Alters-, Perioden- und Kohorteneffekte in Stimmung und Schwerpunkt literarischer Werke

Abschlussbericht

Nicole Schwitter

29. August 2023

1 Einleitung

Literatur nimmt in Deutschland einen hohen Stellenwert ein: Laut einer Befragung von Splendid Research (2017) (mit $n = 1031$) lesen 61 Prozent der Deutschen regelmäßig und pro Jahr erscheinen rund 70000 Bücher neu (Weidenbach und Statista 2021). Lesen (und Schreiben) vergrößert nicht nur den Wortschatz (Wasik u. a. 2016), schützt vor Demenz (Hughes u. a. 2010) und steigert die Konzentration (Sörman u. a. 2018), sondern dient auch dazu, die Welt um einen herum besser zu verstehen oder in andere Welten zu fliehen (Begum 2011; Howard 2011). Worüber gelesen und geschrieben wird, ändert sich über die (Lebens-)Zeit: Es gibt spezielle Kinder- und Jugendbücher, und Rea (2020) fand Unterschiede in den Genre-Präferenzen von Erwachsenen verschiedener Generationen. Weiter ist die Literaturgeschichte generell von Epochen geprägt, in denen bestimmte Themen und Stile besonders präsent sind: Seit 1900 wird literarisch zwischen mehr als zehn Epochen wie dem Im- und Expressionismus, der neuen Sachlichkeit, der Trümmer- und Nachkriegsliteratur und weiteren unterschieden.

An diesem Punkt knüpft dieses Forschungsvorhaben an, welches die zeitliche Komponente von thematischen Trends und Stimmungen genauer untersucht. Zeitliche Unterschiede können Effekte unterschiedlicher Zeitperioden, altersspezifische Veränderungen oder Generationenunterschiede widerspiegeln, wie sie in der demografischen Forschung typischerweise unterschieden werden (Diekmann 2004). Die typische Epochendefinition der Literaturwissenschaft stellt dabei auf Periodeneffekte ab. In diesem Forschungsprojekt werden Schwerpunkte in der Literatur in Bezug auf Stimmung und Kernkonzept über die Zeit hinweg untersucht. Ähnlich wie Martins und Baumard (2020), die die Häufigkeit von Wörtern in englischen und französischen Theaterstücken vor und nach Revolutionen untersuchen, fragt dieses Forschungsvorhaben, welche Kernkonzepte und Stimmungen in der deutschen Literatur seit 1913 über die Zeit hinweg besonders präsent sind. Im Gegensatz zum Epo-

chenbegriff werden keine Autoren[1] als besonders repräsentativ definiert und keine Minderheitenliteratur ausgegrenzt, sondern es soll ein vollständiger quantitativer Überblick geschaffen werden über sowohl Belletristik als auch Sachliteratur. Dafür wird der Katalog der Deutschen Nationalbibliothek (DNB) genutzt: Der Sammelauftrag der DNB erklärt das Ziel, dass der Katalog alle Medienwerke in Schrift, Bild und Ton, die seit 1913 in und über Deutschland oder in deutscher Sprache veröffentlicht wurden, enthalten soll. Der Datenkorpus des Katalogs eignet sich daher ideal als Datengrundlage. Die Daten, die Fischer (1998: 556), Herausgeber von Hansers *Sozialgeschichte der Literatur*, vermisste, gibt es durch den Bibliothekskatalog: "Um eine Sozialgeschichte der Nachkriegsliteratur zu erarbeiten [...], wäre [...] die Beschäftigung mit der realen Buchproduktion und -konsumption – und nicht nur mit den 'bemerkenswerten' literarischen Neuerscheinungen – auf der Basis zuverlässiger Daten vonnöten. Dazu fehlen die einfachsten Materialgrundlagen" (zitiert nach Häntzschel u. a. 2009: 17).

Dieses Forschungsvorhaben konzentriert sich nicht nur auf das Erscheinungsjahr eines Werkes, sondern auch auf mögliche Alterseffekte der Autoren, die in den Literaturwissenschaften mit Ausnahme von vertieften Auseinandersetzungen mit den Lebenswerken von im Kanon als wichtig erachteter Autoren – wie beispielsweise Johann Wolfgang von Goethe (Witte u. a. 1996) oder Friedrich Schiller (Alt 2000) – generell wenig Beachtung finden. Einstellungen und Verhalten von Autoren können sich mit dem chronologischen Alter verändern, sodass vorstellbar ist, dass das Schreiben von der aktuellen Lebensphase geprägt wird. Gleichzeitig ist aber auch zu vermuten, dass verschiedene Geburtsjahrgänge spezifische Einstellungsmuster aufweisen, die sich in ihrem Schreiben äußern, beispielsweise weil Personen unterschiedlicher Kohorten im Kontext anderer zeitgeschichtlicher Ereignisse und gesellschaftlicher Bedingungen unterschiedlich sozialisiert wurden. In der Demografieforschung wird generell angestrebt, die spezifischen Einflüsse von Alter, Kohortenzugehörigkeit und zeitgeschichtlichen Faktoren zu trennen. Dieser Ansatz wird in diesem Forschungsvorhaben verfolgt, um folgende Forschungsfrage mit Daten des DNB-Katalogs zu beantworten:

- Inwieweit lassen sich thematische Schwerpunkte und Stimmungen in der deutschen Literatur durch Perioden-, Alters- und Kohorteneffekte erklären?

Diese Forschungsarbeit möchte insbesondere versuchen, die von Sozialwissenschaftlern ungenutzte Goldmine der Daten der DNB abzugreifen. Das Projekt versteht sich positioniert zwischen den Digital Humanities (DH) bzw. der digitalen Literaturwissenschaft und der literaturwissenschaftlichen Statistik auf der einen Seite und den Computational Social Sciences und der Demografie auf der anderen. Der Bericht ist wie folgt gegliedert: Im nächsten Abschnitt wird bisherige Forschung im gleichen methodischen oder inhaltlichen Bereich diskutiert. Abschnitt 3 beschreibt die Methodik des vorliegenden Forschungsprojekts und erläutert die genutzten Daten, deren Aufbereitung und den statistischen Auswer-

[1] Zur besseren Lesbarkeit wird in diesem Bericht das generische Maskulinum verwendet. Die in diesem Bericht verwendeten Personenbezeichnungen beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Geschlechter.

tungsansatz. Danach werden die Ergebnisse präsentiert und Probleme der Datengrundlage diskutiert. Der letzte Abschnitt diskutiert die Resultate, die Limitation der Analysen und die Chancen und Herausforderungen, die mit der Nutzung des DNB-Katalogs für (sozial-)wissenschaftliche Zwecke einhergehen.

2 Forschungsstand

Dieser Abschnitt diskutiert einerseits bisherige Forschung, die mit Katalogdaten von Bibliotheken gearbeitet hat und dabei einen quantitativen DH-Ansatz genutzt hat. Andererseits wird die Forschung zusammengefasst, die sich mit der Untersuchung der zeitlichen Komponente von literarischen Themen und Stimmungen beschäftigt hat. Durch die Zusammenführung der unterschiedlichen Ansätze wird ein umfassender Einblick in das Forschungsfeld gewährleistet.

2.1 Nutzung der Katalogdaten in der bisherigen Forschung

Die Katalogdaten der DNB sind vergleichsweise ungenutzt und die literaturwissenschaftliche Statistik generell ist nur schwach vertreten im germanistischen Methodenkanon (Häntzschel u. a. 2009: 17). Im internationalen Kontext diskutieren Torres-Salinas und Arroyo-Machado (2020) und Torres-Salinas und Moed (2009) die *Library Catalog Analysis* explizit als Methode, in der bibliometrische Verfahren auf Katalogdaten angewandt werden. Torres-Salinas und Moed (2009) beschreiben in ihrer Studie mit Hilfe von Katalogdaten wissenschaftliche Disziplinen (konkret: die Ökonomie) und ihre Akteure. Blackburn und Heppler (2022) nutzen Daten zu den *Library of Congress Subject Headings* um zu verstehen, wie Werke über Frauen in den Naturwissenschaften beschrieben werden. Weiter werden die Katalogdaten der Library of Congress von Schmidt (2017) auch im Rahmen eines Blogposts diskutiert. International gibt es somit einige Arbeiten, die Bibliotheksdaten nutzen.
Im deutschen Kontext präsentieren Häntzschel u. a. (2009) die fiktionale Literatur der fünfziger Jahre in ihrer ganzen Vielfalt und stützen sich dabei auf Daten der DNB (in Frankfurt am Main) und der Deutschen Bücherei in Leipzig. Häntzschel u. a. (2009: 2) gehen ebenfalls davon aus, dass die Literatur von einer Heterogenität geprägt ist, man sich aber in der Literaturwissenschaft lediglich auf einen engen Kanon konzentriert. In ihrer Arbeit diskutieren sie unterschiedliche Dimensionen statistischer Methodik und exemplarische Untersuchungen zu Verlegern, Autoren, Textsorten und Verlagsstrategien.
Der in dieser Arbeit verfolgt Ansatz ähnelt in mehrerer Hinsicht dem Projekt von Fischer und Jäschke (2018), die ein Framework präsentieren, mit dem der DNB-Katalog untersucht werden kann – ebenfalls mit der Prämisse, dass es – aus geisteswissenschaftlicher Sicht – bis anhin sehr wenig Versuche gab, die Katalogdaten wissensproduktiv zu nutzen. Diese Arbeit fällt in die gleiche Methoden-Tradition und wendet den DNB-Katalog auf demografische Fragestellungen an. Die Arbeit von Fischer und Jäschke (2018) konzentriert sich auf die rund 180'000 Romane in der DNB, während die vorliegende Arbeit sich auf alle Titeldaten

stützt; es werden in dieser Arbeit keine Genreunterschiede gemacht, aber es wird versucht, Belletristik und Sachliteratur zu unterscheiden. Fischer und Jäschke (2018) untersuchen einerseits den literarischen Textumfang über die Zeit und andererseits die Entwicklung von Romantiteln auf Basis häufig genannter Worte. Die wissenschaftliche Betrachtung von Werktiteln wird im nächsten Unterabschnitt weiter diskutiert.

2.2 Was ist in einem Titel?

Für dieses Forschungsvorhaben sollen die Stimmung und die thematischen Schwerpunkte in der deutschen Literatur seit 1913 ausgewertet werden. Die Datengrundlage formen dabei die frei zugänglichen Metadaten im Bibliothekskatalog der DNB. Um die Forschungsfrage zu beantworten, stütze ich mich auf Titel der Bücher. Buchtitel decken zwar nicht den gesamten Inhalt eines Buches ab, erlauben aber einen anderen Blick in das literarische Feld: "half sign, half ad, the title is where the novel as language meets the novel as commodity" (Moretti 2009: 135). Bisherige Forschung hebt die spezielle Rolle von Titel weiter heraus: Er hat metasprachliche, poetische, phatische, appellative, Ausdrucks- und Referenzfunktionen (Rothe 1986; siehe auch Genette und Crampé 1988). Ein Titel muss das Interesse des Lesers wecken und über Inhalt, Gestalt und Gattung aufklären; er steht daher klar in Verbindung mit dem Text, den er repräsentiert. Laut Genette und Crampé (1988) wird ein Titel von Autoren erst gewählt und erst danach der Text geschrieben.
Levin (1977) argumentiert, dass Titel ein Werk definieren und eine Beziehung zwischen Text und Leser herstellen; der Titel selbst ist ohne Text aber bedeutungslos, da dieser die Geschichte erzählt. Er ist aber ebenso als Teil des literarischen Werks zu verstehen (Wilsmore 1987). Shevlin (1999) diskutiert, wie der Titel Pretext, Subtext und/oder Kontext geben kann und im Verhältnis zwischen Verlag, Leser und Schreiber steht. Titel müssen im Kontext ihrer Zeit verstanden werden, wie auch Sullivan (2007) betont, und sie haben sich über die Jahrhunderte hinweg verändert (diskutiert wird die Entwicklung von Titeln in der englischsprachigen Literatur). Bezogen auf den deutschen Kontext untersuchen Fischer und Jäschke (2018) die Entwicklung von Romantiteln über die Zeit und betrachten dafür die am häufigsten vorkommende Worte und Trigramme.
Moretti (2009) stützt sich in seiner Untersuchung auf Fachbibliografien statt auf Katalogdaten und konzentriert sich auf die britische Literatur. In seiner beschreibenden Analyse weist er auf Veränderung der Länge von Titeln über die Zeit, spezifische Titelarten und -formulierungen, aber auch darauf hin, dass Titel Genrezugehörigkeit signalisieren können. In einer aktuellen Arbeit analysieren Mozaheb u. a. (2022) Buchtitel einer ausgewählten Autorin, Agatha Christie. Sie diskutieren wie Titel übersetzt werden um repräsentativ über den Inhalt zu bleiben und Kaufinteresse zu generieren.
Insgesamt nehmen Titel eine spezifische Rolle ein und decken spezielle Informationen ab, die ich in diesem Forschungsprojekt extrahieren möchte. Breiter gedacht fällt diese Arbeit in den Bereich der natürlichen Sprachverarbeitung (NLP); ein Datenverarbeitungsansatz, bei dem natürliche Sprache algorithmisch verarbeitet wird. Durch die fortschreitende Di-

gitalisierung wurden NLP-Projekte in den letzten Jahren zunehmend populär, auch über die Fachgrenzen der Informatik hinaus. In einer Vielzahl von Beispielen wird Text automatisiert verarbeitet: Verfahren des Topic Modellings wurden angewandt, um umfangreiche digitalisierte Zeitungsarchive in Bezug auf Trends und Muster über die Zeit zu untersuchen (Jacobi u. a. 2016), Sentimentenanalyse wurde auf Twitter angewandt, um damit terroristische Tweets zu entdecken und Terrorangriffe vorherzusagen (Al-Shaibani und Al-Augby 2022), und Twitter-Textdaten wurden analysiert, um die Polarisierung im Parlament abzubilden (Green u. a. 2020). Der Versuch, NLP-Algorithmen auf Buchtitel anzuwenden ist jedoch ein Novum.

2.3 Zeitliche Veränderung von Literatur: Alters-, Kohorten und Periodeneffekte

Literatur erlebt Veränderungen über die Zeit – in der klassischen Literaturwissenschaft unterscheidet man verschiedene literarische Epochen, stellt also insbesondere auf *Periodeneffekte* ab. Zeitliche Unterschiede können Effekte unterschiedlicher Zeitperioden, aber auch altersspezifische Veränderungen oder Generationenunterschiede widerspiegeln (Diekmann 2004).

Das Konzept *Periode* bezieht sich auf den Zeitpunkt einer relevanten Beobachtung, was sich je nach Fragestellung unterscheiden kann (z.B. Beobachtungsjahr, Heiratsjahr, Publikationsjahr). Die Periode bezieht sich auf einen Zeitabschnitt mit bestimmten Umweltfaktoren und Ereignissen. Ein Periodeneffekt zeichnet sich durch eine zeitspezifische Regelmäßigkeit im Handeln aller Altersgruppen und Kohorten aus. *Kohorte* wird als Zeitpunkt des relevanten Ausgangsereignisses erfasst; bei der in diesem Projekt relevanten Kohorten ist dies etwa das Geburtsjahr. Zu einer Geburtskohorte gehören alle Personen, die im gleichen Jahr (bzw. im gleichen Intervall) zur Welt kamen. Von einem Kohorteneffekt wird beispielsweise dann gesprochen, wenn ein bestimmtes Ereignis eine bestimmte Kohorte in einem bestimmten Alter so sehr prägt, dass dies sich auf den restlichen Lebensverlauf auswirkt und diese Kohorte sich deswegen von anderen unterscheidet. *Alter* versteht sich sodann als Unterschied zwischen Periode und Kohorte. Alterseffekte können auf das biologische Alter, auf die Lebensphase oder auf Abfolge von bestimmten Phasen zurückgeführt werden (Döring 2018).

Wie Charakteristika der Autoren die geschriebenen Bücher beeinflussen, ist, abgesehen von Analysen einziger, als wichtig erachteter Autoren, relativ unerforscht. Die Unterscheidung von Alters-, Kohorten und Periodeneffekten hat in der Literaturwissenschaft bisher keine Rolle gespielt. Anders in der sozialwissenschaftlichen bzw. demografischen Forschung: Farkas (1977) unterscheidet Alters-, Kohorten- und Periodeneffekte in Bezug auf die Erwerbstätigkeit von Frauen, Park u. a. (2016) untersuchen Suizidraten in Korea und Grasso (2014), Smets und Neundorf (2014) und Stegmueller (2014) unterscheiden die Effekte dieser Zeitdimensionen auf politische Partizipation von Personen.

Alters-, Perioden- und Kohorteneffekte sind eng miteinander verbunden und auf der operationalen Ebene überdeterminiert. Sie sind empirisch schwer auseinanderzuhalten. Dies wird im methodischen Abschnitt 3.3 weiter diskutiert.

3 Daten und Methoden

Im folgenden Abschnitt werden die Daten und das generelle methodische Vorgehen dieses Projekts erläutert. Dabei wird ein besonderer Schwerpunkt darauf gelegt, aufgetretene Probleme und Schwierigkeiten in der Umsetzung zu erläutern. In einem ersten Unterabschnitt werden die genutzten Daten und der technische Zugriff vorgestellt. Anschließend werden die Datenaufbereitung und die Zuweisung von Stimmungen und Themen diskutiert, bevor auf die statistischen Ansätze zur zeitlichen Auswertung eingegangen wird. Alle Skripte sind öffentlich auf GitHub verfügbar: https://github.com/nschwitter/DNB-DH. Ich verwende in diesem Bericht die folgende Terminologie: Ein *Datensatz* bezeichnet die Gesamtheit von Daten in einem bestimmten Zusammenhang (der Datenbestand). Eine *Beobachtung* ist eine Zeile im Datensatz, also ein bestimmter Eintrag. Die Katalogdaten der DNB entsprechen damit einem Datensatz und jedes Werk im Katalog ist eine Beobachtung.

3.1 Datenquellen

Für die Beantwortung der vorliegenden Fragestellung wurden verschiedene Datenquellen genutzt, die im Folgenden vorgestellt werden. Es wird zudem erläutert, wie auf die Daten zugegriffen wurde, bevor im nächsten Abschnitt auf die Datenaufbereitung eingegangen wird. Generell nutzt dieses Forschungsprojekt die kompletten Katalogdaten, die um Daten der Gemeinsamen Normdatei und Wikidata erweitert wurden. Ziel war es, für alle Autoren aller Titel das Geburtsjahr zu erheben.

3.1.1 Titeldaten der DNB

Die vorliegende Arbeit stützt sich auf den Gesamtabzug der Titeldaten der DNB[2], Stand Oktober 2022. Durch den Sammelauftrag der DNB repräsentieren die Katalogdaten die literarische, wissenschaftliche und musikalische Produktion Deutschlands seit 1913 umfassend. Die Daten wurden in ihrer MARC21-XML-Repräsentation bezogen. MARC21 ist ein bibliografisches Datenformat, das von der Library of Congress entwickelt wurde und in der Bibliotheks- und Informationswissenschaft weit verbreitet ist, da es als international anerkannter Standard für die Organisation und den Austausch bibliografischer Daten fungiert. Intern werden die Daten der DNB jedoch im PICA-Format verarbeitet, sodass Informationen im Katalog teilweise reichhaltiger sind als Informationen im MARC21-Datensatz. Generell besteht das MARC21-Format aus Feldern, die Informationen zu verschiedenen

[2]Siehe https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Gesamtabzuege/gesamtabzuege_node.html.

Aspekten einer bibliografischen Einheit wie Titel, Autor, Verlag und Erscheinungsjahr enthalten. Jedes Feld hat eine eindeutige Kennung und eine definierte Struktur, die es ermöglicht, bibliografische Informationen in einem standardisierten Format zu speichern und auszutauschen.

Für die Auswertungen wurden folgende Felder und Informationen aus den MARC21-Daten extrahiert und in einem Tabellenformat gespeichert: Das Leaderfeld für die Satzkennung (Tag 000), die Identifikationsnummer eines Werkes (Tag 001), ein Feld mit fester Länge zur physischen Beschreibung (Tag 008), die ISBN (Tag 020, Code a), den Sprachcode des Werkes (Tag 041, Code a) und des Originals (Tag 041, Code h), der Ländercode der veröffentlichenden/herstellenden Stelle (Tag 044, Code c), die Dewey-Dezimalklassifikation (DDC; Tag 082, Code a), andere Klassifikationsnummern wie die Sachgruppen der Deutschen Nationalbibliographie (Tag 084, Code a), den Titel und gegebenenfalls Untertitel (Tag 245, Codes a und b), die Auflage (Tag 250, Code a), Angaben zum Publikationsjahr (sowohl vom veralteten Tag 260, Code c als auch vom später eingeführten Tag 264, Code c), Angaben zu weiteren Titeln (Tags 700 und 776, Code t), und verschiedene Schlagwörter, die allgemeine Hinweise in Form von GND-Schlagwörtern gemäß Schlagwortkatalog (RSWK) bieten, beispielsweise in Bezug auf Inhalt, Zeitbezug oder Geografika (Code a in folgenden Tags: 648, 650, 651, 653, 655, 688).

Weiter wurden Informationen zu allen am Werk involvierten Personen erhoben (Tag 100 und 700), konkret ihr Name (Code a), ihr Funktionsbezeichner (Code 4), ihre Normdatensatz-Identifikationsnummer bzw. die Identifikationsnummer für die GND (Code 0) und ihre Lebensdaten (Code d). Die Erhebung und Auswertung aller involvierten Personen verlief nicht ganz unproblematisch, da es einige Fehler, Unregelmäßigkeiten und Ausnahmen im MARC21-Datensatz gibt. Teilweise sind die Angaben nicht sauber, indem beispielsweise auch das Geburtsjahr mit ins Namensfeld geschrieben wurde[3] oder (nur) die GND-ID[4]. Das – eigentlich nicht wiederholbare – Feld 100 wird in einigen Fällen wiederholt; teilweise mit neuen Informationen, sodass also mehrere Personen als Haupterschaffer angegeben werden[5], teilweise mit redundanten Informationen, sodass die gleiche Person mehrfach aufgeführt wird[6]. Ich habe mich dazu entschieden, das erste Erscheinen einer Person zu erheben und Wiederholungen zu ignorieren; eine Person wird dabei mit ihrem Personennamen, also mit dem Feld mit Code a identifiziert. Teilweise werden in einem Personenfeld mehrere unterschiedliche Personen aufgeführt[7]. Zum weiteren Datenabgleich und zur Kontrolle wurde auch die weitere Verfasserangabe erhoben (Tag 245, Code c).

Technisch wurde die Vorverarbeitung und Konvertierung mit *Python* und dem Paket *lxml* vorgenommen. Ein eintragsweises Einlesen der "Records" erlaubte das Verarbeiten der sehr umfangreichen Datenmenge mit einem nicht unüblich großen Arbeitsspeicher. Die Infor-

[3] Beispielsweise https://d-nb.info/1002637783.
[4] Beispielsweise https://d-nb.info/920293727.
[5] Beispielsweise https://d-nb.info/202129772.
[6] Beispielsweise https://d-nb.info/57942992X.
[7] Beispielsweise https://d-nb.info/359798268 oder https://d-nb.info/1202658164.

mationen wurden extrahiert und in eine .csv-Tabelle gespeichert, die einerseits schlanker ist und andererseits einfacher und flexibler weiterzuverarbeiten als die Daten im MARC21-Format.

3.1.2 Gemeinsame Normdatei

Die Gemeinsame Normdatei (GND) ist ein Dienst, um Normdaten zu verwalten und kooperativ – von der DNB, Bibliotheksverbünden, Archiven, Museen und weiteren Institutionen – zu nutzen. In ihr werden Entitäten wie Personen, Geografika oder Werktitel beschrieben. Für dieses Forschungsprojekt dient die GND als Quelle von im Katalog eventuell nicht angegebenen Lebensdaten von Autoren.
Jede Entität erhält in der GND einen eindeutigen und stabilen Bezeichner (GND-ID). Diese ID erlaubt es, die Normdaten mit anderen Datensätzen – wie zum Beispiel den Katalogdaten – zu verknüpfen. Der Gesamtabzug der GND mit Stand Oktober 2022, der Personen beschreibt, wurde für dieses Projekt heruntergeladen[8]. Der Datensatz wurde sodann auf gleiche Weise vorverarbeitet wie die Titeldaten. Extrahiert wurden die folgenden Felder: Das Leaderfeld (Code 000), die GND-ID (Tag 001), der Personenname (Tag 100, Code a), alternative Namen der Person (Tag 400, Code a) und die Lebensdaten der Person (Tag 100, Code d). Für einen Datenabgleich und Prüfung der Datenqualität wurden auch der Name und die Lebensdaten von weiteren mit der Beobachtung assoziierten Personen erhoben (Code 700, Tags a und d) – die Erwartung dabei war, dass diese Felder keine neuen Informationen bringen, sondern entweder leer sind oder die Informationen aus dem 100er Feld wiederholen (die Erwartung hat sich als korrekt erwiesen).
Die Personendaten-GND beinhaltet 5'967'059 Beobachtungen. Es ist wichtig, hervorzuheben, dass auch die GND keine perfekt gepflegte Wissensdatenbank ist; auch in der GND gibt es Dubletten, sodass dieselbe Person mehrere GND-IDs aufweisen kann, und Lücken.

3.1.3 Wikidata

Wikidata wurde als weitere Quelle herangezogen, um Lebensdaten der Autoren zu ergänzen. Wikidata ist eine frei bearbeitbare Wissensdatenbank. Sie dient unter anderem dazu, Wikipedia zu unterstützten und allgemeingültige Daten für die Wikimedia-Projekte bereitzustellen. Wie andere Projekte von Wikimedia basiert auch Wikidata auf dem Beitrag von Freiwilligen; sie ist eine nutzergenerierte Datenbank.
Während frühere Forschung (Fischer und Jäschke 2018) den kompletten Wikidata-Datensatz heruntergeladen hat um ihn mit den Titeldaten zu vereinigen, ist das aufgrund des enormen Wachstums von Wikidata nicht mehr durchführbar. Stattdessen wurden die benötigten Informationen mittels einer API-Anfrage (Application Programming Interface) abgefragt. Alle Autoren, für die kein Geburtsjahr im Katalog angegeben ist und die auch keine Angabe in der GND haben, und die entweder eine GND haben oder deren Namen nicht mehrfach

[8] Siehe https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Gesamtabzuege/gesamtabzuege_node.html.

mit und ohne GND in den Titeln vorkommt (da dies als Indiz genommen wird, dass es sich um mehrere unterschiedliche Personen handelt), wurden auf Wikidata nachgeschlagen. Dies betraf 1'365'115 Personen.
Technisch wurde mit der Bibliothek *requests* gearbeitet und eine API-Anfrage an die Mediawiki-API gesendet. Wenn vorhanden, wurden die Personen anhand ihrer GND-ID auf Wikidata nachgeschlagen. War keine vorhanden, wurde der Name benutzt. Wurde auf Wikidata ein Ergebnis eines Menschen gefunden, wurde das Geburtsdatum der gefundenen Person dem Eintrag zugespielt. Wurden mehrere Menschen oder keine gefunden, wurde zur nächsten Person weitergegangen. Das Nachschlagen aller 1'365'115 Personen dauerte mehrere Tage.

3.2 Datenaufbereitung

Im folgenden Abschnitt werden die Schritte der Datenbereinigung protokolliert, getroffene Entscheidungen transparent begründet und Probleme aufgezeigt.

3.2.1 Erfassung der Titeldaten

Diese Arbeit startete mit dem Gesamtabzug der Titeldaten vom 13. Oktober 2022; dieser umfasst rund 26,7 Millionen Titeldaten. Der Gesamtabzug der Katalogdaten beinhaltet auch Werke, die für diese Analyse nicht von Relevanz sind. Die erfassten Titeldaten wurden in mehreren Schritten gefiltert und bereinigt; die einzelnen Schritte werden im Folgenden erläutert. Alle Werke haben einen Titel, was notwendig für die weiteren Analysen ist.
In einem ersten Schritt wurden alle Beobachtungen ausgeschlossen, die nicht deutschsprachig sind ($n = 10'695'369$). Ob ein Werk deutschsprachig ist, wurde festgestellt, indem überprüft wurde, ob "ger" im Sprachcode-Feld vorkommt (Tag 41, Code h; wurde "ger" in diesem Feld nicht genannt, wurde das Werk als nicht-deutschsprachig klassifiziert)[9]. Übersetzung wurden nicht ausgeschlossen; ob ein Werk eine Übersetzung ist, wurde festgestellt, indem überprüft wurde, ob das Werk eine Angabe zur Originalsprache hat (und diese nicht deutsch ist).
Diese Arbeit interessiert sich für Werke, die klar abgegrenzte, eigenständige Werke sind und die von einzelnen Autoren verfasst wurden. Im zweiten Schritt wurden daher Werke ausgeschlossen, die keine oder mehrere Autoren haben (keine oder mehrere Personen mit der Rolle "aut"; $n = 7'619'998$ haben keine Autoren, $n = 762'792$ haben mehrere Autoren). Ebenso wurden Werke ausgeschlossen, die zwar eigentlich eine Autorenangabe haben, bei der aber im Namensfeld des Autors seine Rolle als Herausgeber vermerkt ist ($n = 391$).
Weiter wurden Sammelbänder und Buchreihen ausgeschlossen ("Buchreihen" beziehen sich auf die übergeordnete Reihe; dieses Projekt interessiert sich aber für die einzelnen Wer-

[9] Aufgrund fehlender Angaben im Sprachcodefeld ist es möglich, dass dieses Vorgehen nicht ideal war und die Deutschsprachigkeit besser hätte anders bestimmt werden sollen, um weniger Beobachtungen auszuschließen.

ke[10]). Es wurden daher Werke ausgeschlossen, die im Leaderfeld – das relevante Metadaten enthält – an Position 7 (8. Zeichen) den Wert "b" haben, der einen unselbständigen Teil eines fortlaufenden Sammelwerks auszeichnet, "c", was eine Sammlung auszeichnet, oder "s", was ein fortlaufendes Sammelwerk beschreibt. Zudem wurden Werke ausgeschlossen, die an Position 6 (7. Zeichen) des Felds 008 den Wert "c" oder "d" haben, da dies auf (nicht) abgeschlossene, fortlaufende Ressourcen hindeutet, oder "i", da dies eine Sammlung auszeichnet. So wurden insgesamt $n = 18'688$ Sammelbände und Buchreihen identifiziert, die ausgeschlossen wurden. Das Publikationsjahr im Feld 264 enthält weitere Details, die zum Ausschluss von Buchreihen herangezogen werden: Wenn das Publikationsjahr ein "–" enthält (beispielsweise "1980–1987"), weist dies darauf hin, dass mit dem Werk mehrere Publikationsjahre assoziiert sind bzw. das Werke eine Reihe darstellt[11]. Auch diese Fälle wurden ausgeschlossen ($n = 35'204$). Zusätzlich wurde noch geprüft, welche Einträge weitere Titelangaben haben (Tag 700, Code t)[12]. Wenn mehr als ein weiterer Titel angegeben ist, wurden die Werke ausgeschlossen ($n = 2503$). Weiter wurde jede Art von Software ausgeschlossen, indem nach der Produktform "CD-ROM" oder "DVD-ROM" in einem der Schlagwortfelder gesucht wurde ($n = 3121$, Code 653, Tag a).

Der so reduzierte Datensatz enthält alle Werke, die grundsätzlich relevant für die Analyse sind und mit weiteren Daten angereichert werden können. Die erfassten Titeldaten wurden im nächsten Schritt weiter aufbereitet, da die Daten noch nicht für die Analyse bereit sind. Insgesamt befanden sich nach diesem Schritt 7'603'120 Beobachtungen im Datensatz.

Bevor den Autoren in einem nächsten Schritt zusätzliche Informationen zu den Lebensdaten zugespielt werden konnten, wurden die Autorenfelder noch bereinigt. Jedem Autor wurde eine eindeutige ID zugewiesen. Diese ID soll verschiedene Schreibweisen eines Namens aufgreifen. In einem ersten Schritt wurde der tatsächliche Autor eines Werkes extrahiert; alle weiteren involvierten Personen sind in diesem Forschungsprojekt nicht von Interesse. Um zur eindeutigen ID zu kommen, wurde die GND verwendet: Zu einer eindeutigen GND-ID gehören auch verschiedene Namensvarianten einer Person, sodass unterschiedliche Schreibweisen abgefangen werden können. Hat ein Autor keine Angabe zur GND-ID im Titeldatensatz, wurde in der GND nach der Person gesucht; wurde eine Person mit gleichem Namen gefunden, wurde die entsprechende GND-ID der Person zugewiesen. Gab es mehrere mögliche Matches, wurde *keine* ID zugewiesen. Es ist wichtig, anzumerken, dass dieses Vorgehen eine Annäherung an die Realität ist und Annahmen trifft, die nicht richtig sein müssen (Namensgleichheit garantiert nicht, dass es sich tatsächlich um eine bestimmte Person handelt). Die GND-ID wird als eindeutiger Zuweiser im weiteren Prozess verwendet. Personen ohne GND-ID wurde eine neue, manuell erstellte ID zugewiesen. Dieser Personen-ID Datensatz wurde zudem manuell geprüft, um zu sehen, ob gleichen Personen unterschiedlichen IDs zugewiesen wurde (und solche Einzelfälle zu korrigieren) und auf andere Auffälligkeiten zu untersuchen. Insbesondere wurden Namen mit Auffäl-

[10] Beispielsweise https://d-nb.info/990414353.
[11] Beispielsweise https://d-nb.info/1209136112.
[12] Beispielsweise https://d-nb.info/1231651229 oder https://d-nb.info/1192461541.

ligkeiten in den Schreibweisen (Beginn des Namens mit Sonderzeichen, etc.) geprüft und korrigiert.

3.2.2 Anreicherung mit zusätzlichen Datenquellen

Die Titeldaten weisen teilweise Lebensdaten der Autoren auf. In den Fällen, in denen Angaben zum Geburtsjahr fehlen, wurden die GND und Wikidata herangezogen, um die Daten zu erweitern. Der Titeldatensatz wurde auf Basis der GND-ID mit der GND zusammengeführt, um fehlende Geburtsjahre zu ergänzen. Autoren, die weder in den Titeldaten noch in der GND Angaben zum Geburtsjahr haben, wurden auf Wikidata gesucht. Wenn vorhanden, wurde dafür die GND-ID der Autoren genutzt; wenn nicht, wurde nach dem Personennamen gesucht. Wurden eine Person auf Wikidata gefunden, wurden die Lebensdaten dem Autoreninfo-Datensatz zugespielt. Wurden mehrere Personen gefunden, wurden wiederum *keine* Information zugespielt, da eine eindeutige Zuordnung dann nicht möglich war.
Der Titeldatensatz enthielt nach der vorhin beschriebenen Filterung insgesamt 7'603'120 Werke von 2'583'427 unterschiedlichen Autoren. Von diesen Autoren hatten 727'979 Angaben zum Geburtsjahr (aus den Titeldaten). Diese Zahl erhöhte sich auf 933'604 durch das Anreichern mit der GND. Die restlichen Personen wurden auf Wikidata nachgeschlagen, sofern sie entweder eine GND-ID haben oder ihr Name nicht mehrfach in den Titeldaten mit und ohne GND vorgekommen ist (da dies als Indiz genommen wird, dass es sich dabei um unterschiedliche Personen handelt). Insgesamt wurden 1'365'115 Personen nachgeschlagen. 135'245 dieser 1'365'115 Personen hatten einen Match auf Wikidata, und in 85'556 Fällen wiesen sie ein Geburtsjahr auf.
Der Match auf Wikidata erfolgte über die Wikidatasuchmaske und konnte durch die Suchfunktion von Wikidata, das viele alternative Namen kennt, recht liberal ausfallen. Es wurden somit teilweise durch Wikidata bisher als unterschiedlich wahrgenommene Personen als dieselbe identifiziert, wie beispielsweise "Antonia Haugwitz" und "Rosina Topka", jedoch wurde der auch der nachgeschlagene Name "günther tiede" beispielsweise mit der Person "Gunther Tiedemann" gematcht. Es wurden daher nur exakte Matches übernommen, sowie solche, die sich nur im Rahmen eines Bindestrichs unterschieden, sowie solche, die sich stark unterschieden (ausgedrückt als hohe Levensthein-Distanz[13]), da es sich bei solchen Matches höchstwahrscheinlich um Alternativnamen handelt. Die durch Wikidata herbeigeführten Matches wurden noch stichprobenartig geprüft und gegebenenfalls korrigiert. Unsaubere Matches wurden entfernt.
Nach dem Anreichern durch zusätzliche Datenquellen wiesen 1'001'860 Autoren ein Geburtsjahr auf. Durch das Nachschlagen der Autoren auf Wikidata wurden einige Auto-

[13] Die Levenshtein-Distanz ist eine Maßzahl, die angibt, wie unterschiedlich zwei Zeichenketten voneinander sind. Sie misst die minimale Anzahl an Bearbeitungsschritten, die erforderlich sind, um eine Zeichenkette in eine andere zu transformieren, wobei diese Schritte das Einfügen, Löschen oder Ersetzen von einzelnen Zeichen beinhalten können.

ren als weitere Doppelungen identifiziert. Der Datensatz enthielt dann noch Werke von 2'579'133 unterschiedlichen Autoren, von denen 997'566 ein Geburtsjahr aufwiesen.

3.2.3 Weitere Bereinigung der Titeldaten

Nach dem Filtern der Titeldaten umfasste der Datensatz noch 7'603'120 Werke von 2'579'133 Autoren, wobei nur für 997'566 ein Geburtsjahr vorlag. Bevor Stimmungen und Kernkonzepte zu den Buchtiteln zugewiesen wurden, erfolgte noch der finale Schritt der Vorverarbeitung. In diesem letzten Schritt wurde der Datensatz auf die Werke reduziert, für die alle relevanten Informationen vorliegen, die bereinigt wurden.
In einem ersten Schritt wurde das Publikationsjahr, das bis anhin nicht vorverarbeitet wurde, bereinigt. Das Feld 008 wurde verwendet, um Informationen zum Publikationsjahr zu Erst- (Position 8 bis 11) und eventueller Neuauflage (Position 12 bis 15) zu extrahieren. Das Feld 264 beinhaltet zwar detaillierte, reichere Angaben, da es in der bibliothekarischen Erschließung genutzt wird, jedoch sind diese Details für die vorliegende Analyse irrelevant (wie beispielsweise Hinweise darauf, ob es sich um das Copyrightjahr handelt oder ein auf Monatsbasis ausgewiesenes Erscheinungsdatum). Das Feld 264 weist auch aus, wenn Unsicherheiten im Publikationsjahr bestehen. Generell zeigt ein Vergleich mit Feld 008 keine nennenswerten Abweichungen zwischen Feld 264 und der aus dem Feld 008 extrahierten Jahresangabe. Ich verwendete daher die Angabe im Feld 008 und verstehe sie als beste, wohlbegründete Vermutung.
Werke ohne gültige Angabe im Publikationsjahr wurden ausgeschlossen und unklare Angaben soweit möglich präzisiert (X werden durch "5" ersetzt, Schätzwerte werden als Angabe genommen, etc.). Teilweise wurde das Feld 008 bei der Dateneingabe nicht korrekt befüllt und Angaben zum Publikationsjahr sind nicht in an den vier Positionen angegeben, an denen sie hätten stehen sollen (beispielsweise wurde die Angabe "um 1800" in das Feld 008 übernommen, obwohl im Feld nur vier Zeichen für das Erscheinungsjahr zur Verfügung stehen)[14]. Nach diesem Bereinigungsschritt befanden sich noch 6'803'816 Werke im Datensatz.
Weiter wurden Werke mit gleichem Titel, gleichem Publikationsjahr und gleicher Autoren-ID ausgeschlossen, da es sich dabei um Duplikate handelt bzw. um Mehrfacherscheinungen im gleichen Jahr, die für das vorliegende Forschungsprojekt irrelevant sind (danach waren noch 5'976'975 Werke im Datensatz). Weiter sind nur Erstauflagen eines Werkes von Relevanz. Weitere Auflagen zeigen zwar, wofür sich die aktuelle Bevölkerung interessiert – und was gekauft wird – aber nicht, was geschrieben wird. Es wurden daher alle Beobachtungen ausgeschlossen, die sich auf spätere Auflagen beziehen. Als Neuauflage wird ein Werk verstanden, deren Neuauflagenjahr in der Position 12 bis 15 im Feld 008 nicht leer und nicht "9999" ist. Weiter wurden Werke ausgeschlossen, deren Auflage-Angaben größer als eins ist (sofern eine angegeben ist). Schließlich wurde innerhalb des Datensatzes nach Werken von der gleichen Person mit dem gleichen Titel gesucht, die mehrfach vorkommen. Von diesen

[14] Beispielsweise https://d-nb.info/579233561.

Werken wurde das mit dem frühsten Publikationsjahr behalten. Nach diesem Schritt hat sich der Datensatz auf 5'584'392 Werke reduziert.
Um Perioden-, Kohorten- und Alterseffekte zu untersuchen, ist es notwendig, dass ein Geburtsjahr des Autors vorliegt. Es wurden daher im weiteren Schritt alle Beobachtungen ausgeschlossen, die kein Geburtsjahr vorweisen (danach befanden sich noch 3'166'783 Beobachtungen im Datensatz). Es wurden zudem Beobachtungen ausgeschlossen, bei denen sich das Alter des Autors bei Erscheinung des Werkes nicht in einem bestimmten, plausiblen Rahmen bewegt, nämlich zwischen 18 und 79. Durch diesen Schritt wurden mehrere Fehlerquellen beseitigt: Einerseits wurden posthume Veröffentlichungen entfernt und andererseits half der Schritt dabei, falsch zugewiesene Geburtsjahre zu identifizieren. Zusätzlich wurden, wenn ein Sterbedatum vorhanden ist, Werke ausgeschlossen, die nach dem Sterbedatum publiziert wurden.
Nach diesem Schritt befanden sich noch 2'706'869 Beobachtungen im Datensatz, der einen Zeitraum bis 2023 abdeckt. Der DNB-Katalog enthält Werke, die vor 1913 erschienen sind, auch wenn ihr Sammelauftrag erst 1913 beginnt. Es ist wegen des Sammelauftrags anzunehmen, dass die Sammlung vor 1913 nicht vollständig ist und die Ergebnisse davor bestimmte Verzerrungen enthalten. In einem letzten Schritt wurden daher Werke, die vor 1913 veröffentlicht wurden, ebenfalls ausgeschlossen. Der Datensatz enthält sodann $n = 2'664'047$ Werke, denen Stimmung und Kernkonzept zugewiesen werden soll.

3.2.4 Zuweisung von Stimmungen und Kernkonzept

In den bisherigen Arbeitsschritten wurden Buch- und Autoreninformationen vorverarbeitet. Im nächsten Schritt wurden den Werken Stimmungen sowie Thema zugewiesen. Buchtitel sind ein besonderer Fall: Sie sind extrem kurze Texte mit sehr wenig Kontext.

Stimmungen Stimmungserkennung wird in diesem Projekt auf zwei Arten vorgenommen: Basierend auf lexikalischen Regeln und basierend auf Modellen des maschinellen Lernens. Regelbasierte Stimmungsanalyse ist ein einfacher Ansatz, bei dem ein vordefiniertes Lexikon benutzt wird, das bestimmte Wörtern einen Stimmungswert zuweist; in diesem Projekt wurde *TextBlob-de* verwendet[15].
Maschinelle Lernmodelle basieren auf gekennzeichneten Daten und sind vortrainiert. Für dieses Projekt wurde *DistilBERT* aus dem Paket *torch* verwendet. DistilBERT ist eine kleine, schnelle und leichtere Variante des starken Sprachmodells *BERT*. Trotz weniger Parameter und schnellerer Laufzeit weist es eine ähnliche Performanz auf. Es hat eine limitierte Inputlänge (wie BERT): Die maximale Länge beträgt 512 Tokens. Generell reicht dies für den Anwendungsfall von Titel aus. In den seltenen Fällen, in denen dies nicht der Fall ist, wurde der Titel auf die ersten 512 Tokens gekürzt. Das Zuweisen der Stimmungen

[15] Versuche mit Vader (Valence Aware Dictionary and sEntiment Reasoner) vom gängigen Paket *nltk* lieferten keine guten Ergebnisse. Dies kann daran liegen, dass Vader vor allem auf die Stimmungserkennung in sozialen Medien trainiert ist. Es legt deswegen viel Wert auf repetitive Wörter, Emojis und andere Eigenheiten des digitalen Diskurses.

für den Datensatz nahm etwa vier Tage in Anspruch. Allen Werken wurden somit zwei Sentiment-Scores zwischen -1 und +1 (lexikon-basiert) bzw. 0 und 1 (KI-basiert) zugewiesen. Für die Bestimmungen der Sentiments wurde der Text nur marginal vorbearbeitet; Stoppwörter etc. blieben im Text vorhanden.

Kernkonzepte Die Idee der Kernkonzepte ist es, Werke zu "Themen" zu gruppieren, die verschiedene Buchschwerpunkte repräsentieren. In diesem Projekt wurde versucht, Kernkonzepte bzw. Gruppierungen auf zwei Arten herzuleiten:

1. LDA-Methode, um Themen aus Titeln (und Schlagworten) zu modellieren.
2. Einbettungen der Titel mit Hilfe von Word Embeddings.

Die erste Herangehensweise ist eine Variante des *Topic Modellings*. Beim Topic Modelling werden Themen innerhalb eines Textkorpus identifiziert. Es wird dafür eine statistische Methode verwendet, um die Häufigkeit von Wörtern in Bezug auf bestimmte Themen zu analysieren und Muster zu erkennen. Das Ergebnis ist eine Darstellung des Textkorpus in Form von Themen, die als Gruppen von Wörtern definiert sind, die häufig zusammen auftreten. Einer der mittlerweile populärsten Verfahren des Topic Modelling ist die *Latent Dirichlet Allocation* (Blei 2003) (LDA). Bei der LDA wird die Anzahl der Themen k durch den Nutzer festgelegt. Jedes Dokument enthält mehrere, wenige Themen (die Zugehörigkeit zu einem Thema wird als Wahrscheinlichkeitsverteilung über die k Themen ausgedrückt). Jedes Wort im Dokument ist einem Thema zugeordnet. Diese Themen, deren Anzahl zu Beginn festgelegt wird, erklären das gemeinsame Auftreten von Wörtern in Dokumenten. In kurz: Jedes Dokument besteht aus einer Mischung von Themen und jedes Thema besteht aus einer Sammlung von Wörtern. Topic Models beruhen auf der Vorstellung, dass die Semantik von Dokumenten von latenten Variablen bestimmt wird. Das Ziel der Themenmodellierung besteht darin, diese latenten Variablen – Themen – aufzudecken, die die Bedeutung des Dokuments und Korpus bestimmen. Es gibt verschiedene Ansätze für Topic Modelling, doch LDA ist mittlerweile einer der populärsten. Die Titel wurden vorverarbeitet für effektiveres Topic Modelling. Störende bzw. irrelevante Wörter (insbesondere Stoppwörter und themengenerelle Wörter) wurden entfernt. Diese "irrelevante" Wörter kommen besonders häufig vor (beispielsweise bestimmte Teile der Sprache wie Artikel oder themenbezogen wie die Worte "Buch" oder "Roman") und sind daher informationsarm.
Als zweiten, alternativen Ansatz zum Topic Modelling nutzte ich vortrainierte Word Embeddings. Bei Word Embeddings handelt es sich um eine Einbettung, bei der Worte einem Vektor zugeordnet werden; Worte lassen sich sodann in einem mehrdimensionalen Vektorraum positionieren. Word Embeddings bieten eine abstrakte Darstellung der Bedeutung der Worte bei gleichzeitiger Dimensionsreduktion und berücksichtigen die semantische Bedeutung. Worte, die sich in ihrer Bedeutung ähneln (beispielsweise Synonyme oder Wörter, die sich auf verwandte Konzepte beziehen) sollten im Vektorraum nahe beieinander liegen.

Um Themen zu identifizieren, wurde zunächst die Vektorposition jedes Titels gesucht (gebildet als Durchschnitt der Worte, die im Titel vorkommen) und die Titel dann in Cluster eingeteilt. Die Cluster entsprechen dann verschiedenen Themen. Word Embeddings wurden in diesem Forschungsprojekt nicht weiter trainiert – d.h. die Position von Wörtern im Vektorraum wird nicht weiter auf den spezifischen, vorliegenden Fall angepasst –, sondern es wurden vortrainierte Word-Embeddings von *fasttext* verwendet[16].
Die ursprüngliche Idee war es, Kernkonzepte für Belletristik und Sachliteratur getrennt zu identifizieren. Dies konnte aufgrund fehlender Daten nicht durchgeführt werden und wird in Abschnitt 4 diskutiert.

3.3 Zeitliche Auswertung

Stimmungen und Kernkonzepte wurden auf zeitliche Aspekte hinweg untersucht. Die Unterscheidung von Perioden-, Kohorten- und Alterseffekten in sogenannten APC-Modellen (age-period-cohort-Modellen) ist nicht simpel, sondern wird seit mehreren Jahrzehnten kontrovers diskutiert (siehe beispielsweise Fosse und Winship 2019a,b; Luo und Hodges 2020).
Das Haupthindernis bei den APC-Modellen ist die lineare Abhängigkeit zwischen den drei Zeitskalen: $Alter = Kohorte - Periode$. Mehrere Modellierungsvarianten wurden vorgeschlagen um Punktschätzungen der Effekte zu erhalten, wie beispielsweise die Einführung von Beschränkungen (*Equality contraints*), die Verwendung von Mehrebenenmodellen oder Umwege über Proxy-Variablen. Diese Modelle leiden unter Ad-Hoc-Annahmen und Limitationen in der Robustheit gegenüber Modellspezifikationen. Generell müssen solche Analysen mit Vorsicht interpretiert werden. Um klare Schlüsse ziehen zu können braucht es eine solide theoretische Fundierung (Bell und Jones 2014, 2015). Je nach Anwendungsfall kann so beispielsweise argumentiert werden, dass der Periodentrend auf Null gesetzt wird. Bell und Jones (2015) argumentieren, dass der Mechanismus für langfristige Veränderungen besser durch Kohorten konzeptualisiert werden kann. Von der Literaturgeschichte her kommend ist dieses Argument nicht unbedingt haltbar. Generell fehlt diesem Forschungsprojekt ein solides theoretisches Fundament, da es einen explorativen Charakter hat.
In den Auswertungen wurde daher viel Wert auf deskriptive Analysen gelegt. Ich orientierte mich am beschreibenden Vorgehen in Dassonneville u. a. (2012) und Robinson und Jackson (2001) und an der Auswertungsstrategie von Weigert u. a. (2022). Für die statistische Auswertung wurde die Programmiersprache *R* und insbesondere das Paket *APCtools* (Bauer u. a. 2023) verwendet. In den modellbasierten Analysen wurde ein semiparametrischer Ansatz verfolgt, den die verallgemeinerten additiven Regressionsmodelle (GAMs) bieten. Dieser regressionsbasierte Ansatz umgeht das Problem der linearen Abhängigkeit der APC-Dimensionen, indem er eine flexible zweidimensionale Tensorproduktoberfläche schätzt. Damit trennt er die zugrundeliegenden Effekte von Alter, Periode und Kohorte von zufälligen Schwankungen in den Daten und ermöglicht die anschließende Visualisie-

[16] Siehe https://fasttext.cc/docs/en/pretrained-vectors.html.

rung von marginalen Alters-, Perioden- und Kohorteneffekten. Der Ansatz ermöglicht auch die Berücksichtigung weiterer Kontrollvariablen im Modell; da weitere, sinnvolle Daten im Bibliothekskatalog aber nicht flächendeckend vorkommen und dieses Projekt explorativen Charakter hat, wurden keine Kontrollvariablen aufgenommen. Zudem wurden keine Fixed- oder Random-Effekte aufgenommen, da die Daten zu umfangreich sind (die Schätzung dieser Effekte würde über 6200GB RAM Arbeitsspeicher benötigen). Die Mehrebenenstruktur der Daten wurde daher außen vor gelassen.

4 Datensatzbeschreibung und aufgetretene Probleme

Im folgenden Abschnitt wird einerseits der Datensatz für die darauffolgende Auswertung und andererseits Probleme, die während der Datensatzbeschreibung sichtbar wurden, beschrieben.

4.1 Publikationen über die Zeit

Der Datensatz enthält 2'664'047 Werke, die in 123 verschiedenen Jahren publiziert wurden; das älteste Werk ist, entsprechend der Datenbeschränkung, von 1913, das jüngste von 2040[17] ($Median = 1997$). Die Verteilung der Werke pro Jahr ist in Abbildung 1 abgebildet.

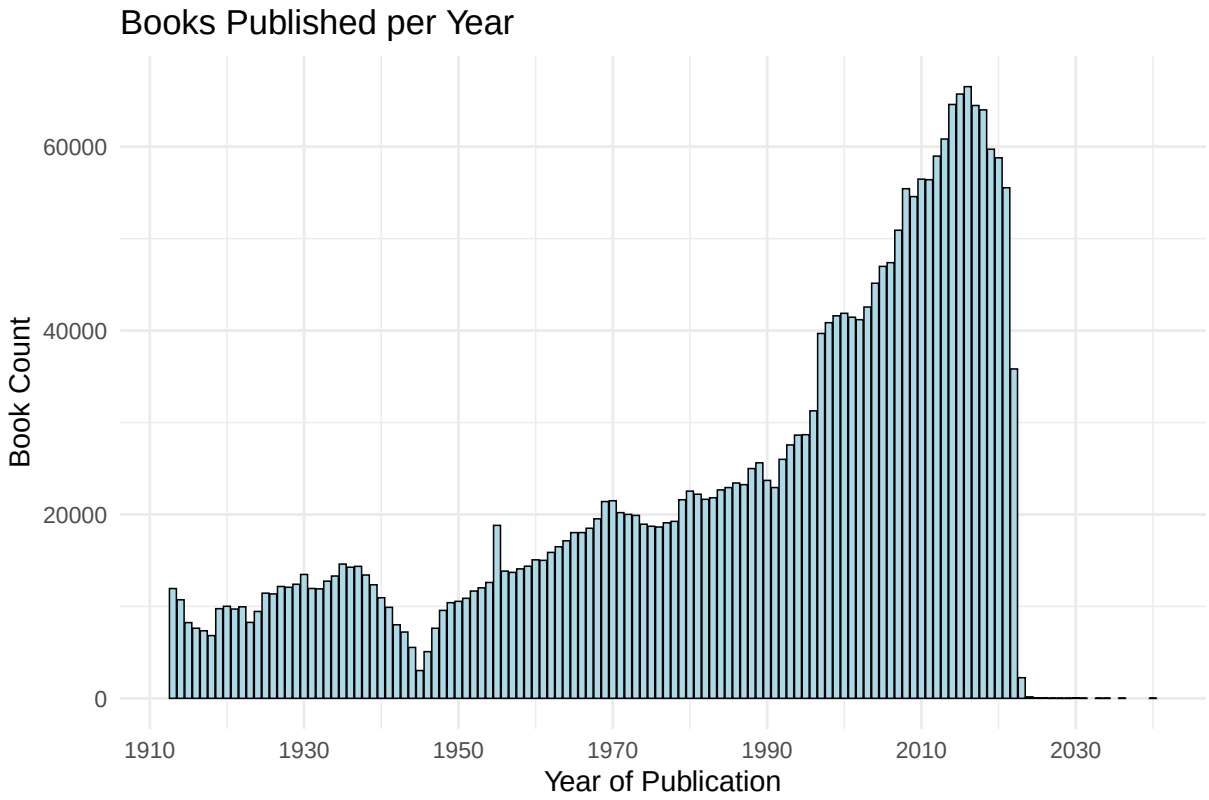

Abbildung 1: Verteilung der Publikationen über die Zeit.

[17] Dabei handelt es sich von Verlagen gemeldete Neuerscheinungen, nicht um Fehler. Siehe beispielsweise https://d-nb.info/1135948402 oder https://d-nb.info/109856605X.

Diese Werke stammen von 912'199 unterschiedlichen Autoren, die im Schnitt 46.01 Jahre alt waren zum Zeitpunkt der Veröffentlichung ($Standardabweichung(SD) = 14.46$); siehe Abbildung 2. Die Häufung junger Autoren bzw. die Veröffentlichungen in jungen Jahren ist auffallend. Dies ist ein Phänomen in einigen, wenn auch nicht allen, Zeitabschnitten. Das Modalwert-Alter – das Alter, das die meisten Autoren von Werken, die in einem bestimmten Jahr erschienen sind, hatten – von publizierenden Autoren ist in Abbildung 3 dargestellt. Vor allem seit 2010 hat das Modalwertalter stark zugenommen, lag in den Jahren davor (ab 1930) aber oft bei ca. 30 Jahren.

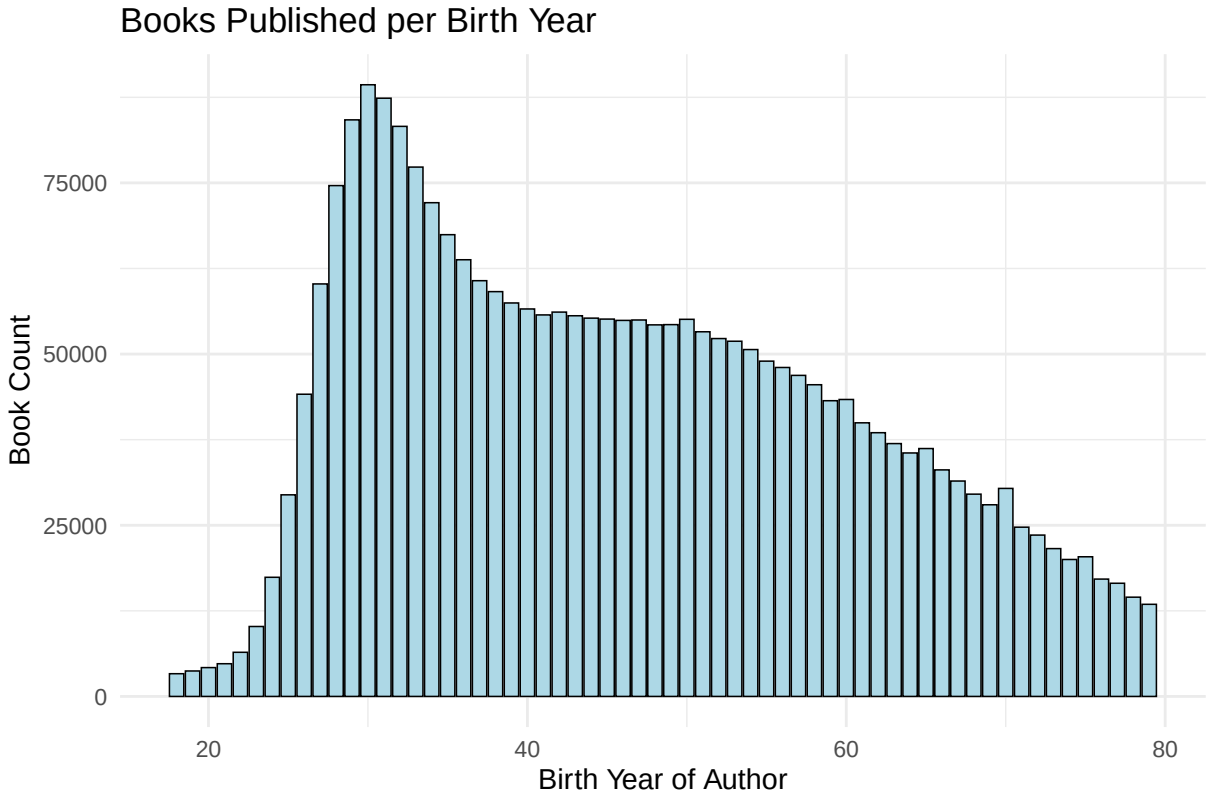

Abbildung 2: Verteilung der Publikationen über das Alter der Autoren.

Die Autoren decken Geburtsjahrgänge von 1834 bis 2004 ab ($Median = 1947$). Die Verteilung der Publikationen über die Geburtsjahre der Autoren ist in Abbildung 4 abgebildet.

Fehlende DDC-/Sachgruppenklassifizierungen Geplant war für die Analyse, Werke getrennt nach Belletristik und Fachliteratur auszuwerten. Die Datensatzbeschreibung hat aber gezeigt, dass von den 2,7 Millionen Werken rund 22%, nämlich 597'409, keine Sachgruppenklassifizierung aufweisen (weder in den Feldern 82, 83, 84, 85 noch 89). Von den Werken mit Sachgruppenklassifizierung gehören 252'087 der Belletristik an. Es ist unklar, ob ein Bias in den Daten vorliegt: Es ist möglich, dass die Werke mit Sachgruppenzuweisung eher einer bestimmten Sachgruppe angehören. Auffallend ist, dass unter den zehn häufigsten Wörtern, die in Belletristiktitel vorkommen, die beiden Wörter "Perry" und "Rhodan" fallen – möglicherweise ein Hinweis darauf, dass die Einordnung in Belletristik nicht ohne

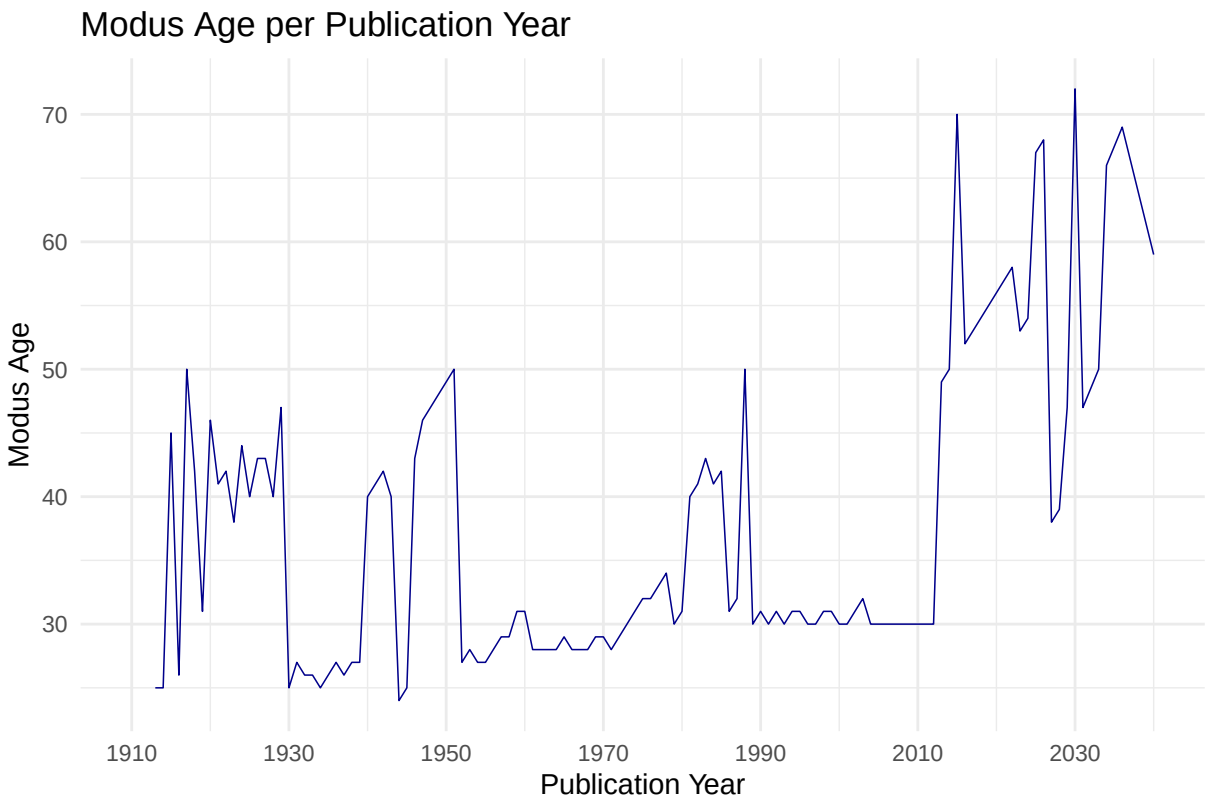

Abbildung 3: Modalwertalter der publizierenden Autoren nach Publikationsjahr.

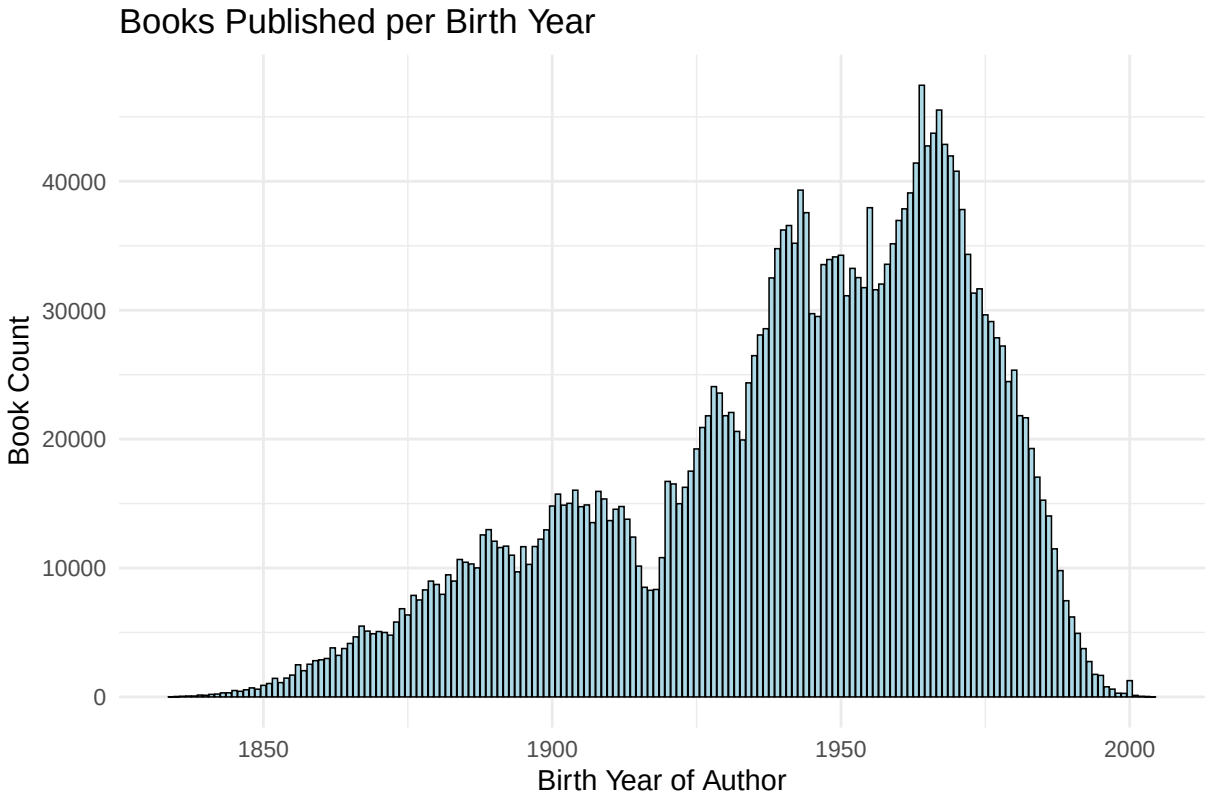

Abbildung 4: Verteilung der Publikationen über Geburtsjahr der Autoren.

Heuristik war (und die Einordnung aller Perry-Rhodan-Heftromane möglicherweise automatisiert erfolgt ist, während dies bei anderen Werken nicht so einfach möglich war). Es ist auch eine offene Frage, ob die nicht klassifizierten Werke eher der Belletristik oder eher der Sachliteratur angehören. Um keine zusätzlichen Verzerrungen in den Datensatz zu bringen, wird im Folgenden davon abgesehen, Sachliteratur und Belletristik zu unterscheiden. Dies führt jedoch zu weiteren Problemen, die in Abschnitt 6 diskutiert werden.

4.2 Titeldaten

Neben der zeitlichen Komponente ist in diesem Forschungsprojekt der Titel eines Werkes relevant, da er benutzt wird, um Stimmung und Kernkonzept abzuleiten. Titel und allfällige Untertitel wurden in der Datenverarbeitung zusammengenommen und werden im Folgenden vereinfacht als "Titel" bezeichnet. Titel sind im Schnitt 9.11 Wörter lang, die Länge streut aber sehr: Die 59'324 kürzesten Titel sind nur ein Wort lang (beispielsweise "Heimat", "Geigenliesel" oder "Goldnüsse"), der längste besteht aus 285 Wörtern[18] ($SD = 5.83$). Die Verteilung der Titellängen ist in Abbildung 5 dargestellt (für bessere Lesbarkeit beschränkt sich die linke Abbildung auf Titel mit einer Länge 30 oder weniger Wörtern). Eine manuelle Prüfung zeigt, dass die extrem langen Titel oft weitere Informationen aufweisen; diese Beobachtungen könnten als Ausreißer entfernt werden, doch da die Titellänge in den folgenden Analysen die Ergebnisse nicht direkt beeinflusst, werden sie nicht ausgeschlossen. Abbildung 6 zeigt die durchschnittliche Länge der Titel über die Jahre. Generell hat die Länge der Titel über die Zeit leicht zugenommen. Die Sprünge in den jüngeren Jahren erklären sich durch eine geringe Anzahl Beobachtungen.

Verschlagwortung im Zeitverlauf Eine Überlegung war es, die in den Metadaten angegebenen Schlagworte zu verwenden um Stimmung und Kernkonzepte abzuleiten. Die

[18] "Music History Writing and National Culture. Proceedings of a seminar Tallinn, December 1-3, 1995. Eesti Muusikaloo Toimetised I [Publications in Estonian Music History I], ed. by Urve Lippus,Tallinn (Eesti Keele Instituut) 1995; Kristel Pappel, Muusikateater Tallinnas XVIII sajandi lõpus ja XIX sajandi esimesel poolel [Musiktheater in Tallinn am Ende des 18. Jahrhunderts und in der ersten Hälfte des 19. Jahrhunderts]. Eesti Muusikaloo Toimetised 2. [Beiträge zur Geschichte der Musik in Estland, 2] Üldtoim. U. Lippus. Vihiku toim. H. Soobik. [Hrsg. U. Lippus. Hrsg. des Bandes H. Soobik]. Tallinn (Eesti Keele Instituut) 1996; Virve Lippus, Eesti pianistliku kultuuri kujunemine [Die Formierung der estnischen pianistischen Kultur], zusammengestellt von U. Lippus, (= Eesti Muusikaloo Toimetised 3 [Beiträge zur Geschichte der Musik in Estland, Band 3]) Üldtoim. U. Lippus, toim. M. Sedrik [Hrsg. U. Lippus, Hrsg. des Bandes M. Sedrik] Tallinn (Eesti Keele Instituut) 1997 [Rezension] : Music History Writing and National Culture. Proceedings of a seminar Tallinn, December 1-3, 1995. Eesti Muusikaloo Toimetised I [Publications in Estonian Music History I], ed. by Urve Lippus,Tallinn (Eesti Keele Instituut) 1995; Kristel Pappel, Muusikateater Tallinnas XVIII sajandi lõpus ja XIX sajandi esimesel poolel [Musiktheater in Tallinn am Ende des18. Jahrhunderts und in der ersten Hälfte des 19. Jahrhunderts]. Eesti Muusikaloo Toimetised 2. [Beiträge zur Geschichte der Musik in Estland, 2] Üldtoim. U. Lippus. Vihiku toim. H. Soobik. [Hrsg. U. Lippus. Hrsg. des Bandes H. Soobik]. Tallinn (Eesti Keele Instituut) 1996; Virve Lippus, Eesti pianistliku kultuuri kujunemine [Die Formierung der estnischen pianistischen Kultur], zusammengestellt von U. Lippus, (= Eesti Muusikaloo Toimetised 3 [Beiträge zur Geschichte der Musik in Estland, Band 3]) Üldtoim. U. Lippus, toim. M. Sedrik [Hrsg. U. Lippus, Hrsg. des Bandes M. Sedrik] Tallinn (Eesti Keele Instituut) 1997 [Rezension] / Kristel Pappel ". Siehe https://d-nb.info/1240846312.

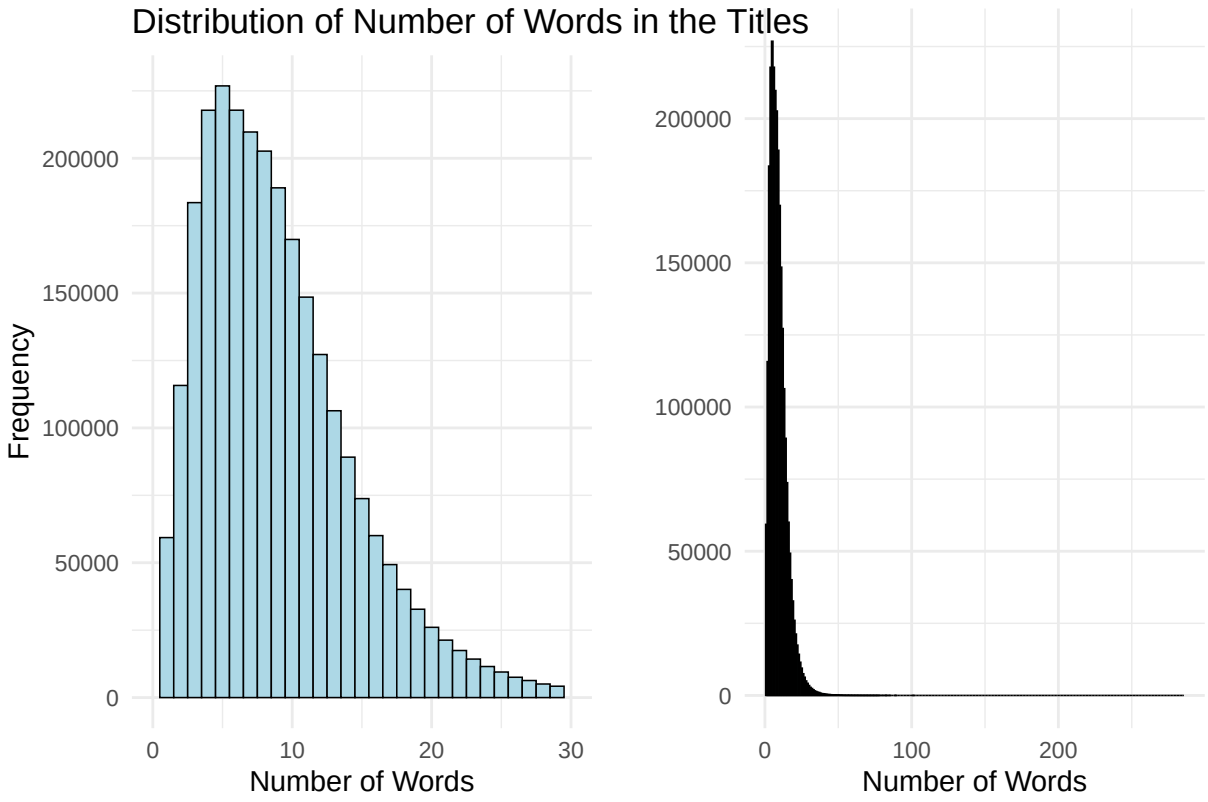

Abbildung 5: Verteilung der Titellängen.

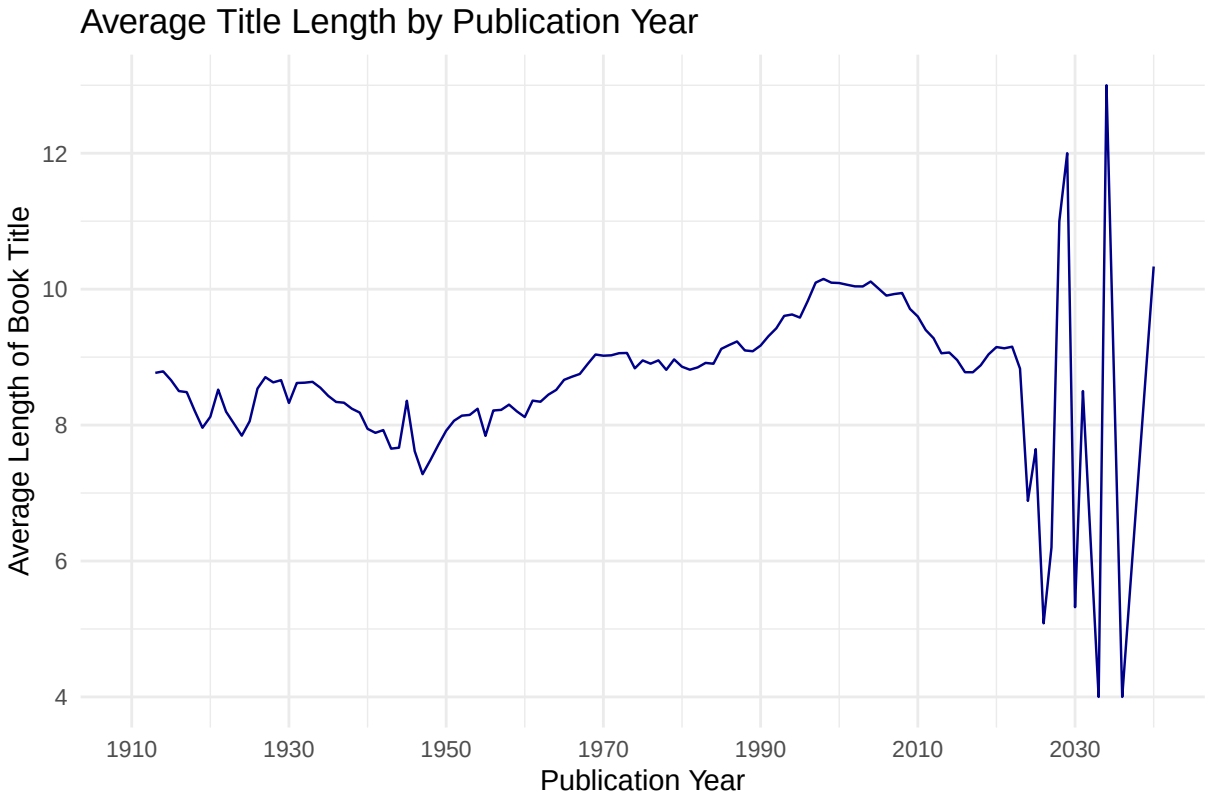

Abbildung 6: Verteilung der Titellängen über das Publikationsjahr.

Beschreibung der Daten zeigt aber, dass sich das Vergeben der Schlagworte über die Zeit stark geändert hat und dies ein neueres Phänomen ist. Die Schlagworte variieren in der Länge von 0 bis 381; im Schnitt werden bei einem Werk 1.60 Schlagworte angegeben ($Median = 0, SD = 3.46$). Ein Vergleich über die Zeit zeigt, dass primär Werke, die ab 1970 publiziert wurden, Schlagworte enthalten, und insbesondere bei Werken seit 2010 hat die Verschlagwortung stark zugenommen (siehe Abbildung 7). Zudem gab es auffällig wenige Schlagworte Anfang der 90er. Da dies auch der Zeitpunkt ist, als es große Veränderungen an den Bibliotheksstandorten in Leipzig und Frankfurt gab, ist vorstellbar, dass in der Zeit des damaligen Aufbruchs und der Veränderung Prioritäten nicht auf der Verschlagwortung lagen. Die Einführung des RSWK – der Katalog, der für eine einheitlichere Verschlagwortung eingeführt wurde – hat die Praxis des Verschlagwortens ebenfalls direkt beeinflusst.
Insgesamt machen diese Veränderungen in der bibliothekarischen Praxis die Schlagworte über die Zeit hinweg nicht vergleichbar. Dies kann zu Verzerrungen führen, deren Konsequenzen nicht einschätzbar sind. Es ist unklar, wie sich diese Änderungen genau auf den Datenbestand auswirken. Es wird deswegen im Folgenden darauf verzichtet, von den Schlagworten Gebrauch zu machen.

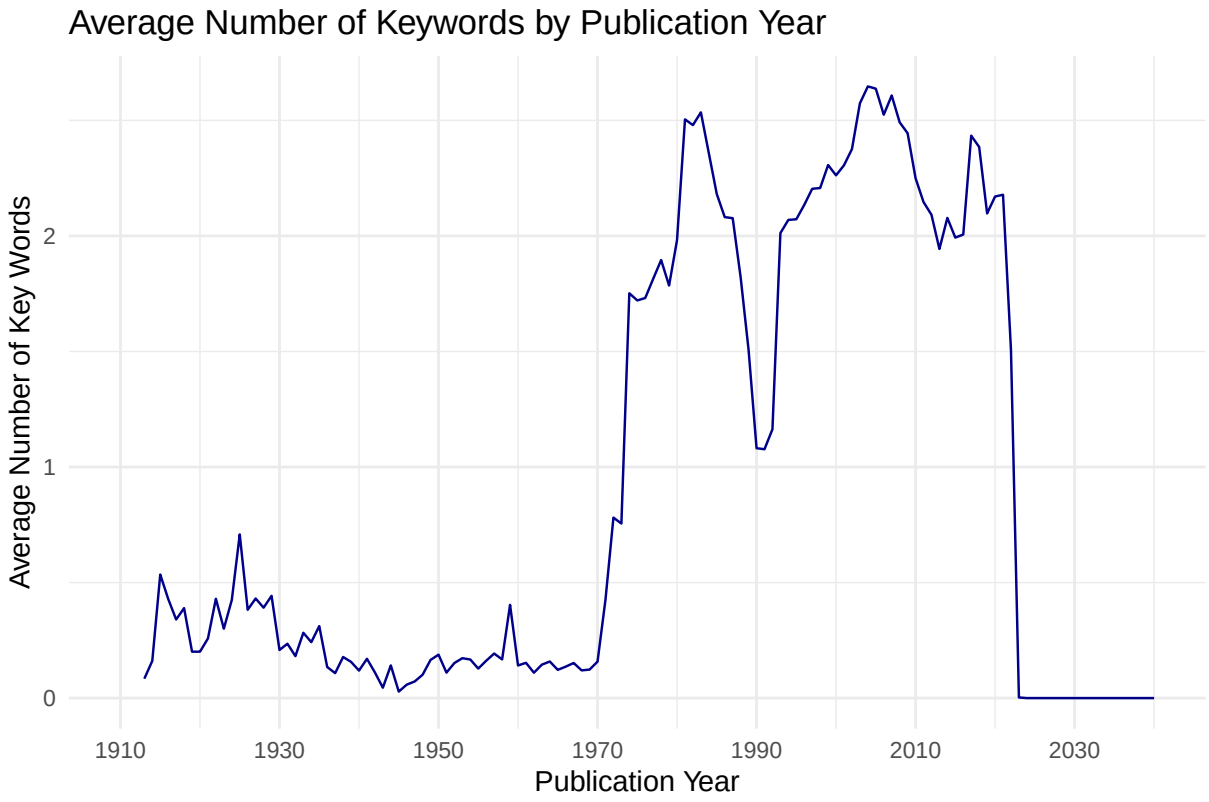

Abbildung 7: Verteilung der Länge der Schlagworte.

Die Schlagworte werden somit nicht genutzt, aber was steht in den Titeln? Die 15 häufigsten Wörter (nach der Entfernung von Stoppwörtern) im Datensatz lauten wie folgt (mit Anzahl Nennungen in Klammern):

1. roman (117204)
2. untersuchungen (70482)
3. geschichte (54453)
4. leben (43332)
5. untersuchung (40405)
6. entwicklung (37066)
7. deutschen (34304)
8. analyse (32863)
9. gedichte (30438)
10. deutschland (27230)
11. patienten (26940)
12. neue (25782)
13. jahre (23826)
14. studie (23573)
15. berücksichtigung (23277)

Diese Liste gibt Aufschluss über die am häufigsten vorkommenden Wörter in den Titel und gibt damit Einblick in die Themen und Inhalte der Werke. Das Wort "roman" kommt am häufigsten vor, was darauf hindeutet, dass viele Titel literarische Werke oder Bücher mit fiktionalen Geschichten behandeln und dies explizit – vor allem auch im Untertitel – angemerkt wird. Ähnliches gilt auch für das Wort "gedichte", das explizit darauf hinweist, dass es sich bei dem Werk um Lyrik handelt. Die Worte "untersuchungen" und "untersuchung", die beide sehr häufig vorkommen, deuten wiederum darauf hin, dass sich eine beträchtliche Anzahl von Titel auf empirischen Forschungen oder Studien beziehen; dies gilt ebenso für "analyse". Worte wie "geschichte", "entwicklung", "deutschen" und "deutschland" könnten auf Titel hinweisen, die sich auf historische Ereignisse, Veränderungen und Entwicklungen beziehen und einen Bezug zur deutschen Kultur, Geschichte oder Politik aufweisen. Das Wort "leben" könnte auf eine Vielzahl von Themen hindeuten, die mit dem menschlichen Leben, der Biografie, der Gesellschaft oder der Lebensqualität zusammenhängen. Insgesamt geben die Worte einen ersten Eindruck über den Datensatz.

4.2.1 Stimmungszuweisung

Den Titeln wurden Stimmungen zugewiesen, die im Folgenden beschrieben werden. Die zugewiesene Stimmung unterscheidet sich je nach Art der Sentimentenanalyse (lexikonbasiert vs. KI-basiert). Die beiden Varianten korrelieren schwach positiv miteinander ($\rho = 0.0077, p < 0.001$). Die lexikonbasierte Variante weist dabei sehr viel mehr Variation auf: Mit einem Minimum von -1, Maximum von 1, Mittelwert von 0.043 und Standardabweichung von 0.288 nutzt sie das Wertespektrum sehr viel stärker aus als der KI-basierte Sentimentenwert, der ein Minimum von 0.50, ein Maximum von 0.582, einen Mittelwert von 0.510 und eine Standardabweichung von 0.00753 aufweist. Tabellen 1 und 2 zeigen beispielhaft die jeweils 15 Werke mit den höchsten und den niedrigsten Sentimentenwerten für die beiden Ansätze der Stimmungszuweisung[19].
Ein Vergleich der Tabellen zeigt, dass die lexikonbasierte Variante intuitiv bessere Ergebnisse liefert, auch wenn beide Varianten gewisse Einschränkungen und Mängel aufweisen. Im folgenden Teil der Ergebnisse wird die lexikonbasierte Variante der Sentimentenanalyse verwendet. Die KI-basierte Variante scheint sich für den besonderen Anwendungsfall der Buchtitel nicht gut zu eignen und hat sensibel auf Großschreibungen reagiert (die nicht entfernt wurde, da in der deutschen Sprache Kapitalisierung nicht irrelevant ist).

4.2.2 Zuweisung der Kernkonzepte

Zwei Verfahren wurden getestet, um Kernkonzepte den Werken zuzuweisen. Beide Verfahren haben nicht zu sinnvollen Ergebnissen geführt. Die Probleme mit beiden Verfahren werden im Folgenden vorgestellt.

Topic Modelling (LDA) Dem ersten Ansatz folgend wurde versucht, mittels Topic Modelling den Werken ein Kernkonzept zuzuweisen. Die Titel wurden vorverarbeitet und ein LDA-Modell geschätzt. Da es keine Theorie oder bestimmtes Vorwissen gibt, die eine bestimmte Anzahl an Themen nahelegen würde, wurde mit einer Zufallsstichprobe (10% der Daten) versucht, die optimale Anzahl an Themen zu finden. Dafür wurden Topic-Modelle mit einer jeweils unterschiedlichen Anzahl von Themen geschätzt und ihr Kohärenzwert ausgegeben. Der Kohärenzwert ist ein Maß dafür, wie zusammenhängend und interpretierbar die Themen sind, die aus einem Textkorpus extrahiert wurden. Er bewertet die Ähnlichkeit der Wörter innerhalb eines Themas und die Unterschiede zwischen den Themen. Ein höherer Kohärenzwert zeigt an, dass die Themen besser definiert und unterscheidbar sind. Die Kohärenzwerte für unterschiedliche Anzahl Themen für die Stichprobe des Datensatzes sind in Abbildung 8 abgebildet. Ein U-Mass Kohärenzwert von 0 zeigt eine optimale Trennung der Themen, wobei niedrigere Werte eine schlechtere Trennung aufzeigen (Mimno u. a. 2011). Der vorliegende Datensatz hat bei einer Unterscheidung von nur einem bzw. vier Themen die beste Kohärenz, die sich mit zusätzlicher Anzahl Themen verschlechtert.

[19] Manche Werke sind doppelt in der Tabelle, da sie nicht als Duplikat erkannt wurden durch Unterschiede in den Angaben in der MARC21-Repräsentation.

Tabelle 1: 15 Titel mit den jeweils höchsten und niedrigsten Sentimentenwerte gemäß lexikonbasierter Analyse.

Titel	Lexikon	KI
Dynamische Kernresonanz an biologischen Modellmembranen Untersuchungen zur Spin-Gitter-Relaxation von 2H- und 31P-Kernspinsonden	1	0.507
Churz vor em Ablösche 21 heitere Geschichten, die das Leben schrieb	1	0.512
Zivilrechtspraktikum Rechtsfälle für Übungen im bürgerl. Recht	1	0.503
Arbeitsbuch systematische Theologie eine Methodenhilfe für Studium und Praxis	1	0.506
Das Forum der Vergebung in der Kirche Studien zum Verhältnis von Sündenvergebung und Recht	1	0.505
Borderlands - unbesiegbar	1	0.514
Borderlands: Unbesiegbar Roman zum Game	1	0.504
Ein herrliches Chaos Roman ; Science Fiction	1	0.520
Bitte, recht freundlich! Schwank in einem Aufzuge	1	0.506
Hansel und Gretchen gratulieren Namenstagsvortrag	1	0.510
Karlchen Huckebein Ein lustiges Kindersp. in 1 Aufz.	1	0.503
Kaviar Heiteres Spiel in 1 Aufz.	1	0.503
Lustige Gratulation Spiel in 1 Aufz.	1	0.507
Mit Liebe! Schwank in 1 Aufz.	1	0.505
Pink und Punk, die lustigen Wandervögel Schwank in 1 Aufz.	1	0.514
Gott strafe England! Vaterländische Erzählung	-1	0.507
Kriegerische Wasserzeichen	-1	0.516
Die Eiszeit war ganz anders das Geheimnis d. versunkenen Brücke nach Amerika	-1	0.510
Die kämpfenden Flotten im Weltkriege Verzeichnis sämtlicher Kriegsschiffe der kriegführenden Staaten mit Angabe der Verluste feindlicher Schiffe nach dem Stande von Ende Juli 1916	-1	0.510
Die traumatische Zwerchfellruptur ihre Ursachen, Diagnostik und Therapie	-1	0.511
Forum Shopping des Gläubigers im Rahmen der Zwangsvollstreckung mißbräuchliches Verhalten oder zu billigende Interessenwahrnehmung?	-1	0.502
Den Mond unterm Arm	-1	0.503
Vergleichende Analyse des Proteoms der Bursa Fabricii des Huhns zu verschiedenen Entwicklungszeitpunkten	-1	0.521
Vergleichende Analyse des Proteoms der Bursa Fabricii des Huhns zu verschiedenen Entwicklungszeitpunkten	-1	0.521
Verhalten in Organisationen organisationale und persönliche Verhaltensanalyse in Abhängigkeit von strukturellen Bedingungen	-1	0.500
Hat das Christentum versagt?	-1	0.515
Der Einfluss des PARY- Agonisten [PPAR-Gamma-Agonisten] Pioglitazon auf den neuronalen Schaden bei der Ratte nach CCI (controlled cortical impact)	-1	0.510
Die widerspüchliche Unterrichtswirklichkeit an unseren Schulen und deren Auswirkung auf das Verhalten von Lehrern und Schülern zur Notwendigkeit der Aufhebung falscher Fronten; Lehrer gegen Schüler und Schüler gegen Lehrer	-1	0.508
Einsatz verschiedener Präparationen eines Extraktes von Solanum glaucophyllum zur Prävention der hypocalcämischen Gebärparese des Rindes	-1	0.506
Experimentelle Untersuchungen zum Einfluss technischer Parameter auf die Gestaltung der Schnittflächen und Schnittränder nach automatischer lamellärer Keratotomie unter Einsatz verschiedener LASIK-Schneidgeräte (Mikrokeratome und Femtosekundenlaser) an Schweinehornhäuten	-1	0.502

Zur weiteren Überprüfung der Validität von Topic-Modellen können zudem die Wörter ausgegeben werden, die am stärksten mit einem bestimmten Thema assoziiert werden. Für das Modell mit vier Themen sieht das wie folgt aus:

1. 'untersuchungen', 'über', 'für', 'leben', 'untersuchung', 'analyse', 'entwicklung', 'patienten', 'berücksichtigung', 'bedeutung', 'einfluss', 'studie', 'beim', 'vergleich', 'besonderer', 'sowie', 'charakterisierung', 'anwendung', 'beitrag', 'praxis'

Tabelle 2: 15 Titel mit den jeweils höchsten und niedrigsten Sentimentenwerte gemäß KI-basierter Analyse.

Titel	Lexikon	KI
Die zwölf Apostel Cf 694 C ; [Texh.]	0	0.582
VOM OBJEKT ZUM SYMBOL	0	0.578
DAS VERWUNSCHENE MUSEUM	0	0.578
DAS VERWUNSCHENE MUSEUM	0	0.578
VERGESSEN	-0.7	0.577
DAS HERZ ZUM VATERLAND	0	0.573
DAS VERGESSENE TAL Abenteuerroman	0	0.573
DAS VERLORENE SYMBOL	0	0.571
DAS METRO KULTURGESCHICHTE EINES WIENER VERGNÜGUNGSORTES	0	0.571
MUSIKMACHEN IM 20. JAHRHUNDERT : AUF DEM WEG ZUM VIRTUELLEN TONSTUDIO IM INTERNET	0	0.571
KEIN ORT ZUM VERWEILEN	0	0.568
DAS GROSSE WANDERN - Brandenburg 100 Touren	0	0.567
EXZESSION - GLÜCK IST EINE ENTSCHEIDUNG ERSCHAFFEN SIE WERTE-WELTEN, DIE IHNEN DAS SCHENKEN, WAS SIE IN DIESEM LEBEN ERWARTEN, GLÜCK IST EIN INSIDE JOB	0	0.565
Neue ertragsteuerliche Vorschriften ab dem VZ 2004 - die wichtigsten Änderungen auf einen Blick	0	0.565
Altes und Neues aus der “Flimmerkiste” - eine Analyse des deutschen TV-Programms für Kinder im Grundschulalter	0	0.565
Erlebnis Glacier- und Bernina-Express	0	0.500
Schwierigkeiten beim Verständnis der Narayama-Lieder Erzählung	0	0.500
Schwierigkeiten beim Verständnis der Narayama-Lieder [Erzählung]	0	0.500
Tierisch	0	0.500
Zur subjektiven Begründung von unterschiedlichen Arbeitsweisen im Heilpädagogischen Voltigieren und Reiten	0	0.500
Mama Margarete, die Mutter Don Boscos	0	0.500
Scarlett	0	0.500
Über die Umsetzung von Dicyclopentadien mit Schwefel und die Eignung des Reaktionsproduktes als Additivkomponente für Getriebeöle hoher Leistungsklassen	0	0.500
Kein Friede mit Deutschland die geheimen Gespräche im Zweiten Weltkrieg 1939 - 1941	0	0.500
Das römische Licht	0	0.500
Verbundbau nach EC 4 Entwurf und Bemessung - mit zahlreichen Beispielen	0.7	0.500
Wegbegleiter Gedichte	0	0.500
Buddhistische Plastik aus China und Japan Bestandskatalog d. Museums f. Ostasiatische Kunst der Stadt Köln	0	0.500
Die letzten 100 Tage Ein Tagebuch vom Kampf d. HJ. an d. Saar	0	0.500
Über die Anwendbarkeit des Farbpyramiden-Tests zur Diagnose der Schizophrenie	0	0.500

2. ’für’, ’über’, ’deutschland’, ’zeit’, ’geschichte’, ’welt’, ’menschen’, ’liebe’, ’deutsche’, ’kleine’, ’stadt’, ’band’, ’einfluß’, ’berlin’, ’kirche’, ’jahre’, ’deutschen’, ’jahrhundert’, ’ende’, ’beitr’

3. ’gedichte’, ’für’, ’grundlagen’, ’geschichte’, ’beiträge’, ’gegenwart’, ’arbeit’, ’politik’, ’kunst’, ’deutschen’, ’europäischen’, ’schule’, ’entwicklung’, ’zukunft’, ’beispiel’, ’gesellschaft’, ’drei’, ’teil’, ’probleme’, ’grenzen’

4. ’für’, ’geschichte’, ’neue’, ’geschichten’, ’deutschen’, ’über’, ’jahre’, ’kriminalroman’, ’einführung’, ’kinder’, ’zwei’, ’studien’, ’recht’, ’literatur’, ’land’, ’frau’, ’frauen’, ’gott’, ’jahrhunderts’, ’heute’

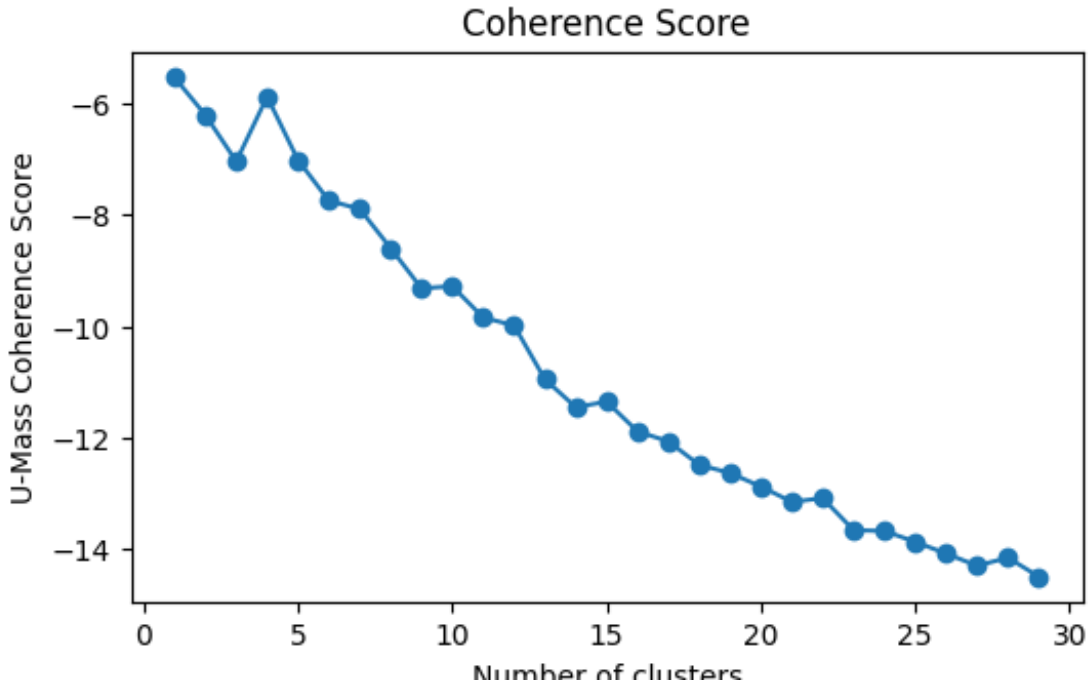

Abbildung 8: Kohärenzwerte für unterschiedliche Anzahl Cluster.

Nur mit genug Fantasie lassen sich daraus grob Überschriften zu den Themen ableiten: Das erste Thema scheint Studien und Untersuchungen, potentiell im medizinischen Bereich, abzudecken – es könnte sich aber auch lediglich auf generelle Sachliteratur beziehen; das zweite Thema bezieht sich auf Aspekte der Geschichte der Welt und Deutschlands; das dritte Thema greift wiederum politische und historische relevante Konzepte auf, beinhaltet aber auch das Wort "Gedichte", was eher aus dem Rahmen fällt; auch im vierten Thema schließlich fällt der Bezug zu Deutschland und Geschichte auf, sowie der Bezug auf zeitliche Bezüge und Frauen. Die Worte, die am stärksten mit Themen verbunden sind, scheinen vor allem aus der Sachliteratur zu stammen. Abgesehen von Worten wie "Kriminalroman" oder "Gedichte" ist die Belletristik nicht klar repräsentiert. Von den Wortlisten her ist auffällig, dass sich Themen eher schwierig unterscheiden lassen und ähnliche Worte in unterschiedliche Themen gruppiert werden (beispielsweise "deutsche" und "deutschen). Bei einer noch größeren Anzahl von Themen werden die Unterscheidungen nicht klarer. Die Wörter, die am stärksten mit zehn Themen assoziiert sind, sind die folgenden:

1. 'neue', 'jahre', 'für', 'neuen', 'beim', 'einfluss', 'recht', 'praxis', 'wege', 'akten', 'aspekte', 'eigenschaften', 'erfahrungen', 'geheimnis', 'lebens', 'struktur', 'fragen', 'vergleichende', 'verfahren', 'friedrich'

2. 'für', 'menschen', 'deutsche', 'erzählungen', 'gott', 'zukunft', 'europa', 'entstehung', 'moderne', 'perry', 'gedanken', 'neuer', 'kindern', 'reich', 'österreich', 'texte', 'licht', 'schatten', 'herz', 'deutscher'

3. 'analyse', 'deutschen', 'patienten', 'untersuchung', 'studie', 'für', 'untersuchungen', 'kunst', 'behandlung', 'mittels', 'bestimmung', 'frau', 'klinische', 'macht', 'anhand', 'krieg', 'beim', 'abenteuer', 'einfluss', 'experimentelle'

4. 'für', 'einführung', 'charakterisierung', 'kleine', 'stadt', 'zwei', 'kirche', 'lernen', 'heute', 'grenzen', 'philosophie', 'sicht', 'politik', 'möglichkeiten', 'warum', 'bilder', 'universität', 'johann', 'perspektiven', 'sozialen'

5. 'entwicklung', 'berücksichtigung', 'beitrag', 'für', 'geschichten', 'studien', 'besonderer', 'theorie', 'jahren', 'ergebnisse', 'darstellung', 'therapie', 'werk', 'ende', 'untersuchung', 'rolle', 'während', 'berücks', 'hilfe', 'bedeutung'

6. 'gedichte', 'für', 'deutschland', 'liebe', 'bedeutung', 'kinder', 'vortrag', 'gesellschaft', 'reise', 'bundesrepublik', 'rede', 'gottes', 'jahr', 'mittelalter', 'peter', 'hamburg', 'grundlage', 'letzte', 'demokratie', 'gehalten'

7. 'über', 'untersuchungen', 'geschichte', 'band', 'beiträge', 'einfluß', 'wirkung', 'jahrhundert', 'arbeit', 'europäischen', 'synthese', 'alten', 'untersuchung', 'kultur', 'schweiz', 'familie', 'funktion', 'politische', 'einsatz', 'mensch'

8. 'beispiel', 'welt', 'für', 'grundlagen', 'berlin', 'drei', 'probleme', 'fall', 'land', 'mehr', 'erzählung', 'sprache', 'auswirkungen', 'unternehmen', 'große', 'deutschen', 'alte', 'bild', 'methode', 'betrachtung'

9. 'zeit', 'geschichte', 'beitr', 'jahrhunderts', 'literatur', 'gegenwart', 'schule', 'frauen', 'ersten', 'musik', 'versuch', 'bildung', 'thriller', 'wandel', 'problem', 'frage', 'vortr', 'einf', 'leipzig', 'modernen'

10. 'leben', 'vergleich', 'kriminalroman', 'teil', 'soziale', 'freiheit', 'jugend', 'beziehungen', 'empirische', 'verstehen', 'psychologie', 'suche', 'gestaltung', 'bereich', 'kommt', 'april', 'tage', 'regulation', 'freunde', 'letzten'

Die Wortlisten wirken ausgesprochen willkürlich und wenig aussagekräftig. Es lassen sich intuitiv keine Themen von diesen Wortlisten ableiten. Das Topic-Modelling hat nicht funktioniert.

Word Embeddings Im zweiten Ansatz wurde versucht, mittels Word Embeddings den Werken ein Kernkonzept zuzuweisen. Dafür wird allen Wörtern ein gelernter Wortvektor zugewiesen. Für dieses Projekt wurde ein vortrainiertes, deutsches Word2Vec-Modell von *fasttext* verwendet[20]. Allen Wörtern in den Titeln wird ein Vektorwert zugewiesen und der Schnitt dieser Werte ist der Wert, der dem Titel als Ganzes zugewiesen wird. Das heißt, der Titel wird durch ein Koordinatenpaar entsprechend der Worten, die im Titel vorkommen, repräsentiert. Anschließend werden die Titel mittels k-Means-Clustering in Gruppen

[20]Siehe https://fasttext.cc/docs/en/crawl-vectors.html.

geclustert. K-Means-Clustering ist ein Algorithmus zur Gruppierung von Datenpunkten in k Cluster, indem er iterativ die Datenpunkte in Gruppen einteilt, basierend auf der Minimierung der Summe der quadrierten Abstände innerhalb jedes Clusters.
Wiederum wurde mithilfe einer Zufallsstichprobe (10% der Daten) versucht, die optimale Anzahl an Clustern zu ermitteln, indem verschiedene Clusteranzahlen ausprobiert und unterschiedliche Verfahren – die Elbow-Methode und die Silhouette-Methode – angewandt wurden, um den optimalen Punkt zu identifizieren.
Die Elbow-Methode beinhaltet die Berechnung der Summe der quadrierten Abstände zwischen den Datenpunkten und ihren jeweiligen Clustermittelpunkten für unterschiedliche Anzahlen von Clustern. Der Punkt, an dem die Veränderung der Summe der quadrierten Abstände abflacht und eine Ellbogenform bildet, wird als optimale Anzahl von Clustern betrachtet. Wie aus Abbildung 9 ersichtlich, scheint der optimale Punkt bei etwa 7 Themen zu liegen. Ein klarer "Ellenbogen" ist in der Abbildung aber nicht ersichtlich.
Der Silhouette-Plot in Abbildung 10 basiert auf der Berechnung der Silhouette-Koeffizienten für jeden Datenpunkt, der ein Maß für die Konsistenz eines Punktes innerhalb seines eigenen Clusters im Vergleich zu anderen Clustern ist. Ein höherer Silhouette-Koeffizient deutet auf eine bessere Clusterbildung hin. Die beste Clusterbildung gibt es in den Katalogdaten bei einer Aufteilung auf nur zwei Cluster, danach sinkt der Silhouttescore dramatisch. Generell ist der Silhouttescore in dieser Anwendung sehr niedrig und nahe Null. Dies deutet darauf hin, dass die Datenpunkte in den Clustern eine geringe Konsistenz aufweisen und möglicherweise nicht gut voneinander getrennt sind. Ein Score von 0 bedeutet, dass die Datenpunkte sich überlappen oder die Zuordnung zu den Clustern unsicher ist. Idealerweise wird ein Silhouette-Score nahe 1 angestrebt, da dies bedeutet, dass die Datenpunkte gut innerhalb ihres Clusters zusammenhängen und klar von anderen Clustern abgegrenzt sind. Diese Ergebnisse zeigen, dass das Zuweisen von Clustern mittels Worteinbettungen ebenfalls nicht gut funktioniert hat.
Beide Varianten des Zuweisens der Kernkonzepte bringen keine sinnvollen, guten Ergebnisse für die Stichprobe. Es wird daher davon abgesehen, diese Analysen weiterzuführen. Im abschließenden Abschnitt 6 wird diskutiert, was die möglichen Probleme waren.

5 Ergebnisse: Alters-, Perioden- und Kohorteneffekte in der deutschen Literatur

Der folgende Abschnitt zeigt die Ergebnisse der Analysen. Die Analysen beschreiben Effekte der zeitlichen Komponenten auf die Stimmung in der deutschen Literatur.
Ein erster Hinweis auf Zeiteffekte lässt sich durch die Betrachtung einiger Datenvisualisierungen gewinnen. Abbildungen 11 und 12 zeigen die durchschnittlichen Sentimente über 1) das Alter der Autoren, 2) das Publikationsjahr des Werkes und 3) das Geburtsjahr der Autoren. Abbildung 11 macht deutlich, dass die Schwankungen in der durchschnittlichen Stimmung sehr gering sind und sich nicht über den vollen Wertebereich erstrecken. Die

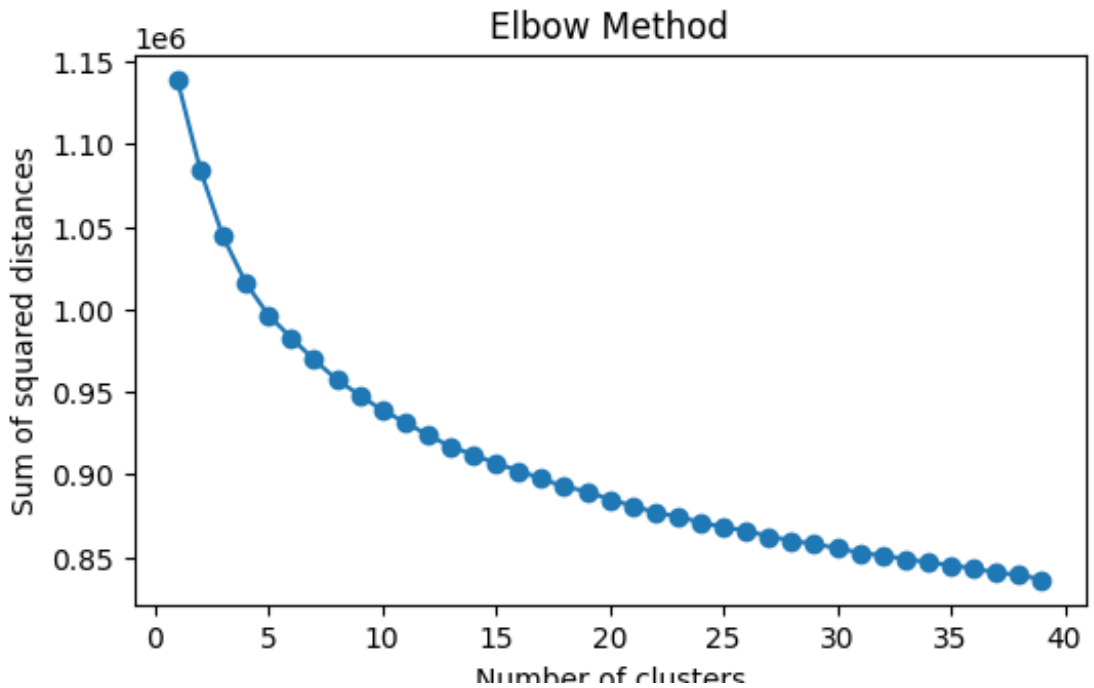

Abbildung 9: Elbow-Scores für unterschiedliche Anzahl Cluster.

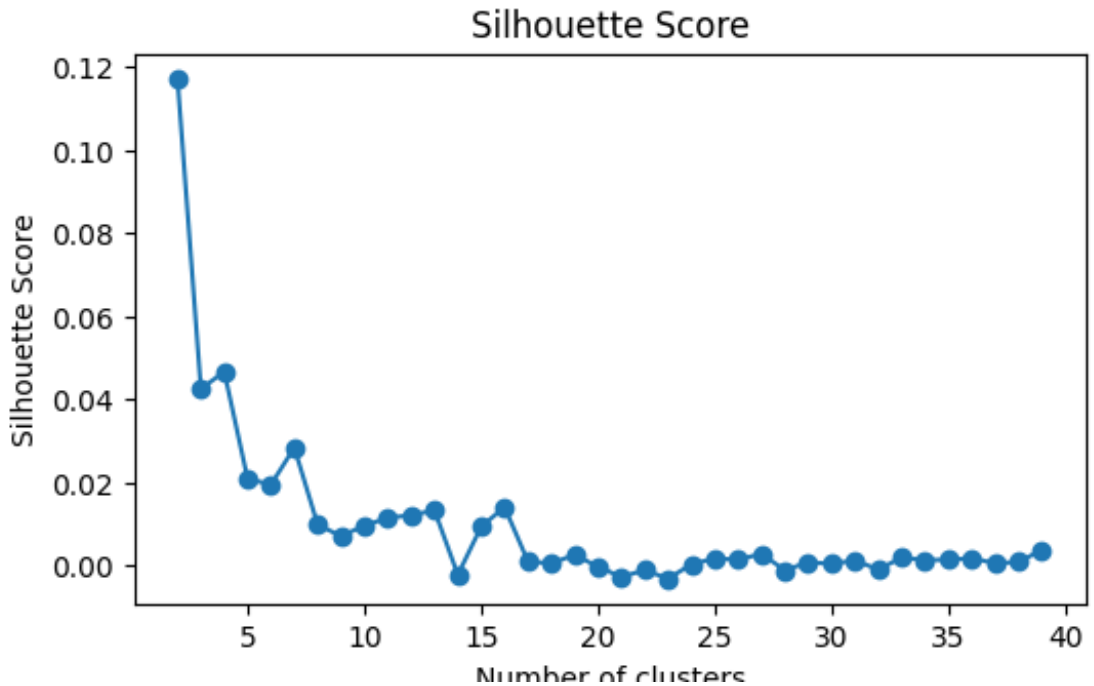

Abbildung 10: Silhoutte-Scores für unterschiedliche Anzahl Cluster.

durchschnittliche Stimmung ist sowohl über Alter, Publikationsjahr als auch Kohorte hinweg recht stabil. Abbildung 12 konzentriert sich auf die beobachteten Schwankungen (siehe den Wertebereich der Ordinate) und zeigt, dass es kleine Schwankungen gibt. Über das Alter hinweg werden Werke generell positiver: Werke, die von Autoren über 30 geschrieben werden, sind im Schnitt generell positiver als Werke von jüngeren Autoren. Im Hinblick auf das Publikationsjahr ist wichtig anzumerken, dass die Ausreißer in den neuen Jahren

durch die geringe Anzahl an Publikationen – vorgemerkte Neuerscheinungen – erklärt wird. Abgesehen von diesen Ausreißern zeigt sich, dass seit ca. 1970 die Werke positiver werden (1970 ist der Tiefpunkt der durchschnittlichen Stimmung). Ebenfalls erscheinen die Werke, die in den Jahren 1935–1945 publiziert wurden, im Schnitt negativer als die Werke in den Jahren davor und danach. Ein Vergleich der Geburtskohorten zeigt, dass Geburtskohorten seit 1950 zu positiveren Titeln neigen als die Kohorten davor.

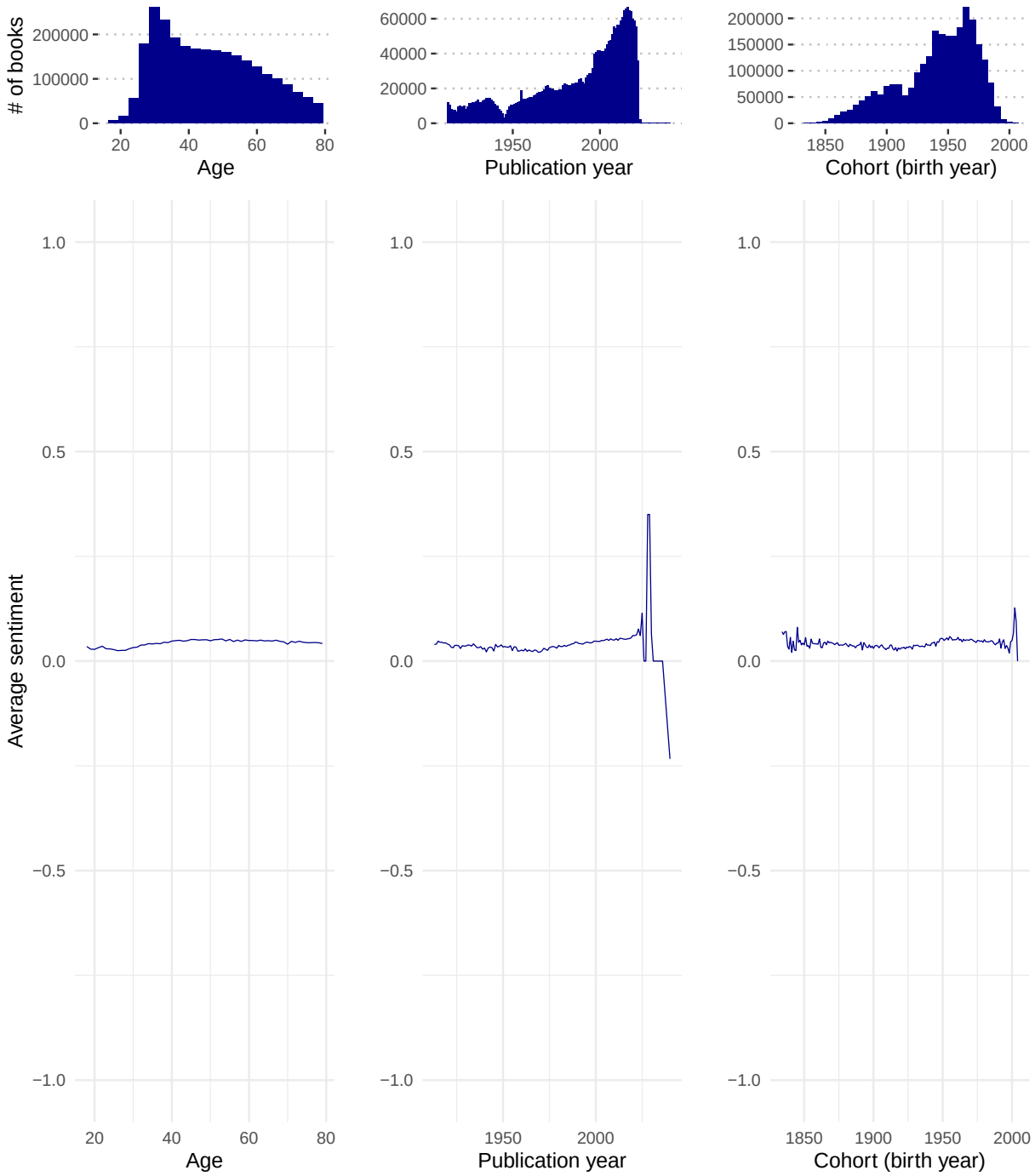

Abbildung 11: Durchschnittliche Stimmung über Alter, Publikationsjahr und Geburtskohorte.

Um alle zeitlichen Dimensionen in ein Diagramm aufzunehmen, wird eine Heatmap erstellt. Eine Heatmap ist eine visuelle Darstellung von Daten, bei der Farben verwendet werden,

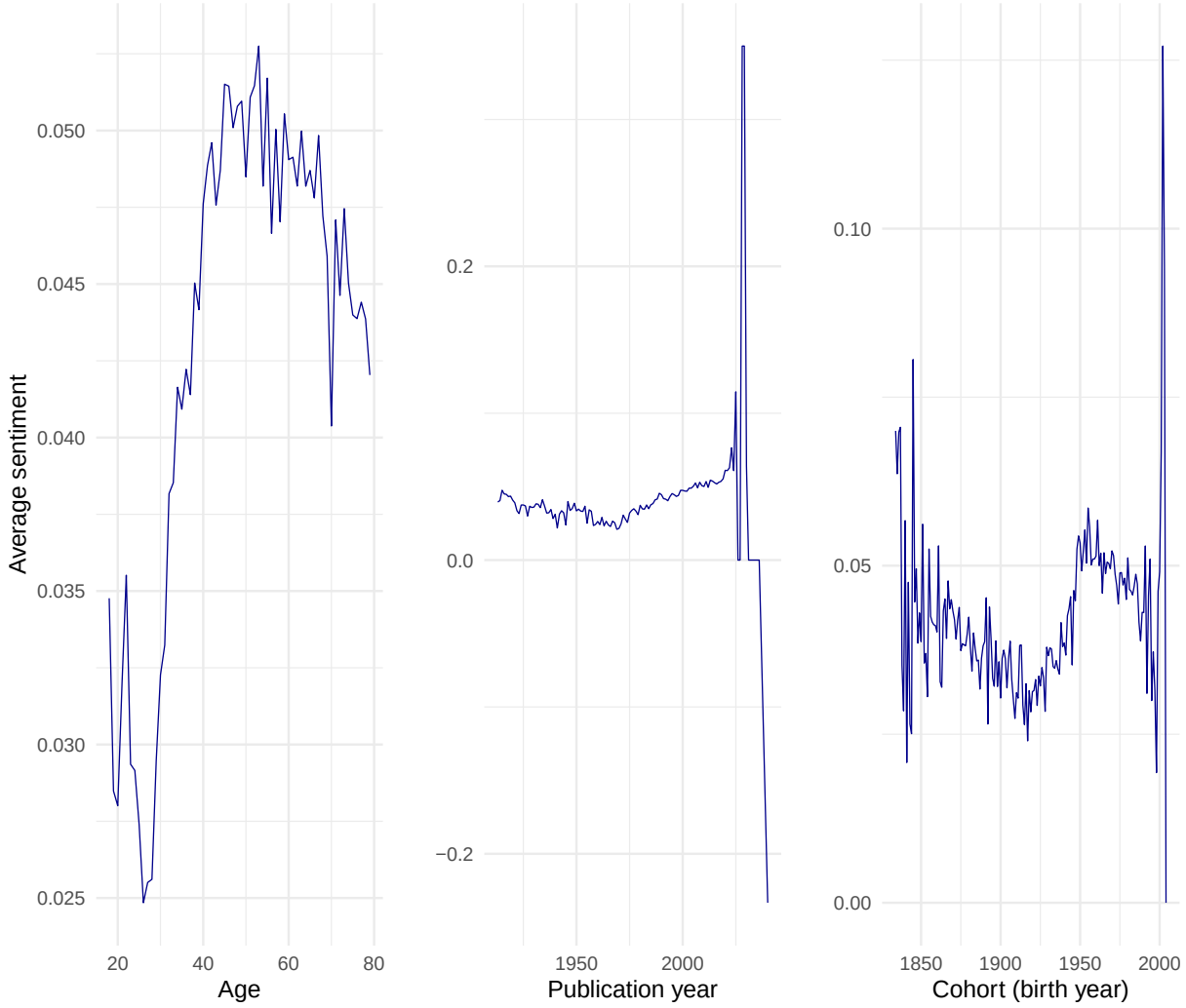

Abbildung 12: Durchschnittliche Stimmung über Alter, Publikationsjahr und Geburtskohorte (Ausschnitt).

um Werte in einer zweidimensionalen Matrix darzustellen. Sie ermöglicht die schnelle Identifizierung von Mustern, Trends oder Unterschieden in den Daten, indem sie verschiedene Farbintensitäten verwendet, um den Wertebereich darzustellen. In Abbildung 13 wird der Mittelwert der Stimmung abgebildet; sie zeigt die beobachteten Durchschnittswerte der metrischen Sentimentenvariable für jede beobachtete Kombination von Alters- und Periodenwerte; Kohorten lassen sich an den Diagonalen ablesen (einige Kohorten sind markiert zur besseren Lesbarkeit). Die Abbildung bestätigt die bisherigen Ergebnisse: Allfällige Kohorten-, Perioden- und Alterseffekte sind schwach, da vielen Werken eine neutrale Stimmung zugewiesen wurde.

Nach dieser grafischen Annäherung folgen modellbasierte Analysen mit GAMs. Da keine Kontrollvariablen geschätzt werden, werden in diesem einfachen Modell nur eine Konstante geschätzt ($\beta = 0.043, p < 0.001$) und eine nichtlineare Schätzung für Alter und Periode gegeben ($edf = 0.043, p < 0.001$, wobei edf für die effektiven Freiheitsgrade stehen und den Grad der Nichtlinearität angeben).

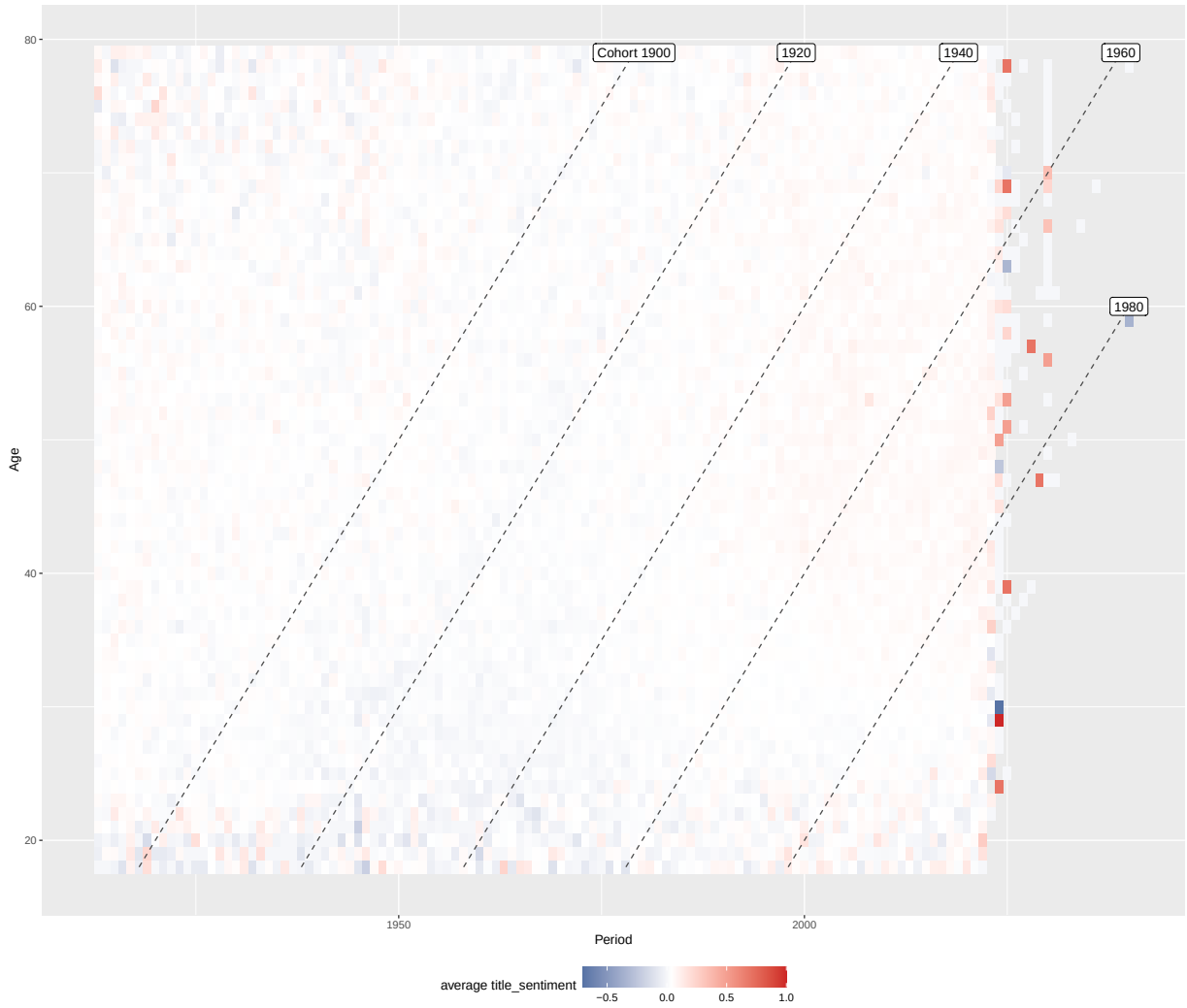

Abbildung 13: Heatmap der Stimmungen.

Ähnlich wie bei der oben beschriebenen deskriptiven Visualisierung kann die mit dem Regressionsmodell geschätzte Tensorproduktoberfläche als Heatmap bzw. als Hexamap visualisiert werden. Abbildung 14 zeigt die Struktur der Sentimentenmittelwerte. In der Abbildung fällt vor allem der negative Effekt des jungen Alters und des späten Publikationsjahres auf.

Für eine besseres Verständnis der Alters-, Kohorten- und Periodeneffekte können auf der Grundlage der durch das Regressionsmodell geschätzten Tensorproduktoberfläche marginale Effekte extrahiert werden, indem der Durchschnitt aller Werte auf der Oberfläche entlang einer Dimension genommen wird. Abbildung 15 zeigt die marginalen Effekte des Alters, des Publikationsjahres und der Kohorte.

Die Ergebnisse der Abbildung 15 ergänzen die Ergebnisse der deskriptiven Beschreibung: Die Modellergebnisse zeigen einen negativen Alterseffekt der jungen Autoren, eine komplexeren Periodeneffekt und einen Nulleffekt der Kohorte (mit Ausnahme der sehr neuen Jahrgänge).

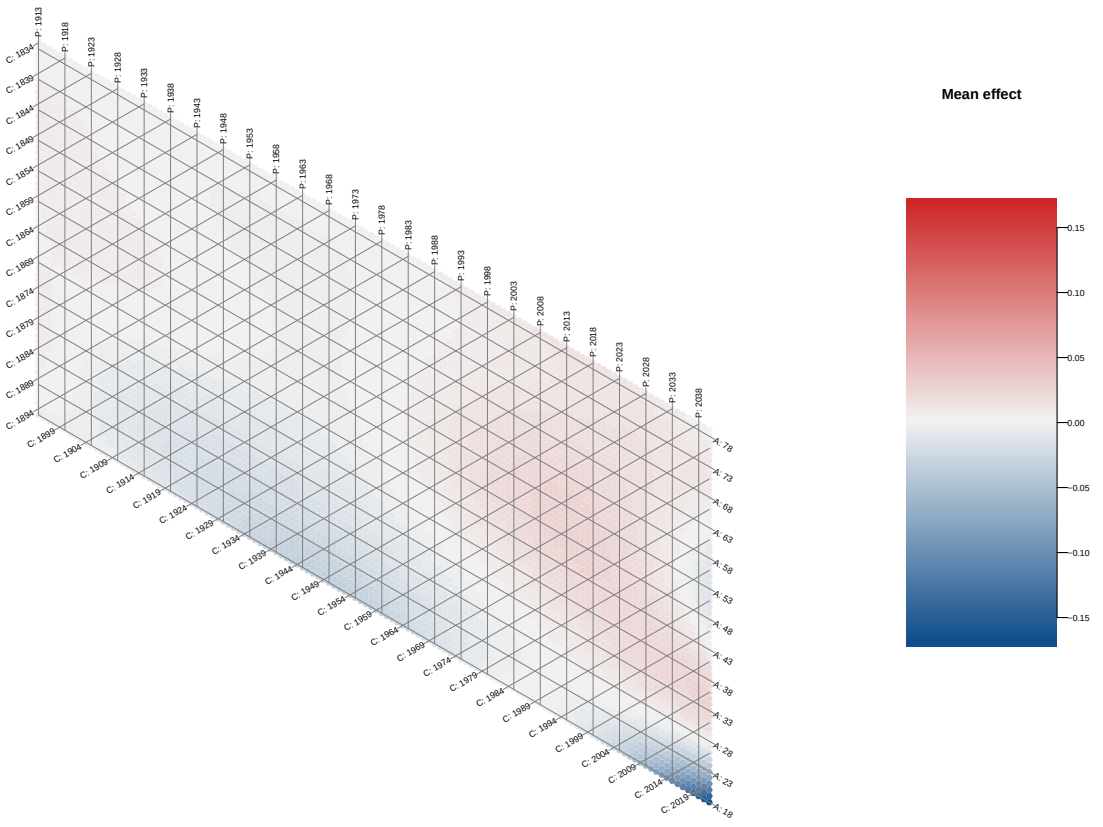

Abbildung 14: Visualisierung der Regressionsergebnisse als Hexamap.

6 Diskussion und Fazit

Die vorliegende Arbeit hat die Katalogdaten der Deutschen Nationalbibliothek genutzt, um Fragestellungen aus dem Feld der demografischen Literaturwissenschaft zu beantworten. Interdisziplinär angesiedelt zwischen den Feldern der Digital Humanities, der digitalen Literaturwissenschaft, den Computational Social Sciences und der Demografie verbindet sie Fragestellungen und Daten unterschiedlicher Traditionen um neues Wissen zu generieren. Konkret hat diese Forschungsarbeit die Fragen gestellt, inwieweit sich thematische Schwerpunkte und Stimmungen in der deutschen Literatur durch die Titel der Werke identifizieren und durch Perioden-, Alters- und Kohorteneffekte erklären lassen. Im Kern hat dieses Forschungsprojekt gezeigt, dass das automatisierte Ableiten von Stimmung und Kernkonzept nicht so einfach möglich ist. Stimmungen lassen sich für Titel zwar herleiten, doch wird dem Großteil der Werke eine neutrale Stimmung zugewiesen, da die wenigen Worte, die in einem Titel vorkommen, oft nicht aussagekräftig genug sind. Die Analyse der Stimmungen zeigt einen leichten Periodeneffekt, was übereinstimmend mit der literaturwissenschaftlichen Herangehensweise ist. Aufgrund der schwierigen Datengrundlage dürfen die Ergebnisse aber nicht überinterpretiert werden. Im Folgenden werden daher insbesondere die Limitationen der vorliegenden Arbeit herausgearbeitet und die Nutzungsmöglichkeiten der Daten der DNB für (sozial-)wissenschaftliche Zwecke kritisch beleuchtet.

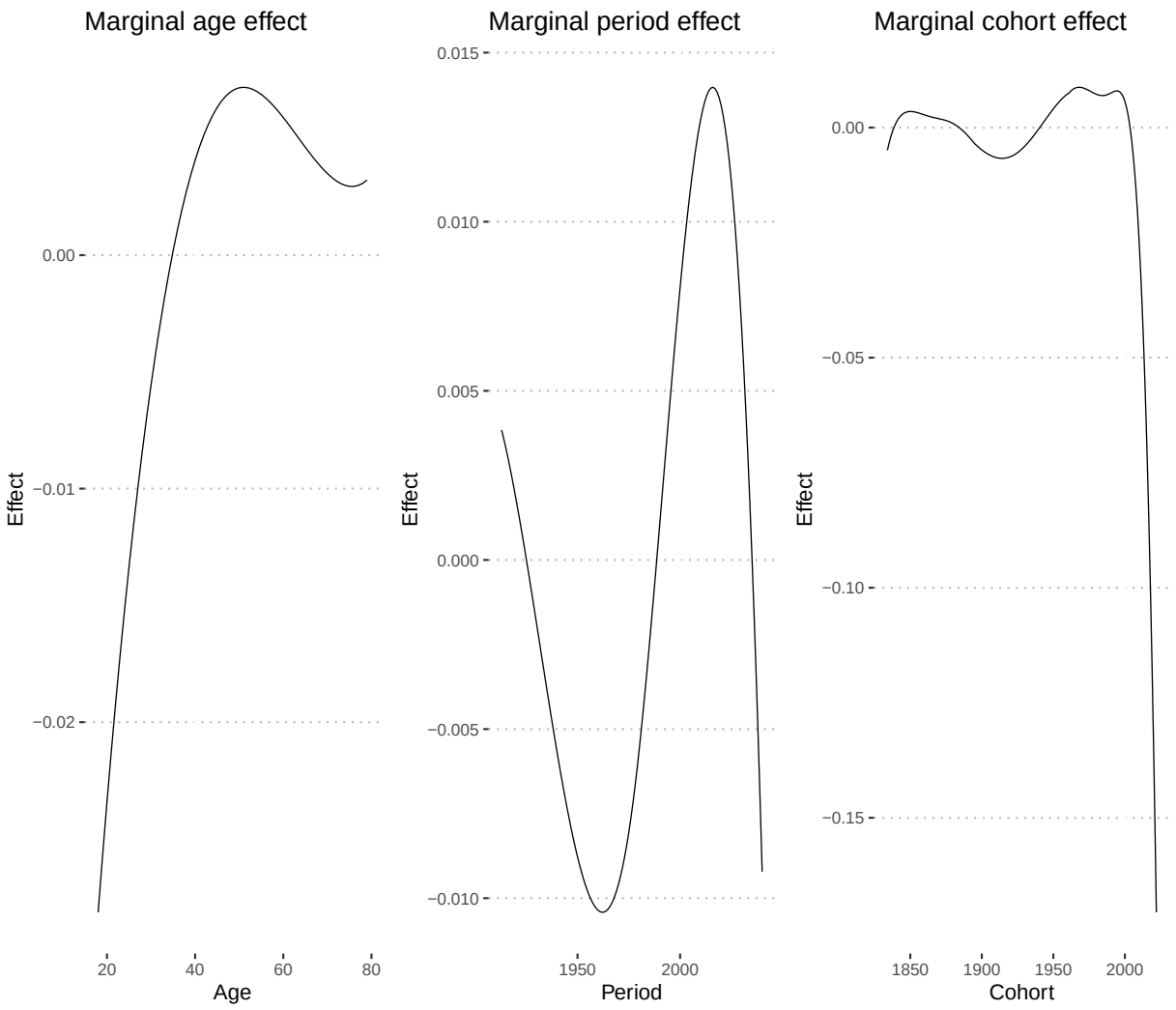

Abbildung 15: Marginale Alters-, Perioden- und Kohorteneffekte.

Automatisierte Verarbeitung von Buchtiteln In dieser Arbeit wurden Buchtitel automatisiert mit Algorithmen der natürlichen Sprachverarbeitung verarbeitet. Zur Stimmungsanalyse und zur Extraktion von Kernkonzepten wurden vordefinierte Lexika und vortrainierte Algorithmen verwendet.

Dieses Vorgehen kommt mit einigen Schwierigkeiten und Problemen. Ein Grundproblem der algorithmischen Verarbeitung von Buchtiteln ist dabei, dass Buchtitel spezielle Sonderfälle von Sprache sind und dass sie kaum Kontext aufweisen. Dies erschwert das Verständnis der tatsächlichen Inhalte und Themen des Buches. Buchtitel sind außerdem oft kreativ und metaphorisch, was die automatische Verarbeitung erschwert. Wortspiele, Anspielungen und kulturelle Referenzen können in Buchtiteln enthalten sein, die für Algorithmen schwierig zu verstehen sind. Da Buchtitel oft kurz und prägnant sind, können sie mehrdeutig sein und verschiedene Interpretationen zulassen. Ein einziger Titel kann verschiedene Themen oder Genres abdecken, was die Zuordnung zu bestimmten Kategorien erschwert. Zusätzliche Herausforderungen ergeben sich dadurch, dass Buchtitel ungewöhnliche oder stilistische Sprachmerkmale aufweisen können, wie etwa unvollständige Sätze, grammatikalische Abweichungen oder symbolische Darstellungen.

Bisherige Algorithmen haben sich nicht auf deutsche Buchtitel bezogen, sondern basieren auf anderen Trainingsdaten – auf längeren Texten, auf Zeitungsartikeln, auf Informationen von Wikipedia oder auf Beiträgen von sozialen Medien. Es ist daher unklar, wie gut sich die Modelle auf Buchtitel anwenden lassen. Dies lässt sich auch nur schwierig feststellen; ich habe keine *Ground Truth* und ich kann daher nicht abschätzen, wie richtig die Anwendung der Algorithmen ist. Ich kann von den Titeln und den Sentimentwerten für eine kleine Anzahl Werke abschätzen, inwieweit die Stimmung der Worte gut erfasst wurde, doch habe ich keine Einschätzung darüber, inwieweit Werke tatsächlich traurig oder fröhlich sind. Um die Korrektheit abschätzen zu können, müsste Beobachtungen manuell ein wahrer Wert zugewiesen werden, mit dem die automatisierte Zuweisung überprüft werden könnte. Dies war im Rahmen dieses Forschungsprojekts leider nicht möglich. Im Idealfall sollten die Algorithmen mit Buchtiteln weitertrainiert und evaluiert werden. Dies benötigt manuelle Zuweisung und manuelle Überprüfung. Optimal wäre es, wenn auch untersucht werden würde, inwieweit Titel und Inhalt von Werken zueinander passen und ob eine Sentimenten- und Themenzuweisung auf Grundlage des Titels und auf Grundlage des Volltexts zueinander passen.

Es hat in dieser Arbeit nicht funktioniert, die Werke in Gruppen einzuteilen bzw. sie einem Thema zuzuordnen, was in einem Ableiten von Kernkonzepten hätte resultieren sollen. Topic-Modelling funktioniert in mehreren Fällen schlechter: Wenn der Datensatz klein ist bzw. nur wenig Text verfügbar ist, wenn der Text unstrukturiert ist und viel Irrelevantes enthält, wenn es keine klaren Themen gibt oder wenn die Daten sehr domänenspezifisch sind und Standardalgorithmen keine optimalen Ergebnisse liefern können. In diesem Projekt habe ich, vor dem Hintergrund literaturwissenschaftlicher Forschung und der Epochendefinitionen, erwartet, dass es klare Themen gibt. Dass keine klare Themen extrahiert werden konnten, kann einerseits an den kurzen und wenig aussagekräftigen Titeln liegen und andererseits an den Algorithmen, die nicht auf diesen Fall hin trainiert sind.

Als weitere Limitation ist anzumerken, dass es ein Problem jedes NLP-Projekts ist, dass ein Gold-Standard in der Disziplin fehlt und die Analysen von subjektiven Entscheidungen abhängen. Im Idealfall sollten die Analysen wiederholt werden und der Effekt unterschiedlicher Spezifizierung auf die Robustheit der Resultate untersucht werden.

Diese Arbeit hat sich auf Titeldaten gestützt, doch ist es vor diesem Hintergrund wichtig, die Aussagekraft von Titel zu reflektieren. Beispielsweise hat Adams (1987: 12) angebracht, dass der "wahre Titel" eines Werks vom Autor gewählt ist – es ist unklar, inwieweit das in der aktuellen publizistischen Praxis durchgesetzt wird, oder ob Mitsprache durch Verlag und Editor die Titelwahl verzerren. Ein weiterer Sonderfall bilden Werke, die mit gleichem Inhalt unter unterschiedlichem Namen erschienen sind. Dies kommt vor allem dann vor, wenn die gleichen Werke später und/oder durch einen neuen Verlag verlegt werden, oder wenn für Hardcover- und Taschenbuchversion unterschiedliche Namen gewählt wurden. Diese Sonderfälle wurden in dieser Arbeit ignoriert, würden sich aber dafür eigenen Ro-

bustheitschecks der Analysen vorzunehmen (es muss jedoch beachtet werden, dass Angaben zu Variantentitel nur teilweise im MARC21-Format verfügbar sind)[21].
Zusammenfassend hat die vorliegende Arbeit gezeigt, dass die automatisierte Verarbeitung von Titeldaten nicht ausreicht um Bücher zu klassifizieren und zu verstehen. Es bleibt weiterhin wichtig, Bücher zu lesen – möglicherweise kann dies automatisiert werden, jedoch brauchen auch Algorithmen dafür die Volltexte oder zumindest Buchbeschreibungen und nicht nur Metadaten. Nur dann kann der Kontext besser verstanden werden. Im Katalog der DNB sind aus urheberrechtlichen und anderen Gründen jedoch nur Titeldaten für alle Werke verfügbar[22]. Der nächste Abschnitt widmet sich den Problemen bei der Nutzung von Katalogdaten.

Nutzung von Katalogdaten Die Nutzung des Katalogs der DNB kommt nicht ohne Schwierigkeiten – der Datensatz ist nicht für (sozial-)wissenschaftliche Forschung kuratiert. Katalogdaten sind über Jahre gewachsen, komplex und nicht prozessgeneriert, sondern werden von Menschen geschrieben und bilden das ab, was in Bücher vermerkt sind – und auch das, was in Büchern vermerkt wird, ist nicht standardisiert.
Geplant war es in diesem Forschungsprojekt, Belletristik und Fachliteratur getrennt auszuwerten. Dies ist sinnvoll: Fachliteratur drückt in Titeln Anderes aus als Romane und Lyrik. Leider war dies mit der aktuellen Datengrundlage so nicht möglich. Die Angaben, welcher Art von Literatur ein Werk angehört, gibt es generell im Katalog: Die DDC und frühere Sachgruppen weisen aus, ob es sich bei einem Werk um Belletristik handelt oder nicht. Leider fehlen diese Angaben bei einem großen Teil der ausgewerteten Werke (bei ca. 22%). Dies ist sehr problematisch, da unklar ist, weshalb diese Klassifikationen fehlen; vorstellbar ist, dass sie damals als implizites Wissen im Zettelkatalog und im Standort des Buchs waren, jedoch im Digitalisierungsprozess nicht explizit übernommen wurden.
Alternative Ansätze, um diese fehlenden Daten zu ergänzen, sind nicht unproblematisch: Im Forschungsprojekt von Fischer und Jäschke (2018) wurden Werke betrachtet, die das Wort "Roman" im Untertitel enthalten. Ein solcher Ansatz hätte hier auch verfolgt werden können, um weitere Werke als Belletristik einzuordnen, jedoch ist ein solches Vorgehen von neuen Verzerrungen betroffen, deren Ausmaß und Effekt unklar sind. Es ist unklar, welche Autoren in welcher Periode und welche Verlage eher dazu neigen, solche Untertitel zu vergeben. Aus diesen Gründen wurde davon abgesehen, diesen Ansatz zu verfolgen.
Eine weitere Lücke im Bibliothekskatalog betrifft fehlende Geburtsjahre der Autoren. Der DNB-Katalog liefert eine enorm große Datenmenge, jedoch müssen die Daten für die sozialwissenschaftliche Forschung generell um sozialwissenschaftlich relevante Informationen

[21] Es ist beispielsweise in der MARC21-Repräsentation von "Sturmfieber" nicht ersichtlich, dass das Werk auch unter dem Titel "Zahn um Zahn" erschienen ist, siehe http://d-nb.info/984344896/about/marcxml. Bei dem Werk "Das Gold von Carcassonne" ist "Die geheimen Worte" jedoch als frührer Titel vermerkt, siehe https://d-nb.info/981946453.

[22] Forschung mit urheberrechtlichen geschützten Werken ist möglich, jedoch anderen Rahmenbedingungen unterlegen. Siehe dafür https://www.dnb.de/DE/Professionell/Services/WissenschaftundForschung/DHCall/dhcall_node.html.

angereichert werden. In diesem Projekt wurden die DNB-Titeldaten mit soziodemografischen Autorendaten von der GND und von Wikidata ergänzt. Beobachtungen mit fehlenden Angaben mussten notwendigerweise exkludiert werden, was die Fallzahl stark reduziert hat. Diese Reduktion ist vergleichbar mit Fischer (1998). Es ist das Angewiesensein auf weitere Angaben, was die Datenmenge stark reduziert. Da möglicherweise vor allem bekanntere Autoren in der GND und auf Wikidata vermerkt sind, ist der Datenverlust nicht ohne Bias: Die Daten sind nicht "missing at random" (Rubin 1976); dies ist ein Problem, das nur schwierig behoben werden kann, welches aber zu möglichen Verzerrungen der Ergebnisse führen kann.
Das Kombinieren mit anderen Datenquellen kommt zudem mit Unsicherheiten, da nicht für jeden Autor eine GND-ID verfügbar ist und Personen deswegen nicht immer eindeutig identifizierbar sind. Es gibt zudem auch Fälle, in denen die gleiche Person mehrere GND-IDs hat, und Fälle, in denen Autor zwar eine GND-ID besitzt, seine Werke aber nicht mit der ID verknüpft sind. Eine weitere Fehlerquelle dieser Auswertung ergibt sich dadurch, dass Erscheinungsjahr des Buches und Lebensalter des schreibenden Autors nicht immer zusammenfallen, wie in dieser Arbeit angenommen wird. Dies fällt insbesondere bei der Veröffentlichung von Werken posthum auf.

Wie können die Katalogdaten der DNB für wissenschaftliche Zwecke noch nutzbarer gemacht werden? Anknüpfend an die diskutierten Limitationen der vorliegenden Analysen wäre es wünschenswert, wenn die Informationen aus dem PICA-Format vollständig übernommen würden und wenn allfällige Fehler behoben werden würden: Eine fehlerfreie Datengrundlage ist in jedem Fall ein wünschenswertes Ziel. Allen voran wünsche ich mir vor dem Hintergrund der sozialwissenschaftlichen Forschung jedoch insbesondere einen verstärkten Umgang mit Zufallsstichproben.
Bibliothekskataloge entstehen und existieren nicht für wissenschaftliche Zwecke sondern für bibliothekarische. Dies kann zu einem Zielkonflikt führen, der eine saubere, wissenschaftliche Nutzung erschwert. Die bibliothekarische Praxis strebt nach Vollerhebung und Komplettübersichten. Dies ist ein ehrenwertes Ziel. Die bibliothekarische Praxis arbeitet systematisch, um dieses Ziel zu erreichen – und diese Systematik ist problematisch für datengetriebene Forschung. Systematische Verarbeitung führt zu systematischen Verzerrungen, die schwierig zu fassen sind. Veränderungen in der Praxis über die Zeit können es unmöglich machen, zeitliche Entwicklungen zu untersuchen, da unklar bleibt, ob sich die Realität verändert hat oder lediglich die Daten. Statt mit einer zeitlichen Systematik zu arbeiten – beispielsweise das Verschlagworten 1000 neuer Werke ab einem Stichtag – wäre es aus Forschungssicht hilfreicher, 1000 zufällig ausgewählte Werke zu verschlagworten. Konfidenzintervalle und Schätzverfahren sind dazu in der Lage, mit Zufallsfehlern und Stichproben umzugehen. Systematische Verzerrungen wiederum wirken sich stets auf die Validität der Ergebnisse aus. Solche systematischen Verzerrungen sind vor allem dann problematisch, wenn sie verdeckt und unklar bleiben – nicht immer sind Veränderungen in den

Daten offensichtlich. Veränderungen der Datengrundlage, speziell Veränderungen über die Zeit hinweg, müssen transparent gemacht werden, um saubere Forschung zu ermöglichen. Nur wenn klar ist, welche Aspekte sich verändert haben – und im Idealfall auch warum und mit welcher Konsequenz – lassen sich systematische Verzerrungen überhaupt diskutieren. Es wäre daher in jedem Fall wünschenswert, wenn der Bibliothekskatalog gründlich dokumentiert wäre.

Systematische Verzerrungen machen Daten für viele Zwecke unbrauchbar, vor allem dann, wenn der Bias nach außen hin nicht bekannt ist. Dies kann auch zu Fehlinterpretationen der Daten führen. Es ist verständlich, dass die DNB bisher nicht mit Zufallsstichproben gearbeitet hat: Logistisch kann dies sehr viel komplizierter bis unmöglich sein, wenn man mit physischen Büchern arbeitet und keinen digitalisierten Katalog hat. In diesem Fall kann die einzig sinnvolle Herangehensweise sein, Stapel bzw. Jahre komplett abzuarbeiten. Mit neuen, digitalen Möglichkeiten sollte das Ziehen von Stichproben jedoch auch in der bibliothekarischen Praxis möglich sein. Wird das Ziel der Vollerhebung und der vollständigen Abdeckung aller Informationen erreicht, so ist es selbstverständlich irrelevant, ob das Ziel durch eine stichprobenartigen oder durch einen systematischen Ansatz erreicht wurde. Aber, bis das Ziel erreicht wurde (und wenn überhaupt!), ist eine gute Zufallsauswahl für viele Zwecke sehr viel brauchbarer.

Trotz dieser Limitationen und Einschränkungen bieten die Katalogdaten der DNB eine einzigartige Übersicht über die deutschsprachige Literatur. Diese Arbeit hat als eine der ersten versucht, die Katalogdaten auf soziologisch-demografische Fragestellungen anzuwenden und mit Titeldaten zu arbeiten. Dies birgt Chancen, kommt aber mit einer Vielzahl an Limitationen und Herausforderungen, die in zukünftigen Projekten beachtet werden müssen. Nicht nur für eine Sozialgeschichte der Nachkriegsliteratur, sondern auch für sozialwissenschaftliche Fragestellungen, die in die zeitliche Breite der Buchproduktion gehen möchte, gilt daher leider weiterhin: "Dazu fehlen die einfachsten Materialgrundlagen" (Fischer 1998: 556; zitiert nach Häntzschel u. a. 2009: 17).

Literatur

Adams, Hazard (1987). Titles, Titling, and Entitlement To. *The Journal of Aesthetics and Art Criticism* 46(1): 7–21.

Alt, Peter-Andre (2000). *Schiller. Leben - Werk - Zeit.* Frankfurt am Main: C.H.Beck.

Bauer, Alexander, Maximilian Weigert und Hawre Jalal (2023). *APCtools: Routines for Descriptive and Model-Based APC Analysis.* Manual.

Begum, Soheli (2011). Readers' Advisory and Underestimated Roles of Escapist Reading. *Library Review* 6(9): 738–747.

Bell, Andrew und Kelvyn Jones (2014). Current Practice in the Modelling of Age, Period and Cohort Effects with Panel Data: A Commentary on Tawfik et al. (2012), Clarke et al. (2009), and McCulloch (2012). *Quality & Quantity* 48(4): 2089–2095.

Bell, Andrew und Kelvyn Jones (2015). Age, Period and Cohort Processes in Longitudinal and Life Course Analysis: A Multilevel Perspective. In: *A Life Course Perspective on Health Tra-*

jectories and Transitions. Hrsg. von Claudine Burton-Jeangros, Stéphane Cullati, Amanda Sacker und David Blane. Cham (CH): Springer.

Blackburn, Heidi und Jason Heppler (2022). Hidden Voices: A Case Study Analysis of Subject Headings for Book Titles on Women in Science. *Science & Technology Libraries* 42(1): 31–49.

Blei, David M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 2: 993–1022.

Dassonneville, Ruth, Marc Hooghe und Bram Vanhoutte (2012). Age, Period and Cohort Effects in the Decline of Party Identification in Germany: An Analysis of a Two Decade Panel Study in Germany (1992–2009). *German Politics* 21(2): 209–227.

Diekmann, Andreas (2004). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Hamburg: Rowohlt.

Döring, Lisa (2018). Analyse von Alters-, Perioden- und Kohorteneffekten mit Hilfe von Standard-Kohorten-Tabellen. In: *Mobilitätsbiografien und Mobilitätssozialisation*. Wiesbaden: Springer Fachmedien: S. 85–126.

Farkas, G. (1977). Cohort, Age, and Period Effects upon the Employment of White Females: Evidence for 1957-1968. *Demography* 14(1): 33–42.

Fischer, Frank und Robert Jäschke (2018). Liebe und Tod in der Deutschen Nationalbibliothek. In: *DHd2018: Kritik der digitalen Vernunft*: S. 261–266.

Fischer, Ludwig (1998). Zur Sozialgeschichte der westdeutschen Literatur. In: *Modernisierung im Wiederaufbau. Die westdeutsche Gesellschaft der 50er Jahre*. Hrsg. von Axel Schildt und Arnold Sywottek. Bonn: Dietz: S. 551–562.

Fosse, Ethan und Christopher Winship (2019a). Analyzing Age-Period-Cohort Data: A Review and Critique. *Annual Review of Sociology* 45(1): 467–492.

Fosse, Ethan und Christopher Winship (2019b). Bounding Analyses of Age-Period-Cohort Effects. *Demography* 56(5): 1975–2004.

Genette, Gérard und Bernard Crampé (1988). Structure and Functions of the Title in Literature. *Critical Inquiry* 14(4): 692–720.

Grasso, Maria T. (2014). Age, Period and Cohort Analysis in a Comparative Context: Political Generations and Political Participation Repertoires in Western Europe. *Electoral Studies* 33: 63–76.

Green, Jon, Jared Edgerton, Daniel Naftel, Kelsey Shoub und Skyler J. Cranmer (2020). Elusive Consensus: Polarization in Elite Communication on the COVID-19 Pandemic. *Science Advances* 6(28): eabc2717.

Häntzschel, Günter, Adrian Hummel und Jörg Zedler (2009). *Deutschsprachige Buchkultur der 1950er Jahre: Fiktionale Literatur in Quellen, Analysen und Interpretationen*. Wiesbaden: Otto Harrassowitz Verlag.

Howard, Vivian (2011). The Importance of Pleasure Reading in the Lives of Young Teens: Self-identification, Self-Construction and Self-Awareness. *Journal of Librarianship and Information Science* 43(1): 46–55.

Hughes, Tiffany F., Chung-Chou H. Chang, Joni Vander Bilt und Mary Ganguli (2010). Engagement in Reading and Hobbies and Risk of Incident Dementia: The MoVIES Project. *American Journal of Alzheimer's Disease & Other Dementias* 25(5): 432–438.

Jacobi, Carina, Wouter Van Atteveldt und Kasper Welbers (2016). Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital Journalism* 4(1): 89–106.

Levin, Harry (1977). The Title as a Literary Genre. *The Modern Language Review* 72(4): xxiii.

Luo, Liying und James S. Hodges (2020). The Age-Period-Cohort-Interaction Model for Describing and Investigating Inter-Cohort Deviations and Intra-Cohort Life-Course Dynamics. *Sociological Methods & Research* 51(3): 1164–1210.

Martins, Mauricio d. J. D. und Nicolas Baumard (2020). The Rise of Prosociality in Fiction Preceded Democratic Revolutions in Early Modern Europe. *Proceedings of the National Academy of Sciences* 117(46): 28684–28691.

Mimno, David, Hanna M. Wallach, Edmund T. M. Leenders und Andrew McCallum (2011). Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Moretti, Franco (2009). Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850). *Critical Inquiry* 36(1): 134–158.

Mozaheb, Mohammad A., Amir Ghajarieh und Zahra Tamizi (2022). Analysis of Novel Title: A Case Study of Agatha Christie's Works Translated from English into Persian. *Journal of Language and Translation* 12(2): 177–190.

Park, Chiho, Yon H. Jee und Keum J. Jung (2016). Age–Period–Cohort Analysis of the Suicide Rate in Korea. *Journal of Affective Disorders* 194: 16–20.

Rea, Amy (2020). Reading through the Ages: Generational Reading Survey. *Library Journal*.

Robinson, Robert V. und Elton F. Jackson (2001). Is Trust in Others Declining in America? An Age–Period–Cohort Analysis. *Social Science Research* 30(1): 117–145.

Rothe, Arnold (1986). *Der literarische Titel. Funktionen, Formen, Geschichte*. Frankfurt am Main: V. Klostermann: S. 479.

Rubin, Donald B. (1976). Inference and Missing Data. *Biometrika* 63(3): 581–592.

Schmidt, Ben (2017). A Brief Visual History of MARC Cataloging at the Library of Congress. URL: http://www.benschmidt.org/sappingattention/a-brief-visual-history-of-marc/ (besucht am 12. 07. 2023).

Al-Shaibani, Hanaa A. und Salam Al-Augby (2022). Terrorist Tweets Detection Using Sentiment Analysis: Techniques and Approaches. In: *2022 5th International Conference on Engineering Technology and Its Applications (IICETA)*: S. 585–590.

Shevlin, Eleanor F. (1999). "To Reconcile Book and Title, and Make'em Kin to One Another": The Evolution of the Title's Contractual Functions. *Book History* 2(1): 42–77.

Smets, Kaat und Anja Neundorf (2014). The Hierarchies of Age-Period-Cohort Research: Political Context and the Development of Generational Turnout Patterns. *Electoral Studies* 33: 41–51.

Sörman, Daniel E., Jessica K. Ljungberg und Michael Rönnlund (2018). Reading Habits among Older Adults in Relation to Level and 15-Year Changes in Verbal Fluency and Episodic Recall. *Frontiers in Psychology*: 1872.

Splendid Research (2017). Studie: 61 Prozent der Deutschen lesen noch immer Bücher. URL: https://www.splendid-research.com/de/statistiken/studie-buecher-deutschland (besucht am 15. 04. 2022).

Stegmueller, Daniel (2014). Bayesian Hierarchical Age-Period-Cohort Models with Time-Structured Effects: An Application to Religious Voting in the US, 1972–2008. *Electoral Studies* 33: 52–62.

Sullivan, Ceri (2007). Disposable Elements? Indications of Genre in Early Modern Titles. *The Modern Language Review* 102(3): 641–653.

Torres-Salinas, Daniel und Wenceslao Arroyo-Machado (2020). Library Catalog Analysis and Library Holdings Counts: Origins, Methodological Issues and Application to the Field of Informetrics. In: *Evaluative Informetrics: The Art of Metrics-Based Research Assessment*. New York: Springer International Publishing: S. 287–308.

Torres-Salinas, Daniel und Henk F. Moed (2009). Library Catalog Analysis as a Tool in Studies of Social Sciences and Humanities: An Exploratory Study of Published Book Titles in Economics. *Journal of Informetrics* 3(1): 9–26.

Wasik, Barbara A., Annemarie H. Hindman und Emily K. Snell (2016). Book Reading and Vocabulary Development: A Systematic Review. *Early Childhood Research Quarterly* 37: 39–57.

Weidenbach, Bernhard und Statista (2021). Buchtitelproduktion: Anzahl der Neuerscheinungen in Deutschland in den Jahren 2002 bis 2020. URL: https://de.statista.com/statistik/daten/studie/39166/umfrage/verlagswesen-buchtitelproduktion-in-deutschland/ (besucht am 01. 06. 2023).

Weigert, Maximilian, Alexander Bauer, Johanna Gernert, Marion Karl, Asmik Nalmpatian, Helmut Küchenhoff und Jürgen Schmude (2022). Semiparametric APC Analysis of Destination Choice Patterns: Using Generalized Additive Models to Quantify the Impact of Age, Period, and Cohort on Travel Distances. *Tourism Economics* 28(5): 1377–1400.

Wilsmore, Susan J. (1987). The Role of Titles in Identifying Literary Works. *The Journal of Aesthetics and Art Criticism* 45(4): 403–408.

Witte, Bernd, Theo Buck, Hans-Dietrich Dahnke, Regine Otto und Peter Schmidt (1996). *Goethe-Handbuch, 4 Bde. in 5 Tl.-Bdn. u. Register, Bd.2, Dramen: Band 2: Dramen*. Stuttgart: J.B. Metzler.