustriert auch noch einmal die Trennschärfe des Modells.

```
1 identdoc = ['women', 'identities', 'gender', 'sex']
2
3 identvec = dictionary.doc2bow(identdoc)
4
5 lda30[identvec]
```

```
1 [(19, 0.8066666666665282)]
```

Ein solches Folding-In kann auch genutzt werden um zu prüfem, wie gut die Wahl der Themenanzahl die beschriebenen Themen abbildet. Übergibt man den Dokumentvektor `gendervec` an das mit zehn Themen initialisierte LDA-Modell, so ergibt sich ein sehr viel uneinheitlicheres Bild. Zwar kristalliert sich auch in diesem Fall ein Thema als Gewinner heraus (Thema 7), jedoch deutet die Überschneidung der Wahrscheinlichkeiten der restlichen Themen darauf hin, dass in diesem Fall eine zu niedrigrangige Repräsentation für das Modell gewählt wurde. Die hohe Ähnlichkeit der Wahrscheinlichkeiten ($0{,}02$) ist ein Zeichen dafür, dass das Modell nicht in der Lage ist den Vektor eindeutig zuzuordnen, weil ein Großteil der Wortwahrscheinlichkeiten gleichmäßig über die Themen

verteilt ist. In diesem Fall ist das theoretisch postulierte Thema sozusagen über den Raum der möglichen Themen verstreut.

```
lda10[gendervec]
```

```
[(0, 0.020002055064330389),
 (1, 0.020000848056885068),
 (2, 0.020001339959204188),
 (3, 0.020002268915145619),
 (4, 0.020001378035562366),
 (5, 0.020000597566554697),
 (6, 0.020001353392014708),
 (7, 0.81998371642600298),
 (8, 0.0200009422066522),
 (9, 0.020005500377647679)]
```

Das letzte Beispiel erinnert zudem noch einmal an die besondere Bedeutung, die der *Wahl der Themenanzahl* im Falle der LDA zukommt. Zwar kann auch hier mit dem gewünschten Auflösunggrad als Entscheidungsgrundlage argumentiert werden, allerdings gibt es in diesem Fall weniger objektive Entscheidungskriterien zur Begründung der Themenanzahl. Zudem bauen, anders als im Falle der LSA, diese Themen nicht sukzessive aufeinander auf. Aus diesem Grunde wurden statistische Verfahren wie zum Beispiel der *Hierarchical Dirichlet Process* (HDP) entwickelt, die zusätzlich noch die optimale Anzahl der Themen schätzen (vgl. Teh et al. 2006; Wang, Paisley und Blei 2011). Allerdings wurde diese Verfahren vor dem Hintergrund praktischer Überlegungen zur weitestgehenden Automatisierung von Indexierungssytemen entwickelt. Da man allerdings davon ausgehen muss, dass sprachliche Äußerungen ein Hierarchie von Bedeutungen auf unterschiedlichsten Ebenen aufweisen, kann es so etwas wie eine einzig „richtige“ Anzahl von Themen nicht geben. Zumindest nicht in dem Sinne, dass die zugrundeliegenden, symbolischen Ordnungen nur durch diese spezifische Menge latenter Dimensionen adäquat beschrieben werden könnte.

Im Endeffekt kommt es auch in diesem Fall sehr stark auf die theoretischen Begründungen und die damit einhergehende Plausibilisierung der Ergebnisse an. Man muss sich allerdings auch vor Augen führen, dass die hier diskutierten Techniken nicht von sich aus auf ein exploratives Vorge-

hen begrenzt sind. Vielmehr stellen sie eine Vielzahl von Möglichkeiten bereit, um die eigene Interpretationsleistung systematisch zu prüfen. Dies scheint notwendig, da gerade bei der Modellierung symbolischer Ordnungen die Tendenz des Menschen zur Mustererkennung kritisch berücksichtigt werden muss. Gerade weil wir keine Anschauungsform für Zufall besitzen, sind systematische, deduktive Tests unserer explorativ gewonnen Erkenntnisse unverzichtbar.

6.5.3 Dimensionen und Themen

Auf den ersten Blick ähneln sich die Verfahren der Analyse latenter, semantischer Dimensionen und die Modellierung probabilistischer Themen. Dies wird insbesondere an einem ähnlichen Aufbau der aus der Anwendung der Verfahren resultierenden Matrizen und deren „Weiterverarbeitung" deutlich. Sowohl zur Bestimmung der Ähnlichkeiten von Dokumenten als auch beim Folding-In werden die Zufallsverteilungen der generativen Modelle als geometrische Räume interpretiert. Auch in der Literatur findet man bereits in den ersten Formulierungen probabilistischer Modelle immer auch eine geometrische Interpretation der resultierenden Themen (vgl. Hofmann 1999; Blei, Ng und Jordan 2003; Steyvers und Griffiths 2007). Der Grund hierfür liegt einerseits sicherlich in der Prominenz der LSA als Wegbereiter der Exploration und Modellierung von Bedeutungen, hat zum anderen seine Wurzeln aber auch in der grundsätzlichen Vorstellung von Bedeutung als Strukturen die durch ihre Relationen bestimmt werden.

Die grundlegenden Gemeinsamkeiten sind jedoch nicht so weitreichend, wie die Metapher des semantischen Raums vermuten lässt. In ihrer Diskussion der LDA als einer Technik der Matrixzerlegung stellen Steyvers und Griffiths eine Reihe von Unterschieden zwischen den Verfahren der LDA und LSA fest:

> In topic models, the word and document vectors of the two decomposed matrices are probability distributions with the accompanying constraint that the feature values are non-negative and sum up to one. In the LDA model, additional a priori constraints are placed on the word and topic distributions. There is no such constraint on LSA vectors, although there are other matrix factorization techniques that require non-negative feature values [...] Second, the LSA decomposition provides an orthonormal basis which is computationally

> convenient because one decomposition for T dimensions will simultaneously give all lower dimensional approximations as well. In the topic model, the topic-word distributions are independent but not orthogonal; model inference needs to be done separately for each dimensionality. (Steyvers und Griffiths 2007: 430)

Diese technischen Unterschiede, die Zerlegung in Wahrscheinlichkeiten anstatt Singulärvektoren sowie die orthonormale Basis der LSA, wurden zum großen Teil bereits in den vorangegangenen Beispielen praktisch verdeutlicht. Hier sollen jedoch deren methodische und methodologische Implikationen noch einmal im Detail diskutiert werden.

Das Resultat der generativen Modelle sind *Wahrscheinlichkeitsverteilungen*. Dies hat zur Folge, dass ihre Interpretation geradliniger und auch einfacher vorstellbar ist als dies bei den Singulärvektoren der LSA der Fall ist. Gewichtiger ist jedoch die Tatsache, dass sich Wahrscheinlichkeiten besser in das methodische Paradigma der Sozialwissenschaften einfügen, d.h. sie leichter in weiterführende Modelle zu integrieren sind.[15] So lässt sich zum Beispiel eine Veränderung der Wahrscheinlichkeiten für thematische Bezüge durch ein Modell gemischter Autorenschaften ergänzen, indem die Verteilung der Autoren über die Dokumente als eine weitere, bedingte Zufallsverteilung angesehen wird (vgl. Rosen-Zvi et al. 2004).

Es muss aber auch festgestellt werden, dass die „arbiträren Achsen" (Steyvers und Griffiths 2007: 426), die das Ergebnis der Singulärwertzerlegung darstellen, sowohl eine klare mathematische Interpretation haben als auch die Folge eines deterministischen Verfahrens sind. Da sie unabhängig von Zufallselementen ist, hat die Singulärwertzerlegung eines spezifischen Raums immer das gleiche Ergebnis zur Folge.[16] Zudem muss hervorgehoben werden, dass die geometrische Repräsentation in diesem Fall und im Gegensatz zu den generativen Modellen keine im nachhinein hinzugefügte Sichtweise ist, sondern essentieller Bestandteil des Verfahrens selbst. Die Auffassung, dass es sich bei Bedeutungen um die laten-

[15] Grundsätzlich lässt natürlich auch die der LSA zugrundeliegende lineare Algebra weitreichende Erweiterungen des Modells zu. Allerdings finden sich in den Sozialwissenschaften nur sehr wenige Modelle die von linearer Algebra Gebrach machen, so dass die Anschlussfähigkeit hier sehr viel geringer ist. Nennenswerte Ausnahmen sind hier unter anderem die Arbeiten von Breiger und Pattison (1986) sowie Borgatti und Everett (Borgatti und Everett 1989). Zumindest implizit liegt hierin auch die Grundlage des bourdieuschen Sozialraums, der in methodischer Hinsicht das Ergebnis einer räumlichen Faktorisierung darstellt.

[16] Wenn man von Rundungsfehlern und algorithmischen Unterschieden einmal absieht.

ten Dimensionen eines Diskurses handelt, ist die grundlegende Prämisse der LSA. Diese Konzeption weist eine große Nähe zu der im Rahmen dieser herausgearbeiten und in Theorie weit verbreiteten Auffassung, dass es sich bei symbolischen Ordnungen um grundlegende Muster in den (räumlichen) Relationen von Symbolen handelt.

Vor dem Hintergrund des Unterschiedes von gesellschaftstheoretischen und individualistischen Perspektiven muss man zudem hinzufügen, dass es sich bei den generativen Modellen streng genommen nicht um Faktorisierungen einer Matrix handelt, auch wenn sie in der Literatur gerne als solche interpretiert werden. Es handelt sich schon deshalb nicht um eine Faktorisierung, weil es keine Möglichkeit der Rückübersetzung der Komponenten in die ursprüngliche Matrix gibt. Der methodologische Unterschied liegt jedoch tiefer. Die generativen Modelle gehen von einer a priori festgelegten, latenten Verteilung der Themen aus, welche die Grundlage für die Auswahl der Worte stellen. Somit sind die Themen *keine linearen Kombinationen* aller Worte des Korpus. Dies hat zur Folge, dass Themen sehr viel klarer voneinenander abgegrenzt werden, solange genug potentielle Themen zur Verfügung stehen.

Nehmen wir beispielsweise an, dass ein Korpus das Resultat zweier sehr verschiedener Diskurse ist, zum Beispiel der Ökonomie und der Soziologie. Wird ein generatives Modell mit einer ausreichenden Anzahl von Themen initialisiert, so würden die Themen beider Diskurse sich im Modell niederschlagen. Allerdings bedeutet dies auch, dass ein „überschätzen" der Themen prinzipiell möglich ist, dass also Themen konstruiert werden, die alleine der Tatsache geschuldet sind, dass das Modell Themen zu vergeben hat. Themen werden hierbei tendenziell von den „individuellen", symbolischen Äußerungen her konstruiert, während die Dimensionen der LSA als eine Reduktion des gesamten Diskurses aufgefasst werden kann.

Die Repräsentation symbolischer Ordnungen als latente Dimensionen geht zudem davon aus, dass es sich tatsächlich um einen gemeinsamen, semantischen Raum handelt. Besteht dieser jedoch aus distinkten Diskursen, so würde dieser Unterschied im Wortgebrauch, je nach Stärke und Häufigkeit, auf eine der vorderen Dimensionen entfallen. Grundsätzlich würden alle anderen Themen im Vergleich dazu orthogonal konstruiert, was zur Folge hätte, das die thematischen Unterschiede zwar dimensional repräsentiert werden würden, nicht aber die entsprechenden Themen innerhalb des jeweiligen Diskurses.

6.6 Sozio-semantische Ansätze

Der hauptsächlich Unterschied zwischen der Netzwerk-Text-Analyse und der Konzeption von symbolischen Ordnungen als latenten Semantiken besteht im zugrundeliegenden Konzept des Raumes. Im Falle der latenten Semantiken wird sowohl in der Konstruktion als auch in der Interpretation von einem euklidischen Raum ausgegangen, in dem das Verhältnis der einzelnen Elemente als eine Funktion der Dimensionen dieses Raumes bestimmt ist. Demgegenüber stellen Netzwerke geodätische Räume dar, die über direkte Relation zwischen den einzelnen Elementen definiert sind, d.h. die Distanzen der Elemente hängen nur von der Position dieser Elemente zueinander ab. Dies hat zur Folge, dass Netzwerke nur Beziehungen und Objekte darstellen können, die relativ unabhängig voneinander existieren. Gleichzeitig macht sie dies wenig geeignet zur Beschreibung von euklidischen Räumen. Prinzipiell wäre es möglich einen euklidischen Raum, als einen vollständig verbundenen Netzwerk-Graphen mit gewichteten Kantenzügen darzustellen, allerdings ist kaum ersichtlich, welchen Vorteil eine solche Darstellung hätte. Da komplette Graphen keine strukturellen Unterschiede zwischen ihren Knoten aufweisen, sind sie aus netzwerkanalytischer Sicht nicht unterscheidbar, was eine schwerwiegende Limitation der Analysemöglichkeiten zur Folge hat.

Während die soziologische Netzwerkanalyse in ihrer formativen Phase, Mitte der Siebziger Jahre, versuchte Strukturen in den Fokus zu rücken und damit eine gewisse Oposition zum Konzept der Kultur vertrat (vgl. Breiger 2010: 37), ist in den letzten Jahren vom *cultural turn* der Netzwerkanalyse die Rede (vgl. White et al. 2007; Fuhse 2009; Fuhse und Mützel 2010). In diesen Ansätzen übernehmen Kultur und Bedeutung oft die Rolle von konstitutiven Elementen für die Formierung und Interpretation von Beziehungen oder werden als essentieller Bestandteil von bestehenden Beziehungsstrukturen aufgefasst. Darin unterscheiden sich diese Auffassungen von der bereits vorgestellten Theorie semantischer Netzwerke, die Bedeutung als ein eigenständiges Phänomen auffasst und Netzwerke zu dessen Modellierung heranzieht. Einfach ausgedrückt kann man sagen, dass es um den Unterschied von Bedeutungen in Netzwerken und Bedeutungen als Netzwerken geht. In jüngster Zeit ist ein Kombination dieser Sichtweisen vorgenommen worden, welche sich auf die Analyse sogenannter „socio-semantic patterns“ konzentriert (vgl. Taramasco, Cointet und Roth 2010).

Die grundlegende Idee ist dabei die Kombination und Analyse von Bedeutungen und anderen Beziehungsmerkmalen in einem umfassenden, netzwerktheoretischen Modell. Um dies zu erreichen werden die unterschiedlichen Ebenen der Beziehungen zwischen Akteuren, Konzepten und anderen Entitäten mittels graphtheoretischer Konzepte zusammengebracht. In seinem Überblick über diese im Entstehen begriffenen Forschungsbereich zieht Camille Roth drei Möglichkeiten für eine solche Modellierung in Betracht: „at the level of a socio-semantic graph $\mathcal{G}$, a socio-semantic hypergraph $\mathcal{H}$ or a socio-semantic lattice $\mathcal{L}$" (Roth 2013: 22). Im Unterschied zum bereits vorgestellten Konzept eines Graphen, handelt es sich bei einem *Hypergraphen* um einen Graphen mit Kantenzügen, die aus n-Tupeln bestehen, welche auch als „Hyperedges" bezeichnet werden. Bei einem *Lattice* (Gitter) handelt es sich um eine regelmäßige Menge von Punkten. Je nachdem welches Kriterium für die Regelmäßigkeit herangezogen wird, ergibt sich ein anderes Gitter.

Sowohl bei Hypergraphen als auch bei Gittern handelt es sich um sehr spezifische und komplexe Modelle, die aufgrund ihrer relativen Neuheit kaum ausgereifte Analyseverfahren anbieten. Dies wird bereits deutlich, wenn man *bipartite Graphen* betrachtet, die ausschließlich Beziehungen zwischen zwei distinkten Mengen von Knoten zulassen, innerhalb dieser aber keine Verbindungen aufweisen. Diese Art von Graphen, die auch als *two-mode Netzwerke* bezeichnet werden, ist bereits seit längerem in der Netzwerkanalyse etabliert. Dennoch sind sie im Vergleich zu eindimensionalen Graphen noch relativ wenig erforscht (vgl. Latapy, Magnien und Vecchio 2008). Was bereits für die bipartiten Graphen gilt ist umso mehr ein Problem für die komplexeren Modelle von Netzwerken. Es bedarf noch eine Reihe weiterer Forschungen, um ein Verständnis für die Bedingungen zu gewinnen unter denen sich diese Strukturen formieren.

Aus diesem Grund, und weil es hier in erster Linie um die Analyse symbolischer Ordnungen und nicht um allgemeine Fragen der Netzwerktheorie geht, wird sich im Folgenden auf die Modellierung der semantischen Elemente im Rahmen eines sozio-semantischen Ansatzes konzentriert. Wie auch Roth (2013: 23) feststellt, besteht eine große Notwendigkeit einer solchen Weiterentwicklung der semantischen Bereiche:

> It also needs to be enriched on the side of information description—our way of appraising mental representations is at best sketchy (n-grams), at worst erroneous (for instance, by generally assuming some sort of perfect copying process in studies focused on contagion). More broadly, while we now

> have good knowledge of social network processes, we still need to enhance our description of local cognition processes.

Diese Kritik schließt auch an die bereits thematisierten Probleme in der Repräsentation von symbolischen Ordnungen durch Netzwerke von Worten oder Phrasen an. Im Resultat führt dies entweder zu einer Gleichsetzung von bestimmten Worten mit symbolischen Ordnungen oder zur Fixierung von Wortbedeutungen unabhängig vom Verwendungskontext.

6.6.1 Konstruktion von Autor-Themen Netzwerken

Als ein erster Lösungsansatz bietet es sich an die Semantiken mittels Verfahren der dimensionalen Modellierung latenter Semantiken zu operationalisieren und die Beziehungsebene durch Netzwerke darzustellen. Im Prinzip kommen hier eine Reihe von Verfahren in Frage, die entsprechend der oben erwähnten Kriterien gewählt werden können. Im Folgenden wurde zu diesem Zweck die LSA herangezogen, die sich für ein exploratives Vorgehen besser eignet, da in diesem Fall die Dimensionen stets gleich bleiben. Unter dem Rückgriff auf das bereits erstellte LSA-Modell können die Beziehungen der Autoren zu diesen Konzepten konstruiert werden, um daran anschließend die Frage zu verfolgen wie sich das Verhältnis von Authoren und Themen aus einer strukturellem Perspektive beschreiben lässt.

Zur Konstruktion des Netzwerkes wurden erst die einzelnen Autoren ausgewählt und deren Texte zu einem gemeinsamen „Werk“ kombiniert. Das Resultat des Folgenden Codes ist eine `pandas.Series` deren Index die Autorennamen darstellen, während die Liste der kombinierten Token die Elemente dieser Serie darstellt.

```
authors = articles.Authors.str.split(' ?; ?',
                                      expand=True)
authors = authors.stack().dropna(None)
authors.name = 'Authors'
authors.index.names = ['key', None]
a_df = pd.DataFrame(authors)

tokens = articles['StopCollTokens']
tokens.name = 'Tokens'
tokens.index.names = ['key']
```

```
t_df = pd.DataFrame(tokens)

author_df = t_df.join(a_df, how='inner')

works = author_df.groupby('Authors').Tokens.apply(sum)
```

Diese Elemente werden mittels des bereits erzeugten `dictionary`-Objekts in eine Sparse Matrix Darstellung überführt. Anschließend wird mittels Folding-In die Bedeutung der einzelnen Dimensionen für das Werk des jeweiligen Autors ermittelt. Dies kann als eine passive Projektion der Textvektoren in den semantischen Raum aufgefasst werden. Passiv deshalb, weil sich die bestehenden Dimensionen dadurch nicht verändern.

```
works_corpus = corpus = [dictionary.doc2bow(doc)
                         for doc in works]

works_tfidf = tfidf[works_corpus]

works_lsa = lsa[works_tfidf]

works_df = pd.DataFrame([dict(doc)
                         for doc in works_lsa],
                        index=works.index)

works_df = abs(works_df)

works_df = pd.DataFrame([dict(doc)
                         for doc in works_lsa],
                        index=works.index).iloc[:, 1:10]
```

Aus dem resultierenden Datensatz wurde die erste Dimension der „allgemeinen“ Begrifflichkeiten entfernt, da diese wie schon ausgeführt, einen Großteil der strukturellen Unterschiede zwischen den Texten umfasst und daher nur kaum mit den anderen Dimensionen vergleichbar ist. Dies wird auch bei der Betrachtung der Σ-Werte (siehe Abbildung 6.5 auf Seite 274) deutlich. Letztlich wurden die neun darauf folgenden Dimensionen – in der Abbildung mittels der grünen und roten Linie kennlich gemacht – für die Konstruktion des sozio-semantischen Netzwerks ausgewählt.

Verglichen mit einem LDA-Modell hat die Verwendung der LSA-Vektoren einen entscheidenden Nachteil. Aufgrund der mathematischen Eigenschaften der SVD hat jeder Dokumentvektor auch immer eine Verbindung zu allen Dimensionen, auch wenn diese relativ gering sein mag. Da ein vollständig verbundenes Netzwerk keine sinnvolle Interpretation im Sinne der Verfahren der Netzwerkanalyse ermöglicht, muss ein Abbruchkriterium für die Existenz eines Kantenzuges zwischen dem Werk eines Autors und den Dimensionen des latenten Raumes gefunden werden. Im Folgenden wird dafür der Durchschnitt der absoluten Werte der jeweiligen Dimension herangezogen. Ist der Wert des jeweiligen Autors höher, so wird eine Beziehung zur semantischen Dimension angenommen.

```
1 works_df = abs(works_df)
2
3 edgelist = []
4 for key in works_df:
5     edges = works_df[works_df[key] > works_df[key].mean()]
6     edges = edges.index.tolist()
7     tuples = list(zip(edges, [key]*len(edges)))
8     edgelist.extend(tuples)
9
10 BG = nx.Graph(edgelist)
```

Der aus der Liste der Kantenzüge gebildete Graph ist ein ungerichtetes, bipartites Netzwerk. Der Umgang mit dieser Art von Netzwerken sowie deren Analyse erfordert spezifische Algorithmen, die im Falle von NetworkX explizit importiert werden müssen.

```
1 import networkx.algorithms.bipartite as bp
```

Zur Visualisierung des Netzwerkes (siehe Abbildung 6.6) wurde das *fruchterman-reingold force directed* Layout verwendet (vgl. Fruchterman und Reingold 1991). Dabei wird versucht das Netzwerk auseinander zu ziehen, während die Verbindungen zwischen den einzelnen Knoten als ein Gegengewicht fungieren. Damit ergibt sich tendenziell eine Darstellung in der nah beeinander liegende Knoten relativ stark und exklusiv miteinander verbunden sind. Allerdings handelt es sich hier immer nur um einen groben Überblick, da der Algorithmus von zufällig gewählten Startposi-

tionen ausgeht und das Endergebnis daher variieren kann. Zudem ist die menschliche, visuelle Wahrnehmung im Allgemeinen nicht gut in der objektiven Einschätzung von Nähe und Distanzen auf zweidimensionalen Flächen. Die Größe der Knoten entspricht in dieser Darstellung der Anzahl der Verbindungen.

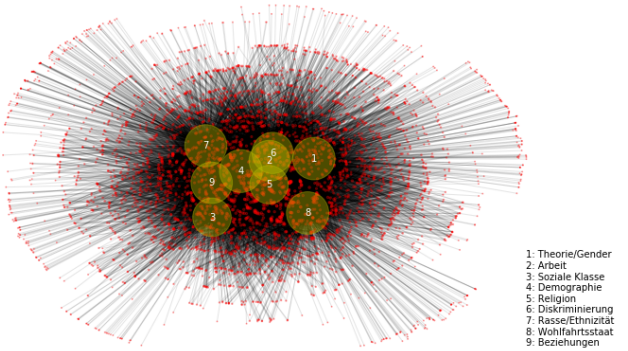

Abbildung 6.6: Bipartiter Graph von Akteuren sowie deren Beziehungen zu den latenten Themen einer LDA.

Auf den ersten Blick weist der Graph auf eine gewisse Sonderstellung für die Dimensionen „Theorie/Gender“ sowie „Demographie“ hin. Sie scheinen in keiner so engen Beziehung zu den anderen Themen zu stehen, wie es zum Beispiel bei den Dimensionen 7 („Rasse/Ethnizität“) und 8 („Wohlfahrtsstaat“) der Fall ist. Das heißt jedoch nicht notwendigerweise, dass sie weniger mit den anderen Dimensionen zu tun haben, es könnte auch der Fall sein, dass sie relativ gleichmäßig verteilte Beziehungen zu den anderen Themen, vermittelt über die Autoren, aufweisen. Gleichzeitig kann die Ursache aber auch in einer relativ hohen Anzahl von exklusiven Autorenbeziehungen liegen, d.h. Autoren die nur ein Thema ansprechen, dies würde die entsprechenden Themen weiter aus der Mitte des Netzwerks „herausziehen“.

6.6.2 Analyse bipartiter, sozio-semantischer Netzwerke

Die bereits angedeuteten Zusammenhänge können nur mit einer eingehenderen Betrachtung der Netzwerkstruktur geklärt werden. Um die Frage der Exklusivität zu klären, können wir die Distanzmatrix des Netzwerks betrachten. Dabei muss jedoch bereits dem Umstand Rechnung getragen werden, dass es sich hier um ein bipartites Netzwerk handelt, dieses somit nur Beziehungen zwischen den zwei distinkten Mengen von Knoten (Autoren und Themen) zulässt. Daher ist es zunächst notwendig je ein `set` für jede der beiden Ebenen zu erzeugen, damit die darauf zugreifenden Funktionen die strukturellen Besonderheiten bipartiter Netzwerke berücksichtigen können.

```
1 import networkx as nx
2 from networkx.algorithms import bipartite as bp
3 BG = nx.read_gpickle('Daten/TopicAuthor.gpkl')
4
5 authors, topics = bp.sets(BG)
```

Um die Frage zu klären, welche Themen die eine vermittelnde Funktion im Diskurs einnehmen kann deren Betweenness-Zentralität eingehender betrachtet werden. Wie schon dargestellt kann damit der vermittelnde Charakter einer Netzwerkposition hinsichtlich der kürzesten Pfade zwischen allen Knoten ermittelt werden. Hieran wird auch deutlich, warum eine Kombination von Themen-Autor und Autor-Autor Beziehungen (z.B.: Koautorenschaften) so problematisch für eine netzwerkanalytische Untersuchung ist. In diesem Fall würden die kürzesten Pfade auf der Ebene der Autoren stark überschätzt werden, weil direkte Beziehungen zwischen den Themen nicht möglich sind. Zudem ergibt sich das konzeptuelle Problem, dass es sich um grundverschiedene Kantenzüge handelt die unterschiedliche Beziehungen abbilden. Geodätische Distanzen sind in so einem Fall nicht sinnvoll interpretierbar.

Verwendet man im hier konstruierten, two-mode Netzwerk eine bipartite Betweenness-Metrik, so ist deren Bedeutung relativ klar, da alle Kantenzüge die selbe Art von Beziehungen abbilden. In diesem Fall die Beziehungen zu einzelnen Themenbereichen von der gesammelten Publikationstätigkeit eines Autors zu den latenten semantischen Dimensionen des Diskurses. Daher deutet die relativ hohe Betweenness eines Themas darauf hin, dass es sich tendenziell mit anderen Themen im Laufe der Publikationstätigkeit verbinden lässt. Auf der Ebene der Themen kann man

daher die Betweenness-Zentralität als ein Maß der Anschlussfähigkeit interpretieren. Würde man die Autoren betrachten, so würde ein hoher Wert auf Autoren mit breit gefächerten thematischen Bezügen hindeuten.

```
1 betweenness = bp.betweenness_centrality(BG, topics)
2
3 for topic in topics:
4     print betweenness[topic], topic
```

```
1 0.125202088617 Diskriminierung
2 0.0999860150374 Soziale Klasse
3 0.167031291593 Rasse/Ethnizität
4 0.160197224698 Demographie
5 0.126920368719 Arbeit
6 0.142850245834 Wohlfahrtsstaat
7 0.0922027268762 Religion
8 0.165700691825 Theorie/Gender
9 0.137584930001 Beziehungen
```

Bei einer Betrachtung der Betweenness-Werte zeigt sich die hohe Bindungskraft der Themen „Theorie/Gender“ und „Demographie“ mit größerer Eindeutigkeit, als dies in der rein grafischen Betrachtung der Fall war. Diese beiden Themen werden jedoch noch von „Rasse/Ethnizität“ übertroffen, welches die höchste Betweenness-Zentralität aufweist. Damit können diese drei Themen als die anschlussfähigsten dieses spezifischen Diskurses angesehen werden, der als eine Repräsentation des soziologischen Mainstreams aufgefasst werden kann.

Eine hohe Anschlussfähigkeit muss jedoch nicht grundsätzlich auf eine dominante Position im Diskurs hindeuten. Es sind auch sehr allgemeine Themen denkbar, wie zum Beispiel eine latente semantische Dimension, die nur aus Standardfloskeln des wissenschaftlichen Schreibens besteht, also keinen Bezug zu inhaltlichen Themen aufweist. Von einem solchen Thema würde man schon aufgrund der hohen Allgemeinheit eine sehr hohe Anschlussfähigkeit erwarten. Gleichzeitig wäre die Existenz solcher Themen aber rein sprachlich und stilistisch begründet, somit wären sie für eine Analyse sozio-semantischer Phänomene kaum von Interesse. Daher lohnt es sich den Bezug zu einzelnen Autoren genauer zu betrachten. Je mehr ein Thema auch für sich stehen kann, d.h. je mehr Autoren

nur dieses Thema anprechen, umso mehr handelt es sich um eine soziosemantische Position in einem Diskurs.

Themen die in der Lage sind Anschlussfähigkeit zu gewährleisten und gleichzeitig dominant genug sind, dass sie alleine ausreichend für die Publikationstätigkeit eines Autors sind, können in sehr grober Anlehnung an die foucaultsche Terminologie als *Dispositive* bezeichnet werden. Eine Begrifflichkeit, die unter den Diskursanalytikern foucaultscher Schule wahrscheinlich die geringste Form der Einigung erzielt hat. Obwohl dieser Begriff in fast allen Forschungsarbeiten Verwendung findet, ist man über die Diskussion der Begrifflichkeit selbst noch nicht hinaus gekommen (vgl. Keller 2008: 98f). Foucault selbst hat den Begriff entweder als selbstverständlich vorausgesetzt (vgl. Foucault 1991: 29) oder sehr ausufernde und widersprüchliche Definitionen dafür gegeben (vgl. Foucault 1978: 119f).

Hier wird folgende Interpretation vorgeschlagen. Foucaults Selbstverständlichkeit der Begriffsverwendung in seinen frühen Arbeiten deutet darauf hin, dass er hauptsächlich an der französischen Bedeutung des Begriffs als einer Metapher interessiert war. Insofern lässt sich „dispositif" als eine Vorrichtung, oder zum Beispiel als ein Maßnahmenkatalog zur Erreichung eines bestimmten Ziels definieren. Im weitesten Sinne also die Vorausbedingungen eines Diskurses. Entkleiden wir diese Begrifflichkeit von ihren normativen Implikationen, so können wir es als die Beschreibung bestimmter Positionen in der Topologie der zugrundeliegenden, semantischen Struktur auffassen. Positionen die besonders günstig für Äußerungen in diesem Diskurs sind, also die Wahrscheinlichkeit erhöhen in einem spezifischen Diskurs weiterhin Aussagen treffen zu können, werden hier als Dispositive aufgefasst.

```
data = bp.biadjacency_matrix(BG, authors)

adj_df = pd.DataFrame(data.todense(),
                      index=authors,
                      columns=topics)

adj_df[adj_df.sum(axis=1) == 1].sum()
```

```
Diskriminierung      31
Soziale Klasse      115
Rasse/Ethnizität     39
```

```
Demographie          78
Arbeit               39
Wohlfahrtsstaat     106
Religion             33
Theorie/Gender       66
Beziehungen          61
dtype: int64
```

Solche günstigen Positionen sind also durch eine hohe Anschlussfähigkeit bei einer gleichzeitg hohen Eigenständigkeit gekennzeichnet. Wie die Analyse der Distanzmatrix zeigt sind es hier vor allem die Themen „Soziale Klasse“ und „Wohlfahrtsstaat“ die eine hohe Eigenständigkeit aufweisen, gefolgt von „Demographie“ und „Theorie/Gender“. Unter der Berücksichtigung der Betweenness-Werte lässt sich sagen, dass vor allem das Thema „Wohlfahrtsstaat“ Dispositivcharakter hat, da es sowohl stark anschlussfähig und gut vernetzt ist, als auch aureichend für alleinige Publikationstätigkeiten zu sein scheint. In gewissem Maße gilt diese Interpretation auch für „Demographie“ und „Theorie/Gender“, die zwar nicht in dem Maße für sich genommen vorkommen, dafür aber relativ stark vernetzt sind. „Soziale Klasse“ scheint hingegen tendenziell ein Außenseiterthema zu sein, dass anscheinend nur mit bestimmten anderen Themen verknüpft ist, aber dennoch ausreichend Möglichkeiten zu Publikation bietet.

Grundsätzlich sollten die vorangegangenen Ausführungen einen kurzen Überblick über das im entstehenden begriffene Forschungsprogramm der Analyse sozio-semantischer Netzwerke geben, sowie deren Möglichkeiten und Probleme aufdecken. Insbesondere wurde gezeigt, dass sich die Verfahren der Analyse latenter Semantiken und Themen gewinnbringend mit Verfahren der Analyse sozialer Netzwerke kombinieren lassen. Jedoch müssen die graphtheoretischen Grundlagen hier in besonderer Weise berücksichtigt werden, da sonst zwar eindrucksvolle Strukturen und Visualisierungen produziert werden können, deren analytischer Wert jedoch extrem gering sein kann.

6.7 Vom Symbol zur symbolischen Ordnung

Ziel der vorangegangenen Ausführungen war es die Modellierung von symbolischen Ordnungen, d.h. Bedeutungen, Wissen, Kultur und dergleichen, auf der Basis des regelmäßigen Gebrauchs von Sprache zu demons-

trieren. Dabei zeigte sich auf methodischer Ebene was auf theoretischer bereits angedacht war, dass eine Bestimmung der symbolischen Ordnungen vom wechselseitigen Bezug des konkreten sprachlichen Aktes zu den anderen sprachlichen Äußerungen, die im selben Kontext getroffen wurden, ausgehen muss. Ausgangspunkt war dabei die Überlegung, dass symbolische Ordnungen als situationsübergreifende Muster des Gebrauchs von Symbolen aufgefasst werden können. Bei der Betrachtung der konkreten Methoden und deren zugrundeliegender Annahmen wurde deutlich, dass sich symbolische Ordnungen sinnvollerweise als Muster in den Relationen von Texten auffassen lassen.

In gewisser Weise lässt sich dieser Ansatz als eine *strukturalistische Auffassung von Kultur* charakterisieren, da versucht wird symbolische Ordnungen über die Regelmäßigkeit und Konstanz der Verknüpfungen von Symbolen zu beschreiben. Des Weiteren wird der Strukturbegriff im Rahmen der Soziologie gerne mit „Statik" assoziiert, während „Wandel" und „Dynamik" eher als kulturelle Phänomene angenommen werden. Eine solche Feststellung übersieht jedoch, dass eine Veränderung, d.h. ein Unterschied zwischen zwei Zeitpunkten, nur durch den Vergleich von („statischen") Messungen zu verschiedenen Zeitpunkten festgestellt werden kann. Die Modellierung des Wandels von symbolischen Ordnungen im Zeitverlauf ist allerdings trotz der Entwicklung neuer Verfahren (z.B.: Blei und Lafferty 2006; Wang, Blei und Heckerman 2012) noch am Anfang. Das hauptsächliche Problem besteht zum gegenwärtigen Zeitpunkt darin, dass in den meisten Fällen die Anzahl möglicher Themen über die Zeit hinweg konstant gehalten werden muss.

Die Betrachtung unterschiedlicher Verfahren und Möglichkeiten symbolische Ordnungen zu messen hat auch gezeigt, dass sich diese vor allem in ihren Vorstellungen des Bezuges von Symbolen und Bedeutung unterscheiden. Zwei Ansätze erscheinen hier grundlegend. Zum einen die Annahme eines direkten Bezuges zwischen Symbolen und Bedeutungen, wie er insbesondere in der Netzwerk-Text-Analyse zum Ausdruck kommt, aber auch in der Bestimmung textueller Ähnlichkeiten durch Vektorraummodelle. Dies erlaubt eine sehr kleinteilige Unterscheidung von Texten und Strukturen in Texten. Im Falle der Netzwerk-Text-Analyse kommt der Vorteil des Vorhandenseins eines ausgereiften Analyseinstrumentariums hinzu. Der generelle Nachteil besteht in der Festschreibung des Bedeutungsgehalts von Worten, unabhängig von möglichen Homonymen und Synonymen. Gerade bei Zeichensystemen mit geringer Eineindeutigkeit kann dies dazu führen, dass stilistische Unterschiede als semantische in-

terpretiert werden. Es lässt sich sagen, dass Bedeutungen hier eher als Element des jeweiligen Textes und dessen Struktur begriffen werden.

Der zweite große Ansatz zur Modellierung von Bedeutungen besteht in der Annahme latenter Dimensionen, welche der Verteilung von Worten über eine Vielzahl an Texten zugrunde liegen. Die Feststellung dieser latenten Themen und Bedeutungen geschieht durch die Reduktion des gesamten Korpus. Dadurch erlauben diese Verfahren eine Rekonstruktion der grundlegenden Strukturen eines semantischen Raumes und eine Analyse der relativen Positionierungen von Texten in diesem. Allerdings hat dies auch zur Folge, dass die Auswahl der Anzahl der Dimensionen und die Festlegung des ursprünglichen Korpus einen starken Einfluss auf das daraus resultierende Modell haben. Verglichen mit anderen Verfahren müssen hier eine Reihe von Entscheidungen vom Forscher getroffen werden. Zudem ist die Repräsentation der Texte als Vektorraum immer eine Voraussetzung der Analyse latenter Bedeutungsstrukturen, was zur Folge hat, dass spezifischere Eigenschaften des Textes, wie zum Beispiel die Reihenfolge der Token, nur begrenzt berücksichtigt werden können. Insofern produzieren diese Verfahren tendenziell eine Beschreibung der Strukturen eines bestimmten Diskurses und eigenen sich weniger für den direkten Vergleich sehr spezifischer Unterschiede von Texten.

Die Klassifikationsverfahren und die Techniken des maschinellen Lernens fallen etwas aus dieser Unterscheidung heraus. Grundsätzlich wird auch hier eher von einer direkten Zuordnung der Attribute des Textes zu spezifischen Bedeutungen und Konnotationen des Textes ausgegangen. Allerdings setzen diese Verfahren immer eine bereits bestehende Zuordnung voraus. Sie sind also eher dazu gedacht entweder etwas über den Einfluss bestimmter Attribute hinsichtlich der Klassifikation des Textes auszusagen oder die trainierte Klassifikation zur Klassifizierung von Texten einzusetzen, die noch keine Klassifikation aufweisen. Da auch hier der Fokus auf den Eigenschaften spezifischer Texte liegt, haben diese Verfahren in methodologischer Hinsicht mehr mit der ersten Gruppe an Verfahren gemeinsam.

Vor dem Hintergrund der Eingangs vorgestellten Kategorisierung von Texten entlang der Dimensionen der Eineindeutigkeit und der Komplexität, lässt sich klären, welche Verfahren sich für die Modellierung bestimmter Symbolsysteme besonders gut eignen (siehe Tabelle 6.9). Demzufolge können Verfahren der Analyse latenter semantischer Dimensionen insbesondere Zeichensysteme von geringerer Komplexität angemessen repräsentieren. Dies liegt vor allem daran, dass komplexere Konstruktionsre-

	Niedrige Komplexität	Hohe Komplexität
Niedrige Eineindeutigkeit	Analyse latenter Dimensionen	Sozio-semantische Ansätze
Hohe Eineindeutigkeit	Maschinelles Lernen	Netzwerk-Text Analysen

Tabelle 6.9: Modelleignung für spezifische Zeichensysteme entsprechend ihrer Eineindeutigkeit und Komplexität.

geln nicht, oder zumindest nur sehr begrenzt, als Vektorräume repräsentiert werden können. Die Eineindeutigkeit ist hingegen kaum problematisch, da diese Verfahren die groben Muster des Gebrauchs modellieren und daher Phänomene wie Polysemie kein großes Problem darstellen.

Aus einem ähnlichen Grund eignen sich die Verfahren der Netzwerk-Text-Analyse insbesondere für den Bereich hoher Komplexität und Eineindeutigkeit. Sie sind in der Lage Konstruktionsregeln relativ gut und detailiert abzubilden, gleichzeitig setzen sie aber einen relativ eineindeutigen Wortgebrauch voraus. In gewisser Weise lässt dies die Feststellung zu, dass es zu einer Betrachtung des Bereichs niedriger Eineindeutigkeit und hoher Komplexität eine sinnvolle Kombination beider Vorgehensweisen im Sinne der sozio-semantischen Verfahren bedarf.

Symbolische Systeme von hoher Eineindeutigkeit und niedrige Komplexität lassen sich tendenziell mit Verfahren des maschinellen Lernens modellieren. Diese eignen sich am besten für die Feststellung kategorialer Unterschiede und klarer abgrenzbarer Eigenschaften. Komplexere Sinnzusammenhänge können so jedoch nur schlecht erfasst werden und überlappende Bedeutungszuordnung von Worten stellen ein fast unumgängliches Problem für das maschinelle Lernen dar, da diese in den meisten Fällen an klaren, eindimensionalen Kategorien trainiert werden müssen.

Abschließend lässt sich festhalten, dass eine direkte Messung von Kultur und Wissen im Sinne symbolischer Ordnungen möglich und sinnvoll ist. Indem Symbole als sozial standardisierte Zeichen mit Objektcharakter aufgefasst werden, ist es möglich Muster und Regelmäßigkeiten in der Verwendung zu analysieren und somit die Intentionen, Bedeutungen und wechselseitigen Bezüge zwischen und innerhalb symbolischer Ordnungen zu bestimmen. Die Voraussetzungen dafür sind jedoch die Berücksichtigung des Kontextes des Symbolgebrauchs und der Besonderheiten von Zeichensystemen sowie ein Verständnis der Limitationen und Möglichkeiten der jeweiligen Modelle.

Im Hinblick auf die dominanten, erkenntnistheoretischen Positionen der Soziologie (individuelles Handeln vs. gesellschaftliche Kultur) lässt sich sagen, dass sich in allen der hier vorgestellten methodischen Ansätzen beide Perspektiven wiederfinden lassen. Zugleich wird auch deutlich, dass eine Betrachtung von symbolischen Ordnungen als individuelle Ideen oder als transzendente Bedeutungen für sich genommen nicht ausreicht. Es lassen sich Muster im Gebrauch von Symbolen erkennen, die nicht zufällig und überindividuell sind, gleichzeitig werden diese nur im Vollzug einzelner symbolischer Akte sichtbar.

7 Abschlussdiskussion

Symbole und symbolische Ordnungen stellen die Grundbausteine sozialer Interaktion und Kommunikation dar. Diese hier vertretene These erscheint auf den ersten Blick sehr weit hergeholt und von ähnlicher Machart, wie die bestehenden Feststellungen eines primären Substrats in unterschiedlichen Theorierichtungen der Soziologie. Die Vertreter des methodologischen Individualismus würden darauf hinweisen, dass es ja immer noch den Abwägungen des Individuums geschuldet sei welche Zeichen wie Verwendung finden. Von Seiten der Gesellschaftstheorie käme wahrscheinlich der Einwand, dass es Institutionen, der Diskurs oder funktionale Teilsysteme sind, die darüber entscheiden ob symbolische Äußerungen Gültigkeit erlangen oder nicht. Im Folgenden soll dargelegt werden, dass eine Methodologie sozialer Symbole nicht nur keinen grundlegenden Widerspruch zu den bestehenden Theorierichtungen darstellt, sondern diese ergänzt. In einem größeren Zusammenhang geht es auch um die Frage, welche Möglichkeiten für eine Neuausrichtung der Sozialwissenschaft, im Sinne einer Naturwissenschaft, damit ermöglicht werden.

Vernachlässigt man für einen Moment die soziologische Tendenz zur Fraktalisierung ihrer Theorien, so lässt sich feststellen, dass soziale Symbole, d.h. sozial standardisierte Zeichen, die notwendige Voraussetzung für eine Reihe von sozialwissenschaftlichen Begriffen sind. Wissen und Kultur sind sicherlich die offensichtlichsten Beispiele. Aber auch Begriffe wie Gruppe, Rolle und Präferenzen und viele andere bestehen aus Symbolen oder können durch sie beschrieben werden. Selbst so zentrale Begrifflichkeiten wie Gesellschaft und Individuum können als eine Funktion von Symbolen und symbolischen Ordnungen aufgefasst werden. Alle sozial relevanten und addressierbaren Entitäten müssen auch symbolisch markiert sein, damit sie Gegenstand sozialen Handelns sein können. Fundamentaler ausgedrückt, erst durch die Bezeichnung mittels sozialer Symbole wird ein sozialer Gegenstand als solcher erfassbar.

Die bloße Feststellung, dass sich soziologische Begrifflichkeiten reformulieren lassen indem man „Symbole“ und „symbolische Ordnungen“ hineininterpretiert ist noch kein besonders überzeugender Erkenntnis-

gewinn. Im Prinzip ist dies nichts anderes als eine Wiederholung des grundlegenden Diktums des Konstruktivismus. Es ist daher wenig überraschend, dass die Aussage: „das Soziale ist symbolisch" von ähnlichen Problemen geplagt wird wie die Aussage: „das Soziale ist eine Konstruktion". Sie erscheint letztlich trivial und wenig hilfreich. Diese Einschätzung ist weitestgehend korrekt, wenn man außer Acht lässt, dass es nicht um eine ontologische Feststellung geht, sondern um die Gesetzmäßigkeiten denen diese Phänomene folgen. Dass Symbole etwas bezeichnen ist trivial, wie dies geschieht, wie es modelliert werden kann und welchen allgemeinen Gesetzen eine solche Bezeichnung folgt ist es absolut nicht.

Zu sagen, dass eine soziale Rolle durch ihre symbolischen Zuschreibungen sowie ihre symbolischen Äußerungen gekennzeichnet ist, schließt keineswegs *alternative Modellierungen* aus. Zum Beispiel lässt sich eine Rolle auch in netzwerkanalytischer Hinsicht definieren, als eine Menge „strukturell äquivalenter" Positionen (vgl. Sailer 1979). Zwei Knoten eines Netzwerkes sind dann strukturell äquivalent, wenn sie beliebig ausgetauscht werden können ohne die Struktur des Netzwerkes zu verändern. Einer solchen Auffassung von Rollen als sozialstrukturellen Positionen lässt sich die Goffmansche Konzeption einer dramaturgischen Rolle gegenüberstellen (vgl. Goffman 2004: 18). Dabei wird die Rolle als sich in fortlaufender Interaktion verfestigende Form symbolischer Zuschreibung verstanden. Rollen werden in dieser Auffassung mit Symbolen gekennzeichnet, damit sie schnell erkenntlich sind und problemlos in die weitere Interaktion eingebunden werden können.

Keine dieser beiden Perspektiven ist jedoch von vorneherein valider. Das heißt allerdings nicht, dass Methodologien beliebig austauschbar wären. Ein bestimmtes soziales Phänomen als symbolische Ordnung zu modellieren mag mehr oder weniger sinnvoll sein. Sobald sich jedoch abgrenzbare, physische Zeichen finden lassen, die sozial standardisiert sind, steht es außer Frage, dass diese als Symbole analysiert werden können.

Weil die hier vorgestellte Methodologie sozialer Symbole und symbolischer Ordnungen hauptsächlich aus formalen Definitionen besteht, steht sie nicht im direkten Widerspruch zu den etablierten soziologischen Theorierichtungen. Wie an vielen verschiedenen Stellen dieses Buches aufgezeigt wurde, kann man sowohl eine individualistische, als auch eine gesellschaftliche Perspektive einnehmen und dennoch mit den gleichen Grundannahmen auskommen. Solange das zu untersuchende Phänomen der Definition sozialer Symbole genügt, kann es auch mittels der hier diskutierten Verfahren modelliert werden.

Geht man über diesen methodologischen Minimalismus hinaus, so kann man soziale Symbole auch als einen eigenständigen Gegenstandsbereich der Soziologie auffassen. Wie bereits in den Ausführungen zu einer Theorie sozialer Symbole dargelegt, finden sich soziale Symbole und damit verwandte Konzepte mehr oder weniger explizit in allen dominanten Theorierichtungen. Darüber hinaus ist jedoch auch ein eigenständiger Erkenntnisgewinn von einer Theorie sozialer Symbole zu erwarten. Ähnlich wie im Falle der Netzwerkanalyse können entweder spezifische Methoden herangezogen werden oder eine eigenständige Theorie verfolgt werden. Diese Flexibilität ist in beiden Fällen dem hohen Grad der Formalisierung geschuldet, welche sich im Bereich der sozialen Symbole vor allem aus den Arbeiten der Linguistik, Informatik und Semiotik, sowie den praktischen Anwendungen speist.

Neben der grundsätzlichen Möglichkeit eine Reihe von Phänomenen zu modellieren, die sich im zentralen Gegenstandsbereich der Soziologie befinden und die bisher nicht zugänglich waren, enthält die theoretische und methodologische Auseinandersetzung mit sozialen Symbolen zusätzlich das Potential die Sozialwissenschaften grundlegend zu verändern. Nicht zuletzt, weil die computergestützte, algorithmenbasierte Analyse von textuellen Daten sich einfügt in einen übergreifenden Wandel des alltäglichen Lebens. Das massive Vordringen digitaler Informationstechnologien macht vor keinem Bereich menschlicher Tätigkeiten halt, auch nicht vor den Sozialwissenschaften. Im Folgenden werde ich meine Ausführungen auf drei Punkte beschränken in denen der mögliche Nutzen einer Methodologie sozialer Symbole für die Sozialwissenschaften zum Ausdruck kommt: die Möglichkeit Sachverhalte als *objektiv* zu behandeln, die bisher nur subjektiv interpretierbar waren, eine stärkere *Formalisierung* soziologischer Aussagen und die Verbreitung von *Digital Literacy* in den Sozialwissenschaften.

Phänomene wie Wissen und Kultur werden in den Sozialwissenschaften meist als subjektive Kategorien betrachtet, die durch Befragung gewonnen und deren Inhalt durch Interpretation entschlüsselt werden muss. Wie in der Gegenüberstellung von qualitativen und quantitativen Verfahren der Textanalyse bereits herausgearbeitet, besteht das eigentliche Problem nicht in der Interpretation oder in der qualitativen Datenerhebung, sondern in der immer stärker werdenden „Subjektorientierung", welche die qualitative Sozialforschung zunehmend im Griff zu haben scheint. Interpretation ist sicherlich unumgänglich, muss sich aber auch auf nachvollziehbare Kriterien gründen. Ansonsten besteht die reale Ge-

fahr, dass der Vorgang des Interpretierens zu einer esoterischen Kunstlehre verkommt.

Bereits in Max Webers methodologischer Aufforderung dem Erklären das Verstehen an die Seite zu stellen werden diese Probleme der mangelnden Konkretheit und ausufernden Subjektbezogenheit kritisch diskutiert. Im Bezug auf die allgmeine Methodologie der Sozialwissenschaften hoffte er diesem Problem mit einer schärferen Begriffsbildung („Idealtypen") begegnen zu können (vgl. Weber 2006: 29). Betrachtet man die gegenwärtige Verfassung der „interpretativen Soziologie", kann man sagen, so zumindest meine Einschätzung, dass die Zunahme von idealtypischen Begrifflichkeiten und Heuristiken anscheinend nicht zwangsläufig mit einer Präzision der Theoriebildung und einer Steigerung der Erklärungskraft sozialwissenschaftlicher Theorien einhergeht. Stattdessen erscheint die Begriffsbildung zunehmend inflationäre Züge zu tragen, während die Nachvollziehbarkeit der Interpretation immer weniger möglich wird.

Eine Betrachtung von sozialen Symbolen und symbolischen Ordnungen als Objekte entbindet keineswegs von der Notwendigkeit des Forschers Entscheidungen zu treffen um zu seinen Schlüssen zu gelangen. Im besten Fall sollten diese Entscheidungen jedoch weitestgehend auf angebbaren und damit kritisierbaren Kriterien basieren. Was eine Interpretation objektiver Gegebenheiten erkenntnistheoretisch überlegen macht ist nicht ihre größere Stabilität, sondern gerade der Umstand, dass sie leichter zu Fall gebracht werden kann. Demgegenüber kann eine zu sehr am subjektiven Empfinden orientierte Interpretation leicht unkritisierbar werden oder sich im eigenen Narrativ verlieren. Des Weiteren macht es die Zunahme objektiverer Verfahren und größerer Datenmengen schwerer rein subjektive Interpretationen zu rechtfertigen: „[s]ociologists will increasingly have to choose between scientifically rigorous but empathetically unsatisfying explanations and satisfying but unscientific stories" (Watts 2014: 344).

Die größere Objektivität der quantitativen Verfahren der Textanalyse wird zudem noch durch deren stärkeren Hang zur Formalisierung unterstützt. Ein Großteil der grundlegenden Theorien ist in mathematischen Formalismen gehalten. Dies gilt beispielsweise für die Informationstheorie (vgl. Hartley 1928; Shannon und Weaver 1976), für das Information Retrieval (vgl. Salton 1979) oder die Linguistik (vgl. Chomsky 1956). Es ist wichtig festzustellen, dass sich diese Formalismen nicht auf die Methoden beschränken, sondern genutzt werden den Gegenstand selbst zu beschreiben. Demgegenüber weist die Soziologie, von bedeutenden Aus-

nahmen wie Rational Choice und der Netzwerkanalyse einmal abgesehen, nur wenige formale Theorien und Modelle auf.

Formale Modellbildung hat zwei entscheidende Vorteile, die auch der Grund für deren weite Verbreitung in den Naturwissenschaften sind. Zum einen ermöglicht die Verwendung einer formalen Sprache die Formulierung eindeutiger Aussagen, die ebenso eindeutig kritisierbar sind. Dadurch wird auch eine vorurteilsfreie Bewertung von Theorien und Gedankengebäuden sowie deren Prüfung anhand von mathematischen und logischen Kriterien ermöglicht. Zum anderen ergibt sich so auch ein nicht zu unterschätzender Gewinn an Präzision für den Wissenschaftler selbst, der sich einer formalen Sprache bedient. Die Konstruktion von mathematischen Aussagen präzisiert das Denken und zwingt den Denkenden damit zur Kongruenz. Zusätzlich enthält dies auch die Möglichkeit zur analytischen Weiterentwicklung der eigenen Theorie.

Mit diesen Entwicklungen verbinde ich auch die Hoffnung einer Erneuerung der Sozialwissenschaften als einer naturwissenschaftlichen Disziplin. In Folge des enormen Bedeutungszuwachs von sozialer Kommunikation und Interaktion durch die digitalen Medien sowie der Möglichkeit diese in prozessgenerierten Daten nachvollziehen zu können, wurden die Sozialwissenschaften bereits zur *Wissenschaft des 21. Jahrhunderts* hinaufstilisiert (vgl. Watts 2007). In der letzten Zeit scheint die fortgesetzte Enttäuschung dieser Hoffnung jedoch in Resignation umgeschlagen zu sein. Einige Autoren haben in Folge dessen, mehr oder weniger stark, für die Etablierung einer neuer Disziplin votiert, die sich an rein naturwissenschaftlichen Gültigkeitskriterien zu orientieren hätte (vgl. Turner 2016; Diekmann 25. September 2016, Uhr).

Wie bereits argumentiert wurde, neigt die Soziologie prinzipiell zur Bildung von epistemologischen Gräben. Es ist daher eine berechtigte Befürchtung, dass eine Verwendung formaler Sprachen als eine Agenda des „Positivismus" aufgefasst werden würde und zu einer weiteren Vertiefung dieser Spaltungen beitragen könnte. Aus diesem Grund möchte ich das grundsätzliche *Integrationspotential* eines solchen Ansatzes herausstellen. Geschichtlich betrachtet hat die Einführung der Mathematik, als Lingua Franca der Wissenschaften, überhaupt erst die modernen Naturwissenschaften hervorgebracht und ihnen nicht zuletzt die grundlegende, interdisziplinäre Kommunikation über ihre Inhalte erlaubt.

Eine gemeinsame Sprache zu sprechen enthält sicherlich noch keine Gewähr für Konsens, aber zumindest die Möglichkeit sich produktiv zu streiten. Ein Streitgespräch, das weder auf den Kreis lebender Personen

noch auf Disziplinen begrenzt wäre. Es ist nicht zu hoch gegriffen, wenn man mit der Verwendung formaler Sprachen in den Sozialwissenschaften dieselbe Hoffnung verbindet, wie Galileo sie am Anbeginn der modernen Naturwissenschaften empfand. Die Hoffnung auf eine Brücke zwischen der abstrakten Theorie und der sinnlichen Erfahrung, auf einen Wegweiser aus einem dunklen Labyrinth:

> Die Philosophie ist in diesem großartigen Buch geschrieben, welches unseren Augen stets offen steht (ich meine das Universum), aber man kann es nicht verstehen, wenn man nicht zuerst lernt die Sprache zu verstehen und die Zeichen zu erkennen, in denen es geschrieben ist. Es ist in der Sprache der Mathematik geschrieben, und die Buchstaben sind Dreiecke, Kreise und andere geometrische Figuren, ohne die es dem Menschen unmöglich ist ein einziges Wort zu verstehen; gleich dem Herumwandern in einem dunklen Labyrinth. (Eigene Übersetzung: Galilei 1832: 13)

Mathematik ist aber nicht die einzige formale Sprache, die einer Beschäftigung mit textanalytischen Verfahren nahe steht. Die Analyse digitaler Texte und anderer symbolischer Ausdrucksformen setzt eine grundlegende Kenntnis digitaler Technologie und abstrakten Konzepten der Informatik voraus. Zwar hat die Verfügbarkeit fertiger Programmpakete in den letzten Jahren stark zugenommen, diese sind aber meistens auf bestimmte Textarten und Kontexte ausgerichtet. Der damit einhergehende Mangel an Flexibilität macht das Erlernen einer Programmiersprache zu einer sinnvollen Alternative und notwendigen Voraussetzung für selbstständiges Forschungshandeln in diesem Bereich.

Die damit zusammenhängenden Einblicke in digitale Technologien sind dabei so fundamental, dass sich das erworbene Wissen auch in einer Reihe anderer Bereiche anwenden lässt. Dies ist sicher nicht nur im Bereich der Textanalyse der Fall. Dennoch führt die enge Verknüpfung der Datenquellen und der Erhebungsverfahren (z.B.: Text Mining) mit den modernen Informationstechnologien dazu, dass hier ein sehr hohes Potential für die Entstehung von grundlegender Expertise im Umgang mit moderner Informationstechnologien besteht. Gerade dies scheint vor dem Hintergrund der rapiden Entwicklung neuer Verfahren und dem Aufkommen neuer Datenquellen auch zwingend notwendig.

In all diesen Möglichkeiten, welche die Methode und Methodologie symbolischer Ordnungen mit sich bringt, liegt aber auch eine fundamen-

tale Herausforderung des soziologischen Selbstbildes. Die Ablehnung des expliziten Einbezugs sozialer Symbole in die soziologische Theoriebildung hatte seine Ursache auch immer im festen Glauben an die *fundamentale Differenz von Kultur und Natur*. Die Trennung des beobachtbaren Universums in eine von anonymen Regeln geleitete Welt und ein Individuum dessen Entscheidungsmaximen in ihm alleine liegen. Solange Denken und Sprechen fundamental unterschieden werden können, solange kann man festhalten an der Vorstellung, dass wir Urheber und Schöpfer unserer eigenen Gedanken sind. Es ist diese fundamentale Kränkung des Menschens und der ihm unterstellten Einzigartigkeit, so meine feste Überzeugung, die eine Anerkennung symbolischer Prozesse und Ordnungen im Rahmen der Soziologie stets behindert hat.

Selbst die Theorien des Strukturfunktionalismus und der Systemtheorie sind letztlich verzeihlicher gegen das Menschenbild der Moderne gewesen. Sie mögen ihn auf eine Rolle im Gewebe des Sozialen reduziert oder ihn gleich ganz weggelassen haben, sie sind aber nie in sein Innerstes vorgedrungen, haben nie Anspruch auf sein Denken und die darin befindlichen Ideen erhoben. Ebenso problembehaftet ist aber auch der Blick in die andere Richtung. Denken und Sprechen als Funktionen symbolischer Ordnungen auf unterschiedlichen Ebenen zu begreifen bedeutet, dass auch unsere höchsten und transzendentesten Begriffe letztlich menschliche Gebilde sind. Mögen die Ontologien und Ideen noch so klar und präzise erscheinen, sie tragen doch den Makel ihrer niederen Herkunft. Auch hier liegt eine fundamentale Kränkung des Menschen vor, allerdings bezogen auf „den Menschen", wie er dem Idealismus entsprang, dessen Vorstellung das Geistige und Symbolische zu transzendenten Wahrheiten erhob. Wie Max Stirner (2005: 41) so trefflich beobachtete, wird aus dem „Geist" so schnell ein „Spuk":

> Nur diese verkehrte Welt der Wesen, existiert jetzt wahrhaft. Das menschliche Herz kann lieblos sein, aber sein Wesen existiert, der Gott, »der die Liebe ist«. Das menschliche Denken kann im Irrtum wandeln, aber sein Wesen, die Wahrheit, existiert: »Gott ist die Wahrheit« usw. Die Wesen allein und nichts als die Wesen zu erkennen und anzuerkennen, das ist Religion: ihr Reich ein Reich der Wesen, des Spuks und der Gespenster.

Symbolische Ordnungen, so die hier vertretene Auffassung, sind nicht bloßer Ausdruck individueller Gedanken, noch sind sie Abbild transzen-

denter Begrifflichkeiten. Vielmehr sind sie das Ergebnis eines evolutionären Prozesses, der von menschlicher Kreativität, Schaffensdrang und Nützlichkeitserwägungen getrieben wird und seine Grenzen in der Selektion durch die umfassende symbolische Ordnung des Sozialen findet. Die Maßstäbe an denen symbolische Äußerungen gemessen werden, sind ebenso Teil dieses Prozesses und den gleichen Gesetzen unterworfen, wie sie für alle anderen Elemente dieses Systems gelten. Der inhärente Pragmatismus dieser Sichtweise mag wenig schmeichelhaft sein, aber er gibt den Weg frei zur Modellierung und empirischen Überprüfung eines fundamentalen Phänomens der sozialen Welt, der ursprünglichen Sozialtechnologie, des ältesten Werkzeugs unserer Spezies, der Sprache.

Literatur

Abbott, Andrew (2001). *Chaos of Disciplines*. University of Chicago Press.

Adorno, Theodor W. (1976). *Positivismusstreit in der deutschen Soziologie*. 5. Aufl. Soziologische Texte ; 58. Neuwied: Luchterhand.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow und Rebecca Passonneau (2011). *Sentiment Analysis of Twitter Data*. In: *Proceedings of the Workshop on Languages in Social Media*. LSM '11. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 30–38.

Albert, Gert (2009). *Sachverhalte in der Badewanne. Zu den allgemeinen ontologischen Grundlagen des Makro-Mikro-Makro-Modells der soziologischen Erklärung*. In: *Das Mikro-Makro-Modell der soziologischen Erklärung*. Hrsg. von Jens Greve, Annette Schnabel und Rainer Schützeichel. VS Verlag für Sozialwissenschaften, S. 21–48.

Ambrogelly, Alexandre, Sotiria Palioura und Dieter Söll (2007). Natural Expansion of the Genetic Code. In: *Nature Chemical Biology* 3.1, S. 29–35.

Amstad, Toni (1978). *Wie verständlich sind unsere Zeitungen?* Zürich: Universität Zürich.

Assmann, Jan (2007). *Das kulturelle Gedächtnis: Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. München: C.H.Beck.

Bail, Christopher A. (2012). The Fringe Effect: Civil Society Organizations and the Evolution of Media Discourse about Islam since the September 11th Attacks. In: *American Sociological Review* 77.6, S. 855–879.

Bail, Christopher A. (2014). The Cultural Environment: Measuring Culture with Big Data. In: *Theory and Society* 43.3-4, S. 465–482.

Ball, Michael S. und Gregory W. H. Smith (1992). *Analyzing Visual Data*. Bd. 24. Newbury Park, London, New Delhi: Sage Publications, Inc.

Bamberger, Richard und Erich Vanecek (1984). *Leichter lesen, leichter lernen*. Wien: Verlag für Jugend und Volk.

Bar-Hillel, Yehoshua (1955). An Examination of Information Theory. In: *Philosophy of Science* 22.2, S. 86–105.

Bearman, Peter S. und Katherine Stovel (2000). Becoming a Nazi: A Model for Narrative Networks. In: *Poetics* 27.2–3, S. 69–90.

Beck, Colin J., Gili S. Drori und John W. Meyer (2012). World Influences on Human Rights Language in Constitutions: A Cross-National Study. In: *International Sociology* 27.4, S. 483–501.

Becker, Gary S. (1985). Human Capital, Effort, and the Sexual Division of Labor. In: *Journal of labor economics* 3.1, S. 33–58.

Bellman, Richard (1957). *Dynamic Programming*. Princeton University Press.

Berger, Peter L. und Thomas Luckmann (1980). *Die gesellschaftliche Konstruktion der Wirklichkeit : Eine Theorie der Wissenssoziologie*. Frankfurt am Main: Fischer.

Berger, Peter und Hansfried Kellner (1964). Marriage and the Construction of Reality An Exercise in the Microsociology of Knowledge. In: *Diogenes* 12.46, S. 1–24.

Bergmann, Jörg R. und Christoph Meier (2000). *Elektronische Prozessdaten und ihre Analyse*. In: Hrsg. von Uwe Flick, Ernst von Kardorff und Ines Steinke. 3. Auflage. Reinbek bei Hamburg: Rowohlt, S. 429–437.

Bernhard, Stefan (2012). *Forschungspragmatische Überlegungen zu einer feldtheoretischen Netzwerkanalyse*. In: *Die Integration von Theorie und Methode in der Netzwerkforschung*. Hrsg. von Marina Hennig und Christian Stegbauer. VS Verlag für Sozialwissenschaften, S. 117–132.

Bertalanffy, Ludwig van (1969). *General System Theory : Foundations, Development, Applications*. New York: Braziller.

Bird, Steven, Ewan Klein und Edward Loper (2009). *Natural Language Processing with Python - Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Blackmore, Susan (1998). Imitation and the Definition of a Meme. In: *Journal of Memetics-Evolutionary Models of Information Transmission* 2.11, S. 159–170.

Blackmore, Susan (2000). *The Meme Machine*. Oxford University Press.

Blackmore, Susan (2001). Evolution and Memes: The Human Brain as a Selective Imitation Device. In: *Cybernetics & Systems* 32.1-2, S. 225–255.

Blei, David M. und John D. Lafferty (2006). *Dynamic Topic Models*. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06, S. 113–120.

Blei, David M., Andrew Y. Ng und Michael I. Jordan (2003). Latent Dirichlet Allocation. In: *Journal of Machine Learning Research* 3, S. 993–1022.

Blossfeld, Hans-Peter und Götz Rohwer (2002). *Techniques of Event History Modeling: New Approaches to Causal Analysis*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Bonacich, Phillip (1987). Power and Centrality: A Family of Measures. In: *American Journal of Sociology* 92.5, S. 1170–1182.

Borgatti, Stephen P. (2005). Centrality and Network Flow. In: *Social Networks* 27.1, S. 55–71.

Borgatti, Stephen P. und Martin G. Everett (1989). The Class of All Regular Equivalences: Algebraic Structure and Computation. In: *Social Networks* 11.1, S. 65–88.

Bourdieu, Pierre (1974). *Zur Soziologie der symbolischen Formen*. Frankfurt am Main: Suhrkamp.

Bourdieu, Pierre (2001). *Die Regeln der Kunst. Genese und Struktur des literarischen Feldes*. Übers. von Bernd Schwibs und Achim Russer. Frankfurt am Main: Suhrkamp.

Bourdieu, Pierre (2003). *Die feinen Unterschiede : Kritik der gesellschaftlichen Urteilskraft*. Frankfurt am Main: Suhrkamp.

Brachmann, Ronald J. (1979). *On the Epistemological Status of Semantic Networks*. In: *Associative Networks. Representation and Use of Knowledge by Computers*. Hrsg. von Nicholas V. Findler. New York, NY: Academic Press, S. 3–50.

Breiger, Ronald L. (2010). *Dualities of Culture and Structure: Seeing Through Cultural Holes*. In: *Relationale Soziologie*. Hrsg. von Jan A. Fuhse und Sophie Mützel. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 37–47.

Breiger, Ronald L. und Philippa E. Pattison (1986). Cumulated Social Roles: The Duality of Persons and Their Algebras. In: *Social Networks* 8.3, S. 215–256.

Breuer, Franz, Arnulf Deppermann, Udo Kuckartz, Günter Mey, Katja Mruck und Jo Reichertz (2014). *All is data – Qualitative Forschung und ihre Daten*. In: *Qualitative Forschung*. Hrsg. von Günter Mey und Katja Mruck. Springer Fachmedien Wiesbaden, S. 261–290.

Bro, Rasmus, Evrim Acar und Tamara G. Kolda (2007). *Resolving the Sign Ambiguity in the Singular Value Decomposition*. SAND2007 -6422. Sandia National Laboratories, S. 135–140.

Brüsemeister, Thomas (2008). *Qualitative Forschung: Ein Überblick*. 2., überarb. Aufl. Hagener Studientexte zur Soziologie. Wiesbaden: VS, Verl. für Sozialwiss.

Bude, Heinz (2004). *Die Kunst der Interpretation*. In: *Qualitative Forschung: Ein Handbuch*. Hrsg. von Uwe Flick, Ernst von Kardorff und Ines Steinke. Reinbek bei Hamburg: Rowohlt.

Burger, Harald (2007). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 3.Auflage. Berlin: Erich Schmidt.

Burrows, Roger und Mike Savage (2014). After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology. In: *Big Data & Society* 1.1, S. 1–6.

Callon, Michel, John Law und Arie Rip (1986). *Mapping the Dynamics of Science and Technology*. Springer.

Carley, Kathleen M. (1997). *Network Text Analysis: The Network Positions of Concepts*. In: *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Hrsg. von Carl W. Roberts. Mahwah, NJ: Laurence Erlbaum, S. 79–100.

Carley, Kathleen M. und David S. Kaufer (1993). Semantic Connectivity: An Approach for Analyzing Symbols in Semantic Networks. In: *Communication Theory* 3.3, S. 183–213.

Carnap, Rudolf und Yehoshua Bar-Hillel (1952). *An Outline of a Theory Semantic Information*. Technical Report 247. Cambridge, Mass.: MIT Research Laboratory of Electronics.

Cassirer, Ernst und Reinold Schmücker (2001). *Gesammelte Werke*. Hamburg: Meiner.

Chomsky, Noam (1956). Three Models for the Description of Language. In: *Information Theory, IRE Transactions on* 2.3, S. 113–124.

Chomsky, Noam (1964). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Chomsky, Noam (2002). *Syntactic Structures*. Walter de Gruyter.

Clauset, Aaron, Cosma R. Shalizi und Mark E. J. Newman (2009). Power-Law Distributions in Empirical Data. In: *SIAM review* 51.4, S. 661–703.

Coffman, Kerry G. und Andrew M. Odlyzko (2002). *Internet Growth: Is There a "Moore's Law" for Data Traffic?* In: *Handbook of Massive Data Sets*. Hrsg. von James Abello, Panos M. Pardalos und Mauricio G. C. Resende. Massive Computing 4. Dordrecht: Springer Science+Business Media, S. 47–93.

Coleman, James S. (1994). *Foundations of Social Theory*. Harvard University Press.

Corman, Steven R., Timothy Kuhn, Robert D. Mcphee und Kevin J. Dooley (2002). Studying Complex Discursive Systems. In: *Human Communication Research* 28.2, S. 157–206.

Creath, Richard (2004). *Logical Empiricism*. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta. Spring 2014. Metaphysics Research Lab, Stanford University.

Croom, Adam M. (2013). How to Do Things with Slurs: Studies in the Way of Derogatory Words. In: *Language & Communication* 33.3, S. 177–204.

Cruttenden, Alan (1997). *Intonation*. Cambridge University Press.

Das, Jishnu, Quy-Toan Do, Karen Shaines und Sowmya Srinivasan (2009). *US and Them: The Geography of Academic Research*. The World Bank.

Davis, Terry C., Michael A. Crouch, Georgia Wills, Sarah Miller und D. M. Abdehou (1990). The Gap between Patient Reading Comprehension and the Readability of Patient Education Materials. In: *The Journal of Family Practice* 31.5, S. 533–538.

Dawkins, Richard. (1996). *Das egoistische Gen*. Reinbek bei Hamburg: Rowohlt.

De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith und Mrinal Kumar (2016). *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media*. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: ACM, S. 2098–2110.

De Waal, Frans B. M. und Michelle L. Berger (2000). Payment for Labour in Monkeys. In: *Nature* 404.6778, S. 563–563.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer und Richard Harshman (1990). Indexing by Latent Semantic Analysis. In: *Journal of the American Society for Information Science* 41.6, S. 391–407.

Denzin, Norman K. (2000). *Reading Films - Filme und Videos als sozialwissenschaftliches Erfahrungsmaterial*. In: *Qualitative Forschung*. Hrsg. von Uwe Flick, Ernst von Kardorff und Ines Steinke. 3. Auflage. Reinbek bei Hamburg: Rowohlt, S. 416–428.

Diaz-Bone, Rainer (2006). Zur Methodologisierung der Foucaultschen Diskursanalyse. In: *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 7.1.

Diaz-Bone, Rainer (2008). Gibt es qualitative Netzwerkanalyse? In: *Historical Social Research / Historische Sozialforschung* 33.4, S. 311–343.

Diekmann, Andreas (25. September 2016, Uhr). Geisteswissenschaften: Die Gesellschaft der Daten. In: *sueddeutsche.de. kultur*.

Diesner, Jana und Kathleen M. Carley (2010). *Extraktion relationaler Daten aus Texten*. In: *Handbuch Netzwerkforschung*. Hrsg. von Christian Stegbauer und Roger Häußling. VS Verlag für Sozialwissenschaften, S. 507–521.

DiMaggio, Paul, Manish Nag und David M. Blei (2013). Exploiting Affinities between Topic Modeling and the Sociological Perspective on Cul-

ture: Application to Newspaper Coverage of US Government Arts Funding. In: *Poetics* 41.6, S. 570–606.
Dörner, Dietrich (2001). *Bauplan für eine Seele*. Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verl.
Dubin, David (2004). The Most Influential Paper Gerard Salton Never Wrote. In: *Library Trends* 52.4, S. 748–764.
Dunning, Ted (1993). Accurate Methods for the Statistics of Surprise and Coincidence. In: *Computational Linguistics* 19.1, S. 61–74.
Durkheim, Émile (1973). *Der Selbstmord*. Suhrkamp.
Durkheim, Émile (1984). *Die Regeln der soziologischen Methode*. 1. Aufl. Suhrkamp-Taschenbuch Wissenschaft 464. Frankfurt am Main: Suhrkamp.
Durkheim, Émile (2007). *Die elementaren Formen des religiösen Lebens*. Frankfurt am Main; Leipzig: Verlag der Weltreligionen.
Durkheim, Émile (2012). *Über soziale Arbeitsteilung: Studie über die Organisation höherer Gesellschaften*. 6. Aufl. Suhrkamp-Taschenbuch Wissenschaft 1005. Frankfurt am Main: Suhrkamp.
Eco, Umberto (1972). *Einführung in die Semiotik*. München: Wilhelm Fink.
Eco, Umberto (1976). Peirce's Notion of Interpretant. In: *MLN* 91.6, S. 1457–1472.
Egghe, Leo (2007). Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. In: *Journal of the American Society for Information Science and Technology* 58.5, S. 702–709.
Elias, Norbert (1939). *Über den Prozeß der Zivilisation*. Bd. 2. Baden-Baden: Suhrkamp.
Elias, Norbert (1988). *Die Gesellschaft der Individuen*. 3. Aufl. Frankfurt (Main): Suhrkamp.
Elias, Norbert (1991). *The Symbol Theory*. Hrsg. von Richard Kilminster. London: Sage.
Erk, Katrin und Lutz Priese (2008). *Theoretische Informatik: eine umfassende Einführung*. 3., erw. Aufl. eXamen.press. Berlin: Springer.
Esser, Hartmut (1990). "Habits", "Framesünd "Rational Choice". Die Reichweite von Theorien der rationalen Wahl (am Beispiel der Erklärung des Befragtenverhaltens). In: *Zeitschrift für Soziologie* 19.4, S. 231–247.
Esser, Hartmut (1996). Die Definition der Situation. In: *Kölner Zeitschrift für Soziologie & Sozialpsychologie* 48.1, S. 1–34.

Etzrodt, Christian (2007). Neuere Entwicklungen in der Handlungstheorie. Ein Kommentar zu den Beiträgen von Kroneberg und Kron. In: *Zeitschrift für Soziologie* 36.5, S. 364–379.

Fitzsimmons, P. R., B. D. Michael, J. L. Hulley und G. O. Scott (2010). A Readability Assessment of Online Parkinson's Disease Information. In: *The Journal of the Royal College of Physicians of Edinburgh* 40.4, S. 292–296.

Flesch, Rudolf (1948). A New Readability Yardstick. In: *Journal of Applied Psychology* 32.3, S. 221–233.

Flick, Uwe (1995). *Handbuch qualitative Sozialforschung: Grundlagen, Konzepte, Methoden und Anwendungen*. Beltz Psychologie-Verlag-Union.

Flick, Uwe (2011). *Triangulation*. Springer.

Flick, Uwe, Ernst von Kardorff und Ines Steinke, Hrsg. (2004). *Qualitative Forschung. Ein Handbuch*. 3. Auflage. rororo ; 55628 : Rowohlts Enzyklopädie. Reinbek bei Hamburg: Rowohlt.

Fligstein, Neil und Adam Goldstein (2010). *The Anatomy of the Mortgage Securitization Crisis*. In: *Markets on Trial: The Economic Sociology of the U.S. Financial Crisis*. Hrsg. von Michael Lounsbury und Paul M. Hirsch. Research in the Sociology of Organizations 30. Emerald Group Publishing, S. 27–68.

Fligstein, Neil und Doug McAdam (2011). Toward a General Theory of Strategic Action Fields. In: *Sociological Theory* 29.1, S. 1–26.

Flory, Steven M., Thomas J. Phillips Jr. und Maurice F. Tassin (1992). Measuring Readability: A Comparison of Accounting Textbooks. In: *Journal of Accounting Education* 10.1, S. 151–161.

Foucault, Michel (1978). *Dispositive der Macht: über Sexualität, Wissen und Wahrheit*. IMD 77. Berlin: Merve.

Foucault, Michel (1981). *Archäologie des Wissens*. Frankfurt am Main: Suhrkamp.

Foucault, Michel (1991). *Sexualität und Wahrheit 1. Der Wille zum Wissen*. Frankfurt am Main: Suhrkamp.

Freeman, Linton C, Douglas Roeder und Robert R Mulholland (1979). Centrality in Social Networks: II. Experimental Results. In: *Social Networks* 2.2, S. 119–141.

Friedl, Jeffrey (2006). *Source of the Famous "Now You Have Two Problems" Quote*. URL: `http://regex.info/blog/2006-09-15/247` (besucht am 21. 04. 2016).

Friedman, Jerome H. (1997). On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. In: *Data Mining and Knowledge Discovery* 1.1, S. 55–77.

Fruchterman, Thomas M. J. und Edward M. Reingold (1991). Graph Drawing by Force-Directed Placement. In: *Software: Practice and Experience* 21.11, S. 1129–1164.

Fuhse, Jan A. (2009). The Meaning Structure of Social Networks. In: *Sociological Theory* 27.1, S. 51–73.

Fuhse, Jan A. und Sophie Mützel, Hrsg. (2010). *Relationale Soziologie. Zur kulturellen Wende der Netzwerkforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Galilei, Galileo (1832). *Opere*. Bd. 2. Nicolò Bettoni.

Gardner, Martin (1970). Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game "Life". In: *Scientific American* 223.4, S. 120–123.

Garfinkel, Harold (1967). *Studies in Ethnomethodology*. 10. print. Englewood Cliffs, N.J.: Prentice-Hall.

Geertz, Clifford (1999). *Dichte Beschreibung: Beiträge zum Verstehen kultureller Systeme*. 16 Aufl. Suhrkamp-Taschenbuch Wissenschaft 696. Frankfurt am Main: Suhrkamp.

Ghaziani, Amin (2014). Measuring Urban Sexual Cultures. In: *Theory and Society* 43.3-4, S. 371–393.

Giddens, Anthony (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Univ of California Press.

Gledhill, Christopher (2000). *Collocations in Science Writing*. Tübingen: Günter Narr Verlag.

Goderis, Benedikt und Mila Versteeg (2011). *The Transnational Origins of Constitutions: An Empirical Investigation*. In: *CentER Discussion Paper Series No. 2013-010*. 6th Annual Conference for Empirical Legal Studies.

Goffman, Erving (1961). *On the Characteristics of Total Institutions*. In: *Symposium on Preventive and Social Psychiatry*, S. 43–84.

Goffman, Erving (1997). *Asyle*. 11. Edition Suhrkamp ; 678. Frankfurt am Main: Suhrkamp.

Goffman, Erving (2004). *Wir alle spielen Theater: die Selbstdarstellung im Alltag*. 2. München [u.a.]: Piper.

Graesser, Arthur C., Sallie E. Gordon und Lawrence E. Brainerd (1992). QUEST: A Model of Question Answering. In: *Computers & Mathematics with Applications* 23.6, S. 733–745.

Greene, Jennifer C. (2008). Is Mixed Methods Social Inquiry a Distinctive Methodology? In: *Journal of mixed methods research* 2.1, S. 7–22.

Habermas, Jürgen (1995). *Theorie des kommunikativen Handelns*. 1. Aufl. Suhrkamp Taschenbuch Wissenschaft. Frankfurt am Main: Suhrkamp.

Halliday, Michael A. K. (1961). Categories of the Theory of Grammar. In: *Word* 17.3, S. 241–292.

Harper, Douglas (2000). *Fotografien als sozialwissenschaftliche Daten*. In: *Qualitative Forschung*. Hrsg. von Uwe Flick, Ernst von Kardorff und Ines Steinke. 3. Auflage. Reinbek bei Hamburg: Rowohlt, S. 402–416.

Hartley, Ralph V. L. (1928). Transmission of Information. In: *Bell System Technical Journal* 7.3, S. 535–563.

Heaps, Harold Stanley (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York: Academic Press.

Hedeker, Donald und Robert D. Gibbons (2006). *Longitudinal Data Analysis*. Bd. 451. John Wiley & Sons.

Heiberger, Raphael H. und Jan R. Riebling (2015). U.S. and Whom? Structures and Communities of International Economic Research. In: *Journal of Social Structure* 16.9.

Heiberger, Raphael H. und Jan R. Riebling (2016). Installing Computational Social Science: Facing the Challenges of New Information and Communication Technologies in Social Science. In: *Methodological Innovations* 9, S. 1–11.

Helfferich, Cornelia (2005). *Die Qualität qualitativer Daten. Manual für die Durchführung qualitativer Interviews*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Herdan, Gustav (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. The Hague: Mouton.

Herman, David, Manfred Jahn und Marie-Laure Ryan (2007). *Routledge Encyclopedia of Narrative Theory*. Routledge.

Hickey-Moody, Anna (2015). Carbon Fibre Masculinity. In: *Angelaki* 20.1, S. 139–153.

Hitzler, Ronald (2014). *Wohin des Wegs?* In: *Qualitative Forschung*. Hrsg. von Günter Mey und Katja Mruck. Springer Fachmedien Wiesbaden, S. 55–72.

Hoffman, Matthew, Francis R. Bach und David M. Blei (2010). *Online Learning for Latent Dirichlet Allocation*. In: *Advances in Neural Information Processing Systems 23*. Neural Information Processing Systems (NIPS). Hrsg. von J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel und A. Culotta. Curran Associates, Inc., S. 856–864.

Hofmann, Thomas (1999). *Probabilistic Latent Semantic Indexing*. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, S. 50–57.

Hofmann, Thomas (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. In: *Machine learning* 42.1-2, S. 177–196.

Homans, George C. (1964). Bringing Men Back In. In: *American Sociological Review*, S. 809–818.

Hülst, Dirk (1999). *Symbol und soziologische Symboltheorie: Untersuchungen zum Symbolbegriff in Geschichte, Sprachphilosophie und Soziologie*. Opladen: Leske + Budrich.

Jaccard, Paul (1912). The Distribution of the Flora in the Alpine Zone. In: *New Phytologist* 11.2, S. 37–50.

Jaynes, Edwin T. (1957). Information Theory and Statistical Mechanics. In: *The Physical Review* 106.4, S. 620–630.

Jeong, Hawoong, S. P. Mason, Albert-László Barabási und Zoltan N. Oltvai (2001). Lethality and Centrality in Protein Networks. In: *Nature* 411.6833, S. 41–42.

Johnson, R. Burke, Anthony J. Onwuegbuzie und Lisa A. Turner (2007). Toward a Definition of Mixed Methods Research. In: *Journal of mixed methods research* 1.2, S. 112–133.

Jordania, Joseph (2006). *Who Asked the First Question? The Origins of Human Choral Singing, Intelligence, Language and Speech*. Logos.

Karlgren, Jussi, Anders Holst und Magnus Sahlgren (2008). *Filaments of Meaning in Word Space*. In: *European Conference on Information Retrieval*. Springer, S. 531–538.

Keller, Reiner (2008). Diskurse und Dispositive analysieren. Die Wissenssoziologische Diskursanalyse als Beitrag zu einer wissensanalytischen Profilierung der Diskursforschung. In: *Historical Social Research / Historische Sozialforschung* 33.1, S. 73–107.

Keller, Reiner (2012). *Das interpretative Paradigma*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Kendall, Maurice G. (1944). On Autoregressive Time Series. In: *Biometrika* 33.2, S. 105–122.

Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers und Brad S. Chissom (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. DTIC Document.

Kleene, Stephen Cole (1951). *Representation of Events in Nerve Nets and Finite Automata*. Project Rand Research Memorandum. Santa Monica, CA: U.S. Air Force, S. 101.

Klemmensen, Robert, Sara Binzer Hobolt und Martin Ejnar Hansen (2007). Estimating Policy Positions Using Political Texts: An Evaluation of the Wordscores Approach. In: *Electoral Studies* 26.4, S. 746–755.

Knorr-Cetina, Karin (2002). *Wissenskulturen: Ein Vergleich naturwissenschaftlicher Wissensformen*. Frankfurt am Main: Suhrkamp.

Knorr-Cetina, Karin (2005). *The Sociology of Financial Markets*. 1. publ. Oxford [u.a.]: Oxford Univ. Press.

Koschorke, Albrecht (2012). *Wahrheit und Erfindung: Grundzüge einer allgemeinen Erzähltheorie*. S. Fischer Wissenschaft. Frankfurt am Main: S. Fischer.

Kroneberg, Clemens (2005). Die Definition der Situation und die variable Rationalität der Akteure. Ein allgemeines Modell des Handelns. In: *Zeitschrift für Soziologie* 34.5, S. 344–363.

Kroneberg, Clemens (2007). Wertrationalität und das Modell der Frame-Selektion. In: *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 59.2, S. 215–239.

Kroneberg, Clemens (2009). *Methodologie statt Ontologie. Das Makro-Mikro-Makro-Modell als einheitlicher Bezugsrahmen der akteurstheoretischen Soziologie*. In: *Das Mikro-Makro-Modell der soziologischen Erklärung*. Hrsg. von Jens Greve, Annette Schnabel und Rainer Schützeichel. VS Verlag für Sozialwissenschaften, S. 222–247.

Krotz, Friedrich (2003). *Zivilisationsprozess und Mediatisierung: Zum Zusammenhang von Medien- und Gesellschaftswandel*. In: *Medienentwicklung und gesellschaftlicher Wandel*. Hrsg. von Markus Behmer, Friedrich Krotz, Rudolf Stöber und Carsten Winter. VS Verlag für Sozialwissenschaften, S. 15–37.

Kuckartz, Udo (2010). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS, Verl. für Sozialwiss.

Kullback, Solomon (1997). *Information Theory and Statistics*. Courier Dover Publications.

Lamnek, Siegfried (2005). *Qualitative Sozialforschung*. 4., vollst. überarb. Aufl. Weinheim [u.a.]: Beltz, PVU.

Landauer, Thomas K. und Susan T. Dumais (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. In: *Psychological review* 104.2, S. 211.

Landauer, Thomas K., D. S. McNamara, S. Dennis und W. Kintsch (2013). *Handbook of Latent Semantic Analysis*. Psychology Press.

Latapy, Matthieu, Clémence Magnien und Nathalie Del Vecchio (2008). Basic Notions for the Analysis of Large Two-Mode Networks. In: *Social Networks* 30.1, S. 31–48.

Lather, Patti und Elizabeth A. St Pierre (2013). Post-Qualitative Research. In: *International Journal of Qualitative Studies in Education* 26.6, S. 629–633.

Lehmann, Fritz (1992). Semantic Networks. In: *Computers & Mathematics with Applications* 23.2, S. 1–50.

Lemke, Matthias und Gregor Wiedemann (2016). *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. In: *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Hrsg. von Matthias Lemke und Gregor Wiedemann. Wiesbaden: Springer Fachmedien, S. 1–13.

Lewis, David D. und William A. Gale (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*. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., S. 3–12.

Leydesdorff, Loet (1989). Words and Co-Words as Indicators of Intellectual Organization. In: *Research Policy* 18.4, S. 209–223.

Leydesdorff, Loet (1997). Why Words and Co-Words Cannot Map the Development of the Sciences. In: *Journal of the American society for information science* 48.5, S. 418–427.

Leydesdorff, Loet und Kasper Welbers (2011). The Semantic Mapping of Words and Co-Words in Contexts. In: *Journal of Informetrics* 5.3, S. 469–475.

Loch, Ulrike und Gabriele Rosenthal (2002). *Das Narrative Interview*. In: *Qualitative Gesundheits- und Pflegeforschung*. Hrsg. von D. Schaeffer und G. Müller-Mundt. Bern, Göttingen, Toronto, Seattle: Hans Huber Verlag, S. 221–232.

Lüders, Christian (2000). *Beobachten im Feld und Ethnographie*. In: *Qualitative Forschung*. Hrsg. von Uwe Flick, Ernst von Kardorff und Ines Steinke. 3. Auflage. Reinbek bei Hamburg: Rowohlt, S. 384–401.

Luhmann, Niklas (1981). *Die Unwahrscheinlichkeit der Kommunikation*. In: *Soziologische Aufklärung 3*. Hrsg. von Niklas Luhmann. VS Verlag für Sozialwissenschaften, S. 25–34.

Luhmann, Niklas (1984). *Soziale Systeme. Grundriß einer allgemeinen Theorie*. 1. Frankfurt am Main: Suhrkamp.

Luhmann, Niklas (1997). *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp.

Luhmann, Niklas (2009). *Soziales System, Gesellschaft, Organisation*. 5. Aufl. Bd. 3. Soziologische Aufklärung. Wiesbaden: VS Verl. für Sozialwiss.

Lynch, Michael (1985). Discipline and the Material Form of Images: An Analysis of Scientific Visibility. In: *Social Studies of Science* 15.1, S. 37–66.

Lynch, Michael (1997). *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge University Press.

Lynch, Michael und Mark Peyrot (1992). Introduction: A Reader's Guide to Ethnomethodology. In: *Qualitative Sociology* 15.2, S. 113–122.

Malinowski, Bronislaw (1979). *Argonauten des westlichen Pazifik*. Schriften / Bronislaw Malinowski ; 1. Frankfurt am Main: Syndikat.

Malvern, David, Hrsg. (2008). *Lexical Diversity and Language Development: Quantification and Assessment*. Nachdr. Basingstoke: Palgrave Macmillan.

Malvern, David D. und Brian J. Richards (1997). *A New Measure of Lexical Diversity*. In: *Evolving Models of Language*. Hrsg. von A. Ryan und A. Wray. Clevedon: Multilingual Matters, S. 58–71.

Manaris, Bill Z., Luca Pellicoro, George Pothering und Harland Hodges (2006). *Investigating Esperanto's Statistical Proportions Relative to Other Languages Using Neural Networks and Zipf's Law*. In: *Artificial Intelligence and Applications*, S. 102–108.

Manning, Christopher D. (2008). *Introduction to Information Retrieval*. 1. publ. Cambridge [u.a.]: Cambridge Univ. Press.

Martin, Carla D. und Mason F. Porter (2012). The Extraordinary SVD. In: *The American Mathematical Monthly* 119.10, S. 838–851.

Martin, John Levi (2003). What Is Field Theory? In: *American Journal of Sociology* 109.1, S. 1–49.

Maturana, Humberto R und Francisco J Varela (1987). *Der Baum der Erkenntnis: Die biologischen Wurzeln menschlichen Erkennens*. Bern: Goldmann.

Maynard-Smith, John und George R. Price (1973). The Logic of Animal Conflict. In: *Nature* 246.5427, S. 15–18.

Mayring, Philipp (2002). *Einführung in die qualitative Sozialforschung*. 5., überarb. und neu ausgestattete Aufl. Beltz-Studium. Weinheim [u.a.]: Beltz.

McCarthy, Philip M. und Scott Jarvis (2007). Vocd: A Theoretical and Empirical Evaluation. In: *Language Testing* 24.4, S. 459–488.
McCarthy, Philip M. und Scott Jarvis (2010). MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. In: *Behavior Research Methods* 42.2, S. 381–392.
McDonnell, Terence E. (2014). Drawing out Culture: Productive Methods to Measure Cognition and Resonance. In: *Theory and Society* 43.3-4, S. 247–274.
McLaughlin, G. Harry (1969). SMOG Grading: A New Readability Formula. In: *Journal of reading* 12.8, S. 639–646.
Mead, Margaret (1965). *Leben in der Südsee. Jugend und Sexualität in primitiven Gesellschaften*. München: Szczesny.
Meade, Cathy D und Cyrus F Smith (1991). Readability Formulas: Cautions and Criteria. In: *Patient Education and Counseling* 17.2, S. 153–158.
Merton, Robert King (1993). *The Sociology of Science*. 4. [print.] Chicago [u.a.]: Univ. of Chicago Press.
Meyer, John W. (2005). *Weltkultur: Wie die westlichen Prinzipien die Welt durchdringen*. Hrsg. von Georg Krücken. Frankfurt am Main: Suhrkamp.
Mihm, Arend (1973). Sprachstatistische Kriterien zur Tauglichkeit von Lesebüchern. In: *Linguistik und Didaktik* 4, S. 117–127.
Mikheev, Andrei (2002). Periods, Capitalized Words, Etc. In: *Computational Linguistics* 28.3, S. 289–318.
Montemurro, Marcelo A. und Damián H. Zanette (2013). Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis. In: *PLoS ONE* 8.6, e66344.
Moreno-Sánchez, Isabel, Francesc Font-Clos und Álvaro Corral (2016). Large-Scale Analysis of Zipf's Law in English Texts. In: *PLOS ONE* 11.1, S. 1–19.
Morse, Janice M. (2007). *Sampling in Grounded Theory*. In: *The SAGE Handbook of Grounded Theory*. Hrsg. von Antony Bryant und Kathy Charmaz. Los Angeles; London: SAGE, S. 229–243.
Münch, Richard (1986). *Die Kultur der Moderne. Ihre Grundlagen und ihre Entwicklung in England und Amerika*. 1. Aufl. Bd. 1. 2 Bde. Frankfurt am Main: Suhrkamp.
Münch, Richard (1998). *Rational Choice - Grenzen der Erklärungskraft*. In: *Norm, Herrschaft und Vertrauen*. Hrsg. von Hans-Peter Müller und Michael Schmid. VS Verlag für Sozialwissenschaften, S. 79–91.
Mützel, Sophie (2007). Marktkonstitution durch narrativen Wettbewerb. In: *Berliner Journal für Soziologie* 17.4, S. 451–464.

Nasukawa, Tetsuya und Jeonghee Yi (2003). *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*. In: *Proceedings of the 2nd International Conference on Knowledge Capture*. ACM, S. 70–77.

Neumann, Hans (1989). "Gerechtigkeit liebe ich...". Zum Strafrecht in den ältesten Gesetzen Mesopotamiens. In: *Das Altertum* 15, S. 13–22.

Newman, Mark E. J. (2002). Spread of Epidemic Disease on Networks. In: *Physical review E* 66.1, S. 016128.

Newman, Mark E. J. (2003). The Structure and Function of Complex Networks. In: *SIAM Review* 45.2, S. 167–256.

Newman, Mark E. J. (2005). Power Laws, Pareto Distributions and Zipf's Law. In: *Contemporary physics* 46.5, S. 323–351.

Ng, Andrew Y. und Michael I. Jordan (2002). *On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes*. In: *Advances in Neural Information Processing Systems 14*. Advances in Neural Information Processing Systems (NIPS). Hrsg. von T. G. Dietterich, S. Becker und Z. Ghahramani. MIT Press, S. 841–848.

Nöth, Winfried (1985). *Handbuch der Semiotik*. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.

Nowviskie, Bethany (2014). On the Origin of "Hackänd "Yack". In: *Journal of Digital Humanities* 3.2.

Pak, Alexander und Patrick Paroubek (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. In: *Proceedings of the International Conference on Language Resources and Evaluation*. LREC 2010. Bd. 10. Valletta, Malta, S. 1320–1326.

Paltoglou, Georgios und Mike Thelwall (2010). *A Study of Information Retrieval Weighting Schemes for Sentiment Analysis*. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, S. 1386–1395.

Pang, Bo und Lillian Lee (2008). Opinion Mining and Sentiment Analysis. In: *Found. Trends Inf. Retr.* 2.1-2, S. 1–135.

Pang, Bo, Lillian Lee und Shivakumar Vaithyanathan (2002). *Thumbs Up?: Sentiment Classification Using Machine Learning Techniques*. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 79–86.

Parsons, Talcott (1964a). Evolutionary Universals in Society. In: *American Sociological Review* 29.3, S. 339–357.

Parsons, Talcott (1964b). *The Social System*. 5. print. New York: Free Press of Glencoe.

Parsons, Talcott (1968). *The Structure of Social Action*. A Free Press paperback. New York: Free Press.
Patterson, Francine G. (1980). Innovative Uses of Language by a Gorilla: A Case Study. In: *Children's language* 2, S. 497–561.
Peirce, Charles S. (1906). Prolegomena to an Apology for Pragmaticism. In: *The Monist* 16.4, S. 492–546.
Peirce, Charles S. (1931–0035). *Collected Papers of Charles Sanders Peirce*. 6 Bde. Cambridge, Mass.: Belknap Press of Harvard Univ. Press.
Pickering, Andrew (2010). *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press.
Plotkin, Joshua B. und Grzegorz Kudla (2011). Synonymous but Not the Same: The Causes and Consequences of Codon Bias. In: *Nature Reviews Genetics* 12.1, S. 32–42.
Popper, Karl R. (2005). *The Logic of Scientific Discovery*. Taylor & Francis e-Library.
Porter, Martin F. (1980). An Algorithm for Suffix Stripping. In: *Program* 14.3, S. 130–137.
Premack, David und Ann James Premack (1994). Levels of Causal Understanding in Chimpanzees and Children. In: *Cognition* 50.1, S. 347–362.
Quillian, M. Ross (1967). Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. In: *Behavioral Science* 12.5, S. 410–430.
Řehůřek, Radim (2011). *Subspace Tracking for Latent Semantic Analysis*. In: *Advances in Information Retrieval*. Hrsg. von Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee und Vanessa Mudoch. Lecture Notes in Computer Science 6611. Berlin Heidelberg: Springer, S. 289–300.
Řehůřek, Radim und Petr Sojka (2010). *Software Framework for Topic Modelling with Large Corpora*. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, S. 45–50.
Riebling, Jan R. (2018). *The Medium Data Problem in Social Science*. In: *Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications*. Hrsg. von Cathleen M. Stuetzer, Martin Welker und Marc Egger. Neue Schriften zur Online-Forschung of the German Society for Online Research (DGOF). Köln: Herbert von Halem, S. 77–103.
Riedl, Rupert (1981). *Biologie der Erkenntnis: Die stammesgeschichtlichen Grundlagen der Vernunft*. Unter Mitarb. von Robert Kaspar. Berlin: Parey.

Riedl, Rupert (1985). *Die Spaltung des Weltbildes: biologische Grundlagen des Erklärens und Verstehens*. Berlin, Hamburg: Parey.

Robertson, Stephen (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. In: *Journal of documentation* 60.5, S. 503–520.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers und Padhraic Smyth (2004). *The Author-Topic Model for Authors and Documents*. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, S. 487–494.

Roth, Camille (2013). Socio-Semantic Frameworks. In: *Advances in Complex Systems* 16 (04n05), S. 1350013.

Rumbaugh, Duane M., Timothy V. Gill und E. C. von Glasersfeld (1973). Reading and Sentence Completion by a Chimpanzee (Pan). In: *Science* 182.4113, S. 731–733.

Russell, Bertrand (1908). Mathematical Logic as Based on the Theory of Types. In: *American Journal of Mathematics* 30.3, S. 222–262.

Sailer, Lee Douglas (1979). Structural Equivalence: Meaning and Definition, Computation and Application. In: *Social Networks* 1.1, S. 73–90.

Salton, Gerard (1979). Mathematics and Information Retrieval. In: *Journal of Documentation* 35.1, S. 1–29.

Saussure, Ferdinand de (2001). *Grundfragen der allgemeinen Sprachwissenschaft*. Walter de Gruyter.

Schegloff, Emanuel A. (1989). Harvey Sacks — Lectures 1964–1965 an Introduction/Memoir. In: *Human Studies* 12.3-4, S. 185–209.

Schulze, Gerhard (2004). *Die beste aller Welten: Wohin bewegt sich die Gesellschaft im 21. Jahrhundert?* 1. Aufl. Frankfurt am Main: Fischer Taschenbuch Verlag.

Schütz, Alfred (1974). *Der sinnhafte Aufbau der sozialen Welt. Eine Einleitung in die verstehende Soziologie*. 1. Aufl. Suhrkamp-Taschenbuch Wissenschaft ; 92. Frankfurt am Main: Suhrkamp.

Senter, R. J. und E. A. Smith (1967). *Automated Readability Index*. DTIC Document.

Shannon, Claude Elwood (1948). A Mathematical Theory of Communication. In: *Bell System Technical Journal* 27, S. 379–423, 623–656.

Shannon, Claude Elwood und Warren Weaver (1976). *Mathematische Grundlagen der Informationstheorie*. Übers. von Helmut Dreßler. München: R. Oldenbourg.

Smith, Dennis (2001). *Norbert Elias and Modern Social Theory*. Sage.

Soeffner, Hans-Georg (2014). *Interpretative Sozialwissenschaft*. In: *Qualitative Forschung*. Hrsg. von Günter Mey und Katja Mruck. Springer Fachmedien Wiesbaden, S. 35–53.

Spärck Jones, Karen (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. In: *Journal of documentation* 28.1, S. 11–21.

Steyvers, Mark und Tom Griffiths (2007). *Probabilistic Topic Models*. In: *Latent Semantic Analysis: A Road to Meaning*. Hrsg. von Thomas K. Landauer, D McNamara, S Dennis und W Kintsch. Bd. 427. Laurence Erlbaum, S. 424–440.

Stirner, Max (2005). *Der Einzige und sein Eigentum*. Erftstadt: Area.

Taramasco, Carla, Jean-Philippe Cointet und Camille Roth (2010). Academic Team Formation as Evolving Hypergraphs. In: *Scientometrics* 85.3, S. 721–740.

Taylor, Ann, Mitchell Marcus und Beatrice Santorini (2003). *The Penn Treebank: An Overview*. In: *Treebanks*. Hrsg. von Anne Abeillé. Text, Speech and Language Technology 20. Springer Netherlands, S. 5–22.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal und David M. Blei (2006). Hierarchical Dirichlet Processes. In: *Journal of the American Statistical Association* 101.476, S. 1566–1581.

Templin, Mildred C. (1957). *Certain Language Skills in Children; Their Development and Interrelationships*. Minneapolis, MN, US: University of Minnesota Press.

Tribus, Myron (1961). *Thermostatics and Thermodynamics: An Introduction to Energy, Information and States of Matter, with Engineering Applications*. Princeton, NJ: Van Nostrand.

Turner, Jonathan H. (2016). Academic Journals and Sociology's Big Divide: A Modest But Radical Proposal. In: *The American Sociologist* 47.2-3, S. 289–301.

Udehn, Lars (2002). The Changing Face of Methodological Individualism. In: *Annual Review of Sociology* 28, S. 479–507.

Uprichard, Emma (2013). Big Data, Little Questions? In: *Discover Society* 1.

Vaisey, Stephen und Andrew Miles (2014). Tools from Moral Psychology for Measuring Personal Moral Culture. In: *Theory and Society* 43.3-4, S. 311–332.

Von Neumann, John und Oskar Morgenstern (1953). *Theory of Games and Economic Behavior*. 3. eds. Princeton, NJ: Princeton University Press.

Wallman, Joel (1992). *Aping Language*. Cambridge University Press.

Wang, Chong, David M. Blei und David Heckerman (2012). Continuous Time Dynamic Topic Models. In: *arXiv preprint arXiv:1206.3298*.

Wang, Chong, John W. Paisley und David M. Blei (2011). *Online Variational Inference for the Hierarchical Dirichlet Process*. In: *International Conference on Artificial Intelligence and Statistics*, S. 752–760.

Wasserman, Stanley und Katherine Faust (1994). *Social Network Analysis: Methods and Applications*. Structural analysis in the social sciences 8. Cambridge ; New York: Cambridge University Press.

Wasserstein, Ronald L. und Nicole A. Lazar (2016). The ASA's Statement on p -Values: Context, Process, and Purpose. In: *The American Statistician* 70.2, S. 129–133.

Watts, Duncan J. (2007). A Twenty-First Century Science. In: *Nature* 445.7127, S. 489–489.

Watts, Duncan J. (2014). Common Sense and Sociological Explanations. In: *American Journal of Sociology* 120.2, S. 313–351.

Watts, Duncan J. und Steven H. Strogatz (1998). Collective Dynamics of Small-World Networks. In: *Nature* 393, S. 440–442.

Weber, Max (2006). *Wirtschaft und Gesellschaft*. Paderborn: Voltmedia.

White, Harrison C., Jan A. Fuhse, Matthias Thiemann und Larissa Buchholz (2007). Networks and Meaning: Styles and Switchings. In: *Soziale Systeme* 13 (1+2), S. 543–555.

Whyte, William Foote (1973). *Street Corner Society. The Social Structure of an Italian Slum*. 2. ed., 14. impr. Chicago: Univ. of Chicago Pr.

Willett, Peter (2006). The Porter Stemming Algorithm: Then and Now. In: *Program: electronic library and information systems* 40.3, S. 219–223.

Young, R. Michael und Johanna D. Moore (1994). *DPOCL: A Principled Approach to Discourse Planning*. In: *Proceedings of the Seventh International Workshop on Natural Language Generation*. Association for Computational Linguistics, S. 13–20.

Zhang, Harry (2004). *The Optimality of Naive Bayes*. In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. FLAIR 2004. Miami Beach, Florida, USA.

Zimek, Arthur, Erich Schubert und Hans-Peter Kriegel (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. In: *Statistical Analysis and Data Mining* 5.5, S. 363–387.

Zipf, George Kinsley (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge, Mass.: Addison Wesley.

Anhang A: TextTools Code

texttools/__init__.py

```
import nltk
from tools import *
```

texttools/measures.py

```
import nltk
import pyphen
from scipy.stats import hypergeom
from math import sqrt

punctuation =
    nltk.tokenize.punkt.PunktSentenceTokenizer.PUNCTUATION

##########################################################
## lexical diversity measures

def ttr(tokens):
    """Type-Token-Ratio for a list of tokens. -> float"""
    return len(set(tokens)) / float(len(tokens))

def hdd(tokens, r=42):
    """Calculates the hd-D measure for a list of tokens"""
    fdist = nltk.FreqDist(tokens)
    N = fdist.N()
    parts = []
    for token in fdist:
        p = hypergeom.pmf(0, fdist.N(), fdist[token], r)
        part = (1. / r) * (1 - p)
        parts.append(part)
    return sum(parts)

```

```
26 def mtld_value(tokens, factor_cutoff=0.72):
27     """McCarthy&Jarvis 2010. One-way calculation of the
28     "measure of lexical and textual diversity".
29     list_of_tokens -> float"""
30     factor_score = float(0)
31     partition = []
32     for token in tokens:
33         partition.append(token)
34         if ttr(partition) < factor_cutoff:
35             factor_score += 1
36             partition = []
37     if partition != []:
38         partial_factor = 1 - ttr(partition)
39         factor_range = 1 - factor_cutoff
40         remainder = partial_factor / factor_range
41         factor_score += remainder
42     value = len(tokens) / factor_score
43     return value
44
45 def mtld(tokens):
46     """Measure of textual, Lexical Diversity;
47     McCarthy & Jarvis 2010. Both ways.
48     list_of_tokens -> float
49     """
50     forward = mtld_value(tokens)
51     reverse = mtld_value(list(tokens)[::-1])
52     mtld_final = (forward + reverse)/2.
53     return mtld_final
54
55 ########################################################
56 ## readability scores
57
58 ## Bestimmung der Parameter
59
60 # Words
61 # Defined as alphabetic tokens.
62
63 def only_words(tokens):
64     return [token for token in tokens
```

```
65              if token.isalpha()]
66
67  # $n_T$
68
69  def word_count(tokens):
70      return len(only_words(tokens))
71
72
73  # $c_T$
74  # Only alphabetic characters, no whitespace.
75
76  def char_count(tokens):
77      words = only_words(tokens)
78      return len(''.join(words))
79
80  # $s_T$
81  # Counting sentences as the amount of punctuation tokens.
82
83  def sent_count(tokens, puncts=punctuation):
84      count = 0
85      for punct in puncts:
86          count += tokens.count(punct)
87      return count
88
89
90  # $\tokens{syl}_T$ and $\tokens{polysyl}_T$
91  # Needs pyphen to work.
92
93  def syl_count(tokens, lang='en_EN'):
94      dic = pyphen.Pyphen(lang=lang)
95      counts = 0
96      words = only_words(tokens)
97      for word in words:
98          syl = dic.inserted(word).count('-') + 1
99          counts += syl
100     return counts
101
102 def polysyl_count(tokens, ge=2, lang='en_EN'):
103     dic = pyphen.Pyphen(lang=lang)
```

```
104     counts = 0
105     words = only_words(tokens)
106     for word in words:
107         if dic.inserted(word).count('-') >= ge:
108             counts += 1
109     return counts
110 
111 ## Formulas
112 
113 def ari(tokens):
114     c = float(char_count(tokens))
115     n = float(word_count(tokens))
116     s = float(sent_count(tokens))
117     return 4.71 * (c / n) + 0.5 * (n / s) - 21.43
118 
119 def fre(tokens, lang='en_EN'):
120     n = float(word_count(tokens))
121     s = float(sent_count(tokens))
122     syl = float(syl_count(tokens, lang=lang))
123     return 206.835 - 1.015 * (n / s) - 84.6 * (syl / n)
124 
125 def fkg(tokens, lang='en_EN'):
126     n = float(word_count(tokens))
127     s = float(sent_count(tokens))
128     syl = float(syl_count(tokens, lang=lang))
129     return 0.39 * (n / s) + 11.8 * (syl / n) - 15.59
130 
131 def smog(tokens, lang='en_EN'):
132     s = float(sent_count(tokens))
133     polysyl = float(polysyl_count(tokens, ge=2, lang=lang))
134     return 1.043 * sqrt((polysyl * (30 / s))) + 3.1291
135 
136 def wst4(tokens, lang='de_DE'):
137     n = float(word_count(tokens))
138     s = float(sent_count(tokens))
139     s_hat = n / s
140     polysyl = float(polysyl_count(tokens, ge=2, lang=lang))
141     return 0.2656 * s_hat + 0.2744 * ((polysyl / n) * 100) -
            1.693
```

texttools/nta.py

```
## Contains functions and methods for
## Network-Text-Analysis

import networkx as nx
import nltk
from itertools import combinations

## CRA - Centering Resonance Analysis

def np_tree(tagged_sent, grammar='NP: {<JJ>*<NN>+}'):
    parser = nltk.RegexpParser(grammar)
    tree = parser.parse(tagged_sent)
    return tree

def np_trees(tagged_sents, grammar='NP: {<JJ>*<NN>+}'):
    trees = []
    for sent in tagged_sents:
        parser = nltk.RegexpParser(grammar)
        tree = parser.parse(sent)
        trees.append(tree)
    return trees

def extract_noun_phrases(tree, identifier='NP'):
    nps = []
    for subtree in tree.subtrees(filter=lambda t:
        t.label()=='NP'):
        nps.append(subtree.leaves())
    return nps

def nps_to_graph(nps_list):
    """Takes a list of noun phrases written as lists and
        returns
    a networkx graph."""
    G = nx.Graph()
    for np in nps_list:
        for nodes in np:
            G.add_node(nodes[0].lower(),
                       pos=nodes[1])
```

```
37          if len(np)!=1:
38              nbunch = [n.lower() for n,_ in np]
39              edges = combinations(nbunch, 2)
40              for u, v in edges:
41                  if G.has_edge(u, v):
42                      G[u][v]['weight'] += 1
43                  else:
44                      G.add_edge(u, v, weight=1)
45      return G
46
47  def cra_graph(tagged_sents, grammar='NP: {<JJ>*<NN>+}',
        identifier='NP'):
48      """From a list of pos-tagged sentences a CRA graph is
            produced."""
49      trees = np_trees(tagged_sents, grammar=grammar)
50      nps_list = []
51      for tree in trees:
52          nps = extract_noun_phrases(tree,
                identifier=identifier)
53          nps_list.extend(nps)
54      graph = nps_to_graph(nps_list)
55      return graph
```

texttools/tools.py

```
1  import nltk
2
3  ## Sentence_word tokenizer
4
5  def sent_word_tokenize(text):
6      """Takes a string and tokenizes it by using the NLTKs
7      sentence tokenizer function. Each sentence is than also
8      tokenized on the level of words. Returns a list of lists
9      of tokens."""
10     L = []
11     sentences = nltk.sent_tokenize(text)
12     for sentence in sentences:
13         tokens = nltk.word_tokenize(sentence)
14         L.append(tokens)
15     return L
```

```
16
17 ## Combine identified collocations using '_'
18
19 def combine_collocations(tokens, collocations, using='_'):
20     before = [first + ' ' + second
21               for first, second in collocations]
22     after = [first + using + second
23              for first, second in collocations]
24     text = ' '.join(tokens)
25     for b, a in zip(before, after):
26         text = text.replace(b, a)
27     return text.split(' ')
```

University
of Bamberg
Press

Gegenstand dieser Arbeit ist die Entwicklung einer sozialwissenschaftlichen Methodologie zur Analyse symbolischer Ordnungen. Darunter werden hier die Strukturen und Systeme von sozial standardisierten Zeichen (Symbolen) verstanden, die sich im Prozess der Kommunikation bilden.

ISBN 978-3-86309-641-0
9 783863 096410
www.uni-bamberg.de/ubp