

Computational Sensitizing. Zum Mehrwert von Textual Data-Mining im Kontext genuin qualitativer Forschung

Martin Siebach & Axel Philipps

Keywords:
qualitative
Forschung;
Textual Data-
Mining;
automatische
Textanalysen;
qualitative
Methoden; Mixed
Methods

Zusammenfassung: Analyse- und Visualisierungsverfahren des *Textual Data-Minings* finden zunehmend auch in der qualitativen Sozialforschung Anwendung. Damit gehen konzeptionelle Überlegungen einher, solche Verfahren direkt in Forschungsprozesse einzubetten. Im Gegensatz zu einer wechselseitigen Kombination von digitalen und qualitativen Methoden argumentieren wir für eine nachgeordnete Nutzung des *Textual Data-Minings*, die sich auf sprachstatistische Analyseschritte zum Zweck eines *Computational Sensitizing* beschränkt. Mit computationaler Sensibilisierung meinen wir eine Perspektivenerweiterung: *Textual Data-Mining* dient dann – wie bisher Vorwissen und Theorien – allein dazu, Fragen an und veränderte Sichtweisen auf das Material anzuregen. Im Beitrag veranschaulichen wir exemplarisch an einem Projekt aus der Bildungsforschung, wie Ergebnisse des *Textual Data-Minings* für Besonderheiten in den qualitativen Daten sensibilisieren können.

Inhaltsverzeichnis

- [1. Einleitung](#)
- [2. TDM und qualitative Sozialforschung](#)
- [3. Von der *Theoretical Sensitivity* zum *Computational Sensitizing*](#)
- [4. Die Arbeit mit dem *interactive Leipzig Corpus Miner* \(iLCM\)
 - \[4.1 Qualitative Teilstudie\]\(#\)
 - \[4.2 TDM-Teilstudie\]\(#\)
 - \[4.3 Ertrag des TDM für die qualitative Teilstudie\]\(#\)](#)
- [5. Fazit zum Einsatz von TDM in der qualitativen Forschung](#)

[Literatur](#)

[Zu den Autoren](#)

[Zitation](#)

1. Einleitung

Als LAZER et al. (2009) die *Computational Social Science* skizzierten, galt ihr Augenmerk der quantitativen Sozialforschung. Sie stellten verschiedene *Data-Mining*-Werkzeuge und Visualisierungsmethoden vor, mit denen sich die exponentiell wachsenden Datenmengen (Stichwort *Big Data*) sozialwissenschaftlich erschließen lassen. *Data-Mining*-Verfahren sind computergestützte Anwendungen, um Trends, Informationen oder Muster in bestehenden Datenbeständen abzuleiten. Sie kommen daher häufig explorativ in *Big Data*-Untersuchungen zum Einsatz (FRANKEN 2022a; IGNATOW & MIHALCEA 2016; RUNKLER 2015). Auf diesem Weg wurde *Data Mining* in die empirische Sozialforschung eingeführt (EVANS & ACEVES 2016; IGNATOW & MIHALCEA 2016; LEMKE & WIEDEMANN 2016) und zugleich der Anspruch formuliert, solche Verfahren auch für theoriegenerierende Verfahren der qualitativen Sozialforschung nutzbar zu machen (z.B. Grounded-Theory-

Methodologie [GTM], Abduktion, Typenbildung; vgl. z.B. BAKER 2012; WIEDEMANN 2016). [1]

Wir beobachten jedoch, dass bei großen Datenmengen der qualitativen Forschung in Kombination mit *Data Mining* (z.B. FÖRSTER 2016; MARCOLIN, DINIZ, BECKER & DE OLIVEIRA 2023; WIEDEMANN 2016) in erster Linie lediglich die Funktion zukommt, innerhalb von *Big Data*-Analysen Parameter zu kalibrieren, Algorithmen zu trainieren und Ergebnisse zu validieren (NELSON 2020; STULPE & LEMKE 2016; WIEDEMANN 2016). In solchen Designs sind die qualitativen Methoden den *Data-Mining*-Analysen untergeordnet. Es gibt hingegen kaum Ansätze, in denen *Data-Mining*-Werkzeuge primär für genuin qualitative Sozialforschung konzipiert werden (Ausnahmen sind FRANKEN 2022a, 2022b; PAPILLOUD & HINNEBURG 2018; PHILIPPS 2018). Doch nur dann sehen wir die Voraussetzung erfüllt, dass bei einer Kombination mit digitalen Methoden eine Untersuchung weiterhin mittels einer qualitativen Logik angeleitet und strukturiert wird. In diesem Beitrag diskutieren wir daher, inwiefern *Data-Mining*-Verfahren qualitative Analysen unterstützen können und zugleich sichergestellt wird, dass ihre Ergebnisse sich der qualitativen Forschung unterordnen und nicht umgekehrt. Damit wollen wir zeigen, wie sich solche Verfahren nutzbar machen lassen, ohne dass die Ergebnisse an den Maßstäben von *Data-Mining*-Analysen gemessen werden. In qualitativen Studien ist der Einsatz von *Data Mining* dann auf die Funktion limitiert, Forschende für Besonderheiten in ihren Daten zu sensibilisieren. [2]

Unsere Überlegungen setzen beim Vorschlag von FRANKEN (2022a, 2022b) an, qualitative Forschung computationell zu unterstützen. Sie demonstriert, wie digitale Methoden eingebunden werden können, um neue Einsichten und Erkenntnisse zu generieren. Wir teilen ihre Einschätzung, dass insbesondere explorative *Data-Mining*-Verfahren wie das *Topic Modelling* "gut geeignet" sind, "erste, induktive Einblicke in große, unbekannte Textmengen" (FRANKEN 2022a, S.171) zu gewähren. Aber aufgrund der fehlenden Passung digitaler Methoden mit qualitativen Gütekriterien sehen wir die Gefahr, dass eine weiterführende Integration von *Data-Mining*-Werkzeugen in die qualitative Forschungslogik letztere untergräbt. Denn die Güte von Ergebnissen, die mit *Data-Mining*-Verfahren generiert sind, wird stets an quantitativen Maßstäben gemessen, auch in Kombination mit qualitativen Methoden. Den Zugewinn von computationell generierten Ergebnissen für die qualitative Studien sehen wir daher funktionell ausschließlich als Anregung und Perspektiverweiterung. Diese Art der computationellen Sensibilisierung (*Computational Sensitizing*)¹ hilft vergleichbar zu Vorkenntnissen und Theorien (TAVORY & TIMMERMANS 2014) oder den *Sensitizing Concepts* (BLUMER 1954), Eigenheiten in digital aufbereiteten Daten zu erfassen, ohne den qualitativen Auswertungsprozess zu überformen. Für unsere Herangehensweise orientieren wir uns am oft diskutierten Verhältnis von theoretischen Vorannahmen und qualitativer Forschungspraxis (CHARMAZ 2006;

1 Axel PHILIPPS hat dieses Konzept 2022 in seinem Vortrag "Qualitative Sozialforschung und Computational Sensitivity" auf dem Kongress der Deutschen Gesellschaft für Soziologie vorgestellt. Mit der Umformulierung soll das aktivierende Moment stärker betont und eine Verwechslung mit dem Begriff der *Computational Sensitivity* (z.B. AUGUSTIN & MAURER 2001) in den Natur- und Ingenieurwissenschaften vermieden werden.

GLASER & STRAUSS 2006 [1967]; TAVORY & TIMMERMANS 2014). Hier wird die Rolle von Konzepten und Theorien oft klar limitiert: Theorien sollen weder anleiten, noch bestätigt werden. Sie sollen vielmehr ausschließlich für Eigenheiten in den Daten des Untersuchungsgegenstandes sensibilisieren. [3]

Im Beitrag gehen wir zunächst auf Verfahren des *Textual Data-Minings* (TDM) und ihre bisherige Nutzung in den Sozialwissenschaften ein (Abschnitt 2). Anschließend legen wir unseren Vorschlag für die Verwendung von TDM in der qualitativen Sozialforschung dar (Abschnitt 3). Dazu leiten wir das Konzept der computationellen Sensibilisierung aus dem Verständnis der *Theoretical Sensitivity* ab (GLASER & STRAUSS 2006 [1967]) und zeigen im Abschnitt 4 an einem Beispiel aus der Bildungsforschung (SIEBACH 2022), wie in diesem Sinne das TDM in einer qualitativen Untersuchung genutzt werden kann. Wir diskutieren daran, welchen Mehrwert diese Art der Sensibilisierung für die qualitative Forschung haben kann (Abschnitt 5). [4]

2. TDM und qualitative Sozialforschung

Für die qualitative Sozialforschung sehen wir zwei sehr unterschiedliche Strategien des Einsatzes computergestützter Verfahren: die Verwendung von qualitativer Analysesoftware (z.B. ATLAS.ti, MAXQDA oder NVivo) und den Einsatz von TDM-Werkzeugen. Für Erstere ist ein zentrales Kennzeichen, dass (zumindest ohne Nutzung der KI-Assistenz) Kodierprozesse technisch unterstützt werden, ohne Forschungsprozesse zu überformen (KUCKARTZ 2010). Durch TDM-Verfahren werden hingegen computergestützt sprachstatistische Textanalysen über Bedeutungsstrukturen und -muster in großen und häufig unstrukturierten Textmengen erstellt (EVANS & ACEVES 2016; FRANKEN 2022a; RUNKLER 2015). Während sich in einem Einzeltext oder einer Textpassage eine Bedeutungsstruktur beim Lesen verhältnismäßig gut erkennen lässt, ist dies in größeren Textkorpora nicht ohne Weiteres möglich. Mit den TDM-Werkzeugen hingegen können auf Grundlage bestimmter Muster im realen Sprachgebrauch semantische Zusammenhänge aufgedeckt werden. HEYER, QUASTHOFF und WITTIG (2006, S.3) sprachen deshalb von computergestützten Verfahren für "die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten". In erster Linie lassen sich damit explorative Analysen durchführen. Forschende können so beispielsweise spezifische Wörter und Wortkombinationen identifizieren, semantische Relationen berechnen, statistische Ähnlichkeiten und Differenzen aufzeigen oder vergleichbare Dokumente hinsichtlich bestimmter Kriterien identifizieren (BIEMANN, HEYER & QUASTHOFF 2022). [5]

Die letztgenannten Verfahren haben im vergangenen Jahrzehnt zunehmend zur Wissensproduktion in den Sozialwissenschaften beigetragen. Mittlerweile ist eine Reihe von Einführungsartikeln und Lehrbüchern erschienen (EVANS & ACEVES 2016; IGNATOW & MIHALCEA 2016; FRANKEN 2022a) und es liegen viele Forschungsarbeiten über die Nutzung unterschiedlicher TDM-Verfahren für eine große Bandbreite an Themen vor. Exemplarisch zeigten NELSON, GETMAN und HAQUE (2022), dass für Wikipedia-Beiträge zur Frauenbewegung nur

unvollständige Daten genutzt wurden. VOYER, KLINE, DANTON und VOLKOVA (2022) demonstrierten, wie durch Migration geteilte Vorstellungen von sozialer Inklusion verändert wurden, und BONIKOWSKI und NELSON (2022) beschrieben, wie rechtsradikale Kandidat*innen diskursive Rahmungen aus der Mainstream-Politik übernahmen. [6]

Mit der Einführung von TDM in die Sozialwissenschaften finden sich auch Überlegungen, diese Verfahren für qualitative Textanalysen zu nutzen (EVANS & ACEVES 2016; FRANKEN 2022a; WIEDEMANN 2016), wobei in der Regel die Kombination von sprachstatistischen und qualitativen Verfahren vorgeschlagen wurde. STULPE und LEMKE (2016) führten beispielsweise mit dem Begriff *Blended Reading* ein mehrstufiges modularisiertes methodisches Setting ein, in dem größere Textmengen computationell strukturiert werden und qualitative Analysen helfen, die automatisiert generierten Ergebnisse zu verbessern. Einige Autor*innen stellten sogar in Aussicht, mit TDM die Aussagekraft qualitativer Forschung durch objektive Ergebnisse zu verbessern (CHEN, DROUHARD, KOCIELNIK, SUH & ARAGON 2018; KOBAYASHI, MOL, BERKERS, KISMIHÓK & DEN HARTOG 2018; WIEDEMANN 2013, 2016). Für solche Vorschläge ist kennzeichnend, dass zwar qualitative Verfahren und TDM verknüpft werden, aber Ersteren in der Regel ausschließlich die Aufgabe zugewiesen wird, die Aussagekraft und Robustheit quantitativ erzeugter Ergebnisse zu erhöhen. Es gibt wenige Untersuchungen, in denen von eigenständigen qualitativen Analysen ausgegangen und *im direkten Anschluss* mit TDM-Verfahren fortgefahren wird, um die Aussagekraft qualitativ gewonnener Befunde zu verallgemeinern oder zu ergänzen. So hat etwa PAJO (2015) in ihrer Untersuchung zum öffentlichen Diskurs um Atommüll bezüglich Umweltbelastungen und Gesundheit in den USA Beobachtungen zusammengefasst und daraus Annahmen abgeleitet, die sie im Anschluss mit TDM-Werkzeugen empirisch überprüfte (PAJO 2019). Vergleichbar haben BAKER et al. (2008) Verfahren der Computerlinguistik mit der kritischen Diskursanalyse nach FAIRCLOUGH und WODAK (1997) kombiniert und typische, in den Massenmedien genutzte sprachliche Konstrukte für Flüchtlinge und Zugewanderte identifiziert. Mit der Diskursanalyse rekonstruierten die Forschenden abwertende sprachliche Darstellungen und durch computerlinguistische Verfahren konnte diese Topoi erweitert werden. Herauszustellen ist, dass in beiden Untersuchungen durch den Einsatz von TDM-Werkzeugen nachträglich eine ergänzende Perspektive auf die qualitativ gewonnenen Resultate ermöglicht wurde. [7]

Bisher lassen sich nur wenige Ansätze identifizieren, bei denen TDM *im Vorfeld* oder *im Verlauf* mit qualitativen Methoden kombiniert wird. So hatten beispielsweise JANASIK, HONKELA und BRUUN (2009) oder PAPILLOUD und HINNEBURG (2018) TDM-Werkzeuge verwendet und demonstriert, wie sich damit qualitativ erhobene Interviewdaten auswerten ließen. FRANKEN (2022a, 2022b) schlug die Erweiterung und Veränderung qualitativer Forschung mit verschiedenen computergestützten Verfahren vor. Als Bedingung formulierte sie, dass die mit TDM erzeugten Ergebnisse kritisch betrachtet werden müssten, "um fortlaufend Grenzen und Möglichkeiten für die qualitative Sozialforschung auszuloten" (FRANKEN 2022b, Zusammenfassung). Sie legte dies am Beispiel

einer wissenssoziologischen Diskursanalyse dar und formuliert den Anspruch, "eine kreative Integration neuer digitaler Elemente in den qualitativen Forschungsprozess zu etablieren", bei der "die qualitativen Perspektiven weder aufgegeben noch untergeordnet werden" (a.a.O.). Das bedeute, dass z.B. qualitatives Sampling und Erhebungsverfahren um digitale Methoden wie *Web Crawler* ergänzt und die Möglichkeiten der Auswertung von Textdaten um Verfahren des *Distant Reading* (MORETTI 2013) wie Kookkurrenzanalysen und *Topic Modelling* (für Details siehe FRANKEN 2022a) erweitert werden könnten. In dieser Kombination blieben Schritte wie die "manuelle Annotation", die "Markierung von zentralen Textteilen und induktive Kategorienentwicklung" unberührt (FRANKEN 2022b, §23). FRANKEN betonte, dass ein gewichtiger Teil der Forschung im Kern weiterhin qualitativ bleiben müsse, weil "nicht alle Schritte des Forschungsprozesses mit digitalen Verfahren erweitert werden (können). Die epistemologisch bestimmten Anteile sind weiterhin von den Forschenden manuell zu bearbeiten" (§26). [8]

Wir sehen bei dieser Konzeption allerdings die Gefahr, dass in der Praxis genuin qualitativ generierte Ergebnisse gegenüber computationell erzeugten nachrangig behandelt werden könnten. Dies mag nicht in der Absicht der Forschenden liegen, wenn aber qualitative Analysen im *Blended Reading* (STULPE & LEMKE 2016) dazu genutzt werden, die Aussagekraft *quantitativ* erzeugter Ergebnisse zu prüfen und zu erhöhen, dann steigert dies deren Plausibilität, was sie zugleich stärker als eigenständig herausstellt. Vergleichbar nehmen in der *Computational Grounded Theory* (NELSON 2020) qualitative Analyseschritte eine prominente Stellung ein, durch die vor allem sprachstatistische Analysen verbessert werden sollen. Doch auch wenn bei diesen Herangehensweisen qualitative und quantitative Verfahren als gleichwertig behandelt werden, dürfte es qualitativ Forschenden schwerfallen, insgesamt an einem qualitativ orientierten Vorgehen festzuhalten. [9]

Wir schlagen deshalb eine – vorsichtige und konservative – Verwendung von TDM-Werkzeugen vor und präferieren eine Hierarchisierung, bei der digitale Methoden einen ausschließlich dienenden Charakter für das im Kern qualitative Vorgehen haben. In diesem Sinn sollen sie nur dafür genutzt werden, für Besonderheiten der zu untersuchenden Daten zu sensibilisieren und die Möglichkeit zu geben, Designs daran auszurichten. [10]

Dafür sehen wir zwei Gründe: Erstens werden durch TDM-Werkzeuge Visualisierungen erzeugt, mit denen Muster in den Daten identifiziert werden können. So werden beim Latent Dirichlet Allocation (LDA)-*Topic Modelling* (BLEI 2012) unüberwacht Dokumente in thematische Cluster klassifiziert. Im Detail wird abgebildet, wie sich die Themen in jedem der vorliegenden Dokumente verteilen. Diese Themen wiederum stellen verschiedene Wahrscheinlichkeitsverteilungen über das Vokabular dar. Darin unterscheidet sich die Erzeugungsweise des *Topic Modellings* von anderen Gruppierungsverfahren. In den Sozialwissenschaften werden für die Identifikation distinkter Gruppen Verfahren wie die Clusteranalyse genutzt (EVERITT, LANDAU, MORVEN & STAHL 2011), durch die unterschiedliche Variablenausprägungen möglichst trennscharf gruppiert werden,

oder wie die Korrespondenzanalyse (BLASIUS 2001), bei der Variablen räumlich nah oder fern voneinander platziert werden. Dies kann überraschende Einblicke und Ansätze liefern, Daten zu strukturieren und fokussiert und vertieft zu untersuchen. Allerdings lässt sich die Gültigkeit und Plausibilität der identifizierten Muster nicht vollständig statistisch sichern, sondern die Bewertung muss durch die Forschenden erfolgen. Natürlich können sich diese bei der Anzahl von zu bildenden Clustern an Koeffizienten wie der Fehlerquadratsumme orientieren, sie brauchen aber stets Feldkenntnisse, um einschätzen zu können, durch welche und wie viele Cluster (*Topics*) eine Datenmenge allgemein und spezifisch beschrieben wird. Vor diesem Hintergrund erscheinen uns TDM-Verfahren durchaus wertvoll, um Muster in Daten zu identifizieren. Sie liefern aber lediglich Ansatzpunkte und es bedarf qualitativer Untersuchungen, um eine Untersuchungsfrage adäquat beantworten zu können. [11]

Zweitens ist uns für den Einsatz von TDM in der qualitativen Sozialforschung wichtig, dass die Nutzung an deren Spezifik und Gütekriterien ausgerichtet bleibt (FLICK 2022; PRZYBORSKI & WOHLRAB-SAHR 2014; SCHNEIJDERBERG 2023; STRÜBING, HIRSCHAUER, AYAß, KRÄHNKE & SCHEFFER 2018). Ähnlich wie PRZYBORSKI und WOHLRAB-SAHR (2014) sehen wir die Gefahr, dass bei der zusätzlichen Verwendung quantitativer Verfahren (einschließlich der TDM-Werkzeuge) Paradigmen und Gütekriterien qualitativer Sozialforschung zugunsten quantitativer Ansätze vermischt oder sogar aufgegeben werden. Die Relevanz des Einsatzes von TDM in qualitativen Studien sollte daher an Gütekriterien – wie sie von STRÜBING und Kolleg*innen zusammengefasst wurden – bemessen werden:

"*Gegenstandsangemessen* ist eine Weise der Herstellung des Forschungsgegenstandes, die das empirische Feld ernst nimmt und Methoden, Fragestellungen und Datentypen einer fortlaufenden Justierungsanforderung unterwirft. *Empirische Sättigung* reflektiert die Güte der Verankerung von Interpretationen im Datenmaterial. *Theoretische Durchdringung* markiert die Qualität der Theoriebezüge, in die das Forschen eingespannt ist, und arbeitet an deren Irritationspotential. *Textuelle Performanz* bezeichnet die Leistung, die Texte als Güte konstituierende Kommunikation gegenüber Rezipientinnen der Forschung zu erbringen haben. *Originalität* schließlich ist das Kriterium, an dem die Einlösung des Neuigkeitsanspruchs wissenschaftlichen Wissens zu prüfen ist" (2018, S.83) [12]

Im Mittelpunkt steht die datengetriebene Interpretation des Untersuchungsmaterials, welche sich daran bemisst, reflexiv und nachvollziehbar originelle Erkenntnisse zu generieren. TDM-erzeugte Daten mögen zwar helfen, überraschende Muster offenzulegen, ihre Relevanz zeigt sich aber erst in vertiefenden (qualitativen) Analysen und der theoretischen Durchdringung. Erst im Durchlaufen des qualitativen Forschungsprozesses lässt sich der Beitrag von TDM-Verfahren für das Endergebnis ermessen. [13]

3. Von der *Theoretical Sensitivity* zum *Computational Sensitizing*

Nach unserem Verständnis können TDM-Verfahren in qualitativen Studien die gleiche Stellung wie Theorien und Vorwissen einnehmen. Um dies zu verdeutlichen, gehen wir im Folgenden kurz auf die GTM (GLASER & STRAUSS 2006 [1967]) ein. Sie entstand in dem Versuch, sich methodisch von Großtheorien für die Analyse der sozialen Wirklichkeit abzugrenzen. Dazu wurde als notwendig erachtet, bei den Daten zu beginnen und in der Auswertung für deren Eigenheiten offen zu sein. Ebenso gehörte dazu, theoretisches Vorwissen und theoriebezogene Annahmen auszuklammern. Während GTM-Techniken wie Memoschreiben oder Kodieren bis zur theoretischen Sättigung in den Kanon qualitativer Forschung eingingen, wird die Stellung der Theorie und des Vorwissens bis heute kontrovers diskutiert (BRYANT 2002; CHARMAZ 2006; TAVORY & TIMMERMANS 2014). [14]

Bei GLASER und STRAUSS (2006 [1967], S.46f.) finden sich in "The Discovery of Grounded Theory" einige Ausführungen zur theoretischen Sensibilisierung (*Theoretical Sensitivity*). Darin heißt es, dass im Auswertungsprozess Anregungen aus dem vorgängigen Wissensstand und aus den Erfahrungen eigener früherer Untersuchungen genutzt werden könnten. Es gehe allerdings nicht darum, Vorannahmen und Theorien zu überprüfen und zu bestätigen, sondern sich für Gemeinsamkeiten und Unterschiede im Untersuchungsfeld zu sensibilisieren. Konzepte und Theorien liefern demnach weniger eine Grundlage, den Untersuchungsgegenstand zu verstehen oder zu erklären, sondern bieten vor allem eine Kontrastfolie, um daran dessen Eigenheiten herauszuarbeiten (siehe auch TAVORY & TIMMERMANS 2014). [15]

In Anlehnung daran möchten wir TDM im Kontext qualitativer Forschung als eine weitere Möglichkeit empfehlen, sich für Besonderheiten eines Untersuchungsgegenstandes – *computationell* – zu sensibilisieren. In einer solchen Herangehensweise stellen Ergebnisse durch unüberwachtes und nicht trainiertes TDM weder ein vorläufiges noch das finale Ergebnis der Untersuchung dar. Wir gehen mit PÄÄKKÖNEN und YLIKOSKI (2021) bei solchen Verfahren ausschließlich von einem instrumentellen Einsatz aus. Entgegen der Annahme, dass *Topic Modellings* die Wirklichkeit repräsentierten, können solche Quellen – instrumentell gedacht – nur zu vertiefenden Interpretationen anregen: Sie dienen dazu, eine sprachstatistisch informierte zusätzliche Perspektive auf die untersuchten Daten zu erhalten und solche Anregungen in das weitere Design einzubinden. [16]

Dies kann beispielsweise bei der Fallauswahl oder bei der thematischen Sichtung der Daten hilfreich sein. So erfolgt in qualitativ angelegten Untersuchungen in der Regel *erstens* eine Annäherung an den Forschungsgegenstand über einzelne Texte. Durch deren vertiefende Analysen werden die Spezifik der Textgattung und inhaltliche Dimensionen erschlossen. Aus den ersten punktuellen Materialsondierungen werden dann Kriterien für die Auswahl des Textkorpus bestimmt bzw. eine Samplingstrategie entwickelt. Eine solche Vorgehensweise bleibt jedoch stets selektiv und punktuell. Unsicherheiten bei der Auswahl und bei

der Durchdringung des Materials lassen sich nie vollständig beseitigen. Die Zweifel dürften mit steigenden Textmengen angesichts zunehmender (digital verfügbarer) Daten sogar noch zunehmen. Durch den begleitenden TDM-Einsatz kann dann eine zusätzliche Perspektive z.B. für die Auswahl von minimalen und maximalen Kontrasten ermöglicht werden. [17]

Durch Visualisierungen mittels TDM-Verfahren können *zweitens* relativ schnell und mit geringem Aufwand thematische Strukturierungen der Textdaten offengelegt werden (FRANKEN 2022a; IGNATOW & MIHALCEA 2016; NIEKLER, KAHMANN, BURGHARDT & HEYER 2023). Mit dem *Topic Modelling* lassen sich beispielsweise Themencluster in Textkorpora identifizieren. Konzentriert man sich auf bestimmte Schlüsselworte, liefern Frequenzanalysen Aufschluss über ihre Verteilungen in den Dokumenten, Kookkurrenzanalysen und Similaritätsanalysen zeigen häufig auftretende Wortkombinationen und inhaltliche Konnotationen und eine Volatilitätsanalyse kann veranschaulichen, wie gemeinsam auftretende Worte im Längsschnitt verändert, neu kombiniert oder durch andere ersetzt werden. [18]

Im zweiten Teil des Artikels gehen wir auf die Verwendung von TDM-Werkzeugen im Rahmen einer wissenssoziologischen Diskursanalyse ein (SIEBACH 2022). Martin SIEBACH hatte mit digitalen Methoden Schwerpunkte der fachlichen Debatte zum Thema Identität in der erziehungswissenschaftlichen Subdisziplin Didaktik des Sachunterrichts sichtbar gemacht und, durch dieses Vorwissen sensibilisiert, das kanonisierte Wissen über Identität in Lehr- und Einführungsbüchern zum Sachunterricht kritisch analysiert. [19]

4. Die Arbeit mit dem *interactive Leipzig Corpus Miner* (iLCM)

SIEBACH (2022) verwendete in seiner Untersuchung einen Teil der TDM-Werkzeuge auf der Plattform [iLCM](#). Hierbei handelt es sich um eine browserbasierte, von der Universität Leipzig bereitgestellte und interaktive "integrated research environment for the analysis of text data" (NIEKLER et al. 2023, S.327). Auf der Plattform werden unterschiedliche Mining- und Visualisierungswerkzeuge für die textbasierte Sozialforschung – für die Vorverarbeitung über Standardtextanalysen bis zu Visualisierungen – bereitgestellt, die häufig genutzt werden und manuell sehr zeitaufwändig wären. [20]

In dem hier vorgestellten Projekt untersuchte SIEBACH (2022) wissenssoziologisch den Diskurs zum Thema Identität in wissenschaftlichen Veröffentlichungen zum Sachunterricht. Sachunterricht ist zum einen ein Fach in der Grundschule, welches im Kontext der sozial-liberalen Bildungsreformen zu Beginn der 1970er Jahre eingeführt wurde und sukzessive (jedoch nicht in allen Bundesländern gleichermaßen) das Fach Heimatkunde ersetzte. Zum anderen wurde Sachunterricht als wissenschaftliche Subdisziplin an denjenigen Hochschulen etabliert, in denen ein Studiengang Lehramt an Grundschulen angeboten wird. Hier werden inhaltliche (fachwissenschaftliche) und (fach-) didaktische Positionierungen zum Sachlernen in der Kindheit verhandelt. Im fachlichen Diskurs werden dazu bildungsrelevante Themen aus anderen

Wissenschaftsdisziplinen und aus dem Alltag aufgegriffen und in den Lehrkanon integriert. Dies gilt beispielsweise für so Unterschiedliches wie physikalische Eigenschaften von Stoffen, Nachhaltigkeit oder Demokratiebildung. [21]

Identität wird in den Erziehungs- und Sozialwissenschaften zwar sehr breit behandelt, hat aber wenig Eingang in den Sachunterricht gefunden. Vor diesem Hintergrund arbeitete SIEBACH die Identitätsvorstellungen im wissenschaftlichen Diskurs der Disziplin Sachunterricht heraus. Im Fokus dieser wissenssoziologisch informierten Diskursanalyse standen differierende Begriffsverständnisse und Identitätsaspekte sowie die Herausarbeitung dominierender und marginaler Diskursstränge. Die Analyse erfolgte in zwei Schritten mit jeweils unterschiedlich umfangreichen Korpora. [22]

4.1 Qualitative Teilstudie

Im Mittelpunkt der an den Prämissen der kritischen Diskursanalyse (TITSCHER; WODAK; MEYER & VETTER 1998) orientierten qualitativen Feinanalyse stand ein Textkorpus aus Lehr- und Einführungsbänden sowie einigen weiteren Monografien, die eine bedeutsame konzeptionelle Neuorientierung in der Sachunterrichtsdidaktik markierten. Dieses Korpus repräsentiert den als kanonisiert geltenden wissenschaftlichen Wissens- und Diskussionsstand im Sachunterricht und bildet die jeweils aktuelle inhaltliche Grundlage des Studiums des Sachunterrichts ab. Diese Texte wurden in Bezug auf den formalen und inhaltlichen Aufbau der Publikationen im Allgemeinen analysiert, besonders aber daraufhin, welche Bedeutung die Diskursakteur*innen der Identität einräumten, wie sie den Begriff einführten und definierten, welches Verständnis von Identität und Identitätsentwicklung darin deutlich wurde und mit welchen Sachthemen dies verknüpft wurde. Ziel war es also, 1. die Bedeutsamkeit des Gegenstandes Identität einzuschätzen, 2. die Identitätsbegriffe und -vorstellungen in den Texten zu erschließen und zu klären, ob und wie an gängige Identitätsbegriffe, -konzepte oder -theorien aus den Erziehungs- und Sozialwissenschaften angeknüpft wurde, und 3. zu untersuchen, in welchen Kontexten des Sachunterrichts Identität thematisiert wurde. [23]

Als Ergebnis der Analyse stellte sich heraus, dass die untersuchten Bände eine große inhaltliche Breite, aber zugleich strukturelle Ähnlichkeiten aufwiesen. So wurde etwa die *Geschichte der Disziplin* als Beitrag zur Klärung der Ziele und grundlegender Strukturen der Disziplin behandelt. Außerdem wurden zentral *grundsätzliche Leitideen und Prinzipien der Disziplin* diskutiert, beispielsweise die Balance zwischen der Orientierung an Bedürfnissen und Entwicklungsaufgaben von Kindern oder der Orientierung an Wissen aus Wissenschaftsdisziplinen und -kulturen. Dies ging mit unterschiedlichen didaktischen Konzeptionen einher, zum Beispiel dem wissenschaftsorientierten oder dem vielperspektivischen Sachunterricht.² Des Weiteren wurden *fachliche Inhalte* thematisiert, also z.B.

2 Der wissenschaftsorientierte Sachunterricht ist eine zuerst in den 1970er-Jahren als Reaktion auf den Erfolg der russischen bemannten Raumfahrt ("Sputnikschock") entwickelte Konzeption, die stark an Wissensbeständen aus den naturwissenschaftlichen Bezugsdisziplinen des Fachs ausgerichtet wurde. Beim vielperspektivischen Sachunterricht hingegen wird erwartet, Unterrichtsgegenstände sowohl bezüglich kindlicher Sichtweisen als auch mit Blick auf

Orientierung in der Zeit (unterschiedliche Zeitkonzepte sowie Umgang mit Geschichte), Orientierung im Raum (geografische Aspekte) sowie naturwissenschaftliche (zum Beispiel ökologische Systeme, Körper und Körperfunktionen, physikalische Eigenschaften von Stoffen) und sozialwissenschaftliche Fragen (etwa Zusammenleben in der Gesellschaft). In der Regel geschah das im direkten Zusammenhang mit *fachdidaktischen Inhalten*, beispielsweise Kriterien der Auswahl von Inhalten für den Unterricht, Stufen des Wissensaufbaus sowie fachbezogenen Methoden. [24]

Für die Untersuchungsfrage war zentral, dass Identität innerhalb dieses Korpus kaum thematisiert, sondern wenn, dann in der Regel als nicht näher bestimmtes Schlagwort genutzt wurde. In den Lehr- und Einführungsbänden wurden zudem keine Bezüge zu gängigen Identitätskonzepten hergestellt. Die Autor*innen dieser in sich relativ geschlossenen Werke boten als Diskursakteur*innen daher kaum Anschlussmöglichkeiten zum Identitätsdiskurs in den Sozial- und Erziehungswissenschaften. Hierfür gäbe es aber durchaus naheliegende Möglichkeiten, beispielsweise wenn Identitätsentwicklung von Kindern und die daraus resultierenden Bedürfnisse als ein didaktisches Leitprinzip formuliert oder wenn Bezüge zu Identität bei Themen des Sachunterrichts hergestellt würden, die in den Sozial- und Erziehungswissenschaften meist im Kontext von Identität bzw. Identitätsentwicklung besprochen werden – wie Interkulturalität, Sexualität, Geschlecht und Gender oder Beschäftigung mit der eigenen Biografie. [25]

4.2 TDM-Teilstudie

Der qualitativen Analyse ging ein TDM eines größeren Datenkorpus voraus, der alle wissenschaftlichen Veröffentlichungen im Fach Sachunterricht im Zeitraum von 1992 bis 2017 umfasste.³ Dieses Korpus bestand aus 1.587 Dokumenten und schloss unterschiedliche Formate wie Handbuchartikel, Einzelbeiträge in Zeitschriften und Sammelbänden ebenso ein wie die schon erwähnten Lehr- und Einführungsbände und Monografien. Die Textquellen wurden, wenn nicht schon in maschinenlesbarer Form verfügbar, in eine solche überführt und mittels verschiedener TDM-Analysetools sprachstatistisch erschlossen. Zu den verwendeten Verfahren zählten unter anderem eine Kookkurrenzanalyse und ein *Topic Modelling* (zu weiteren Verfahren siehe SIEBACH 2022, S.73-79). Während durch die Analyse von Kookkurrenzen aufgezeigt werden sollte, mit welchen Worten das Schlagwort Identität statistisch häufig zusammen auftritt, sollte ein *Topic Modelling* einen Einblick in die Breite der behandelten Themen geben, einschließlich jener, in deren Textkontext Identität thematisiert wird (Näheres zu den Verfahren weiter unten in diesem Abschnitt). [26]

sozialwissenschaftliche, naturwissenschaftliche, technische, historische und geografische Perspektiven zu analysieren und auszubalancieren.

3 In das Korpus wurden einbezogen alle Publikationen der Fachgesellschaft (Gesellschaft für die Didaktik des Sachunterrichts – GDSU), alle sachunterrichtsbezogenen Beiträge in den grundschuldidaktischen und grundschulpädagogischen Publikationsorganen (ermittelt über Titel und Abstracts) sowie alle Einführungsbände zum Fach und die Handbücher zur Didaktik des Sachunterrichts. Ausgeschlossen wurden alle unterrichtspraktischen Publikationen bzw. Beiträge, die einen überwiegenden Anteil unterrichtspraktischer Inhalte aufweisen.

Im Zuge des TDM wurden die Ergebnisse inhaltlich dahingehend geprüft, ob die *Topics* tatsächlich die behandelten Themen in Einzeltexten wiedergaben (STULPE & LEMKE 2016). Beispielsweise wurden Texte eines *Topics* mit geschichtswissenschaftlichen Begriffen (z.B.: Quelle, Geschichtsbewusstsein, Nationalsozialismus) herangezogen und geklärt, worum es in den Beiträgen inhaltlich ging und was es für das Verständnis des *Topics* bedeutete. Für diesen Analyseschritt wurde also ein quantitatives Verfahren (*Topic Modelling* als sprachstatistisches Verfahren) mit einer qualitativen Inhaltsanalyse (MAYRING 2015) kombiniert. Die mit TDM-Verfahren generierten Ergebnisse eröffneten dabei eine weitere Perspektive für die qualitative Teilstudie des kanonisierten Wissens zum Sachunterricht; die Texte konnten gezielt bezüglich des Vorkommens bestimmter Begriffe und mit ihnen verbundener Themen analysiert werden. [27]

Im Folgenden zeigen wir exemplarisch, was durch die Kookkurrenzanalyse und das *Topic Modelling* sichtbar gemacht wurde und wie auf diese Weise der qualitative Fokus für Spezifisches geschärft werden konnte. Wir beginnen mit der Analyse der Kookkurrenzen, also der Untersuchung des gemeinsamen Vorkommens von Wörtern in Sätzen. Die sprachstrukturalistische Grundannahme ist, dass beim gemeinsamen Auftreten von Wörtern in Sätzen eine syntagmatische Relation vorliegt, beide sich also funktional und inhaltlich ergänzen (HEYER et al. 2006, S.20-25). [28]

Abbildung 1 zeigt die statistisch signifikanten Kookkurrenzen zu Identität im Datenkorpus (NIEKLER et al. 2023, S.334). Deutlich wird, dass das gemeinsame Auftreten von "identität" und "alterität" (das auf Identität zurückwirkende [Andere](#)) im Korpus die größte Signifikanz aufwies. Die Stärke der Verbindungslinien verdeutlicht die Signifikanz des gemeinsamen Auftretens von zwei Begriffen: je stärker, desto höher ist der Dice-Koeffizient (Signifikanzmaß im Spektrum von 0, zwei Begriffe treten niemals miteinander auf, bis 1, sie treten immer gemeinsam auf). Geringer fällt die Signifikanz für die Worte "kultur", "kulturelle", "kultureller" und "kulturellen" aus. Die Summe der Signifikanzmaße aller Wörter mit dem Wortstamm "kultur" ist jedoch größer als die von "alterität". Die Visualisierung legt demnach nahe, dass Identität im gesamten Datenkorpus signifikant gemeinsam mit Kultur und Alterität verwendet wird. Andere relativ signifikant gemeinsam mit Identität auftretende Worte verweisen auf Prozesse ("dauer", "aufbauen", "wandel", "erzählung") und individuelle Zustände ("bewusstsein"). Zusammen mit den mittelbar gemeinsam auftretenden Worten werden unterschiedliche semantische Felder aufgespannt, in denen Identität zusammen mit "pluralität", "fremdheit", "moral", "gesellschaft", "gedächtnis", "entwicklung" und anderen Begriffen vorkommt (zu weiteren Ergebnissen siehe SIEBACH 2022, S.122).

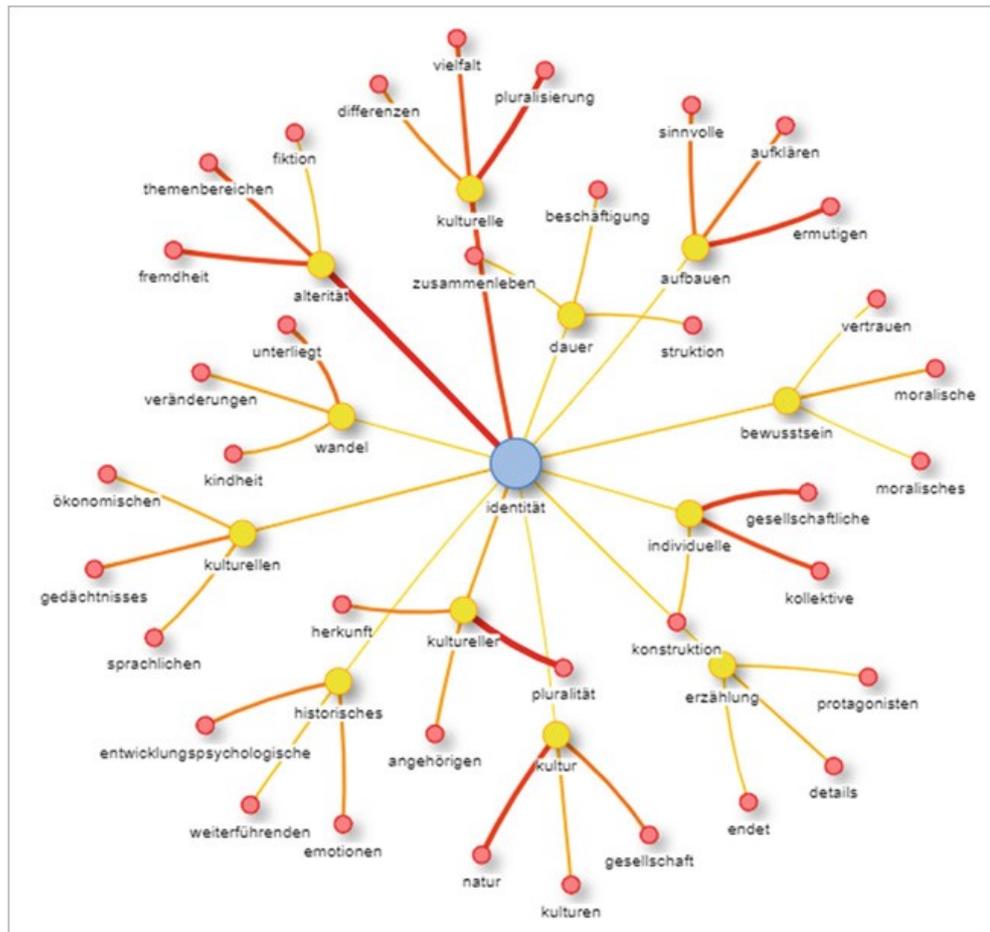


Abbildung 1: Kookkurrenzen für "Identität" [29]

Das Resultat der Kookkurrenzanalyse sensibilisierte für gemeinsame Wortnutzungen der Diskursakteur*innen und welche dieser Wörter im Kontext von Identität häufig und damit erwartbar verwendet wurden. Das half, die Thematisierung von Identität in den Lehr- und Einführungsbänden zu kontextualisieren und spezifisch zu qualifizieren. Vor dem Hintergrund dieser TDM-Analyse war auffällig, dass im Gegensatz zum signifikanten gemeinsamen Auftreten von Identität mit Kultur und Alterität im untersuchten Korpus (wissenschaftlicher Gesamtdiskurs) dies kaum in den Einführungsbänden und Monografien aufgegriffen wurde. Zudem steht der thematischen Breite, in der Identität im wissenschaftlichen Diskurs verhandelt wurde (siehe Abb. 1), eine Engführung im kanonisierten Wissen gegenüber. Kennzeichnend für die Monografien und Handbücher zum Sachunterricht war, dass Identität entweder als unbestimmtes Schlagwort oder im Kontext von Geschlecht, Sexualität, Körper und regionaler Verortung sowie als abhängig von Zeitbewusstsein thematisiert wurde. Viele andere Anschlüsse zum Diskurs über Identität (in der Subdisziplin Sachunterricht) wurden hier nicht behandelt. [30]

Das genutzte *Topic Modelling* bietet einen unüberwachten Ansatz zum Clustern von Textsammlungen. Die Methode basiert auf dem LDA-Algorithmus. Mit LDA

werden die Wortzusammensetzungen modelliert und entsprechend der Wortverwendungen in zusammenhängende Gruppen organisiert, die als Themenfelder zu identifizieren sind (NIEKLER et al. 2023, S.335). Die *Topics* werden durch Wortlisten repräsentiert (generiert aus einer Kombination von Häufigkeit des Vorkommens und dem ausschließlichen Vorkommen in einem *Topic*). Im Koordinatensystem auf der linken Seite von Abb. 2 werden die *Topics* durch Kreise symbolisiert. Ihre Größe korreliert mit dem Anteil, den sie am Gesamtdiskurs, also im Datenkorpus haben. Im rechten Feld wird die Wortliste der häufigsten Wörter eines *Topics* (hier *Topic 18*) aufgeführt. Der rote Balken steht für die konstitutive Häufigkeit eines Wortes für das *Topic*. Die blauen Balken zeigen das Vorkommen des Wortes in anderen *Topics* im Gesamtdiskurs. Das Verhältnis von rot zu blau gibt also wieder, inwiefern das Wort spezifisch für dieses *Topic* ist.

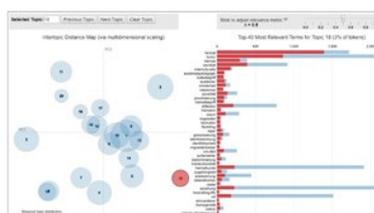


Abbildung 2: sozialwissenschaftlich geprägtes *Topic* Heimat, Kultur und interkulturelle Bildung (*Topic 18*). Bitte klicken Sie [hier](#) oder auf die Abbildung für eine Vergrößerung. [31]

Mit Blick auf die Verteilung der *Topics* wurde zum einen ersichtlich, welche Themen im untersuchten wissenschaftlichen Diskurs insgesamt vorkamen. Beispielsweise konnten *Topics* identifiziert werden zum allgemeinen Diskurs des Sachlernens (*Topic 1*), über Heimat, Kultur und interkulturelle Bildung (*Topic 18*), zur Umweltbildung (*Topic 20*), zu Geschlecht und Gender (*Topic 9*), zu naturwissenschaftlichen Inhalten des Sachunterrichts (*Topic 11*) oder zu raumbezogener Bildung, Verkehrserziehung und Bildung für nachhaltige Entwicklung (*Topic 12*). Zum anderen ließen sich verschiedene Topoi mit Bezug zu Identität ausmachen. Dabei zeigte sich beispielsweise, dass dieses Wort zwar in einer Reihe von *Topics* vorkam, letztlich aber in nur wenigen Themenfeldern des Sachunterrichts eine nennenswerte Rolle spielte. Dies war beispielsweise in *Topic 1* (allgemeiner Diskurs des Sachlernens) der Fall. Beim Lesen zugehöriger Texte zu diesem *Topic* stellte sich jedoch heraus, dass viele sich auf die Identität des Fachs Sachunterricht bezogen und nicht im Sinne persönlicher oder kollektiver Identität oder mit Bezug auf die Entwicklung von Identität gebraucht wurden. Außerdem gab es nennenswerte Wortverwendungen in *Topic 5* (sozialwissenschaftliche Bildung im Sachunterricht) und in den spezifischen Themenfeldern Geschlecht und Sexualität (*Topic 9*, Schwerpunkt Mädchen/Frauen) sowie Geschichtsdidaktik (*Topic 13*, Schwerpunkt Nationalsozialismus). Das größte Gewicht kam Identität jedoch in der Konstruktion des Themenfeldes Heimat, Kultur und interkulturelle Bildung zu (*Topic 18*). [32]

Diese Auffächerung der verhandelten Themen sensibilisierte in der qualitativen Teilstudie der Lehr- und Handbücher dafür, dass weder die Breite noch die Gewichtungen des Fachdiskurses im kanonisierten Wissensbestand gespiegelt wurden. Charakteristisch für das Thema Identität war, dass es in den Lehr- und Handbüchern zum einen nur marginal vorkam, zum anderen auch Anschlussmöglichkeiten zum Diskurs in den Erziehungs- und Sozialwissenschaften weitgehend fehlten. Eine vollständige Darstellung der Ergebnisse bietet SIEBACH (2022, S.134-157). [33]

4.3 Ertrag des TDM für die qualitative Teilstudie

Wir möchten nun kurz zusammenfassen, welche Anregungen sich aus dem TDM für die qualitative Untersuchung ergaben. Bezüglich des Analyseziels, die Bedeutsamkeit des Identitätsthemas im kanonisierten Wissen herauszuarbeiten, zeigte sich beispielsweise, dass Identität und Identitätsentwicklung im Großteil der untersuchten Literatur keine oder nur eine marginale Rolle spielten. Die sprachstatistischen Verteilungen und Muster im gesamten wissenschaftlichen Diskurs generierten ein überblicksartiges Vorwissen für die folgende qualitative Analyse. [34]

Während durch das TDM eine Breite an Themen rund um Identität aufgezeigt werden konnte, wurden durch die qualitative Analyse nur einige wenige Teilidentitäten (Geschlechts-, sexuelle, Körper- und regionale Identität) herausgearbeitet. Der Vergleichshorizont half, deren Spezifik einzuordnen. So wurden etwa geschlechts- und genderspezifische Fragen in den Lehr- und Handbüchern lediglich am Rande behandelt. Für KAHLERT (2016, S.36) beispielsweise diente die Thematisierung von Geschlechtsidentität lediglich zur Illustration des Verhältnisses von didaktischer Planung und Unvorhersehbarem im Unterrichtsverlauf (bei einem Experiment zur unterschiedlichen Wärmeleitfähigkeit von Materialien). Für die Behandlung von Identität im Kontext von Sexualität zeigte sich ein etwas anderes Bild: Sexuelle Identität wurde in den Lehr- und Handbüchern ebenso wie im wissenschaftlichen Diskurs kaum thematisiert. Allerdings fand sich im Einführungsband von HARTINGER und LANGE-SCHUBERT (2017) ein Artikel von LANDWEHR (2017) zur Sexualbildung, in welchem sexuelle Identität als bedeutsam diskutiert wurde (allerdings mit wenig aussagekräftigen Referenzen). Mit dem Vorwissen aus den TDM-Analysen wird die Bedeutung dieser Veröffentlichung noch einmal offensichtlicher, bietet sie doch einen Anknüpfungspunkt für eine didaktische Aufbereitung des Themas Identität. [35]

Die größte Irritation lösten die TDM-Analysen durch die Ausweisung erwartbarer Thematisierungen aus. Im untersuchten Korpus des wissenschaftlichen Diskurses war Identität in erheblichem Maße im Kontext von Kultur und Interkulturalität verhandelt worden. Mit diesem Vorwissen war es umso auffälliger, dass sich in den Lehr- und Handbüchern solche Bezüge von Identität und Kultur bzw. von kulturellem oder interkulturellem Lernen kaum fanden. GLUMPLER (1996) legte beispielsweise eine Monografie zum interkulturellen Lernen vor,

ohne das Wort Identität zu verwenden oder die individuelle Identitätsentwicklung zu thematisieren. [36]

Das aus den TDM-Analysen resultierende überblicksartige Wissen sensibilisierte also für die verhandelten Themen und vor allem für Nichtthematisiertes, mithin Marginalisiertes in den Einführungsbänden (welche das kanonisierte Wissen der wissenschaftlichen Subdisziplin repräsentieren). Das half, die Spezifik der thematischen und strukturellen Aufbereitung der Werke und damit die Diskurspositionen der Autor*innen herauszuarbeiten. Mit den TDM-Werkzeugen wurde folglich die Aufmerksamkeit (*Computational Sensitizing*) auf bestimmte Aspekte der Verwendung des Wortes Identität in den Lehr- und Handbüchern gelenkt. Sicher hätte eine umfassende Lektüre der Fachdebatte ebenfalls Einblicke geliefert, möglicherweise in differenzierterer Weise als der computationell erzeugte Überblick. Das vollständige Lesen und Erschließen eines umfangreichen Diskurses ist aber schon aus forschungspragmatischen Erwägungen nicht immer empfehlenswert und zudem selten leistbar. Anders ausgedrückt: TDM-Analysen können das Eintauchen und gezielte Lesen nicht ersetzen. Mit ihnen können auch keine Erkenntnisse generiert werden, *wie* Diskursakteur*innen bestimmtes Wissen erzeugen und aufbereiten. Mittels Häufigkeitsauszählungen und thematischen Clusterungen können aber Hinweise auf sprachliche Besonderheiten und thematische Verteilungen generiert werden. Dies bietet eine zusätzliche Perspektive auf die Daten. [37]

5. Fazit zum Einsatz von TDM in der qualitativen Forschung

Mit *Computational Sensitizing* sprechen wir uns für eine produktive Verbindung von TDM-Werkzeugen und qualitativer Forschung aus. Wir leiten sie aus der *Theoretical Sensitivity* nach GLASER und STRAUSS (2006 [1967]) bzw. TAVORY und TIMMERMANS (2014) ab, in der Erwartung, dass mit computergenerierten Visualisierungen qualitativ Forschende für Besonderheiten und Eigenheiten der untersuchten Textdaten sensibilisiert werden und diese dazu beitragen, qualitative Textinterpretationen zu unterstützen. Es geht aber nicht darum, Ergebnisse aus Letzteren durch TDM zu prüfen. Wir plädieren vielmehr dafür, wie am Beginn von WEBERs (2017 [1905]) Abhandlung "Die protestantische Ethik und der 'Geist' des Kapitalismus" durch quantitative Daten Fragen anzuregen. In seiner berühmten Studie stellte er seiner Argumentation statistische Daten aus dem Zensus über die Verteilung von Konfessionen, Berufs- und Bildungshintergründen voran. Mit ihnen wurde aber nicht der Beweis für die zentrale These des Buches angetreten, sondern diese bildeten vielmehr den Ausgangspunkt dafür, sich mit dem Phänomen, dass die Entstehung der kapitalistischen Wirtschaftsform gemeinsam mit der protestantischen Ethik in Europa auftrat, eingehender auseinanderzusetzen und es dann anhand von Fallbeispielen zu erklären. [38]

Wenn der Gebrauch von TDM-Werkzeugen in der qualitativen Sozialforschung auf *Computational Sensitizing* beschränkt wird, sehen wir die Voraussetzung erfüllt, dass der qualitative Forschungsprozess durch sprachstatistische Verfahren nicht überformt wird. Die computergenerierten Visualisierungen bleiben

Hilfsmittel. Sie stehen am Anfang des Forschungsprozesses und liefern Hinweise zur Erhebung und für vertiefenden Auswertungen. Sie sind kein zentrales Untersuchungsergebnis, welches aufgrund seiner Erzeugungsweise den Gütekriterien quantitativer Forschung folgen müsste. Die Analyse im Ganzen bleibt qualitativ ausgerichtet, und ihre Aussagekraft bemisst sich an den spezifischen Gütekriterien qualitativer Forschung. [39]

Literatur

- Augustin, Dirk & Maurer, Helmut (2001). Computational sensitivity analysis for state constrained optimal control problems. *Annals of Operations Research*, 101, 75-99.
- Baker, Paul (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247-256.
- Baker, Paul; Gabrielatos, Costas; Khosravinik, Majid; Kryzanowski, Michal; Mc Enery, Tony & Wodak, Ruth (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3), 273-306.
- Biemann, Chris; Heyer, Gerhard & Quasthoff, Uwe (2022). *Wissensrohstoff Text. Eine Einführung in das Text Mining*. Wiesbaden: Springer VS.
- Blasius, Jörg (2001). *Korrespondenzanalyse*. Berlin: Oldenbourg.
- Blei, David M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84, <https://doi.org/10.1145/2133806.2133826> [Datum des Zugriffs: 16. September 2024].
- Blumer, Herbert (1954). What is wrong with social theory?. *American Sociological Review*, 19(1), 3-10.
- Bonikowski, Bart & Nelson, Laura (2022). From ends to means: The promise of computational text analysis for theoretically driven sociological research. *Sociological Methods & Research*, 51(4), 1484-1539.
- Bryant, Antony (2002). Re-grounding grounded theory. *The Journal of Information Technology Theory and Application*, 4(1), 25-42, <https://aisel.aisnet.org/jitta/vol4/iss1/7> [Datum des Zugriffs: 10. September 2024].
- Charmaz, Kathy (2006). *Constructing grounded theory*. London: Sage
- Chen, Nan-Chen; Drouhard, Margaret; Kocielnik, Rafal; Suh, Jina & Aragon, Cecilia (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems*, 9(4), Art. 39.
- Evans, James & Aceves, Pedro (2016). Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42, 21-50, <https://doi.org/10.1146/annurev-soc-081715-074206> [Datum des Zugriffs: 10. September 2024].
- Everitt, Brain S.; Landau, Sabine; Morven, Leese & Stahl, Daniel (2011). *Cluster analysis*. Chichester: Wiley.
- Fairclough, Norman & Wodak, Ruth (1997). Critical discourse analysis. In Teun A. van Dijk (Hrsg.), *Discourse studies: A multidisciplinary introduction* (Bd. 2, S.258-284). London: Sage.
- Flick, Uwe (2022). Gütekriterien qualitativer Sozialforschung. In Nina Baur & Jörg Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S.533-547). Wiesbaden: Springer VS.
- Förster, Annette (2016). Der Folterdiskurs in den deutschen Printmedien. In Matthias Lemke & Gregor Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (S.139-166). Wiesbaden: Springer VS.
- Franken, Lina (2022a). *Digitale Methoden für qualitative Forschung: Computationelle Daten und Verfahren*. Münster: Waxmann.
- Franken, Lina (2022b). Digitale Daten und Methoden als Erweiterung qualitativer Forschungsprozesse: Herausforderungen und Potenziale aus den Digital Humanities und Computational Social Sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 23(2), Art. 12, <https://doi.org/10.17169/fqs-22.2.3818> [Datum des Zugriffs: 17. Juni 2024].

- Glaser, Barney G. & Strauss, Anselm L. (2006 [1967]). *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick, NJ: AldineTransaction.
- Glumpler, Edith (1996). *Interkulturelles Lernen im Sachunterricht*. Bad Heilbrunn: Klinkhardt.
- Hartinger, Andreas & Lange-Schubert, Kim (Hrsg.) (2017). *Sachunterricht – Didaktik für die Grundschule*. Berlin: Cornelsen.
- Heyer, Gerhard; Quasthoff, Uwe & Wittig, Thomas (2006). *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. Herdecke: W3L.
- Ignatow, Gabe & Mihalcea, Rada (2016). *Text mining: A guidebook for the social sciences*. London: Sage.
- Janasik, Nina; Honkela, Timo & Bruun, Henrik (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods*, 12(3), 436-460.
- Kahlert, Joachim (2016). *Der Sachunterricht und seine Didaktik* (4. Aufl.). Stuttgart: Klinkhardt.
- Kobayashi, Vladimer; Mol, Stefan; Berkers, Hannah; Kismihók, Gábor & Den Hartog, Deanne (2018). Text classification for organizational researchers: A tutorial. *Organizational Research Methods*, 21(3), 766-799, <https://doi.org/10.1177/1094428117719322> [Datum des Zugriffs: 10. September 2024].
- Kuckartz, Udo (2010). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Landwehr, Brunhild (2017). Sexualbildung. In Andreas Hartinger & Kim Lange-Schubert (Hrsg.), *Sachunterricht – Didaktik für die Grundschule* (S.189-203). Berlin: Cornelsen.
- Lazer, David; Pentland, Alex; Adamic, Lada; Aral, Sinan; Barabási, Albert Laszlo; Brewer, Devon; Christakis, Nicholas; Contractor, Noshir; Fowler, James; Gutmann, Myron; Jebara, Toni; King, Gary; Macy, Michael; Roy, Deb & Alstyn, Van Marchall (2009). Computational social science. *Science*, 323(5915), 721-723.
- Lemke, Matthias & Wiedemann, Gregor (2016). Einleitung: Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. In Matthias Lemke & Gregor Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (S.1-16). Wiesbaden: Springer VS.
- Marcolin, Carla Bonato; Diniz, Eduardo Henrique; Becker, João Luiz & de Oliveira, Henrique Pontes Gonçalves (2023). Who knows it better? Reassessing human qualitative analysis with text mining. *Qualitative Research in Organizations and Management*, 18(2), 181-198.
- [Mayring, Philipp](#) (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Beltz.
- Moretti, Franco (2013). *Distant reading*. London: Verso Books.
- Nelson, Laura K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3-42.
- Nelson, Laura K.; Getman, Rebekah & Haque, Syed Arefinul (2022). And the rest is history: Measuring the scope and recall of Wikipedia's coverage of three women's movement subgroups. *Sociological Methods & Research*, 51(4), 1788-1825, <https://journals.sagepub.com/doi/full/10.1177/00491241211067514> [Datum des Zugriffs: 10. September 2024].
- Niekler, Andreas; Kahmann, Christian; Burghardt, Manuel & Heyer, Gerhard (2023). The interactive Leipzig corpus miner: An extensible and adaptable text analysis tool for content analysis. *Publizistik*, 68, 325-354, <https://link.springer.com/article/10.1007/s11616-023-00809-4> [Datum des Zugriffs: 10. September 2024].
- Pääkkönen, Juho, & Ylikoski, Petri (2021). *Humanistic interpretation and machine learning*. *Synthese*, 199(1-2), 1461-1497, <https://doi.org/10.1007/s11229-020-02806-w> [Datum des Zugriffs: 10. September 2024].
- Pajo, Judi (2015). Danger explodes, space implodes: The evolution of the environmental discourse on nuclear waste, 1945-1969. *Energy, Sustainability and Society*, 5(36), 1-13, <https://doi.org/10.1186/s13705-015-0064-6> [Datum des Zugriffs: 10. September 2024].
- Pajo, Judi (2019). Quantitative falsification for qualitative findings: Falsifying an ethnographic theory of American public discourse on nuclear waste with text mining in R. *Social Science Computer Review*, 37(3), 315-332.

- Papilloud, Christian & Hinneburg, Alexander (2018). *Qualitative Textanalyse mit Topic-Modellen*. Wiesbaden: Springer VS.
- Philipps, Axel (2018). Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung. *Sozialer Sinn. Zeitschrift für hermeneutische Sozialforschung*, 19(2), 367-387, <https://doi.org/10.15488/9930> [Datum des Zugriffs: 10. September 2024].
- Przyborski, Aglaja & Wohlrab-Sahr, Monika (2014). Forschungsdesigns für die qualitative Sozialforschung. In Nina Baur & Jörg Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S.123-142). Wiesbaden: Springer VS.
- Runkler, Thomas (2015). *Data Mining. Modelle und Algorithmen intelligenter Datenanalyse*. Wiesbaden: SpringerNature.
- Schneijderberg, Christian (2023). Konventionen von Gütekriterien empirischer Sozialforschung qualitativ verstehen. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 24(3), Art. 1, <https://doi.org/10.17169/fqs-24.3.3994> [Datum des Zugriffs: 13. September 2024].
- Siebach, Martin (2022). *Identität als Diskursgegenstand der Didaktik des Sachunterrichts. Eine historisch-diskursanalytische Untersuchung*. Wiesbaden: Springer VS.
- Strübing, Jörg; Hirschauer, Stefan; Ayaß, Ruth; Krähnke, Uwe & Scheffer, Thomas (2018). Gütekriterien qualitativer Sozialforschung. Ein Diskussionsanstoß. *Zeitschrift für Soziologie*, 47(2), 83-100, <https://doi.org/10.1515/zfsoz-2018-1006> [Datum des Zugriffs: 10. September 2024].
- Stulpe, Alexander & Lemke, Matthias (2016). Blended Reading. In Matthias Lemke & Gregor Wiedemann (Hrsg.), *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse* (S.17-61). Wiesbaden: Springer VS.
- Tavory, Iddo & Timmermans, Stefan (2014). *Abductive analysis: Theorizing qualitative research*. Chicago, IL: University of Chicago Press.
- Titscher, Stefan; Wodak, Ruth; Meyer, Michael & Vetter, Eva (1998). *Methoden der Textanalyse. Leitfaden und Überblick* (S.178-203). Opladen: Westdeutscher Verlag.
- Voyer, Andrea; Kline, Zachary; Danton, Madison & Volkova, Tatiana (2022): From strange to normal: Computational approaches to examining immigrant incorporation through shifts in the mainstream. *Sociological Methods & Research*, 51(4), 1540-1579, <https://journals.sagepub.com/doi/full/10.1177/00491241221122596> [Datum des Zugriffs: 10. September 2024].
- Weber, Max (2017[1905]). *Die protestantische Ethik und der "Geist" des Kapitalismus*. Stuttgart: Reclam.
- Wiedemann, Gregor (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 14(2), Art. 23, <https://doi.org/10.17169/fqs-14.2.1949> [Datum des Zugriffs: 17. Juni 2024].
- Wiedemann, Gregor (2016). *Text mining for qualitative data analysis in the social sciences: A study on democratic discourse in Germany*. Wiesbaden: Springer VS.

Zu den Autoren

Martin SIEBACH (Dr. phil.) ist wissenschaftlicher Mitarbeiter am Institut für Schulpädagogik und Grundschuldidaktik der Martin-Luther-Universität Halle-Wittenberg. Seine Forschungsschwerpunkte sind Professionalisierung in der politischen Bildung, Sachunterrichtsdidaktik sowie Lern- und bildungstheoretische Diskursforschung.

Kontakt:

Dr. Martin Siebach

Institut für Schulpädagogik und Grundschuldidaktik
Martin-Luther-Universität Halle-Wittenberg
Franckeplatz 1 HS 31, 06110 Halle/Saale

E-Mail: martin.siebach@paedagogik.uni-halle.de

URL: https://schulpaed.philfak3.uni-halle.de/grundschule_bereiche_mitarbeiter/sac/hu/personal/1939560_2394147/

Axel PHILIPPS (PD Dr. phil.) ist Forschungsreferent und wissenschaftlicher Mitarbeiter am Institut für Hochschulforschung (HoF) an der Martin-Luther-Universität Halle-Wittenberg. Seine Forschungsschwerpunkte sind Hochschul- und Wissenschaftsforschung, Bildwissenschaften und Methoden der qualitativen Sozialforschung.

Kontakt:

PD Dr. Axel Philipps

Institut für Hochschulforschung (HoF)
Martin-Luther-Universität Halle-Wittenberg
Collegienstr. 62, 06886 Wittenberg

Tel.: +49 3491 466 254

E-Mail: axel.philipps@hof.uni-halle.de

URL: <https://www.hof.uni-halle.de/institut/mitarbeiter/axel-philipps/>

Zitation

Siebach, Martin & Philipps, Axel (2024). Computational Sensitizing. Zum Mehrwert von Textual Data-Mining im Kontext genuin qualitativer Forschung [39 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 25(3), Art. 13, <https://doi.org/10.17169/fqs-25.3.4246>.