https://doi.org/10.21248/ka-notizen.85.16

# Nachnutzung von Forschungsdaten für qualitative Forschungen. Text Mining als Ansatz zur Exploration transkribierter Interviews

**Lina Franken, Nils Egger, Luis Fischer, Katharina Lillich, Florian Schmid**

**Zusammenfassung**

Forschungsdaten werden zunehmend in digitalen Repositorien gespeichert und Dritten zur Nachnutzung zugänglich gemacht, teils beschränkt auf wissenschaftliche Anliegen. Es stellt sich die Frage, wie entsprechende Datensätze genutzt werden können, nachdem alle ethischen Dimensionen berücksichtigt sind. Dies wird anhand der exemplarischen Bearbeitung eines umfangreichen Datensatzes aufgezeigt, der lebensgeschichtliche Interviews mit Sozialwissenschaftler:innen des 20. Jahrhunderts enthält. Während der Schwerpunkt in der Erhebung auf Lebenserzählungen lag, wird in der Nachnutzung nach Veränderungen der Berufsbiografien im 20. Jahrhundert gefragt. Der Beitrag beschreibt, wie die Datensätze explorativ mit digitalen Methoden untersucht werden und geht der Frage nach, welche Erkenntnisse aus dieser Nachnutzung von Interviewtranskripten gezogen werden können. Wie viel Aussagekraft haben die Interviewpassagen, welche zur eigenen Fragestellung identifizierbar sind? Welche Leerstellen bleiben in der Interpretation? Hierbei werden verschiedene Standardverfahren des Text Mining im Rahmen von digitalen Methodenanwendungen, wie Wortfrequenzen, n-Gramme und Topic Modeling, auf ihre Nutzbarkeit und Nützlichkeit für die Bearbeitung der Datensätze hin befragt.

**Schlagwörter:** digitale Methoden, Sekundäranalyse, Forschungsdaten, Text Mining, Digital Humanities

Prof. Dr. Lina Franken, Digital Humanities in den Kulturwissenschaften, Universität Vechta, Deutschland

Nils Egger, Zentrum für interdisziplinäre Risiko- und Innovationsforschung, Universität Stuttgart, Deutschland

Luis Fischer, Institut für Soziologie, Ludwig-Maximilians-Universität München, Deutschland

Katharina Lillich, Institut für Soziologie, Ludwig-Maximilians-Universität München, Deutschland

Florian Schmid, Institut für Soziologie, Ludwig-Maximilians-Universität München, Deutschland

## Forschungsdaten mit digitalen Methoden nachnutzen

Forschungsdaten werden zunehmend digital hinterlegt, um sie für andere Forschende verfügbar und transparent zu machen (Bishop & Kuula-Luumi 2017; Hollstein & Strübing 2018; Corti & Thompson 2004). Dies erfordert gutes Forschungsdatenmanagement (vgl. Imeri et al. in diesem Band) und geschieht in strukturierten Datenspeichern. Ein solches sogenanntes Repositorium ist etwa das Forschungsdatenzentrum Qualiservice Data Sharing, welches spezialisiert qualitative Forschungsdaten bereithält. Die in ethnografischer Forschung erhobenen, oftmals sensiblen Daten können, anders als in den Naturwissenschaften, nicht im großen Stil offen zugänglich gemacht werden. Repositorien schützen vor unbefugtem Zugriff durch Begründungsanforderungen und Auflagen (Hollstein & Strübing 2018). Zudem sind qualitative Daten besonders vielfältig und eng mit den Forschenden als Personen verbunden (vgl. Positionspapier der DGEKW 2018, aktuell in Revision). Fragen nach Dokumentation, Schutz und Transparenz stellen sich also ganz neu (vgl. Imeri 2018; 2019; zu Forschungen mit historischer Dimensionierung Timm 2020; vgl. aus Perspektive von Qualiservice Heuer et al. 2020 oder für die empirische Sozialforschung Watteler & Kinder-Kurlanda 2015).

Neben den notwendigen Schutzmaßnahmen und den oft diskutierten Problemen bietet diese neue Verfügbarkeit von Quellen großes Potenzial für Sekundäranalysen im Sinne einer Nachnutzung von Daten anderer Forschenden (zur Systematik unterschiedlicher Nachnutzungen vgl. Schöch 2017, zu Möglichkeiten der Umsetzung von Sekundäranalysen aus der qualitativen Sozialforschung Ruggiano & Perry 2017). Was bisher höchstens mit persönlichen Anfragen oder mit großem zeitlichem Abstand über Archive und Nachlässe verfügbar war, ist nun mit entsprechenden Metadaten im besten Fall systematisch recherchierbar und strukturiert zugänglich.

Für synchrone und diachrone Vergleiche oder für Arbeiten in schwer zugänglichen Feldern entstehen enorme Potentiale. So werden vulnerable Gruppen von Forschungspartner:innen geschützt, da keine zusätzlichen Interviews stattfinden müssen (vgl. Gladstone et al. 2007: 440). Voraussetzung ist allerdings die Vergleichbarkeit der Erkenntnisinteressen. Besonders relevant ist die Beantwortung ganz neuer Fragestellungen mit bestehenden Daten (vgl. Heaton 2008: 510). Bei den Datensätzen in Repositorien kann davon ausgegangen werden, dass diese hinsichtlich der ethischen und rechtlichen Dimensionen geprüft wurden und erst nach Abwägung zur Nachnutzung freigegeben werden. Digitale Methoden der Nachnutzung sind hierdurch auch für jene relevant, die ihre Forschungsdaten selbst für die Nachnutzung aufbereiten möchten. In der Folge ergibt sich ein reichhaltiger Fundus an qualitativen Quellen, insbesondere Interviewtranskripten, die potentiell in Repositorien liegen. Diese sind auch über die Empirische Kulturwissenschaft/Kulturanthropologie/Europäische Ethnologie (EKW/EE/KA) hinaus weit verbreitete Forschungsdaten, die bereits heute in großer Zahl in Repositorien liegen. Im Methodenspektrum der EKW/EE/KA sind Nachnutzungen bisher wenig verbreitet, in klassisch ethnografischen Arbeiten wird Material in der Regel selbst erhoben, um Kontexte nachvollziehen zu können (vgl. Bischoff et al. 2014). Aber auch die qualitative Sozialforschung profitiert von den genannten Potentialen.

Wer mit Forschungsdaten von anderen arbeitet, hat schnell große unbekannte Datenmengen vorliegen, die mit manuellen Ansätzen kaum überschaubar sind. Bisher kommt in qualitativer Forschung in der Regel digitale Handarbeit (zum Begriff siehe Franken 2020b) zum Einsatz, um einzelne Arbeitsschritte zu vereinfachen. Dabei wird häufig Qualitative Data Analysis (QDA) Software genutzt, in der codiert bzw. annotiert wird (vgl. für die EKW/EE/KA Sattler 2014). Deshalb bieten sich digitale Methoden mit automatisierten, durch

den Computer durchgeführten – computationellen – Verfahren an. Bei den von uns unten beschriebenen Verfahren werden quantitative Ansätze, wie etwa die Auswertung statistischer Verteilungen von Wortkombinationen, nutzbar gemacht. Sie alleine reichen aber noch nicht aus, um ethnografische Fragestellungen zu beantworten, sondern ergänzen das bisherige Methodenrepertoire vielmehr (Franken 2022).

Durch die Kombination digitaler Methoden und qualitativer Analyse können die vorhandenen Daten tiefgreifend untersucht werden und auch fehlende Kontexte und andere Lücken im Korpus systematisch identifiziert werden, die dann zu einer weiteren Erhebung führen können (vgl. z.B. Watters et al. 2018). Damit kann etwa explorativ die eigene Fragestellung präzisiert werden. Neue Erkenntnisse ergeben sich dabei nicht nur aus der großen Menge der verwendeten Datensätze, sondern auch aus dem vielfältigen Spektrum der berücksichtigten Transkripte und deren Untersuchung mit Methodenkombinationen verschiedener digitaler Verfahren.

Der vorliegende Beitrag erörtert, wie computationelle Unterstützungen aussehen können, die über digitale Handarbeit hinausgehen. Es geht uns dabei nicht (nur) um die Analyse von einzelnen Quellen, sondern um die Möglichkeit der Kombination mit neuen methodischen Zugängen, die sich aus der Entwicklung digitaler Methoden auch für qualitative Forschung ergeben. Sie ermöglichen erste Momente des Verstehens und der Orientierung in großen Datensätzen. So diskutieren wir, welche digitalen Methoden sich für diese Sichtung und Sortierung von Forschungsdaten eignen, welche Probleme dabei bestehen, wie sich Perspektiven auf das Material verändern und welche Mehrwerte sich aus dieser Nutzung ergeben. Wir gehen davon aus, dass diese neuen Methoden besonderes Potential für die Nachnutzung von zunächst unbekannten Forschungsdaten haben. Die Frage, ab wann sich der damit verbundene Aufwand lohnt, ist allerdings nicht eindeutig zu beantworten. Im unten vorgestellten Fallbeispiel wurden die Transkripte von 56 mehrstündigen lebensgeschichtlichen Interviews mit Sozialwissenschaftler:innen untersucht. Unsere Fragestellung fokussiert die Veränderung wissenschaftlicher Arbeitssituationen im 20. Jahrhundert und die Rolle von Flexibilisierung in den jeweiligen Berufsbiografien. Lineares Durchlesen würde die Transkripte nur mit sehr hohem Zeitaufwand erschließen. Ab welcher Datengröße sich digitale Methoden lohnen, ist eine Einzelfallentscheidung: Wie viel Zeit ist vorhanden, um sich in die Verfahren einzudenken? Werden die digitalen Methoden im Anschluss auch für andere Untersuchungen genutzt, so dass sich die Vorarbeit besonders lohnt? Wer sich längerfristig mit digitalen Methoden beschäftigen möchte, kann deren Realisierung auch für kleinere Datenmengen umsetzen. Zudem ist wichtig, bereits zu Beginn zu bedenken, welche Arbeitsschritte für das eigene Vorgehen notwendig sind und ob Tools existieren, die diese Schritte erledigen. Diese machen viele Schritte zwar einfacher, aber auch weniger umfassend im Erkenntnisgewinn, da nicht jeder Schritt nachvollziehbar bleibt. Bestehen eigene Programmierkenntnisse oder kann mit Menschen zusammengearbeitet werden, die solche Kenntnisse haben?

Der Beitrag beschreibt unser methodisches Vorgehen Schritt für Schritt. Nach der Vorstellung unserer Fragestellung und Daten gehen wir auf die Datenerhebung und (computationelle) Vorbereitung der Datensätze ein. Dann zeigen wir die digitalen Methoden Wortfrequenzen, *Term Frequency – Inverted Document Frequency* und *Keyword in Context-Suche*, n-Gramme sowie Topic Modeling am Material. Abschließend folgt eine methodische Reflexion der Veränderungen durch diese Schritte in der Nachnutzung von Forschungsdaten und wichtige Punkte für andere Forschungsprojekte.

## Nachnutzung qualitativer Daten

Die Nachnutzung qualitativer Daten in der Sozialforschung hat sich in den letzten Jahren etabliert (Hollstein & Strübing 2018). Es bestehen bereits etliche Sekundäranalysen aus qualitativer Perspektive mit unterschiedlichen Herangehensweisen, die viele Anknüpfungspunkte für ethnografische Forschungen bieten. Nicht nur die Fragestellungen sind in den Sekundäranalysen vielfältig, auch die verwendeten Methoden unterscheiden sich teils stark voneinander und reichen von textnaher Inhaltsanalyse bis hin zu automatisierten Suchabfragen im Material.

Die Quellen müssen dabei nicht zwangsläufig aus Repositorien kommen, auch wenn dies sicherlich der einfachste Weg ist, gerade für große Datenmengen und Vergleiche. Ebenso können selbst erhobene Daten erneut untersucht oder informell geteilt werden (vgl. Heaton 2008: 509). Da sich Repositorien auch aufgrund der Anforderungen der Wissenschaftspolitik (vgl. Positionspapier der DGEKW) zunehmend etablieren, wird diese Form der Nachnutzung in näherer Zukunft jedoch wahrscheinlich zunehmen.

Ob Fragestellungen allein durch ein Korpus an erneut ausgewerteten Interviewdaten beantwortet werden können, ob also eine Sättigung eintritt, hängt von der Fragestellung ab: „The *study purpose* influences saturation" (Hennink et al. 2017: 606, Hervorh. i. Orig.). In der Regel wird eine Sekundäranalyse kulturanalytische Fragestellungen kaum abschließend beantworten können, weil Sättigung ohne eigene Kenntnis der Kontexte schwer zu erreichen ist. Sie unterstützt vielmehr den explorativen Ansatz, der in der zufälligen Entdeckung von Materialien im (digitalen) Archiv ebenso zu Überraschungsmomenten führen kann wie in der ethnografischen Erhebung (vgl. Star 2010: 605; zur Entdeckung im Archiv vgl. auch Fenske 2006). Wenn sich aus dem Datensatz die eigene Fragestellung nicht ausreichend beantworten lässt, helfen die besten (digitalen) Methoden nicht weiter, sondern dann muss auf andere Verfahren zur Generierung von Daten zurückgegriffen werden. Nicht nur selbst zu erheben, sondern die Forschungsdaten anderer nachzunutzen, ist somit eine bewusste Forschungsentscheidung, die veränderte Perspektiven nach sich zieht und genauso reflektiert werden muss wie andere Feldzugänge.

**Infobox 1: Beispiele für Sekundäranalyen von qualitativen Daten.** Erste konkrete Einblicke in bereits durchgeführte Sekundäranalysen qualitativer Daten liefern unter anderem folgende Studien, die nicht nur unterschiedliche Herangehensweisen, sondern auch theoretische und methodologische Überlegungen widerspiegeln:

- Dargentas & Le Roux 2005 diskutieren Potenziale und Herausforderungen der Sekundäranalyse bei explorativen und spezifischen Fragestellungen vor dem Hintergrund von Forschungen eines französischen Stromanbieters.
- Boris 2015 rekonstruiert die Perspektiven HIV-positiver Frauen in Kenia und untersucht, wie die Diagnose ihre Identitäten beeinflusst.
- Holubek 2017 ergründet auf der Grundlage von Grounded Theory Methodology, wieso sich Eltern für ein zweites Kind entscheiden.
- Watters et al. 2018 kombinieren in ihrer Studie zu alleinerziehenden Müttern in Kanada bestehende und eigens neu erhobene qualitative Daten.

Unabhängig davon, wie ausgefeilt digitale Methoden sind, wird qualitative und ethnografische Forschung immer den Schwerpunkt auf tiefgehende Analysen kleinerer Teile des Korpus legen und von großen zu kleinen Datenmengen wechseln (Franken 2020a). So kann beispielsweise in Anlehnung an die Grounded Theory (Glaser & Strauss 1967; Charmaz 2014) eine qualitative Analyse mit jenen Transkripten oder Teilen davon durchgeführt werden, die durch den Einsatz digitaler Methoden als relevant ermittelt werden. Das theoretische Sampling (Morse 2007; Draucker et al. 2007) wird dann durch computationelle Methoden realisiert, aber das Codieren (Holton 2007) zur Erkenntnisproduktion kann nur mit einzelnen Verfahren unterstützt werden und muss weiterhin manuell von den Forschenden durchgeführt werden. Der Forschungsprozess wird also durch den Einsatz (teil-)automatisierter Verfahren erweitert, während die genaue, manuelle Textanalyse weiterhin zentraler Bestandteil des Forschungsprozesses bleibt. Insgesamt ist ein iterativer Analyseprozess notwendig, der zwischen textnahen und textfernen Verfahren wechselt (Franken 2022), also fokussierte Analysen an einzelnen Textstellen mit Überblicksphasen über das Gesamtkorpus kombiniert. Wie menschliche und maschinelle Annotationen dabei ineinandergreifen können, haben wir an anderer Stelle bereits dargestellt (Egger et al. 2023; Franken et al. 2020). Nicht immer ist ein Annotieren im engeren Sinne notwendig. Beispielsweise ist gerade in der Nutzung digitaler Methoden das Verfassen von auch Methoden-übergreifenden Memos (Lempert 2007) besonders wichtig, um den Erkenntnisprozess beim Ausprobieren und Explorieren der Korpora festzuhalten.

**Infobox 2: Beispiele für Repositorien mit qualitativen Forschungsdaten.** Die wichtigsten Repositorien für deutsch- und englischsprachige Datensätze aus qualitativer Forschung sind aktuell:

- Qualiservice Data Sharing explizit für qualitative Forschungsdaten
- UK Data Service für quantitative und qualitative Datensätze aus Großbritannien
- Austrian Social Science Data Archive (AUSSDA) für österreichische Datensätze
- GESIS – Leibniz-Institut für Sozialwissenschaften für deutsche Daten aus quantitativer und qualitativer Sozialforschung
- Datenkatalog des Consortium of European Social Science Data Archives (CESSDA) als „Dach" verschiedener Repositorien
- eLabour Forschungsdatenzentrum mit Daten zur Arbeits- und Wirtschaftsforschung
- Online-Archiv Deutsches Gedächtnis und Zwangsarbeit Archiv für Oral History Interviews (ein Institutionen-übergreifendes Dach-Portal für Oral History steht noch aus)
- Forschungsdatenzentrum für die Hochschul- und Wissenschaftsforschung
- Europeana und Deutsche Digitale Bibliothek listen nicht nur Forschungsdaten, enthalten aber auch Daten, die sich zur Nachnutzung eignen können
- Zenodo und Google DataSet Search sammeln Forschungsdaten ohne Fachspezifika, in der Suche lohnt sich hier jedoch ebenfalls eine Abfrage
- Die European Open Science Cloud ist ebenfalls nicht fachspezifisch, erst im Entstehen begriffen und sammelt auch Publikationen etc.

## Digitale Methoden in der Analyse von Interviewtranskripten am Anwendungsbeispiel

Anhand einer Fallstudie zeigen wir nun, wie wir verschiedene computationelle Verfahren anwenden, um eine große Zahl Interviewtranskripte nachzunutzen. Der Schwerpunkt liegt dabei auf den methodischen Schritten und weniger auf der inhaltlichen Fragestellung. Für die Bearbeitung der Datensätze haben wir eigenen Computercode in der Programmiersprache Python geschrieben. Denn auch wenn für einige Schritte Tools verfügbar sind, die wir im Folgenden aufzeigen, sind durch die Setzungen der jeweiligen Entwickler:innen gewisse Grenzen gegeben, sowohl was die Möglichkeiten der Verarbeitung als auch was die Nachvollziehbarkeit der Funktionsweise angeht (vgl. Franken 2023: 211–214). Oft können bestehende Tools bereits erste gute Erkenntnisse liefern, so dass auch Forschende ohne eigene Programmierkenntnisse die beschriebenen Methoden realisieren können.

In einem explorativen Vorgehen, wie es für die EKW/EE/KA üblich ist, steht nicht nur der Output einzelner Verfahren im Mittelpunkt, sondern es sollte auch reflektiert werden, wie dies Perspektiven auf das eigene Forschungsfeld verändert. Welche Gedanken in der Auseinandersetzung mit Material und Methoden aufscheinen, sollte deshalb wie in jeder ethnografischen Erhebungsmethode mittels Feldnotizen festgehalten werden, um die Interpretationen ebenso wie die Zugänge nachvollziehbar zu halten.

### Fragestellung der Beispielstudie

In den letzten Jahrzehnten haben sich die Arbeitsregime und damit das Erwerbsleben vieler Menschen umfassend verändert. So wurde der Fordismus in vielen Bereichen durch prekäre Arbeit abgelöst (Dörre 2019; Sutter 2013). Die Phänomene Selbstständigkeit und Leiharbeit treten häufiger auf (für frühe Ergebnisse siehe Atkinson 1984; Kalleberg 2001), ebenso wie atypische Beschäftigungsverhältnisse und der Wandel von Arbeitsanforderungen hin zum unternehmerischen Selbst (Bröckling 2016). Frauen sind öfter von nicht-standardisierter Beschäftigung betroffen (Pernicka & Stadler 2006; Götz & Rau 2017). Gleichzeitig besteht Normalarbeit mit ihren unbefristeten Verträgen und finanzieller Sicherheit in einigen Arbeitsbereichen fort (Muckenhuber et al. 2018). Nichtsdestotrotz ist das Arbeitsleben stärker um die Individuen herum organisiert, weshalb sich uns nicht allein die Frage nach den Erfahrungen der Arbeitenden stellt, sondern auch die nach deren Handlungsmacht.

Der Bereich Arbeitskulturen eignet sich für Sekundäranalysen, da er bereits intensiv beforscht ist. Auch erste Sekundäranalysen bestehen (Dunkel et al. 2019). Zudem existieren viele Interviewdaten, in denen Arbeit Gesprächsgegenstand ist, obwohl das primäre Forschungsinteresse ein anderes war. Interviewtranskripte eignen sich für die gewählten Fragen besonders, weil in ihnen biografische Erzählungen vorhanden sind, die Zugang zur erlebten Alltagskultur ermöglichen. Bei vielen in den Repositorien verfügbaren Datensätzen war unsere Fragestellung nicht Thema der Primärforschung, das Material zeigt jedoch Perspektiven der Interviewpartner:innen zu diesem Bereich auf. Generell bieten sich für die computationelle Analyse Themen an, die entweder allgemein auch ohne explizite Nachfrage oft in Erzählungen angeschnitten werden oder solche, zu denen schon lange Forschung betrieben wird. Außerdem kommen Forschungen zur Vergangenheit oder Untersuchungen mit historischer Dimension, beispielsweise mit Fokus auf längerfristige Transformationsprozesse, für eine Nachnutzung besonders infrage.

Unsere inhaltliche Fragestellung haben wir bewusst allgemein gehalten. Denn zu Beginn einer Sekundäranalyse wissen wir nicht, welche Fragen mit den verfügbaren Datensätzen tatsächlich beantwortet werden können und welche nicht. Die Tatsache, dass die Daten nicht selbst erhoben wurden und somit relevantes Kontextwissen fehlt (Heaton 2008: 511) stellt eine zentrale Herausforderung bei der Nachnutzung von Datensätzen dar. Es ist wahrscheinlich, dass Datensätze nicht alle Informationen enthalten, die wir benötigen – ein Problem, das grundsätzlich bei Sekundäranalysen besteht (vgl. Heaton 2008: 511; Tate & Happ 2018). Dargentas & Le Roux (2005: § 23) differenzieren deshalb einerseits zwischen dem Kontext der Erhebung, der berücksichtigt werden muss, und andererseits dem Kontext der Textstelle, an der ein relevantes Thema festgemacht wird.

## Schritte der Datenerhebung und allgemeinen Vorbereitung

Es gibt zahlreiche Repositorien, die relevantes Quellenmaterial enthalten oder tendenziell enthalten könnten. Für die eigene Suche empfiehlt es sich, mit übergreifenden Portalen zu beginnen und dort den Links zu konkreten Datengebern zu folgen (vgl. Infobox 2). Repositorien sind dezentral organisiert und damit unübersichtlich, eine zentrale Anlaufstelle fehlt bisher. Dementsprechend sollten verschiedene Portale angesteuert werden. In den Suchabfragen hilft die Verwendung möglichst generischer und beschreibender Suchbegriffe verbunden mit der Nutzung der Filter-Funktionen der jeweiligen Suche, die oft auch eine kontrollierte Verschlagwortung umfassen. So können die interessierenden Datensätze zielgerichteter gefunden werden. Auch das eigene Wissen um das Forschungsthema und entsprechend relevante Institutionen ist hilfreich für die Suche nach potenziell nachnutzbaren Forschungsdaten. Gerade bei Studienergebnissen jüngeren und jüngsten Datums empfiehlt sich zudem eine explizite Recherche oder Anfrage nach dem erzeugten Material, denn nicht alles ist online verzeichnet. Mit unserem Forschungsinteresse haben wir gezielt nach passenden Datensätzen gesucht und wurden auch an unterschiedlicher Stelle fündig. Wir zeigen die methodischen Schritte im Folgenden anhand des Datensatzes „Pioneers of Social Research" (PoSR, Thompson 2019) aus Großbritannien. Den Datensatz haben wir im Zuge einer thematischen Recherche im Repositorium UK Data Service gefunden, wo er öffentlich zugänglich ist.

Im zweiten Schritt sollte mittels Sichtung der Metadaten sowie vorhandenem Kontextwissen entschieden werden, ob die Datensätze potentiell für die eigene Fragestellung relevant sein könnten. Nicht immer ist es dafür notwendig, bereits die kompletten Datensätze einzusehen. Qualiservice bietet beispielsweise zu jedem Datensatz einen Studienreport, der zentrale Informationen enthält (Heuer et al. 2020). Zusätzliche Informationen können durch das Lesen der aus diesem Kontext publizierten Ergebnisse und im Einzelfall auch durch Kontakt zu den Autor:innen der Originalstudie (vgl. z.B. Dargentas & Le Roux 2005: § 34) gewonnen werden. Da gerade qualitative Daten meist erst nach einer Registrierung zugänglich sind, spielt dieser Schritt eine wichtige Rolle, um unnötigen Aufwand auf beiden Seiten zu vermeiden und nur die Datensätze anzufragen, die potentiell tatsächlich interessant sind. Der Kontakt zu den Repositorien muss aber auch für kleinere Arbeiten, etwa im Studium, nicht gescheut werden.

Für den von uns verwendeten Datensatz steht nur ein kurzer Textabschnitt im Repositorium als Beschreibung zur Verfügung, ein Report existiert nicht. Schon mit der Kurzbeschreibung hatten wir jedoch einen Eindruck einer Passung in Bezug auf unser Forschungsinte-

resse: Zwischen 1996 und 2018 wurden in dem Projekt „Pioneers of Social Research“ 56 Interviews geführt, in denen Sozialwissenschaftler:innen mit unterschiedlichen thematischen Schwerpunkten über ihre Forschung, aber auch Hintergründe wie Familie und Bildungsverlauf Auskunft geben. Das Material wurde hinsichtlich der Wissenschaftsgeschichte der britischen Sozialwissenschaften analysiert (Thompson et al. 2021). Die vielfältigen Wandlungsprozesse von Arbeitskulturen lassen sich auch an der akademischen Arbeit aufzeigen, die in diesen Interviews detailliert beschrieben und reflektiert wird. Uns interessiert, anders als die Primärforschenden, besonders die Perspektive der Akteur:innen auf ihre Arbeitsbedingungen.

## Download der Daten aus Repositorien

Sobald Interviewtranskripte identifiziert sind, die für die eigene Fragestellung interessant sein könnten, geht es darum, die Daten von den entsprechenden Repositorien auf ein eigenes Speichermedium zu übertragen. Der Zugang kann dabei je nach Repositorium sehr unterschiedlich ausfallen – von frei zugänglichen Korpora bis hin zu registrierungspflichtigen Zugängen, die mit einer Genehmigung der eigenen geplanten Nachnutzung einhergehen. Der von uns genutzte PoSR-Datensatz ist unter der Creative Commons Lizenz „Attribution 4.0 International (CC BY 4.0)“ veröffentlicht, die es erlaubt, den Datensatz uneingeschränkt zu teilen und zu verändern, insofern immer auf den ursprüngliche Datensatz verwiesen wird und Veränderungen klar vermerkt werden. Unsere Auswahl war dabei allerdings weniger von der Zugänglichkeit geprägt, als durch die geschlossene Berufsgruppe, die hier befragt wurde, was thematische Befunde wahrscheinlich macht.

Wie auch bei primär erhobenen Interviews werden die heruntergeladenen Interviewtranskripte auf einer sicheren, eigenen Dateninfrastruktur gespeichert. Dies bedeutet in der Regel die Nutzung von universitätseigenen Servern – falls solche nicht vorhanden sind, stellen auch lokale, ggf. passwortgeschützte Ordner auf dem eigenen Computer eine Lösung dar (vgl. den Beitrag von Imeri et al. in diesem Band). Für die Weiterverwendung von größeren Datensätzen aus Forschungsdaten anderer ist es zudem besonders wichtig, sich eine strukturierte Datenablage zu überlegen und die Daten entsprechend einheitlich benannt in Ordnern sortiert abzulegen. Zudem ist es notwendig, bei der Bearbeitung verschiedene Versionen der Daten abzulegen, um bei Fehlern oder neuen Erkenntnissen auf die vorherigen Versionen zurückgreifen zu können.

## Forschungsethik prüfen

Parallel zur praktischen Zugänglichkeit sowie der grundsätzlichen Erschließung des Kontextes der Datensätze gehört zur Vorbereitung einer Sekundäranalyse qualitativer Daten auch die Prüfung forschungsethischer Dimensionen (Medjedović 2020). Zu den ethischen Standards der qualitativen Forschung, die genauso auch für Sekundäranalysen gelten, gehören als zentrale Prinzipien die informationelle Selbstbestimmung der untersuchten Subjekte und die Schadensminimierung, welche durch eine informierte Einwilligung der Subjekte sowie eine zumeist damit verbundene Pseudonymisierung persönlicher Daten gesichert werden sollen (von Unger 2014). Für Sekundäranalysen bedeutet dies, dass die informierte Zustimmung der Teilnehmenden, die auf die Möglichkeit der Weiterverwendung der Daten durch andere hinweist, innerhalb der ursprünglichen Studie eingeholt werden muss – eine Situation, die dazu führt, dass Interviewdaten für Dritte oft nicht verfügbar sind (Hollstein

& Strübing 2018; Tate & Happ 2018; Weller & Kinder-Kurlanda 2017). Werden Interviewtranskripte in institutionellen Repositorien zur Verfügung gestellt, kann grundsätzlich davon ausgegangen werden, dass die verfügbaren Datensätze diesen Standards entsprechen – ansonsten würden sie nicht in diesen Zusammenhängen geteilt. Dennoch sollten die forschungsethischen Dimensionen für die weitere Verwendung geprüft werden um zu hinterfragen, ob das eigene Forschungsvorhaben ethisch unbedenklich für die beteiligten Personen ist (Edmond et al. 2018; Imeri 2018; von Unger 2018).

Bei dem von uns verwendeten PoSR-Datensatz haben alle interviewten Sozialwissenschaftler:innen der öffentlich zugänglichen Archivierung und Nachnutzung ihrer Interviews unter Vollnamen beim UK Data Service zugestimmt. Im Datensatz steht ein Beispiel für eine Einverständniserklärung zur Verfügung, auf dessen Grundlage die Interviews veröffentlicht wurden. Zudem ist mit der genutzten Creative Commons Lizenz die Nachnutzung und Änderung mit Zitation und Änderungshinweisen uneingeschränkt erlaubt. Die Interviewtranskripte dieses Datensatzes können also, in Kombination mit einer eigenen forschungsethisch sensibilisierten Praxis, ohne weitere forschungsethische und datenschutzrechtliche Bedenken für eine Sekundäranalyse genutzt werden. Bei anderen Datensätzen muss hingegen in der weiteren Analyse beachtet werden, dass von der Sekundäranalyse aus geteilte und veröffentlichte Interviewausschnitte konsequent pseudonymisiert sind (Thomson et al. 2005). Unsere Erfahrungen zeigen, dass in Repositorien zugängliche Datensätze nicht zwingend in bereits forschungsethisch unbedenklicher Form vorliegen müssen und eine eigene ethische Abwägung notwendig ist.

## Vorverarbeitung der Daten für digitale Analyseschritte

Sind die Daten auf eigenen Geräten gespeichert und forschungsethische Standards geprüft bzw. reflektiert, kann es an die engere Vorbereitung der Daten gehen. Text ist für computationelle Verfahren zunächst unstrukturierte Information. Zur Vorbereitung für den Einsatz digitaler Methoden müssen alle Texte einer Vorverarbeitung, einem sogenannten Preprocessing, unterzogen werden, durch das die Unstrukturiertheit in gewissem Maße reduziert wird. Um allerdings relevante Textstellen später lesen und tiefergehend analysieren zu können, sollte der Datensatz auch in seiner menschenlesbaren Form, also in unserem Fall die unveränderten Transkripte, gespeichert werden. Schon hier sollte eine einheitliche Benennung der Datensätze vorgenommen werden. Wir haben einen Ordner ‚Datensatz original' und verschiedene Ordner mit den bearbeiteten Versionen der Dateien angelegt. Den Programmiercode für die digitalen Auswertungsverfahren haben wir über einen GitLab-Server der LMU München verwaltet. Zur Nachvollziehbarkeit ist der Code auf GitHub veröffentlicht.

Interviewtranskripte sind singulär und werden von den einzelnen Forschenden, die sie produzieren, strukturiert. Dementsprechend unterschiedlich sehen sie aus, auch wenn in formaler Hinsicht ihre Frage-Antwort-Struktur vergleichbar ist. Zusätzlich handelt es sich bei Interviewtranskripten um ein Surrogat gesprochener Sprache mit inkonsistenten Sprachstrukturen und Abweichungen von standardisierter Schriftsprache. Auch wenn die computationelle Verarbeitung und Auswertung von Texten in den letzten Jahren einige Fortschritte gemacht hat, entspricht die Sprache in Interviewtranskripten nicht der Sprache, die in schriftlicher Kommunikation üblich ist, etwa in Wikipedia-Artikeln. Allerdings sind Daten in dieser standardisierten Schriftsprache oft Grundlage der Entwicklung von Methoden des

**Infobox 3: Bearbeitung der Datensätze mit digitalen Verfahren und eigenes Programmieren.** Um Textdaten computationell zu analysieren, müssen diese zunächst aufbereitet werden. Dies erfolgt im Zuge eines Preprocessing, das in der Regel mit einem selbst zusammengestellten Skript durchgeführt wird. Auch wenn diese Schritte auf den ersten Blick sehr technisch und komplex erscheinen mögen, so sind sie doch in der Regel mit Aneignung von rudimentären Programmierkenntnissen oder in Zusammenarbeit mit Programmierer:innen einfach zu realisieren. Für viele beschriebenen Funktionalitäten stehen Programmierbibliotheken zur Verfügung, die man lediglich richtig in den eigenen Code einbinden muss.

Die Hürde zum Erlernen grundlegender Programmierkenntnisse ist niedriger, als oft angenommen wird. Für die in den Geistes-, Sozial- und Kulturwissenschaften verbreitete Programmiersprache Python gibt es zahlreiche Online-Tutorials, mit deren Hilfe Grundlagen schnell erlernt sind. Etwa beim Portal Programming Historian bestehen Einführungskurse, die auf die Bedürfnisse der Textanalyse zugeschnitten sind. Es lohnt sich außerdem, sich in Kleingruppen zusammenzuschließen und die Hilfe der online bestehenden Communities zu nutzen.

Für die Vorverarbeitung von natürlichem Text im Rahmen des Natural Language Processing sind in Python die Libraries spaCy und NLTK besonders weit verbreitet. Für Analyseverfahren, wie sie unten vorgestellt werden, sind außerdem die Programmierbibliotheken Gensim und MALLET zielführend. Im Ausprobieren ist es hilfreich, auf die bestehenden Skripte von anderen zurückzugreifen und diese für den eigenen Datensatz anzupassen. So bestehen mittlerweile Jupyter Notebooks, als Kombination von Computercode und Notizen in einer klickbaren Ansicht, die entsprechende Schritte zusammenstellen und mit Anleitungen versehen sind, wie etwa für die Arbeit von Historiker:innen, die gut übertragbar sind.

Wer nach einer Alternative zum eigenen Programmieren sucht, wird in den verschiedenen bestehenden Listen von Tools fündig. Besonders zu empfehlen ist für die Vorverarbeitung das Tool Nopaque. Mit diesem können die meisten Schritte auch ohne eigenes Programmieren umgesetzt werden, das Tool baut auf der Bibliothek spaCy auf. Zudem bestehen kostenpflichtige Angebote, von denen jedoch abzuraten ist. Mit der Nutzung von bestehenden Tools geht immer eine geringere Nachvollziehbarkeit der Prozesse des Tools einher. Wissenschaftliche Arbeiten sind zwar leichter reproduzierbar, wenn öffentliche und gut dokumentierte Tools genutzt werden. Jedoch ist bei kommerziellen Tools ohne Zugriff auf den Code zu beachten, dass durch die Entwickler:innen bestimmte Entscheidungen getroffen wurden, die undurchsichtig sind und die Tools somit an Nachvollziehbarkeit einbüßen.

In der Analyse hilft das Tool AntConc, mit dem Arbeitsschritte fundiert und nachvollziehbar umgesetzt werden können, auch ohne dass eigene Programmierung notwendig ist. Ein Datensatz kann hier als Sammlung von TXT Dateien mit wenigen Klicks eingespielt werden. Außerdem ist Voyant empfehlenswert, allerdings sind die hier getätigten Operationen weniger nachvollziehbar und komplizierter zu speichern. Auch kommerzielle QDA-Software wie etwa MaxQDA bietet mittlerweile einige einfache Text Mining-Verfahren, wobei die Verfahren jedoch wenig transparent bleiben. Ein umfangreiches Glossar mit Tools, Überblickslisten und Einschätzungen liegt bereits vor (Franken 2023).

Text Mining. Daher sind computationelle Verarbeitungen von Interviewtranskripten, die nicht an das spezifische Korpus angepasst sind, problematisch, wann immer sie über einzelne Begriffe hinaus gehen. Dennoch wurden digitale Methoden bereits in einzelnen Studien erfolgreich angewendet (Sherin 2013; Karlgren et al. 2020; Shrader et al. 2021). Es ist also wichtig, dass bei der Anwendung computationeller Verfahren gewisse Besonderheiten berücksichtigt werden. Ein für die verschriftlichte gesprochene Sprache angepasstes Preprocessing ist insbesondere für solche Methoden relevant, die nicht nur einzelne Wörter, sondern Wortgruppen oder Satzstrukturen verarbeiten.

In einem ersten Schritt des Preprocessing von Interviewtranskripten geht es um die Entfernung von inhaltlich nicht relevantem Text aus den Dokumenten, wie Infos zur interviewten Person, Zeitstempel im Textverlauf oder Sprechendenbezeichnungen: Diese für Interviewtranskripte typischen Textteile sind für unsere digitale Analyse nicht interessant, da sie nicht zum eigentlichen Gespräch gehören. Je nach Fragestellung und gewählter Methode können solche Metadaten allerdings durchaus relevant sein und dementsprechend verarbeitet werden (für einen tieferen Einstieg etwa zur Verwendung von *Structural Topic Modeling* vgl. Tonidandel et. al. 2022).

Zum Preprocessing für Natural Language Processing im engeren Sinne gehört die Tokenisierung, d.h. die Zerlegung des Textes in einzelne Wörter (Jentsch & Porada 2020), und die Lemmatisierung, welche die Wörter anhand ihres Vorkommens im Dokument auf die Grundform abstimmt, z.B. erste Person Singular und Präsens bei Verben (HaCohen-Kerner et al. 2020). So können Wörter unabhängig von ihrer Verlaufsform gefunden werden, die Treffermenge von Suchbegriffen erhöht sich durch die Lemmatisierung signifikant. Mit der Entfernung von Stoppwörtern, also Wörtern ohne inhaltsanalytischen Wert, werden die darauffolgenden Analyseschritte von sehr häufigen Wörtern (wie etwa „und“, „der / die / das“) entlastet, die für Ergebnisse inhaltlich nicht relevant sind.

Für die konkrete Umsetzung haben wir selbst programmiert, da das Preprocessing jeweils auf die Spezifitäten der Korpora zugeschnitten werden muss. Weiterhin verlangt die jeweilige Analysemethode meist unterschiedliche Grade und Arten der Vorverarbeitung, sodass diese in den Einzelheiten auf die Analysemethode zugeschnitten werden sollte. Unser Preprocessing am PoSR-Datensatz war zugeschnitten auf unser Topic Modeling. In Abbildung 1 ist ein beispielhafter Ausschnitt aus einem Interview dargestellt. In den Abbildungen 2 und 3 wird derselbe Interviewausschnitt so dargestellt, wie er nach der computationellen Verarbeitung durch die nachfolgenden Preprocessing-Schritte aussieht.

1. Umwandlung des Dateiformats: Da die Interviewtranskripte bei diesem Datensatz nur im RTF-Dateiformat herunterladbar sind, wurden die Transkripte in das einfacher einzulesende Format TXT konvertiert. Es zeigte sich außerdem, dass die Dateien in ihrer heruntergeladenen Form nicht ganz einheitlich benannt sind, weshalb die Dateien einheitlich und nachvollziehbar umbenannt wurden.
2. Entfernen von Metadaten um das Transkript: Nach diesen Vorarbeiten auf Dateiformatsebene wurden vor Beginn und nach Ende des eigentlichen Interviews Metadaten wie etwa Informationen zu interviewter Person, Ort und Zeit entfernt. Dabei mussten wir uneinheitliche Metadaten, wie unterschiedliche Formate des Interviewdatums oder verschiedene Bezeichnungen für das Ende des Interviews, regelbasiert erkennen, um sie automatisch aufzulösen. Theoretisch könnte man diesen Schritt auch in Fleißarbeit manuell erledigen, dies ist jedoch bei größeren Datensätzen zeitlich kaum realisierbar.

**Paul:** And so you spend quite a lot of time on this political activity?

**Tirril:** Not now.

**Paul:** No, but you did?

**Tirril:** Yes. Yes. Well ...

**Paul:** I'm just wondering how you had time to at all, really ... with children and psychotherapy.

**Tirril:** Well, I think the answer is, not that much time.

**Tirril:** I mean, right at the very beginning, before we had children, when we were still students, you know, Nigel getting his doctorate, and we ... because between me finishing my second undergraduate degree, and Nigel finishing his doctorate, I didn't get another job, because we were going to go and work abroad, so I did silly things like work in toy shops and so on. So in those years, we've been a sort of total of ... sort of a couple of years while we were students in London. We had time, you know. And later, although I did go to branch meetings, and I never sold a paper at six o'clock, because I had to give the kids breakfast, you know, but I did go and sell the paper. And the children got quite good at going round Council flats in Camden and, "Hello! We've brought you the Socialist Worker paper this week, instead of mummy!" You know, they'd be on one balcony, and I'd be on the one above! (LAUGHS)

**Paul:** And do you think that's had any influence the perspectives you bring to the research?

**Tirril:** Probably, yes. You know, probably, if I had been a conservative, I wouldn't have been happy when we found that there was a higher proportion of women with depression in the working class. You know, that probably wouldn't have suited me, so I might have ... I don't know, I might have changed my politics to meet up with my findings. But I might have moved on to change my work to stick in coherence with my politics! Who can say!

**Paul:** Now, I think that's probably all about the two university experiences. So shall we go on to you going abroad?

Abbildung 1: Textabschnitt aus Interview 18 in seiner menschenlesbaren Form vor dem Preprocessing.
*Quelle: Eigene Darstellung.*

3. Tokenisierung: Die bereinigten Interviewtranskripte wurden dann mit Hilfe der spaCy Library tokenisiert, d.h. die einzelnen Wörter wurden im Text der Transkripte ‚für die Maschine' identifiziert.
4. Segmentierung: Weiterhin wurde der Text zunächst anhand der Sprechendenwechsel mit unserem Skript in Abschnitte unterteilt, sodass für eine zusammenhängende Aussage einer Person ein Segment vorlag. An dieser Stelle konnte die eine vorhandene Struktur in unserem Testdatensatz ausgenutzt werden: Eine Leerzeile zeigte einen zusammenhängenden Redebeitrag an und vor einem zusammenhängenden Redebeitrag stand (meist) der Name des Sprechenden. Somit konnte regelbasiert ein Sprechendenwechsel festgestellt werden.
5. Entfernen von Metadaten im Interviewtranskript: Wir haben außerdem Segmente innerhalb der Transkripte identifiziert und entfernt, die inhaltlich nicht zum Interview gehören. Dies waren Meta-Anmerkungen eingeleitet mit Begriffen wie „INTERRUPTION" oder „TELEPHONE", die Unterbrechungen des normalen Interviewablaufs beschreiben.
6. Identifizieren und Entfernen der Sprechendenbezeichnungen: Darauf aufbauend wurden die Sprechendenbezeichnungen in den Transkripten entfernt. Dies geschah durch die Entfernung des ersten Tokens pro Segment, denn dieses ist im PoSR-Datensatz stets der Name, sofern es sich um einen Redebeitrag handelte. So blieben als Text nur die Aussagen der Interviewenden und Interviewten im Transkript. Zudem

wurde regelbasiert identifiziert, welche Person befragt hat und welche Person befragt wurde. Diese Informationen wurden für die einzelnen Segmente separat gespeichert, sodass sie die weitere Textverarbeitung zwar nicht stören, aber nicht verloren gehen.

7. Lemmatisierung: Darauf folgte die Lemmatisierung, also die Rückführung von Wörtern auf ihre Grundformen, ebenfalls mittels spaCy Library.
8. Stoppwortentfernung: Weiterhin wurden die Interviews mit Hilfe der in spaCy enthaltenen Sprachmodelle von gewöhnlich inhaltlich unbedeutenden und häufig auftretenden Wörtern wie „therefore“, „thus“ oder „I“ bereinigt. Die dabei zum Einsatz gekommene und zunächst auf spaCy basierende Stoppwortidentifizierung wurde iterativ erweitert, denn Stoppwortlisten sollten an die jeweilige Textsorte, bei Transkripten also an gesprochene Sprache, angepasst werden. Für die Erstellung und Anpassung einer Stoppwortliste ist Vorwissen zum Inhalt des Textes notwendig – wir empfehlen, den Umfang der Stoppwortliste je nach Methode in den nächsten Schritten auch auszuprobieren, da sie die Ergebnisse deutlich beeinflussen kann. In diesem Schritt wurden außerdem die Satzzeichen entfernt.
9. Verbesserung der Tokenisierung mit n-Grammen: Zur Qualitätserhöhung der Ergebnisse des Topic Modeling wurden häufig vorkommende Bi- und Trigramms in die Tokenisierung miteinbezogen, wie es auch von Heiberger und Galvez (2021) empfohlen wird. Beispielsweise wurden Vorkommnisse von „social mobility" zu einem

| speaker | "interviewer" |
|---|---|
| list_of_words | ["spend", "time", political", "activity"] |

| speaker | "interviewee" |
|---|---|
| list_of_words | [] |

| speaker | "interviewer" |
|---|---|
| list_of_words | [] |

| speaker | "interviewee" |
|---|---|
| list_of_words | [] |

| speaker | "interviewer" |
|---|---|
| list_of_words | ["wonder", "time", "child", "psychotherapy"] |

| speaker | "interviewee" |
|---|---|
| list_of_words | ["answer", "time"] |

| speaker | "interviewee" |
|---|---|
| list_of_words | ["beginning", "child", student", nigel", "doctorate", "finish", second", "undergraduate", "degree", "nigel", "finish", doctorate", job", "work", "abroad", silly", "work", "toy", "shop", "total", "couple", "student", "london", "time", "branch", "meeting", "sell", "paper", o", "clock", "kid", "breakfast", "sell", "paper", "child", "good", "round", "council", "flat", "camden", "hello", "bring", "socialist", "worker", paper", "week", "instead", "mummy", "balcony"] |

| speaker | "interviewer" |
|---|---|
| list_of_words | ["influence", "perspective", "bring", "research"] |

| speaker | "interviewee" |
|---|---|
| list_of_words | ["probably", probably", "conservative", "wouldn", "happy", "high", "proportion", "woman", "depression", "working_class", "probably", wouldn", "suit", "change", "politic", "meet", finding", "move", "change", "work", stick", "coherence", politic"] |

| speaker | "interviewer" |
|---|---|
| list_of_words | ["probably", "university", "experience", "shall", "abroad"] |

Abbildung 2: Textabschnitt aus Abbildung 1 nach Preprocessing-Schritt 9, dargestellt als Tabelle. Jede Tabelle stellt ein Segment dar. Bei leeren Segmenten wurden alle Wörter durch Schritt 8 entfernt.
*Quelle: Eigene Darstellung.*

Token zusammengefasst, um „social mobility“ von einzelnen Verwendungen von „social“ und „mobility“ zu unterscheiden.

10. Zusammenlegen von Segmenten: Im letzten Schritt wurde die Aufteilung der jeweiligen Interviews in Segmente angepasst, um die Menge an Text innerhalb der Segmente anzugleichen, aber gleichzeitig die Information über die zusammenhängende Äußerung durch den Sprechendenwechsel nicht zu verlieren. Diese Angleichung ist notwendig für digitale Methoden in denen ein Segment in einem *Bag of Word*-Ansatz nur durch die enthaltenen Wörter repräsentiert wird und diese Segmente gleichbedeutend in das Verfahren einfließen. Unser Topic Modeling würde bei sehr heterogenen Textlängen der Segmente eine verzerrte Datengrundlage erhalten. Deshalb wurden die vorhandenen Segmente so zusammengelegt, dass eine Mindestzahl an Wörtern pro Segment erfüllt war und die neu entstehenden Segmente immer mit einer Äußerung des Befragenden, also meist mit einer Frage, beginnen.

Über alle Schritte hinweg gingen wir iterativ vor, indem wir immer wieder die Ergebnisse überprüften und auf der Basis der Evaluation Verbesserungen am Code vornahmen. Auch wenn das Preprocessing der eigentlichen Analyse vorgelagert werden sollte, ist es nicht vollkommen getrennt. Gewisse Variablen können während der weiteren Analyse angepasst werden, etwa die weitere Aufnahme von unbedeutenden und die Analyse störenden Wörtern in die Stoppwortliste oder eine Veränderung der Tokenisierung aufgrund der Ergebnisse einer Analyse der n-Gramme, wie wir sie im Folgenden vorstellen. Das Preprocessing sollte immer fallspezifisch auf die jeweiligen Spezifitäten des Korpus angepasst werden. Selbst bei einer kombinierten Analyse verschiedener Datensätze in einem Projekt muss dieser Schritt getrennt auf der Ebene der Korpora stattfinden.

| | |
|---|---|
| list_of_words | [spend", "time", political", "activity", “wonder", time", "child", "psychotherapy", answer", "time", "beginning", "child", student", nigel", "doctorate", "finish", second", "undergraduate", "degree", "nigel", "finish", doctorate", job", "work", "abroad", silly", "work", "toy", "shop", "total", "couple", "student", "london", "time", "branch", "meeting", "sell", "paper", “o", "clock", “kid", "breakfast", “sell", "paper", "child", "good", "round", "council", "flat", "camden", "hello", "bring", "socialist", "worker", paper", "week", "instead", "mummy", "balcony", "influence", "perspective", "bring", "research", "probably", probably", "conservative", "wouldn", "happy", "high", "proportion", "woman", "depression", "working_class", "probably", wouldn", "suit", "change", "politic", "meet", finding", “move", “change", "work", stick", "coherence", politic"] |

| | |
|---|---|
| list_of_words | ["probably", "university", "experience", "shall", "abroad",<br>[...]<br>] |

Abbildung 3: Textabschnitt aus Abbildung 1 und 2 nach Preprocessing-Schritt 10, dargestellt als Tabelle. Jede Tabelle stellt ein Segment dar, das zweite Segment ist gekürzt.
*Quelle: Eigene Darstellung.*

## Anwendung digitaler Methoden

Die vorbereiteten Datensätze können mit einer Reihe etablierter Text-Mining-Verfahren analysiert werden. Koch und Franken (2020) weisen darauf hin, dass sich die qualitative Analyse mit den bereits verfügbaren computergestützten Methoden auf die Filterung, also verfahrensgeleitete Auswahl, konzentrieren sollte, die bei großen, menschlich nicht überschaubaren Datenmengen einen schnelleren Zugriff auf die relevantesten Datensätze und Textabschnitte ermöglicht. In unserer Analyse der Interviewtranskripte haben wir die aus unserer Sicht interessantesten Ansätze aus den interdisziplinär bereits etablierten Methoden verwendet, um sie einzeln und in Kombination auf ihr Potenzial zur Filterung zu untersuchen: Wortfrequenzen zeigen die vorherrschenden Repräsentationen in Texten, n-Gramme die statistische Wahrscheinlichkeit von Wortkombinationen. Beim Topic Modeling werden Themen auf der Basis von gemeinsam im Text stehenden (ko-okkurrierenden) Wörtern vermutet. Im Zusammenspiel dieser unterschiedlichen Methoden kann ein erstes Verständnis und eine inhaltliche Zusammenfassung großer Datensätze erreicht werden, indem die Aufmerksamkeit auf Muster und Anomalien gelenkt wird. Wenn Interviewtranskripte mit vielen Metadaten hinterlegt sind oder in Segmente unterteilt werden, die Anzahl von Datensätzen also wächst, könnten die Elemente auch in einer relationalen Datenbank untersucht werden (vgl. Beitrag Vepřek in diesem Band). Das Filtern und Vergleichen mit computationellen Methoden erweitert damit bestehende Forschungsprozesse (Franken 2022) und ermöglicht eine qualitative, durch gezielte Auswahl geleitete Analyse.

**Infobox 4: Digitale Methoden.** Im Folgenden liefern wir einen Überblick über die Methoden, deren Anwendung wir anschließend am praktischen Beispiel demonstrieren (zur Vertiefung vgl. Franken 2023).

- Frequenzanalysen erstellen einfache Darstellungen und Klassifizierungen der Anzahl der in einem Dokument oder einer Reihe von Dokumenten verwendeten Wörter. Während Worthäufigkeiten das Vorkommen von Wörtern in Relation zur Gesamtzahl im Korpus bestimmen (Bag-of-Words-Ansatz), ermitteln Dokumenthäufigkeiten, wie viele Dokumente einen bestimmten Begriff enthalten (Ignatow & Mihalcea 2018; Lemke & Wiedemann 2016).
- Beide Maße werden im tf-idf-Score (*term frequency – inverted document frequency*) kombiniert, der Wörter identifiziert, die für ein oder mehrere Dokumente in einer größeren Sammlung besonders sind (Jurafsky & Martin 2021).
- Bei der Suche nach n-Grammen wird die statistische Wahrscheinlichkeit des gemeinsamen Auftretens von Wörtern oder Buchstaben innerhalb eines bestimmten Fensters untersucht. Die Variable ‚n' stellt dabei die Anzahl von Wörtern dar, die dieses Fenster bilden (Evert 2009; Bubenhofer 2017).
- Für Topic Modeling wird eine Form des maschinellen *Soft Clustering* realisiert, bei dem eine bestimmte Anzahl und Verteilung von Themen in einem Text auf der Basis von Wort-Kookkurrenz angenommen wird (Blei 2012).

Weitere digitale Methoden, wie etwa automatische Annotationen oder Netzwerkanalysen, haben wir für unsere Fragestellung als wenig relevant ausgeschlossen. Dies bedeutet jedoch nicht, dass sie in anderen Forschungen der EKW/EE/KA nicht durchaus relevant werden können.

Die von uns genutzten digitalen Methoden stellen etablierte Vorgehensweisen in den Digital Humanities und Computational Social Sciences dar. Wie beim Preprocessing haben wir sie einzeln auf unsere Datensätze angewandt, da das Vorgehen immer wieder individuell zugeschnitten werden muss. Außerdem haben wir in der Analyse erst einzelne Methoden umgesetzt und die Ergebnisse dann in Kombination betrachtet – sowohl über die Datensätze als auch über die Methoden hinweg. Dabei ist es besonders zielführend, die einzelnen Schritte, besonders aber die eigenen Forschungsentscheidungen und Erkenntnisse in der Exploration, in Form von Memos aufzuschreiben, um flüchtige Ideen und Interpretationsansätze festzuhalten. Für eigene Forschungsvorhaben können je nach Forschungsinteresse und Struktur der vorliegenden Korpora nur eine Auswahl der vorgestellten Methoden oder auch eine andere Reihenfolge ihrer Anwendung sinnvoll sein. Die Erkenntnisse aus unserer Anwendung der digitalen Methoden sowie deren Potenzial für eine Erweiterung qualitativer Forschungsprozesse stellen wir nun entlang der einzelnen Methoden dar, wobei wir zunächst mit den Methoden beginnen, deren Ergebnisse möglicherweise eine weitere Runde Preprocessing bedingen.

## Wortfrequenzen

Als erstes betrachteten wir die Wortfrequenzen des Datensatzes. Anhand entsprechender Listen können wichtige Wörter im Gesamtkorpus und für einzelne Interviews erschlossen werden. Begriffe werden dabei nach Häufigkeit sortiert ausgegeben. Da auf Ebene von Wortfrequenzen die genaueren Bedeutungen der Wörter jedoch nur über die allgemeine Kenntnis des Kontextes erschließbar sind, stellen Erkenntnisse aus Wortfrequenzanalysen nicht erste Analyseergebnisse, sondern einen Einstieg in große Textmengen dar. Mit einfachen Mitteln zeigen sie uns, ob sich der aus den dominanten Wörtern schließbare Inhalt des Korpus mit den Erwartungen deckt, die wir aufgrund der Kontextinformationen zum Datensatz haben – das Datenkorpus also für eine weitere Untersuchung mit Blick auf die eigene Fragestellung interessant bleibt. Ist dies der Fall, zeigen sich mit den verschiedenen Wortfrequenzlisten durch den eigenen analytischen Blick möglicherweise erste Clusterungen als vorläufige Analysekategorien, die weiterverfolgt, validiert oder transformiert werden können. Die in den Frequenzlisten erkannten Muster können als erste thesenhafte Findeheuristiken (Adelmann et al. 2019) dienen, als überblickender und strukturierender Blick im Prozess der Filterung der großen Datenmengen auf relevante Interviews und Interviewausschnitte. Da die Wörter der Wortfrequenzanalysen sehr oft in den entsprechenden Texten vorkommen, macht es aber wenig Sinn, diese selbst als Filter zu nutzen, da mit einer Keyword-Suche sehr viele Textstellen angezeigt werden würden, die qualitativ wiederum nicht zu überblicken sind.

In unserem Anwendungsfall haben wir mit einem einfachen Counter und einer Liste von allen Wörtern aus dem Korpus (die wir per Tokenisierung im Preprocessing erstellt haben) die absolute Worthäufigkeit berechnet. Durch unser Skript lassen wir uns Listen ausgeben, welche die häufigsten Wörter für das Gesamtkorpus bzw. für einzelne Interviews ausweisen. Auch Tools wie MaxQDA oder AntConc geben solche Listen aus, unsere Liste ist allerdings durch die eigenen, angepassten Stoppwörter bereinigt.

Die hundert häufigsten Wörter im PoSR-Datensatz sind in Abbildung 4 dargestellt. Die Wörter „people“ (8.991 mal), „work“ (7.095 mal), „time“ (6.470 mal), „good“ (3477 mal), „write“ (3307 mal) und „course“ (3140 mal) stellen also, unter Ausschluss der Stoppwörter im Preprocessing, die im gesamten Korpus sechs häufigsten Wörter dar. In Kombination mit

„('people', 8991), ('work', 7095), ('time', 6470), ('good', 3477), ('write', 3307), ('course', 3140), ('book', 3091), ('talk', 2991), ('research', 2929), ('school', 2882), ('social', 2697), ('study', 2629), ('different', 2281), ('family', 2270), ('woman', 2235), ('call', 2195), ('university', 2070), ('interview', 2003), ('read', 1971), ('child', 1944), ('idea', 1891), ('important', 1808), ('man', 1778), ('day', 1749), ('interesting', 1732), ('great', 1621), ('feel', 1611), ('student', 1608), ('long', 1607), ('early', 1549), ('term', 1541), ('mother', 1539), ('give', 1508), ('interested', 1503), ('group', 1499), ('point', 1488), ('sense', 1481), ('friend', 1469), ('class', 1462), ('new', 1455), ('end', 1436), ('sociology', 1424), ('father', 1418), ('teach', 1416), ('question', 1406), ('job', 1389), ('person', 1365), ('big', 1365), ('old', 1354), ('young', 1323), ('history', 1284), ('little', 1274), […]“

Familie Arbeit

Abbildung 4: Gekürzte Liste der 100 häufigsten Wörter aufgrund der globalen Wortfrequenzanalyse des PoSR-Datensatzes, Einfärbungen zeigen mögliche thematische Cluster.
*Quelle: Eigene Darstellung.*

unserem Kontextwissen können anhand dieser Schlagworte erste, vorläufige Schlüsse gezogen werden. „People“ als häufigstes Wort verweist vor dem Hintergrund der Interviews mit Sozialwissenschaftler:innen auf den in den Interviews besprochenen Forschungsbereich der Interviewten, in dem Menschen eine zentrale Stellung einnehmen. Das zweithäufigste Wort entstand wahrscheinlich durch die zentrale Fragestellung der Interviews nach „research work“ (Thompson 2019). Es wird also – wenig überraschend, aber für unsere Fragestellung und somit die weitere Beschäftigung mit dem Korpus grundlegend – viel über Arbeit gesprochen.

Die Häufigkeit von „time“ und „good“ stammt möglicherweise von häufig verwendeten Ausdrücken im gesprochenen Englisch wie „at this time“ oder schlicht „good“ als motivierende Erwiderung in einem Gespräch. Wenn Wörter keine inhaltliche Bedeutung für die Fragestellung haben, kann hier über eine Anpassung der Stoppwortliste nachgedacht werden, Wortfrequenzanalysen also zur Optimierung des Preprocessing genutzt werden. Jedoch könnten in unserem Falle die Wörter „time“ und „good“ in gewissen Kontexten gerade auch mit Inhalten verknüpft sein, die für unsere Fragestellung des Wandels von Arbeitskulturen interessant wären, also keine Stoppworte sein sollten. Rein über die Wortfrequenzen ist dies nicht zu erschließen und eine Betrachtung aller Textstellen, in denen diese Wörter vorkommen, ist bei häufigen Wörtern wie dem 6.470 mal vorkommenden „time“ manuell kaum machbar. Hier zeigt sich das Defizit der fehlenden sinnvollen textimmanenten Kontextualisierung oder Erschließung von Zusammenhängen über einzelne Wörter hinaus. Wortfrequenzen können also nur einen ersten Zugang leisten, der mit anderen Methoden kombiniert werden sollte.

In den weiteren häufigen Wörtern lassen sich bereits gewisse Muster oder Cluster an Wörtern erkennen, wie z.B. „familiy“, „child“, „mother“, „father“ als Cluster zum Thema

Familie, oder „write“, „course“, „research“, „university“, „teach“ u.a. als Cluster wissenschaftlicher Arbeitspraxis. Die Wortfrequenzen sagen jedoch nichts darüber aus, ob diese Wörter in den Interviews auch tatsächlich zusammen auftauchen und textimmanent ein Cluster bilden. Die Muster werden also rein über unsere analytischen Vermutungen, also durch von Fragestellung und eigenem Vorwissen geleitete Sinnerschließungen hergestellt.

Wortfrequenzen auf der Ebene einzelner Interviews, also Dokumenthäufigkeiten, zeigen inhaltlich dominante Wörter einzelner Interviews und Unterschiede zwischen einzelnen Interviews auf. Die Ausgabe der häufigsten Wörter pro Interview als TXT-Datei ist in Abbildung 5 dargestellt.

Bei der Betrachtung der Dokumenthäufigkeiten lassen sich bei einzelnen Interviews Fachzugehörigkeiten wie „sociology“ (Interviews 1 und 3) oder „anthropology“ (Interview 7), Familienbezüge mit „child“ oder „family“ (insbesondere bei Interview 9) sowie auch Geschlechteraspekte mit „woman“ und „man“ (insbesondere bei Interview 6) erkennen. Für gewisse Themen können wir also interessante Interviews identifizieren, die dann genauer betrachtet werden können. Jedoch verbleiben auch hier die Wörter eher unbestimmt, da sie je nach Kontext in den Interviews unterschiedliche Bedeutungen haben können. Wird das Vorkommen der Wörter wie „course“, „work“, „research“ und „write“, die als Begriffe akademischer Arbeitspraxis eingeordnet werden können, über die Interviews hinweg verglichen, zeigt sich, dass diese in den meisten Interviews vorkommen. Beispielsweise taucht „work“ in 55 der 56 Interviews unter den häufigsten Wörtern auf – es handelt sich also nicht nur global um ein absolut häufiges Wort, sondern es ist über fast alle Dokumente verteilt.

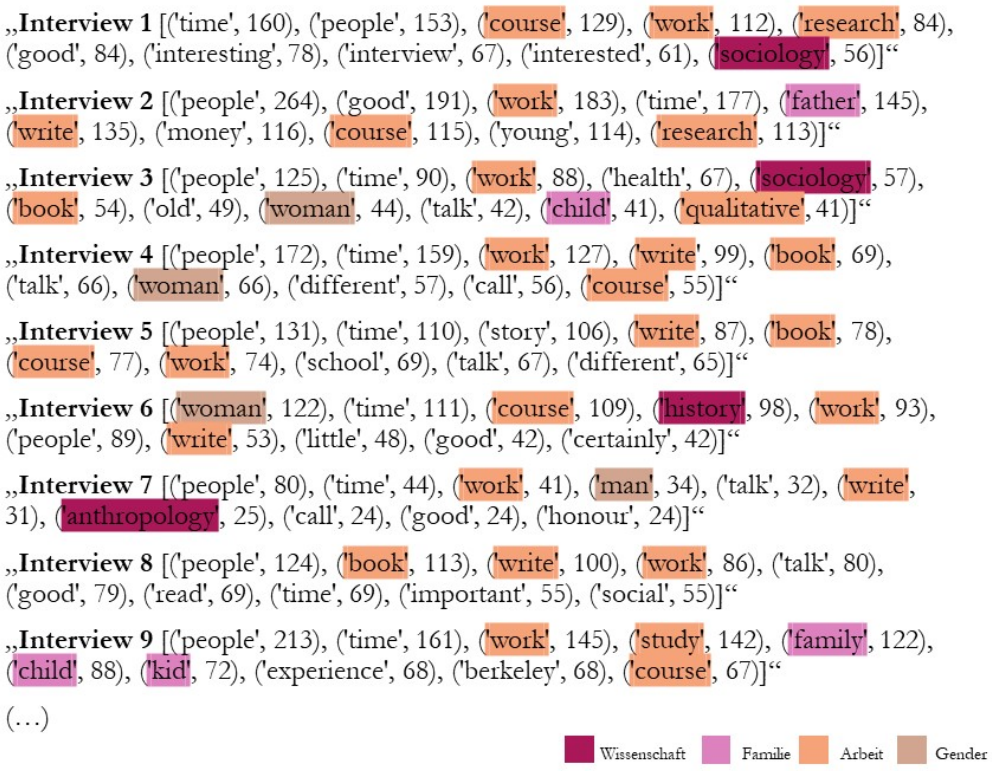


Abbildung 5: Gekürzte Liste der Dokumentfrequenzanalyse aufbauend auf dem PoSR-Datensatz, Einfärbungen zeigen mögliche thematische Cluster und Unterscheidungskategorien. *Quelle: Eigene Darstellung.*

Für unsere Fragestellung lassen sich über die Dokumentfrequenzen also keine bedeutenden Unterschiede identifizieren und somit keine Unterscheidungen besonders interessanter von uninteressanten Interviews vornehmen. Je größer die Menge an Interviewtranskripten wird, desto weniger ist ein händischer Vergleich der Frequenzlisten auf Interviewebene praktikabel. Beim PoSR-Datensatz mit 56 Interviews ist dies noch einigermaßen machbar, andere Datensätze mit teils mehr als 100 Interviews sind dafür nicht mehr geeignet. Auch bei den Wortfrequenzen auf Interviewebene zeigt sich, dass es Sinn macht, diese für erste überblickende Eindrücke zu nutzen, für eine tatsächliche Auswahl von Interviews oder Interviewpassagen als Einstieg in die qualitative Analyse jedoch auf eine Kombination mit weiteren Methoden zu setzen.

## Term Frequency – Inverted Document Frequency und Keyword in Context-Suche

Der tf-idf Score, die *term frequency – inverted document frequency*, verrechnet die Wortfrequenzen auf der Korpus- und auf Dokumentenebene miteinander, um pro Dokument (in unserem Fall pro Interviewtranskript) Wörter herauszuarbeiten, die im Dokument häufig, im Datensatz insgesamt aber eher selten vorkommen. Damit können Besonderheiten einzelner Interviews herausgearbeitet werden. Da die Begriffe mit hohem tf-idf Score per definitionem nicht oft vorkommen, ist es hier im Gegensatz zu den dominanten Wortfrequenzen erkenntnisversprechend, einzelne Textstellen interessanter Wörter in den Interviews herauszusuchen und genauer anzuschauen. Dafür ist eine Keyword in Context (KWIC) Suche sinnvoll, bei der das gesuchte Wort im Kontext der Textstelle angezeigt wird. Dazu bietet sich etwa das Tool AntConc an. Wird ein Begriff im Suchfeld eingegeben, gibt AntConc alle Textstellen des Wortes mit einem über die Anzahl an Wörtern zu definierenden linken und rechten Kontext sowie die zugehörige Datei an. So können also die Begriffe wieder den einzelnen Dokumenten, dabei aber auch ihrem Verwendungskontext, zugeordnet werden.

Die tf-idf Scores für den PoSR-Datensatz wurden von uns mit Hilfe der Gensim Library erstellt. Die Resultate der Berechnung werden auch bei tf-idf als nach Text aufgeteilte Listen ausgegeben, in denen pro Interview Wörter nach ihrem Score von 0 (gar nicht spezifisch) bis 1 (sehr spezifisch) geordnet werden. Ein Durchsehen dieser Listen nach relevanten Wörtern kann zu interessanten Funden führen, ist aber aufgrund der großen Menge an Resultaten wiederum wenig praktikabel für eine systematische Auswahl. Stattdessen haben wir die spezifischen Wörter aller Interviews in eine Liste mit absteigendem tf-idf Score gesammelt, um so die rechnerisch besonders auffälligen Wörter einzelner Interviews über das Gesamtkorpus vergleichend durchzugehen. Dies ist wiederum eine von uns getroffene Entscheidung, die den Score so anpasst, dass er einen Überblick zu interviewspezifischen Spezialthemen über das Material hinweg ermöglicht, welche dann bei näherem Interesse textnah betrachtet werden können. Es wäre auch denkbar, Begriffe in ihrer Häufigkeit z.B. über Dokumente hinweg zu vergleichen. Eine Liste von tf-idf im PoSR-Datensatz zeigt Abbildung 6.

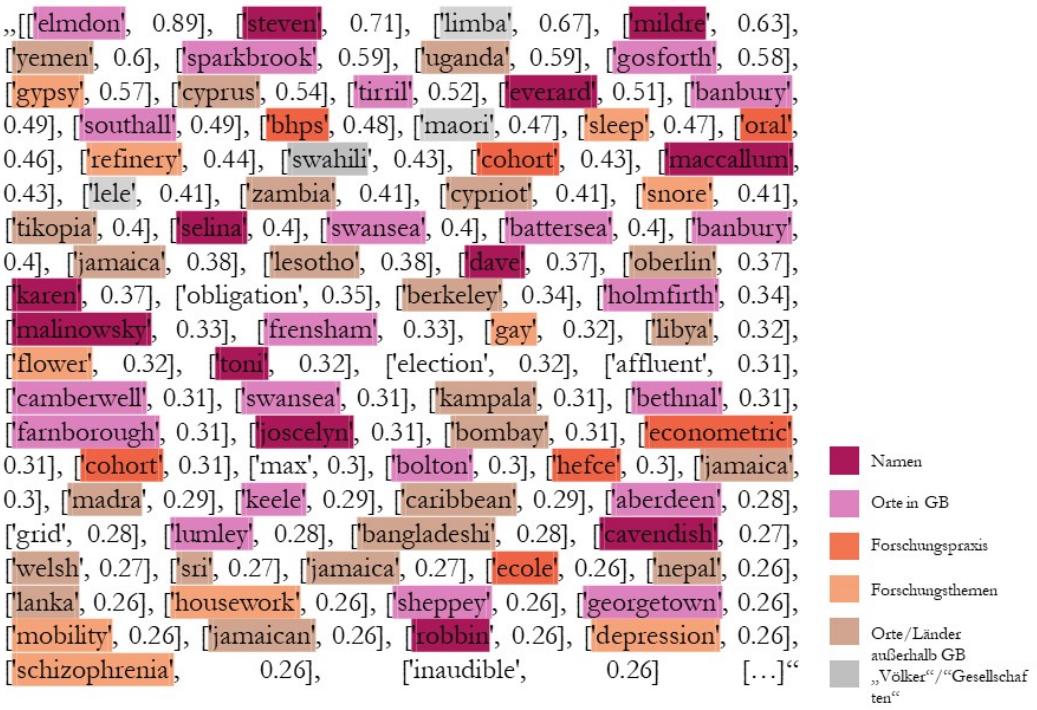


Abbildung 6: Zusammenstellung der tf-idf Analyse, aufbauend auf dem PoSR-Datensatz, absteigender tf-idf Score. Einfärbungen zeigen mögliche thematische Cluster.
*Quelle: Eigene Darstellung.*

Bei dieser Liste lassen sich, wie bei den Wortfrequenzen, mit dem eigenen Forschungsinteresse gewisse Cluster herausarbeiten, wie bspw. Orte in Großbritannien wie „Elmdon“ oder „Gosforth“, die in der biographischen Erzählung der Interviewpartner:innen anscheinend eine wichtige Rolle spielen. Auch Bezüge zu anderen Ländern oder Städten außerhalb der UK, wie „Kampala“ oder zu unspezifischen Vornamen und bekannten Namen wie „Malinowski“ sind in einigen Interviews stark vertreten. Die genauere Bestimmung der Bedeutung gewisser Wörter bleibt aber auch hier ohne Blick ins Material schwierig.

Werden mit KWIC Wörter mit hohem tf-idf-Score wie „ecole“, „bhps“, „hefce“ oder „oral“ gesucht, zeigt die Durchsicht der Textstellen, dass in mehreren Interviews über die „Ecole Normale Supérieure“ (v. a. in Interview 2), das „British Household Panel Survey“ (bhps, v. a. in Interview 51), den „Higher Education Funding Council for England“ (hefce, v. a. in Interview 24) oder „oral history“ (v. a. in Interview 40) gesprochen wird. Diese Wörter beziehen sich neben „cohort“ und „econometrics“ also auf Spezifitäten der Forschungspraxis der jeweils Interviewten. Wörter wie „gypsy“, „sleep“, „refinery“ oder „flower“ lassen sich mit einem Blick in ihren Kontext als wichtige Forschungsthemen der jeweiligen Sozialwissenschaftler:innen identifizieren. Dazu gehört auch „housework“, bei dem sich in der genaueren Betrachtung gerade für unsere Fragestellung der Flexibilisierung von Arbeit zeigt, dass „housework“ im Interview 25 nicht nur als Gegenstand sozialwissenschaftlicher Forschung der Interviewten behandelt wird, sondern auch im Verhältnis zu „employment work“, in

Bezug auf die Vereinbarkeit von beidem sowie vor dem Hintergrund wahrgenommener Unterschiede für die Zeit des Interviews und der erzählten Zeit diskutiert wird. Die tf-idf Analyse in Kombination mit KWIC ermöglicht also einen konkreten Einstieg in die qualitative Analyse relevanter Textstellen.

## n-Gramme

Mit dem Suchen nach n-Grammen finden wir Begriffe, die mehrfach zusammen im Text auftauchen. Durch diese häufigen Wortfolgen lassen sich neue, spezifischere Perspektiven auf das Material eröffnen, denn es zeigen sich Zusammenhänge und feststehende Begriffe. N-Gramme helfen so, den Text weiter aufzubrechen und an spezifischen Stellen einer weitergehenden Interpretation zugänglich zu machen. So können etwa Redewendungen oder Institutionen gefunden werden.

Technisch haben wir die n-Gramme in unserem Fallbeispiel mit Hilfe der Bibliotheken NLTK und spaCy umgesetzt. Für unsere Fragestellung ist es unwahrscheinlich, mehr als drei zusammengehörige Begriffe zu finden, weshalb wir nach Bigrams (also zwei mehrfach zusammenstehende Wörter) und Trigrams (drei gemeinsam auftauchende Wörter) gesucht haben. Abbildung 7 zeigt den entsprechenden Output.

| | |
|---|---|
| ('people', 'like') | 347 |
| ('write', 'book') | 321 |
| ('social', 'science') | 315 |
| ('long', 'time') | 306 |
| ('work', 'class') | 298 |
| ('spend', 'time') | 268 |
| ('middle', 'class') | 266 |
| ('year', 'ago') | 261 |
| ('people', 'work') | 208 |
| ('year', 'later') | 201 |
| ('grammar', 'school') | 198 |

Abbildung 7: 10 häufigste Bigrams im PoSR-Datensatz.
*Quelle: Eigene Darstellung.*

Die recht hohe Anzahl der Bigrams zeigt, dass in unserem Textkorpus einige Begriffe sehr häufig zusammenstehen. Tatsächlich handelt es sich bei allen Ausgaben um inhaltliche Aussagen. In den Bi- und Trigramms treten die Forschungsbereiche der Interviewpartner: innen mit Begriffen wie „middle class" noch deutlicher hervor als bei den bisher vorgestellten Verfahren. Den Output für Trigramms zeigt Abbildung 8.

Bei den Trigramms ist auffällig, dass es sich häufig um Eigennamen von Institutionen und Personen handelt, wobei diese wie bei der „child poverty action" durchaus inhaltlich relevant sein können. Allerdings sind bei „break recording" und „telephone interruption" auch Hinweise auf Interviewabläufe statt auf Interviewinhalte zu erkennen. Diese Fälle

könnten durch Ergänzung der Stoppwortliste oder – bei entsprechender automatisierter Erkennung der Metadaten – durch Ausschluss der Textstellen im Preprocessing verhindert werden. Mit Bigramms wird dieses Korpus also am besten inhaltlich erschlossen, da auf potentiell relevante Begriffskonstellationen aufmerksam gemacht wird. Auch hier wäre es möglich, sich die gefundenen n-Gramme mittels Tools wie AntConc genauer anzuschauen und etwa über die Keywords in Context in die Interpretation der Zusammenhänge einzusteigen.

| | |
|---|---|
| ('social', 'science', 'research') | 39 |
| (' ', 'break', 'recording') | 38 |
| ('take', 'long', 'time') | 37 |
| ('middle', 'class', 'family') | 35 |
| ('london', 'school', 'economic') | 35 |
| ('child', 'poverty', 'action') | 33 |
| (' ', 'telephone', 'interruption') | 32 |
| ('brian', 'abel', 'smith') | 32 |
| ('poverty', 'action', 'group') | 32 |
| ('household', 'panel', 'study') | 31 |
| ('science', 'research', 'council') | 29 |
| ('institute', 'community', 'study') | 29 |

Abbildung 8: 10 häufigste Trigrams im PoSR-Datensatz.
*Quelle: Eigene Darstellung.*

## Topic Modeling

Im Topic Modeling wird angenommen, dass jeder Text verschiedene Themen enthält, die durch gemeinsames Vorkommen (Kookkurrenz) einzelner Wörter in Wahrscheinlichkeitsrechnung gebildet werden. Im Forschungsprozess fungiert Topic Modeling dann als Filter (Koch & Franken 2020), um aus dem großen Korpus die für die Fragestellung relevanten Themen und Textstellen zu finden. Die inhaltliche Arbeit geschieht am Text bzw. an über die mit dieser digitalen Methode erschlossenen Teilen des Textes. Dabei besteht im Vorgehen eine gewisse Ähnlichkeit zu qualitativen Codierverfahren: Beim Topic Modeling findet eine Segmentierung in kleinere Sinnabschnitte statt, denen hier allerdings über rechnerische Verfahren gewisse Topics zugewiesen werden. Die Ergebnisse des Topic Modelings ergeben ein weitaus detaillierteres Bild als die bisher dargestellten Methoden.

Technisch haben wir hier auf die Anwendung von MALLET zurückgegriffen. Da wir in Python programmierten, MALLET aber standardmäßig in der Programmiersprache Java vorhanden ist, verwendeten wir die Python-Wrapper-Bibliothek Little-Mallet-Wrapper, die die Übersetzungsleistung übernimmt. Für lange Texte wie Transkripte teils mehrstündiger Interviews bietet es sich für das Topic Modeling an, diese in kleinere Einheiten (*Chunks*) zu unterteilen, wie bereits im Preprocessing beschrieben. Diese Segmentierung ermöglicht eine

| Mindestwortanzahl pro Segment | Anzahl Topics | Einbezug der Interviewer-Segmente |
|---|---|---|
| **10** | 47 | Nein |
| **10** | 47 | Ja |
| **10** | 71 | Ja |
| **20** | 59 | Ja |
| **30** | 52 | Ja |
| **40** | 48 | Nein |
| **510** | 16 | Ja |

Abbildung 9: Tiefer analysierte Modelle des Topic Modeling.
*Quelle: Eigene Darstellung.*

feingliedrigere Analyse nach *Topics*, als dies der Fall wäre, wenn ganze Interviews als Texteinheit in die Verarbeitung gegeben werden. Dabei muss die Segmentierung der Interviews in einzelne Analyseeinheiten sowie die Vorgabe der Anzahl auszugebender Topics iterativ ausprobiert werden, bis für den eigenen Analysezweck sinnvolle Topics vorliegen. Für die Anzahl an Topics sowie die Größe der Chunks gibt es keinen Goldstandard. Die Einstellung beider Variablen lässt sich zudem nicht getrennt voneinander vornehmen, da sich ihr Optimum gegenseitig bedingt (Sbalchiero & Eder 2020), was wir in unsere Modellevaluation miteinfließen ließen. So müssen am Material die unterschiedlichen Parameter in unterschiedlichen Kombinationen iterativ ausprobiert und das dann am besten zum Forschungsinteresse passende Modell händisch validiert werden. Indem wir die Mindestzahl der Wörter in einem Segment, die Anzahl der Topics und das Mitaufnehmen der Segmente, die von dem Interviewer gesprochen wurden, variierten, ließen wir automatisiert über 200 verschiedene Topic Models berechnen. Die Library gibt zum Vergleich verschiedener Modelle auch rechnerische Maße aus. Diese können als erste Annäherung an unterschiedlich spezifizierte Modelle, jedoch nicht als eindeutige Kriterien zur Bewertung und Auswahl eines Modells genutzt werden. Deshalb wählten wir anhand der Kennzahlen sowie einiger manueller Betrachtungen sieben Modelle zur genaueren Bewertung aus, deren Eigenschaften in Abbildung 9 dargestellt sind. Für inhaltlichere Unterteilungen wäre eine Lektüre der Texte notwendig gewesen, was wiederum zu aufwendig wäre.

Diese sieben Modelle wurden von uns jeweils individuell intensiv mit Blick auf die Brauchbarkeit für unser Forschungsinteresse bewertetet und dann in der Gruppe diskutiert. Unser Ziel war es dabei, dass ein Modell möglichst viele und zugleich differenzierte Bezüge zu unserem Forschungsinteresse aufweist. Die erstellten Topics eines Modells haben wir uns

| Topic | Anteil | Terms |
|---|---|---|
| 0 | 0,05254 | theory sociology read marx book social influence class idea theoretical marxist history time marxism sociological weber empirical economic write sociologist |
| 1 | 0,09205 | school teacher teach university level college education cambridge history good english oxford form science scholarship degree call exam influence latin |
| 2 | 0,09015 | family study research interview work project qualitative material case method different group kinship question quantitative survey social sample idea approach |
| … | | |
| 36 | 0,08738 | woman man work child wife young gender husband old marriage marry role time male domestic occupation female pay family help |
| … | | |
| 55 | 0,10263 | job ph.d lse student apply degree work time university sociology finish interview cambridge department day decide professor post thesis offer |
| 56 | 0,19151 | time day london week stay work long month meet marry child spend university wife home period got leave move england |
| 57 | 0,07536 | village man cyprus sit morning day drink greek house night car coffee pub tea meet evening shop drive eat meal |
| 58 | 0,11879 | friend meet important close influence time good person talk great cambridge touch work mention group intellectual friendship relationship man student |

Abbildung 10: Ausschnitt einer Topic-Term-Matrix des PoSR-Datensatzes. Die zweite Spalte gibt die Gewichtung des Topics innerhalb des Korpus an.
*Quelle: Eigene Darstellung.*

dabei über die Begriffslisten der Topics (siehe Topic-Term-Matrix in Abbildung 10) erschlossen. Je mehr Topics Bezüge zu unserer Frage nach Wandlungsprozessen von Arbeit aufweisen und umso differenzierter verschiedene Facetten des Forschungsinteresses beleuchtet werden, als desto besser geeignet haben wir ein Modell bewertet, da wir damit fundierter für uns interessante Textstellen aus dem umfassenden Korpus herausfiltern können. Der Bewertungsprozess von Topic Models findet vorzugsweise in einer Gruppenkonstellation statt, um die finale Auswahl eines Topic Models intersubjektiv zu validieren.

Beim PoSR-Korpus ergaben schlussendlich 59 Topics bei Segmentierung mit mindestens 20 Wörtern in einem Chunk (nach der Stoppwortentfernung) und einem Verzicht auf die Entfernung des Texts der Interviewenden die besten Ergebnisse. Einen Ausschnitt der Topic-Term-Matrix des gewählten Modells zeigt Abbildung 10.

Die Darstellung besteht in diesem Beispiel aus den zwanzig stärksten gewichteten Begriffen (Terms) des jeweiligen Topics. Durch eine Betrachtung der Zusammensetzung dieser Begriffe lässt sich erahnen, von was die Topics bzw. die den Topics zugewiesenen Textstellen handeln könnten und somit in einer ersten Annäherung festhalten, inwiefern ein Topic für das eigene Forschungsinteresse von Relevanz ist. In Bezug auf das uns interessierende Thema des Wandels von Arbeit lassen sich in dem Modell zwölf relevante Topics identifizieren, deren tatsächliche Relevanz durch die Ausgabe der zehn am höchsten gewichteten Chunks bzw. Textstellen dieser Topics exemplarisch überprüft wurde. Von der Wortliste wurde also mithilfe des Verfahrens auf exemplarische Textstellen geschlossen und diese ausführlicher untersucht. Es zeigt sich in diesen Textstellen, dass sich der Bezug zu „work“ als

Begriff eines Topics häufig durch Erzählungen über die Forschungsthemen der Befragten ergibt, es also nicht um Erzählungen über die eigenen Arbeitsbedingungen und deren Erfahrung geht, sondern um das eigene Forschungsgebiet, in dem wiederum Arbeit relevant ist.

Mit den Topics 36, 55 und 56 (siehe Abbildung 10) bleiben nach einer solchen exemplarischen Sichtung jedoch drei übrig, deren Bezug zu Arbeit im Sinne unseres Forschungsinteresses sich auch in der Betrachtung konkreter Textstellen zeigt. In den ausgegebenen Textstellen des Topics 36 werden zwar auch Forschungsarbeiten über das Verhältnis von Geschlecht und Arbeit besprochen. An zwei Beispielen wird hier jedoch deutlich, dass sich in diesem Topic auch Aussagen zum Wandel der Arbeit in der Wissenschaft finden: Einerseits wird berichtet vom eigenen Unterbrechen der Wissenschaftskarriere als ‚the director's wife' und dem späteren Wiedereinstieg in die akademische Karriere. Die andere Textstelle enthält

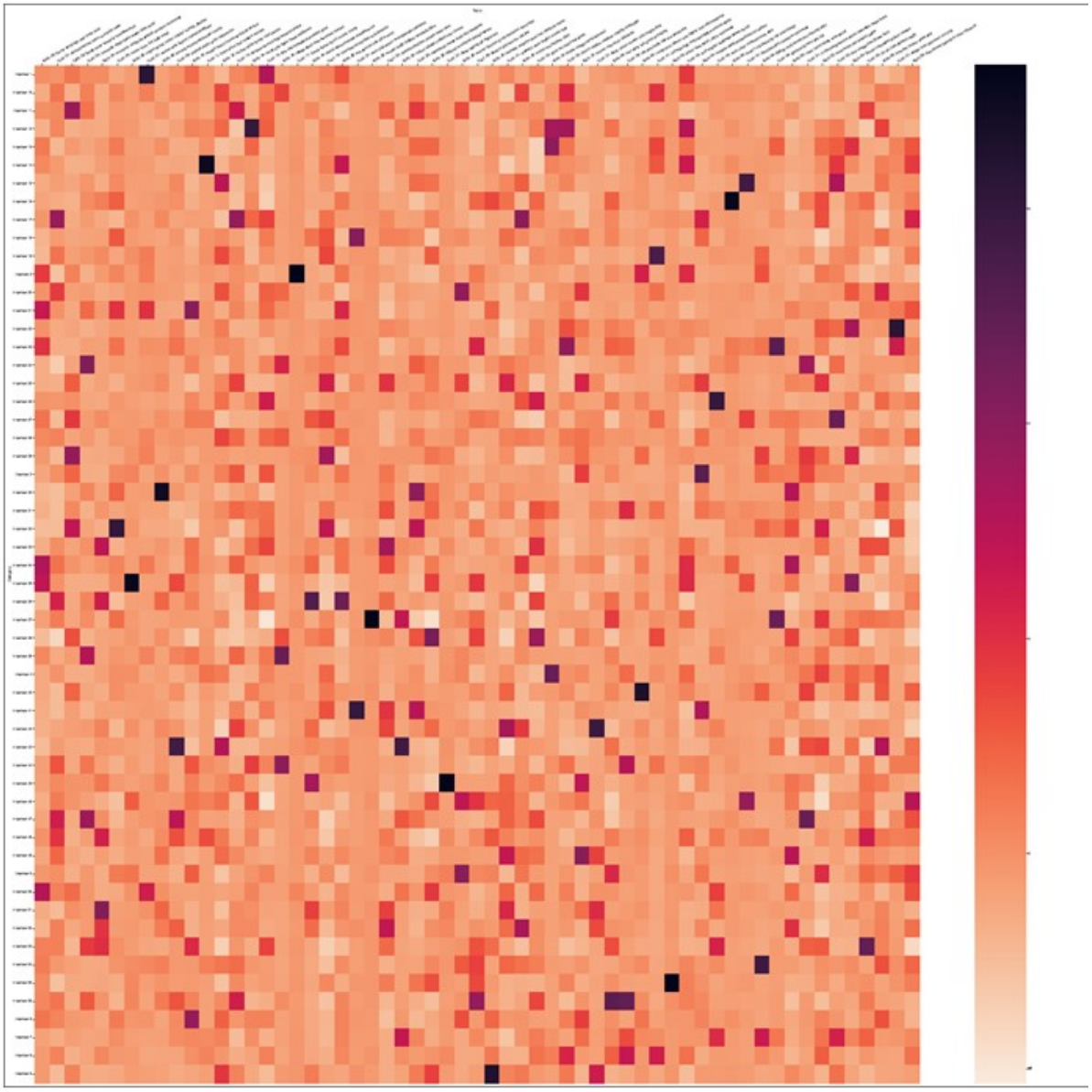

Abbildung 11: Topic-Document-Heatmap: Auf der X-Achse befinden sich die Interviews, auf der Y-Achse die Topics. Je dunkler die Einfärbung, desto stärker ist ein Topic im Interview vertreten. *Quelle: Eigene Darstellung.*

die Schilderung der prekären Rentensituation einer Mitforscherin, die nach der Heirat lange nicht gearbeitet hatte und nun nach der Scheidung fast ohne Rente dasteht. In Topic 55 finden sich Erzählungen über die eigenen akademischen Karrierewege und in Topic 56 wird vom Verhältnis von Arbeit und Familie in der eigenen Biografie berichtet.

Wir haben also für die für unsere Fragestellung relevanten Topics im Korpus die zehn am höchsten gewichteten Textstellen identifiziert. Dadurch sind wir explorativ auf erste interessante Interviews und Textstellen gestoßen. Von da aus kann weiter und tiefergehend in bestimmte Topics, Interviews und Textausschnitte geschaut werden. So kann zum Beispiel eine so genannte Topic-Document-Matrix ausgegeben werden, die oft in Form einer Heatmap wie in Abbildung 11 dargestellt wird. In dieser wird anzeigt, welche Topics in welchen Dokumenten wie stark vorkommen.

Neben einer Topic-Document-Matrix sind auch Topic-Chunk-Matrizen möglich, die pro Interview die Gewichtung der Topics im Verlauf der vorab festgelegten Textabschnitte anzeigen. Die Arbeit mit entsprechenden Tabellen und Visualisierungen kann also feingliedriger gestaltet werden, wenn dies für die eigene Analyse als hilfreich angesehen wird. So wird Material nicht einfach reduziert, sondern es können gezielt einzelne Textpassagen herausgefiltert werden, die für die Fragestellung besonders interessant sind. Darüber ist eine Annäherung an den konkreten Text und der Einstieg in eine weitergehende, qualitative Analyse möglich.

Insgesamt zeigt sich, dass eine konkrete Fragestellung wichtig ist, um das Modell passend zu spezifizieren und die Ergebnisse zu fokussieren. Beim Topic Modeling ist zudem abschließend zu beachten, dass die errechneten Topics immer über textimmanente Zusammenstellungen von Wörtern laufen. Interessiert eine analytische Perspektive, die in den Interviews nicht explizit als solche behandelt wurde (wie hier der Wandel von Arbeit), tauchen die interessierenden Dimensionen vermutlich nicht explizit in den Topics auf. Die Forschenden müssen dann also Topics bestimmen, die für die eigenen analytischen Kategorien (zur Unterscheidung vgl. Rapp 2017: 258–9) interessant sein könnten, dies am Text valideren und inhaltlich ausarbeiten.

## Zum Nutzen digitaler Methoden in der Nachnutzung von Forschungsdaten

Wenn wir uns eine digitale Analyse von Interviewtranskripten, wie wir sie im zweiten Teil des Beitrags beschrieben haben, im Überblick anschauen, so wird schnell deutlich, dass neue Perspektiven auf das Datenmaterial entstehen, die wir mit einer hermeneutischen Analyse oder mit Verfahren etwa nach Grounded Theory nicht, oder zumindest nicht in einem vertretbaren Zeitrahmen, gewonnen hätten. In der Anwendung digitaler Methoden gibt es bisher keine goldene Regel und kein einheitlich zu verfolgendes Schema. Es geht vielmehr um ein Ausprobieren und um die Überprüfung des praktischen Nutzens von Erkenntnissen für die eigene Analyse. Digitale Methoden ermöglichen durch eine veränderte Triangulation auf Analyseebene bzw. auf Ebene der Auswertungsmethoden (Flick 2019; Strübing et al. 2018) neue Perspektiven in der Nachnutzung von Forschungsdaten.

Wir haben verschiedene Methoden vorgestellt, mit denen ein Korpus unbekannter Interviewdaten erschlossen werden kann. Unserer Erfahrung nach hat jedes Verfahren dabei seine Vorteile. Mittels Frequenzanalysen konnten wir uns einen ersten Eindruck davon verschaffen, welche Themen häufig im Korpus auftreten. Zudem haben wir einen Überblick

darüber gewonnen, welche Begriffe in welchem Umfang in den einzelnen Interviews zur Sprache kommen. Tiefere Einblicke in die Daten gewähren n-Gramme, mit denen oft gemeinsam auftretende Wortgruppen ausgegeben werden können. Es hat sich gezeigt, dass diese in den meisten Fällen auch in einem inhaltlichen Zusammenhang stehen. Wenn eine oder mehrere dieser Kombinationen in der eigenen Forschung thematisiert werden sollen, dann können auf dieser Grundlage Texte für die nähere Analyse ausgewählt werden. Mithilfe des Topic Modeling ist es schließlich möglich, diejenigen Transkripte bzw. Textteile zu identifizieren, die offenbar relevante Themen enthalten. Gerade in der Beschäftigung mit verschiedenen Modellen und den in den Topics zentral stehenden Begriffen gewinnt man weitergehende Einblicke in das Material, die für die Analyse einzelner Textteile hilfreiche Kontextualisierung anbieten.

Wie wir immer wieder deutlich gemacht haben, kann die Verwendung der derzeitigen Standardansätze digitaler Methoden nur ein Ausgangspunkt sein, um große Datenmengen für die qualitative Analyse zu erschließen. Für die eigentliche Beantwortung der Fragestellung braucht es in der Regel viele weitere Schritte. Andere Punkte werden erst durch die Kenntnisse der Forschenden und durch einen menschlichen Blick in die Materialien und den Kontext einzelner Textstellen deutlich: Es ist zum Beispiel notwendig, zwischen dem persönlichen Erleben einer Person und – im Falle von Forschenden wie im PoSR-Datensatz – dem eigenen Arbeitsbereich zu unterscheiden. Ob sich häufige Wörter wie „Vater" auf die eigene oder eine untersuchte Situation beziehen, kann in der Regel am leichtesten durch eine eigene Lektüre der betreffenden Textpassage herausgefunden werden.

Die berücksichtigten Daten beantworten womöglich Fragen, die sich erst aus dem Material ergeben und durch die Sekundäranalyse können neue Fragen aufkommen, wie Conlon et al. (2020: 953–954) berichten. Nach unserer Analyse lässt sich unsere Ausgangsfragestellung beispielswiese auf die Frage hin spezifizieren, wie sich Erfahrungen zum Wiedereinstieg in die z.B. durch unbezahlte familiäre Sorgearbeit unterbrochene Berufsarbeit durch den postfordistischen Wandel der Arbeitsregime verändern. Die vorgestellten digitalen Verfahren schließen also gut an Forschungsdesigns aus der EKW/EE/KA an, die explorativ und offen vorgehen und im Studienverlauf Methode und Fragestellung anpassen und präzisieren. Sowohl in digitalen Analyseverfahren als auch mit bestehenden Forschungsdaten werden nur ganz bestimmte Fragen beantwortet und die Daten in spezifische Formen zugeschnitten. Rob Kitchin (2014) weist zu Recht auf die Gefahren einer solchen Data Driven Science ohne theoretische Perspektivierung und Fragestellung hin und kommt zu dem Ergebnis, dass eine breitere und tiefere Analyse mit digitalen Methoden prinzipiell möglich ist, diese aber theoretisch informiert umgesetzt werden muss. Deshalb können digitale Methoden genutzt werden „to use guided knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing" (ebd.: 6). Ähnlich argumentieren Marres und Gerlitz (2016), dass die mit digitalen Methoden möglichen Forschungszugänge experimentell und performativ sind und nicht auflösbare Ambivalenzen in ihrer Nutzung beinhalten (vgl. auch Rogers 2019).

Digitale Methoden sind besonders in Kombination und Iteration verschiedener einzelner Verfahren hilfreich für die analytischen Zugänge der EKW/EE/KA und bleiben gleichzeitig ein Teilbereich der eigentlichen Analyse. Es ist unumgänglich, für tiefere Analysen in das Datenmaterial selbst einzusteigen und das für die Analyse vorhandene Methodenrepertoire wieder sehr textnah zu verwenden, wenn Filterprozesse stattgefunden haben. Damit schließen wir an Zugänge an, die in der qualitativen Sekundäranalyse bereits etabliert sind.

Wenn die Erkenntnisse aus den untersuchten Datensätzen nicht ausreichen, kann im nächsten Schritt nach thematisch ähnlichen Datensätzen gesucht und diese ebenfalls mit digitalen Methoden bearbeitet werden. Besonders vielversprechend ist es außerdem, die Sekundäranalyse mit eigenen Erhebungen zu verbinden (vgl. Andrews et al. 2012; Watters et al. 2018: § 26). Mit Kenntnis der Forschungsdaten anderer können wir informierter in eigene Erhebungen gehen und so gezielt ergänzen, was aus den vorhandenen Quellen noch nicht deutlich wird, uns aufgrund unserer Fragestellung aber besonders interessiert. Zudem können insbesondere historische Perspektiven aufgenommen werden, wenn die Forschungsdaten anderer ausgewertet werden – ein neues Forschungsfeld, dessen Potenzial erst noch erarbeitet werden muss. Ob der Schwerpunkt des jeweiligen Forschungsprojektes auf der eigenen Erhebung oder auf der Auswertung von Forschungsdaten liegt, hängt von der Fragestellung und den damit verbundenen Forschungsentscheidungen ab.

Eine kulturanalytische Fragestellung wird also mit den vorgestellten Verfahren kaum allein zu beantworten sein, sie können jedoch für diese genutzt und kombiniert werden. Methoden des Text Minings weisen uns den Weg zu interessanten Textstellen, die wir dann selbst erkunden und hinsichtlich ihres Quellenwertes einschätzen müssen. Computationelle Methoden ermöglichen es qualitativ-kulturanalytisch Forschenden dadurch, mit großen Datenmengen fundiert zu arbeiten. Denn um in dem schier unendlichen Meer von Forschungsdaten zu navigieren, sind digitale Methoden besonders hilfreich, da sie quantifizierende Ansätze aufnehmen und deshalb Datenmengen besser händelbar machen. Dadurch tragen digitale Methoden einen Teil dazu bei, dass bestehende Wissensbestände nicht in Vergessenheit geraten (vgl. z.B. Watters et al. 2018: § 29).

Die Nachnutzung von qualitativen Forschungsdaten unter Einsatz computationeller Methoden ist jedoch in vielen Bereichen noch nicht einfach umzusetzen. Es fehlt oftmals schlichtweg eine Übersicht zu vorhandenen Daten und gerade bei nicht öffentlich sichtbaren Datensätzen muss schon für die Sichtung oft ein Account beantragt werden. Für die Vorverarbeitung und Analyse von Datensätzen ist ein technisches Wissen notwendig, das in der Regel nicht in Studiengängen der EKW/EE/KA oder der breiteren Methodenlehre der qualitativen Sozialforschung vermittelt wird. Zudem stellen die Spezifika von Transkripten gesprochener Sprache einige der Standardverfahren vor bisher ungelöste Herausforderungen. Insgesamt sind also die Hürden zum Einstieg in die Anwendung digitaler Methoden noch hoch, auch wenn es für viele Schritte gut nachnutzbare Tools gibt. Es muss viel ausprobiert werden, bis tatsächlich Ergebnisse von Verfahren des Text Minings in die eigene Analyse einbezogen werden können. Es ist deshalb ratsam, entsprechend Zeit einzuplanen und sich in Teams zusammenzuschließen.

Dort, wo wir bereits heute selbstverständlich einen Forschungsstand recherchieren und zusammentragen, welche Publikationen zu unserer Fragestellung bereits existieren und zu welchen Ergebnissen andere gekommen sind, sollte diese Recherche künftig auf Forschungsdaten ausgeweitet werden. Die Suche nach bestehenden und (begrenzt) zugänglichen Datensätzen kann dabei helfen, das eigene Quellenspektrum zu ergänzen oder auch dazu dienen, die eigene Fragestellung so weiterzuentwickeln, dass neue Erkenntnisse generiert werden können. Insofern ist die wissenschaftspolitische Forderung nach vermehrter Archivierung und vor allem Nachnutzung von Forschungsdaten, unterstützenswert. Für die reine grobe Durchsicht von bestehenden Datensätzen sind oft schon die vorhandenen Metadaten ausreichend. Die Nutzung von digitalen Methoden würde dabei Aufwand und Ertrag in

kein angemessenes Verhältnis setzen. Sie ist vielmehr sinnvoll, wenn intensiv mit vorhandenen, potentiell interessanten Datensätzen gearbeitet werden soll und Textmengen so umfangreich sind, dass sie in manueller Durchsicht kaum zu überschauen sind.

## Literatur

Adelmann, Benedikt, Melanie Andresen, Anke Begerow, Lina Franken, Evelyn Gius & Michael Vauth (2019): Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics. In: Digital Humanities 2019 Conference Paper.

Andrews, Lorraine, Agnes Higgins, Michael Waring Andrews & Joan G. Lalor (2012): Classic Grounded Theory to Analyse Secondary Data: Reality and Reflections. In: Grounded Theory Review 11/1, 12–26.

Atkinson, John (1984): Manpower Strategies for Flexible Organisations. In: Personnel Management 16/8, 28–31.

Bischoff, Christine, Karoline Oehme-Jüngling & Walter Leimgruber (Hgs.) (2014): Methoden der Kulturanthropologie. Bern: Haupt.

Bishop, Libby & Arja Kuula-Luumi (2017): Revisiting Qualitative Data Reuse. A Decade On. In: SAGE Open 7/1, 1–15. https://doi.org/10.1177/2158244016685136.

Blei, David M. (2012): Probabilistic Topic Models. Surveying a Suite of Algorithms That Offer a Solution to Managing Large Document Archives. In: Communications of the ACM 55/4, 77–84. https://doi.org/10.1145/2133806.2133826.

Boris, Lenore L. (2015): Hearing the Voices of HIV Positive Women in Kenya. Secondary Analysis of Interview Data Using Dialogic/Performance Analysis. Dissertation. University of Wisconsin-Milwaukee.

Bröckling, Ulrich (2016): The Entrepreneurial Self. Fabricating a New Type of Subject. London: SAGE.

Bubenhofer, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In: Kersten Sven Roth, Martin Wengeler & Alexander Ziem (Hgs.), Handbuch Sprache in Politik und Gesellschaft. Berlin, New York, 69–93.

Charmaz, Kathy (2014): Constructing Grounded Theory. Introducing Qualitative Methods. 2nd edition. Los Angeles, London, New Delhi: SAGE.

Conlon, Catherine, Virpi Timonen, Catherine Elliott-O'Dare, Sorcha O'Keeffe & Geraldine Foley (2020): Confused About Theoretical Sampling? Engaging Theoretical Sampling in Diverse Grounded Theory Studies. In: Qualitative Health Research 30/6, 947–959. https://doi.org/10.1177/1049732319899139.

Corti, Louise & Paul Thompson (2004): Secondary Analysis of Archived Data. In: Clive Seale, Giampietro Gobo, Jaber F. Gubrium & David Silverman (Hgs.), Qualitative Research Practice. Reprinted. London: SAGE, 297–313.

Dargentas, Magdalini & Dominique Le Roux (2005): Potentials and Limits of Secondary Analysis in a Specific Applied Context. The Case of EDF-Verbatim. In: Forum Qualitative Social Research 6/1. https://doi.org/10.17169/fqs-6.1.505.

DGEKW (2018): Positionspapier zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten. Deutsche Gesellschaft für Empirische Kulturwissenschaft. Online verfügbar unter https://www.dgekw.de/wp-content/uploads/2019/04/dgv-Positionspapier_FDM.pdf. Letzter Zugriff: 28.08.2023.

Dörre, Klaus (2019): Die Neuen Vagabunden. Prekarität in Reichen Gesellschaften. In: Uwe H. Bittlingmayer, Alex Demirović & Tatjana Freytag (Hgs.), Handbuch Kritische Theorie. Wiesbaden: Springer Fachmedien Wiesbaden, 981–1003.

Draucker, Claire B., Donna S. Martsolf, Ratchneewan Ross & Thomas B. Rusk (2007): Theoretical Sampling and Category Development in Grounded Theory. In: Qualitative Health Research 17/8, 1137–1148. https://doi.org/10.1177/1049732307308450.

Dunkel, Wolfgang, Heidemarie Hanekop, Nicole Mayer-Ahuja (Hgs.) (2019): Blick Zurück Nach Vorn. Sekundäranalysen zum Wandel von Arbeit nach dem Fordismus. Frankfurt, New York: Campus.

Edmond, Jennifer, Nicola Horsley, Elisabeth Huber, Rihards Kalnins, Joerg Lehman, Georgina Nugent-Folan, Mike Priddy & Thomas Stodulka (2018): Big Data & Complex Knowledge. Observations and Recommendations for Research from the Knowledge Complexity Project. Trinity College Dublin. Dublin.

Egger, Nils, Lina Franken, Dennis Möbus & Florian Schmid (2023): Oral History auf dem Weg zu Big Data: menschliche und maschinelle Annotation lebensgeschichtlicher Interviews im Vergleich. In: DHd2023 (Hg.), Open Humanities. Open Culture. Abstracts zur 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Luxemburg/Trier. https://doi.org/10.5281/zenodo.7715317.

Evert, Stefan (2009): Corpora and Collocations. In: Anke Lüdeling und Merja Kytö (Hgs.), Corpus Linguistics. An International Handbook. Berlin: de Gruyter, 1212–1248.

Fenske, Michaela (2006): Mikro, Makro, Agency. Historische Ethnografie als kulturanthropologische Praxis. In: Zeitschrift für Volkskunde 102, 151–177.

Flick, Uwe (2019): Gütekriterien qualitativer Sozialforschung. In: Nina Baur & Jörg Blasius (Hgs.), Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer Fachmedien Wiesbaden, 473–488.

Franken, Lina (2020a): Methodologie der Zukunft? Automatisierungspotentiale in kulturwissenschaftlicher Forschung. In: Dagmar Hänel, Ove Sutter, Ruth Dorothea Eggel, Fabio Freiberg, Andrea Graf, Victoria Huszka & Kerstin Wolff (Hgs.), Planen. Hoffen. Fürchten. Zur Gegenwart der Zukunft im Alltag (Bonner Beiträge zur Alltagskulturforschung, 13). Münster, New York: Waxmann, 217–233.

Franken, Lina (2020b): Kulturwissenschaftliches Digitales Arbeiten. Qualitative Forschung als ‚Digitale Handarbeit'? In: Berliner Blätter. Ethnographische und ethnologische Beiträge 82, 107–118.

Franken, Lina (2022): 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 22/2. https://doi.org/10.17169/fqs-22.2.3818.

Franken, Lina (2023): Digitale Methoden für qualitative Forschung. Computationelle Daten und Verfahren. Münster: UTB, Waxmann.

Franken, Lina, Gertraud Koch & Heike Zinsmeister (2020): Annotationen als Instrument der Strukturierung. In: Julia Nantke & Frederik Schlupkothen (Hgs.), Annotation in Scholarly Editions and Research. Oldenburg, 89–108.

Gladstone, Brenda M., Tiziana Volpe & Katherine M. Boydell (2007): Issues Encountered in a Qualitative Secondary Analysis of Help-Seeking in the Prodrome to Psychosis. In: The Journal of Behavioral Health Services & Research 34/4, 431–442. https://doi.org/10.1007/s11414-007-9079-x.

Glaser, Barney G. & Anselm L. Strauss (1967): The Discovery of Grounded Theory. Strategies for Qualitative Research. Hawthorne, N.Y.: de Gruyter.

Götz, Irene & Alexandra Rau (Hgs.) (2017): Facetten des Alter(n)s. Ethnografische Porträts über Vulnerabilitäten und Kämpfe älterer Frauen. Herbert Utz Verlag. München: Herbert Utz Verlag.

HaCohen-Kerner, Yaakov, Daniel Miller & Yair Yigal (2020): The Influence of Preprocessing on Text Classification Using a Bag-of-Words representation. In: PloS one 15/5. https://doi.org/10.1371/journal.pone.0232525.

Heaton, Janet (2008): Secondary Analysis of Qualitative Data. In: Pertti Alasuutari, Leonard Bickman & Julia Brannen (Hgs.), The SAGE Handbook of Social Research Methods. Los Angeles: SAGE, 506–519.

Hennink, Monique M., Bonnie N. Kaiser & Vincent C. Marconi (2017): Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? In: Qualitative Health Research 27/4, 591–608. https://doi.org/10.1177/1049732316665344.

Heiberger, Raphael H. & Sebastian Munoz-Najar Galvez (2021): Text Mining and Topic Modeling. In: Uwe Engel, Anabel Quan-Haase, Sunny Liu & Lars Lyberg (Hgs.), Handbook of Computational Social Science, Volume 2: Routledge, S. 352–365.

Heuer, Jan-Ocko, Susanne Kretzer, Kati Mozygemba, Elisabeth Huber & Betina Hollstein (2020): Kontextualisierung Qualitativer Forschungsdaten für die Nachnutzung. Eine Handreichung für Forschende zur Erstellung eines Studienreports. Unter Mitarbeit von Universität Bremen.

Hollstein, Betina & Jörg Strübing (2018): Archivierung und Zugang zu Qualitativen Daten. In: Doris Bambey, Louise Corti, Michael Diepenbroek, Wolfgang Dunkel, Heidemarie Hanekop, Betina Hollstein, Sabine Imeri, Hubert Knoblauch, Susanne Ketzer, Christian Meyermann, Maike Porzelt, Marc Ritterberger, Jörg Strübing, Hella von Unger & René Wilke (Hgs.), Archivierung und Zugang zu Qualitativen Daten. RatSWD Working Paper 267/201, 1–13. https://doi.org/10.17620/02671.35.

Holton, Judith A. (2007): The Coding Process and Its Challenges. In: Antony Bryant und Kathy Charmaz (Hgs.), The SAGE Handbook of Grounded Theory. Los Angeles: SAGE, 265–289.

Holubek, Stefan (2017): Motive für das Zweite Kind. Eine Qualitative Sekundäranalyse Problemzentrierter Interviews. In: Journal of Family Research 29/3, 319–339. https://doi.org/10.3224/zff.v29i3.04.

Ignatow, Gabe & Rada F. Mihalcea (2018): An Introduction to Text Mining. Research Design, Data Collection, and Analysis. Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE.

Imeri, Sabine (2018): Archivierung und Verantwortung. Zum Stand der Debatte über den Umgang mit Forschungsdaten in den Ethnologischen Fächern. In: Doris Bambey, Louise Corti, Michael Diepenbroek, Wolfgang Dunkel, Heidemarie Hanekop, Betina Hollstein, Sabine Imeri, Hubert Knoblauch, Susanne Ketzer, Christian Meyermann, Maike Porzelt, Marc Ritterberger, Jörg Strübing, Hella von Unger & René Wilke (Hgs.), Archivierung und Zugang zu Qualitativen Daten. RatSWD Working Paper 267/201, 69–79. https://doi.org//10.17620/02671.35.

Imeri, Sabine (2019): ‚Open Data' Zum Umgang mit Forschungsdaten in den ethnologischen Fächern. In: Jens Klingner & Merve Lühr (Hgs.), Forschungsdesign 4.0. Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive. Dresden: Institut für Sächsische

Geschichte und Volkskunde, Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden, 45–59.

Jentsch, Patrick & Stephan Porada (2020): From Text to Data. Digitization, Text Analysis and Corpus Linguistics. In: Silke Schwandt (Hg.), Digital Methods in the Humanities. Challenges, Ideas, Perspectives. Bielefeld: Transcript, 89–128.

Jurafsky, Daniel & James H. Martin (2021): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Stanford: Stanford University Press.

Kalleberg, Arne L. (2001): Organizing Flexibility: The Flexible Firm in a New Century. In: British Journal of Industrial Relations 39/4, 479–504. https://doi.org/10.1111/1467-8543.00211.

Karlgren, Jussi, Renee Li & Eva M. Meyersson Milgrom (2020): Text Mining for Processing Interview Data in Computational Social Science. In: arXiv:2011.14037. https://doi.org/10.48550/arXiv.2011.14037.

Kitchin, Rob (2014): Big Data, New Epistemologies and Paradigm Shifts. In: Big Data & Society 1/1, 1–12. https://doi.org/10.1177/2053951714528481.

Koch, Gertraud & Lina Franken (2020): Filtern als Digitales Verfahren in der Wissenssoziologischen Diskursanalyse. In: Samuel Breidenbach, Peter Klimczak & Christer Petersen (Hgs.), Soziale Medien. Interdisziplinäre Zugänge zur Onlinekommunikation. Wiesbaden: Springer, 121–138.

Lemke, Matthias & Gregor Wiedemann (Hgs.) (2016): Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen Qualitativer und Quantitativer Diskursanalyse. Wiesbaden: Springer.

Lempert, Lora Bex (2007): Asking Questions of the Data. Memo Writing in the Grounded Theory Tradition. In: Antony Bryant & Kathy Charmaz (Hgs.), The SAGE Handbook of Grounded Theory. Los Angeles: SAGE, 245–264.

Marres, Noortje & Carolin Gerlitz (2016): Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. In: The Sociological Review 64/1, 21–46. https://doi.org/10.1111/1467-954X.12314.

Medjedović, Irena (2020): Forschungsdatenmanagement und Sekundärnutzung qualitativer Daten. In: Martin Wazlawik & Bernd Christmann (Hgs.), Forschungsdatenmanagement und Sekundärnutzung qualitativer Forschungsdaten, Bd. 6. Wiesbaden: Springer Fachmedien, 9–43.

Morse, Janice M. (2007): Sampling in Grounded Theory. In: Antony Bryant und Kathy Charmaz (Hgs.), The SAGE Handbook of Grounded Theory. Los Angeles: SAGE, 229–244.

Muckenhuber, Johanna, Josef Hödl & Martin Griesbacher (Hg.) (2018): Normalarbeit. Nur Vergangenheit oder auch Zukunft? Bielefeld: Transcript.

Pernicka, Sussanne & Bettina Stadler (2006): Atypische Beschäftigung. Frauensache? In: Östereichische Zeitschrift für Soziologie 31/3, 3–21. https://doi.org/10.1007/s11614-006-0023-8.

Rapp, Andrea (2017): Manuelle und automatische Annotation. In: Fotis Jannidis, Hubertus Kohle & Malte Rehbein (Hgs.), Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler, 253–267.

Rogers, Richard (2019): Doing Digital Methods. London: SAGE.

Ruggiano, Nicole & Tam E. Perry (2019): Conducting Secondary Analysis of Qualitative Data. Should We, Can We, and How? In: Qualitative Social Work 18/1, 81–97. https://doi.org/10.1177/1473325017700701.

Sattler, Simone (2014): Computergestützte qualitative Datenbearbeitung. In: Christine Bischoff, Karoline Oehme-Jüngling & Walter Leimgruber (Hgs.), Methoden der Kulturanthropologie. Bern: Haupt, 476–487.

Sbalchiero, Stefano & Maciej Eder (2020): Topic Modeling, Long Texts and the Best Number of Topics. Some Problems and solutions. In: Quality & Quantity 54/4, 1095–1108. https://doi.org/10.1007/s11135-020-00976-w.

Schöch, Christof (2017): Wiederholende Forschung in den Digitalen Geisteswissenschaften. In: Michael Stilz (Hg.), Digitale Nachhaltigkeit. Abstracts zur 4. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. Bern, 207–212. https://doi.org/10.5281/zenodo.277113.

Sherin, Bruce (2013): A Computational Study of Commonsense Science. An Exploration in the Automated Analysis of Clinical Interview Data. In: Journal of the Learning Sciences 22/4, 600–638. https://doi.org/10.1080/10508406.2013.836654.

Shrader, Charles B., Sue Pickard Ravenscroft, Jeffrey B. Kaufmann & Kyle Hansen (2021): Collusion Among Accounting Students. Data Visualization and Topic Modeling of Student Interviews. In: Decision Sciences Journal of Innovative Education 19/1, 40–62. https://doi.org/10.1111/dsji.12226.

Star, Susan Leigh (2010): This is Not a Boundary Object. Reflections on the Origin of a Concept. In: Science, Technology, & Human Values 35/5, 601–617. https://doi.org/10.1177/0162243910377624.

Strübing, Jörg, Stefan Hirschauer, Ruth Ayaß, Uwe Krähnke & Thomas Scheffer (2018): Gütekriterien qualitativer Sozialforschung. Ein Diskussionsanstoß. In: Zeitschrift für Soziologie 47/2, 83–100. https://doi.org/10.1515/zfsoz-2018-1006.

Sutter, Ove (2013): Erzählte Prekarität. Autobiographische Verhandlungen von Arbeit und Leben im Postfordismus. Frankfurt a.M.: Campus.

Tate, Judith A. & Mary B. Happ (2018): Qualitative Secondary Analysis. A Case Exemplar. In: Journal of Pediatric Health Care 32/3, 308–312. https://doi.org/10.1016/j.pedhc.2017.09.007.

Thompson, Paul (2019): Pioneers of Social Research, 1996-2018. Data Collection. 4. Aufl. Essex: UK Data Service. http://doi.org/10.5255/UKDA-SN-6226-6.

Thompson, Paul, Ken Plummer & Neli Demireva (2021): Pioneering Social Research. Life Stories of a Generation. Bristol: Policy Press.

Thomson, Denise, Lana Bzdel, Karen Golden-Biddle, Trish Reay & Carole A. Estabrooks (2005): Central Questions of Anonymization. A Case Study of Secondary Use of Qualitative Data. In: Forum Qualitative Social Research 6/1. https://doi.org/10.17169/fqs-6.1.511.

Tonidandel, Scott, Karoline M. Summerville, William A. Gentry & Stephen F. Young (2022): Using Structural Topic Modeling to Gain Insight into Challenges Faced by Leaders. In: The Leadership Quarterly 33(5): 101576. https://doi.org/10.1016/j.leaqua.2021.101576.

Timm, Elisabeth (2020): Forschungsdatenmanagement in der Europäischen Ethnologie. Eine kurze Kritik des dgv-Positionspapiers. In: Zeitschrift für Volkskunde 116/1, S. 88–89.

von Unger, Hella (2014): Forschungsethik in der qualitativen Forschung. Grundsätze, Debatten und offene Fragen. In: Hella von Unger, Petra Narimani & Rosaline M´Bayo (Hgs.), Forschungsethik in der qualitativen Forschung. Reflexivität, Perspektiven, Positionen. Wiesbaden: Springer VS, 15–39.

von Unger, Hella (2018): Archivierung und Nachnutzung Qualitativer Daten aus Forschungsethischer Perspektive. In: Doris Bambey, Louise Corti, Michael Diepenbroek, Wolfgang Dunkel, Heidemarie Hanekop, Betina Hollstein, Sabine Imeri, Hubert Knoblauch, Susanne Ketzer, Christian Meyermann, Maike Porzelt, Marc Ritterberger, Jörg Strübing, Hella von Unger & René Wilke (Hgs.), Archivierung und Zugang zu Qualitativen Daten. RatSWD Working Paper 267/201, 91–100. https://doi.org/10.17620/02671.35.

Watteler, Oliver & Katharina E. Kinder-Kurlanda (2015): Anonymisierung und sicherer Umgang mit Forschungsdaten in der empirischen Sozialforschung. In: Datenschutz und Datensicherheit 39/8, 515–519. https://doi.org/10.1007/s11623-015-0462-0.

Watters, Elizabeth C., Sara Cumming & Lea Caragata (2018): The Lone Mother Resilience Project. A Qualitative Secondary Analysis. In: Forum Qualitative Social Research 19/2. https://doi.org/10.17169/fqs-19.2.2863.

Weller, Katrin & Katharina Kinder-Kurlanda (2017): To Share or Not to Share? Ethical Challenges in Sharing Social Media-Based Research Data. In: Michael Zimmer & Katharina Kinder-Kurlanda (Hgs.), Internet Research Ethics for the Social Age. New York: Peter Land, 115–129.

## Autor:inneninformation

Lina Franken, Dr. phil., ist Professorin für Digital Humanities in den Kulturwissenschaften an der Universität Vechta und vertrat zuvor die Professur für Computational Social Sciences an der LMU München. Studium der Volkskunde, neueren Geschichte und Medienwissenschaft in Bonn, Promotion in der Vergleichenden Kulturwissenschaft Regensburg. Forschungsschwerpunkte: Methodologie und digitale Methodenentwicklung, Technisierung und Digitalisierung in Alltag und Wissenschaft, Bildungskulturen und -politik, Immaterielles Kulturerbe, Arbeits- und Nahrungskulturen. Publikationen zu digitalen Methoden für qualitative Forschung (UTB Verlag 2023) und Unterrichten als Beruf (Campus Verlag 2017).

Nils Egger, M.A., ist wissenschaftlicher Mitarbeiter am Zentrum für interdisziplinäre Risiko- und Innovationsforschung an der Universität Stuttgart. Studium der Soziologie und Volkswirtschaftslehre sowie Environmental Studies in Zürich und München. Forschungsschwerpunkte: Soziologie der Digitalisierung, Soziologie der sozial-ökologischen Nachhaltigkeit, Critical Computational Studies und partizipativ-transformative Methodologien.

Luis Fischer, B.A., Studium der Soziologie und Politikwissenschaft an der Ludwig-Maximilians-Universität München. Von 2021 bis 2022 studentische Hilfskraft am Lehrbereich Computational Social Sciences des Instituts für Soziologie. Er interessiert sich besonders für die Forschung zu Methoden der quantitativen und qualitativen Sozialforschung, auch in ihrer Kombination.

Katharina Lillich, B.A., ist Studentin der Soziologie und Informatik an der Ludwig-Maximilians-Universität München und war von 2021 bis 2022 studentische Hilfskraft am dortigen Lehrbereich Computational Social Sciences.

Florian Schmid, B.A., ist Associate Consultant an einer international tätigen Unternehmensberatung mit den Schwerpunkten IT und Management und war zuvor studentische Hilfskraft am Lehrbereich Computational Social Sciences des Instituts für Soziologie in München. Studium der Soziologie und Informatik an der Ludwig-Maximilians-Universität München. Forschungsinteressen: Computational Social Science, Text Mining, Spieltheorie und Attraktivitätsforschung.