ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen

Autor/in:
Christof Schöch

Kontakt: schoech@uni-trier.de
Institution: Universität Trier, Trier Center for Digital Humanities
GND: 135594480 ORCID: 0000-0002-4557-2753

Autor/in:
Frédéric Döhl

Kontakt: f.doehl@dnb.de
Institution: Deutsche Nationalbibliothek
GND: 13895500X ORCID: 0000-0003-3493-8585

Autor/in:
Achim Rettinger

Kontakt: rettinger@uni-trier.de
Institution: Universität Trier, Computerlinguistik und Digital Humanities
GND: 143299247 ORCID: 0000-0003-4950-1167

Autor/in:
Evelyn Gius

Kontakt: evelyn.gius@tu-darmstadt.de
Institution: Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft
GND: 1084241307 ORCID: 0000-0001-8888-8419

Autor/in:
Peer Trilcke

Kontakt: trilcke@uni-potsdam.de
Institution: Universität Potsdam, Theodor-Fontane-Archiv
GND: 139145117 ORCID: 0000-0002-1421-4320

Autor/in:
Peter Leinen

Kontakt: P.Leinen@dnb.de
Institution: Deutsche Nationalbibliothek
GND: 1079692932 ORCID: 0000-0002-3014-000X

Autor/in:
Fotis Jannidis

Kontakt: fotis.jannidis@uni-wuerzburg.de
Institution: Julius-Maximilians-Universität Würzburg, Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte
GND: 114523525 ORCID: 0000-0001-6944-6113

Autor/in:
Maria Hinzmann

Kontakt: hinzmannm@uni-trier.de
Institution: Universität Trier, Trier Center for Digital Humanities
GND: 1220315826 ORCID: 0000-0001-7199-1436

Autor/in:
Jörg Röpke

Kontakt: roepke@uni-trier.de
Institution: Universität Trier, Universitätsbibliothek
GND: 111990093X ORCID: 0000-0002-1575-6105

DOI des Artikels:
10.17175/2020_006

Nachweis im OPAC der Herzog August Bibliothek:
1726146014

Erstveröffentlichung:
05.11.2020

Letzte Überprüfung aller Verweise:
30.10.2020

GND-Verschlagwortung:
Data Mining | Literaturwissenschaft | Open Science | Text Mining | Urheberrecht |

Zitierweise:
Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis: Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel 2020. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2020_006.

Christof Schöch, Frédéric Döhl, Achim Rettinger , Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis

Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen

Abstracts

Das *Text* und *Data Mining* (TDM) mit urheberrechtlich geschützten Texten unterliegt trotz der TDM-Schranke (§ 60d UrhG) weiterhin Einschränkungen, die u. a. die Speicherung, Veröffentlichung und Nachnutzung der entstehenden Korpora betreffen und das volle Potenzial des TDM in den *Digital Humanities* ungenutzt lassen. Als Lösung werden *abgeleitete Textformate* vorgeschlagen: Hier werden urheberrechtlich geschützte Textbestände so transformiert, dass alle wesentlichen urheberrechtlich relevanten Merkmale entfernt werden, verschiedene einschlägige Methoden des TDM aber weiterhin zum Einsatz kommen können. Mehrere abgeleitete Textformate werden aus Sicht der *Computational Literary Studies*, der Informatik, der Gedächtnisinstitutionen und der Rechtswissenschaften beleuchtet.

Despite the TDM exception in German copyright law, *Text* and *Data Mining* (TDM) with copyrighted texts is still subject to restrictions, including those concerning the storage, publication and follow-up use of the resulting corpora, leaving the full potential of TDM in the *Digital Humanities* untapped. We propose *derived text formats* as a solution: here, copyrighted textual materials are transformed in such a way that copyright-relevant features are removed, but that the use of various relevant methods of TDM remains possible. Several derived text formats are examined from the perspectives of *Computational Literary Studies*, Computer Science, memory institutions and Law.

1. Einleitung

Es ist ein offenes Geheimnis in den *Digital Humanities* (DH), dass es für die *Computational Literary Studies* (CLS) bezüglich der verfügbaren Textbestände ein *window of opportunity* gibt, das sich um 1800 öffnet und um 1920 wieder schließt. Es öffnet sich um 1800, weil für Materialien vor dieser Zeit die technischen Herausforderungen im Bereich *Optical Character Recognition* (OCR) und Normalisierung von orthographischer Varianz immer noch so groß sind, dass deutlich weniger umfangreiche beziehungsweise qualitativ weniger hochwertige Textsammlungen zur Verfügung stehen als für die Zeit nach 1800. Und es schließt sich um 1920, weil für Texte, die später erschienen sind, in sehr vielen Fällen (abhängig vom Todesdatum der Autor*innen) das Urheberrecht nach wie vor greift und sowohl das Erstellen als auch das Teilen von Textsammlungen mit Dritten damit deutlich erschwert sind. Dieser Umstand hat bedauerlicherweise zur Folge, dass die Setzung von Forschungsschwerpunkten häufig nicht primär von den Erkenntnisinteressen und Zielen der Forschung selbst, sondern wesentlich von technischen und rechtlichen, also dieser Forschung externen Faktoren, bestimmt wird. Als Konsequenz daraus ist eine Forschung auf dem methodischen *state of the art* mit neueren Textbeständen nur begrenzt, teilweise sogar überhaupt nicht möglich. Die Forschung in den CLS verwendet zwar aktuelle, oft aus Informatik, Computerlinguistik und Statistik adaptierte Verfahren, kann sie aber in den meisten Fällen nicht auch auf diejenigen Textbestände anwenden, die unsere zeitgenössische literarische Kultur ausmachen.[1]

Allerdings verbessert sich die Lage seit einigen Jahren deutlich, sodass es Anlass zu Optimismus gibt: Auf der einen Seite wird derzeit verstärkt in neue Verfahren für OCR investiert, bei denen die Texterkennung auf neuronalen Netzen beruht und deutliche Verbesserungen auch für Materialien vor 1800, bis hin zu frühen Drucken und Handschriften, erreichen konnte.[2] Auf der anderen Seite haben sich die rechtlichen Rahmenbedingungen in den letzten Jahren stark zum Positiven verändert, in Deutschland insbesondere mit dem März 2018 in Kraft getretenen Gesetz zur Angleichung des Urheberrechts an die aktuellen Erfordernisse der Wissensgesellschaft (UrhWissG)[3], welches das Gesetz über Urheberrecht und verwandte Schutzrechte (UrhG) modifizierte. So enthält das derzeit geltende nationale Recht in § 60d UrhG eine Schranke[4] von Urheber- und Leistungsschutzrechten zugunsten des Text und Data Mining.[5] Diese Schranke gestattet nun Vervielfältigungen zum Zwecke des Text und Data Mining (TDM) in der Forschung, erlaubt allerdings nur unter engen Voraussetzungen die langfristige Speicherung und die Weitergabe der hierfür erstellten Korpora an Dritte. So darf beispielsweise die langfristige Speicherung (und fallweise Weitergabe) der Korpora nur durch Institutionen wie Bibliotheken und Archive erfolgen. Und es ist zwar gestattet, ein Korpus für die wissenschaftliche Qualitätssicherung (*peer review*) an einzelne Dritte weiterzugeben, nicht aber für die ebenso wichtige Anschlussforschung. Auch

[1] Der vorliegende Beitrag ist im Kontext der Workshopreihe »Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte« am 28.11.2019 und 17.01.2020 in Trier, organisiert von Benjamin Raue (IRDT) und Christof Schöch (TCDH) entstanden. Wir danken den Teilnehmer*innen des Workshops für die engagierten und produktiven Diskussionen und der DFG für die Förderung der Veranstaltung im LIS-Programm. Ein wesentlicher Impuls für die Workshopreihe war die Trierer Tagung Text und Data Mining – in Recht, Wissenschaft und Gesellschaft 2018.
[2] Vgl. zu großen Initiativen wie OCR-D oder Werkzeugen wie OCR4all u. a. Neudecker et al. 2019, passim; Reul et al. 2019, passim.
[3] Hierbei handelt es sich um die Ergänzungen in §§ 60a–60h UrhG.
[4] Eine ›Schranke‹ bezeichnet im rechtswissenschaftlichen Kontext eine gezielte Einschränkung oder eine Ausnahme von einer allgemeineren rechtlichen Regelung.
[5] Einführend dazu Raue 2019, passim.

auf der europäischen Ebene bringt die Europäische Richtlinie 2019/790 zum Urheberrecht im digitalen Binnenmarkt (DSM-RL) von 2019 erneut eine Öffnung für das TDM, die auch in deutsches Recht umgesetzt werden wird; doch auch diese Regelungen werden nicht alle Bedarfe des TDM erfüllen.

Der vorliegende Beitrag macht Vorschläge für einen pragmatischen Umgang mit der derzeitigen rechtlichen Situation bei der Nutzung von Methoden des TDM[6] in den DH und speziell in den CLS.[7] Es sollen Perspektiven und Möglichkeiten für die Erstellung, Analyse und Anschlussforschung an solchen Textsammlungen, die auf der Grundlage von urheberrechtlich geschützten Textbeständen entstanden sind, eröffnet werden. Ziel ist es, die offene Publikation und freie Nachnutzbarkeit von *abgeleiteten Textformaten* – zur Nachvollziehbarkeit von Analyseergebnissen für Dritte auch außerhalb formaler Qualitätssicherungsprozesse wie auch zur Anschlussforschung ohne rechtliche Einschränkungen – zu ermöglichen. Dies ist ein Bereich, den weder das geltende noch das anstehende neue Recht gesondert behandeln, wodurch die rechtsuchenden Forschenden weitgehend auf allgemeine Regeln und Grundsätze des Urheberrechts verwiesen bleiben.

Grundsätzlich kann man vier Ansätze zur Lösung der beschriebenen Problematik unterscheiden:

- den Zugang zu lizenzierten Inhalten über eine API
- die Nutzung von Analyseplattformen
- die Forschung im *closed room*
- und die Arbeit mit abgeleiteten Textformaten

Erstens kann man also auf die lizenzierten Angebote beispielsweise von großen Verlagen setzen, die es über Schnittstellen, sogenannte *Application Programming Interfaces* (APIs), ermöglichen, gezielt ausgewählte, große Mengen an Text-/Datenbeständen herunterzuladen, um sie dann selbst mit Methoden des TDM zu analysieren. Nachteilig hieran ist, dass man auf die Bestände der Verlage beschränkt ist und in der Regel die auf diese Weise erstellten Korpora nicht weitergeben darf. Zweitens kann man auf Plattformen von Datenanbietern setzen, die bestimmte Textbestände und Analyseverfahren zur Verfügung stellen und so Forschung gewissermaßen vom Browser aus ermöglichen. Nachteilig an diesem Modell ist, dass die Forschenden dabei die Daten selbst nicht einsehen, herunterladen, modifizieren oder ergänzen oder Datensätze aus verschiedenen Quellen kombinieren können. Außerdem können die Forschenden die Analyseverfahren selbst nicht implementieren oder grundlegend modifizieren. Drittens kann das *closed room-Modell* zum Einsatz kommen, bei dem Forschende in einem technisch abgeschotteten Raum in einer bestimmten Institution Zugang zu Textbeständen der Institution erlangen. Dies können sie dann auf dem vor Ort verfügbaren Arbeitsplatzrechner auch mit individuell entwickelten Analyseprogrammen untersuchen. Hier liegen die Nachteile in erster Linie in der Ortsgebundenheit der Forschungstätigkeit, die einen entsprechenden praktischen Aufwand mit sich bringt und in Zeiten der Digitalisierung wenig zeitgemäß erscheint; zudem skaliert dieses Modell nicht gut, können Datensätze aus mehreren Institutionen nicht kombiniert und an Dritte weitergegeben werden.

Die hier entwickelten Perspektiven beziehen sich in erster Linie auf den vierten Ansatz, die abgeleiteten Textformate. Abgeleitete Textformate sind auf der Grundlage eines Ausgangstextes systematisch generierte Repräsentationen des Textes, welche die Anwendung bestimmter Verfahren des Text und Data Mining erlauben, wobei urheberrechtlich geschützte Bestandteile des Ausgangstextes im abgeleiteten Format aber nicht mehr repräsentiert sind.[8] Hieraus ergeben sich neue Möglichkeiten, aber auch neue Herausforderungen: Die Forschenden können dadurch ihre Forschungsfragen anhand größerer und vor allem aktuellerer Textbestände bearbeiten und damit auch neue Forschungsfragen erschließen; allerdings müssen die eingesetzten Verfahren gegebenenfalls auf die abgeleiteten Textformate angepasst werden. Ebenso werden durch die freie Verfügbarkeit der abgeleiteten Textformate Transparenz und Reproduzierbarkeit von Forschung und die unbeschränkte Zirkulation und Nachnutzbarkeit von Forschungsdaten bei urheberrechtlich geschützten Ursprungsmaterialien überhaupt erst umfassend ermöglicht; allerdings ergibt sich für alle Forschenden eine hohe Abhängigkeit von der Transparenz, Plausibilität und Korrektheit der Erstellungsprozesse der abgeleiteten Textformate.

Zunächst und noch grundlegender stellt sich aber die Frage, welche abgeleiteten Textformate überhaupt für einen solchen Ansatz geeignet sind. Mit dem Ziel, erste Antworten auf diese Frage zu entwickeln und eine Diskussion in der Fachgemeinde der DH anzuregen, wird im Folgenden zunächst die aktuelle rechtliche Situation zusammengefasst. Anschließend werden das Prinzip der abgeleiteten Textformate sowie die Anforderungen an abgeleitete Textformate für verschiedene relevante Verfahren des TDM diskutiert. Es folgt die Beschreibung und kritische Begutachtung mehrerer konkreter abgeleiteter Formate, und zwar

[6] Vgl. zu TDM u. a. Hotho et al. 2005, passim; Allahyari et al. 2017, passim.
[7] Vgl. zu CLS u. a. Jannidis et al. 2017, passim; Jockers 2013, passim; Schöch 2017a, passim.
[8] Die rechtlichen Kriterien, die ein Urteil hinsichtlich der Frage erlauben, ob ein abgeleitetes Textformat dem Urheberrecht unterliegt oder nicht, sind Gegenstand eines aktuellen Beitrags von Karina Grisse (Grisse 2020, passim); zu den relevanten Faktoren gehören u. a. die Möglichkeit des Werkgenusses, die Wiedererkennbarkeit und die Rekonstruierbarkeit des Ausgangstextes. Die durch das Urheberrechtsgesetz vermittelten Schutzmöglichkeiten spezifisch für Texte sowie die Schutzvoraussetzungen erläutert Florian Jotzo (Jotzo 2020, passim). Beide Beiträge sind als rechtswissenschaftliche *companion papers* zum vorliegenden Beitrag konzipiert.

sowohl bezüglich ihrer Erstellung, Publikation und Anwendung als auch hinsichtlich ihrer urheberrechtlichen Einordnung. Abschließend kommen im Sinne eines Ausblicks auf die nächsten erforderlichen Schritte einige weiterführende Punkte zur Sprache: Diese beleuchten insbesondere die neue Rolle von Bibliotheken und Archiven als Datenanbieter, skizzieren aber auch eine Agenda für Informatik, Computerlinguistik und DH. Ziel des Beitrags ist es in der Summe, eine Diskussion nicht nur innerhalb der DH, sondern auch mit den Bibliotheken und Archiven einerseits sowie den Rechtswissenschaften andererseits anzuregen. Am Ende einer solchen Diskussion könnte ein Konsens über ein Inventar geeigneter abgeleiteter Textformate stehen, für die standardisierte Lösungen zur Erstellung und Bereitstellung entwickelt und angeboten werden. Dies wird auch Grundlage für weitergehende Diskussionen z. B. mit Vertreter*innen der Rechteinhaber (u. a. der Verlage) sein können.

2. Die aktuelle Rechtslage zum Text und Data Mining mit urheberrechtlich relevanten Textbeständen

Bevor auf die im Beitrag zentrale Strategie der abgeleiteten Textformate eingegangen wird, soll an dieser Stelle zunächst die aktuelle Rechtslage zum Thema Text und Data Mining (TDM) mit urheberrechtlich relevanten Textbeständen zusammenfassend dargestellt werden. Unter TDM versteht das Urheberrecht das automatisierte Auswerten einer Vielzahl von Werken[9] beziehungsweise, wie es nun in Art. 2 DSM-RL EU-weit korrespondierend, aber etwas ausführlicher legaldefiniert ist, die »Technik für die automatisierte Analyse von Texten und Daten in digitaler Form, mit deren Hilfe Informationen unter anderem – aber nicht ausschließlich – über Muster, Trends und Korrelationen gewonnen werden können.«[10]

Das TDM selbst ist in der Regel keine urheberrechtsrelevante Handlung, sehr wohl aber all die Vervielfältigungs- und Bearbeitungsschritte, die mit der Korpusbildung einhergehen, und gegebenenfalls auch das, was nach der Analyse in der Ergebnispräsentation von der Verfasstheit und ästhetischen Identität der analysierten Werke wieder sichtbar wird. Hier entstehen Konflikte zwischen Forschungsinteressen einerseits und einem in seinem Kern vordigitalen, seit 1966 in Kraft befindlichen UrhG andererseits. Die Sache liegt zusätzlich schief, weil es beim TDM nicht um Werkintegrität und Werkgenuss im klassischen Sinne geht, also das, was das Urheberrecht zu regulieren sucht. Vielmehr handelt es sich um Informationen, die sich quasi in geschützten Werkhüllen finden, auch wenn sie oftmals für sich genommen selbst gar nicht schutzfähig sind. Zusätzlich verkompliziert wird die Lage dadurch, dass es im TDM vielfach um Massenanalysen geht, was Einzellizenzierungen der analysierten Werke und damit diese Art Forschung an geschützten Inhalten praktisch unmöglich macht.

TDM war daher vor dem UrhWissG heikel und nur mit aufwendigen argumentativen Konstruktionen als erlaubnisfrei rechtlich herzuleiten. Die Sache war also mindestens rechtsunsicher, das Ausweichen auf klar rechtsfreie Analysegegenstände die Regel. Aus wissenschaftlicher Sicht war daher durch die technische und forschungspolitische Entwicklung ein Bedarf für eine Beschränkung der beim TDM berührten Urheber- und Leistungsschutzrechte entstanden, die einen Spielraum für rechtssicheres erlaubnisfreies TDM schafft. Dieser Einschätzung folgte auch der Gesetzgeber, zumal man darauf verweisen konnte, dass die grundrechtlich abgesicherten berechtigten persönlichkeitsrechtlichen und vor allem ökonomischen Interessen der Rechteinhaber durch TDM in der Regel überhaupt nicht nennenswert tangiert werden. Das derzeit geltende nationale Recht wurde vor diesem Hintergrund 2018 in Deutschland unter Bezugnahme auf die allgemeinen Ausnahmen und Beschränkungen zugunsten von Wissenschaft[11] und die 2001 in der sogenannten InfoSoc-RL 2001/29/EG EU-weit gewährten und vollharmonisierten Vervielfältigungsrechte von Urhebern und Leistungsschutzberechtigten[12] eingeführt. Es sieht seit nunmehr zwei Jahren in § 60d UrhG eine Schranke von Urheber- und Leistungsschutzrechten zugunsten des Text und Data Mining vor.[13] Nach diesem seit März 2018 geltenden Recht ist nun insbesondere Folgendes erlaubnisfrei zulässig:[14] das Erstellen von Korpora aus Werken jedweder Art[15] für das TDM einschließlich aller für Korpusbildung und das anschließende TDM erforderlichen Bearbeitungs- und Vervielfältigungshandlungen etwa des Digitalisierens, Normalisierens, Strukturierens, Kategorisierens, Sortierens, Annotierens, Kombinierens aus verschiedenen Quellen,[16] vorausgesetzt, man verfolgt nicht-kommerzielle, wissenschaftliche Forschung und hat rechtmäßigen Zugang zum Ursprungsmaterial. Für alle Arbeitsschritte der Vorbereitung und Durchführung dürfen sich die Forschenden dabei auch von Dritten, etwa Gedächtnisinstitutionen, unterstützen lassen; des Weiteren dürfen Forschende

[9] Vgl. § 60d Abs. 1 UrhG.
[10] Art. 2 Nr. 2 DSM-RL.
[11] Vgl. Art. 5 Abs. 3 Buchst. a InfoSoc-RL 2001/29/EG.
[12] Vgl. Art. 2 InfoSoc-RL 2001/29/EG.
[13] Vgl. zu UrhWissG und DH/TDM Berger 2017, passim; Durantaye 2017, passim; Pflüger / Hinte 2018, passim; Raue 2017a, passim; Raue 2017b, passim; Schack 2017, passim; Specht 2018, passim; Spindler 2016, passim; Spindler 2018, passim.
[14] Vgl. insoweit einführend die Standardkommentierungen zu § 60d UrhG in Dreier 2018, § 60d, Rn 116; Nordemann 2018, § 60d, Rn 1–13; Anton 2019, § 60d, Rn 1–23; Bullinger 2019, § 60d, Rn 1–21; Hagemeier 2020, § 60d, Rn 1–21.
[15] §§ 2 UrhG: Texte, Töne, Bilder, Filme usw., sogar Datenbankwerke, § 60d Abs. 2 UrhG.
[16] Vgl. § 23, S. 3; § 44a; § 60d, Abs. 1, S. 1, Nr. 1 UrhG.

das Korpus bis zur Grenze eines fest bestimmten, aber doch schon rechtlich als öffentlich geltenden Personenkreises zur gemeinsamen Arbeit daran zugänglich machen; und sie dürfen nach Projektabschluss zur Sicherung von Referenzierbarkeit und Qualitätsprüfung das Korpus privilegierten Gedächtnisinstitutionen zur dauerhaften Aufbewahrung übergeben.

Das ist viel, verglichen mit der Rechtslage zuvor. Doch so positiv diese mit dem UrhWissG vorgenommenen Änderungen für die wissenschaftliche Arbeit im Bereich des TDM in mancherlei Hinsicht auch sind, sie stellen weiterhin entscheidende Hürden für die Forschung dar, die den Mehrwert und die Praktikabilität der Regelung für die Forschung spürbar einschränken:

- Rechtmäßiger Zugang zu den verwendeten Materialien wird z. B. vorausgesetzt, nicht unter Verweis auf TDM zum Zwecke wissenschaftlicher Forschung automatisch gewährt
- Es besteht kein Recht auf Umgehung etwaiger technischer Schutzmaßnahmen, nur das Recht, vom Beschränkenden Mittel zur Aufhebung des Schutzes zu erhalten,[17] aber auch nur, wenn dies nicht bei Online-Inhalten durch vertragliche Vereinbarung, auch als AGB, ausgeschlossen ist[18]
- Die TDM-Schranke gilt nur im Rahmen von nichtkommerzieller Forschung
- Das Korpus ist nach Ende des Forschungsprojektes an eine Gedächtnisinstitution abzugeben oder zu löschen
- Das Korpus kann der Wissenschaftscommunity jenseits des engen Bereichs der Qualitätsprüfung nicht für Anschlussforschung zugänglich gemacht werden
- Änderungsverbot, Vergütungspflicht und Pflicht zu Quellenangaben sind zu beachten
- Die Arbeit mit rechtswidrigen Werken (Plagiaten, geleaktem Material, jugendgefährdendem Material usw.) ist nicht geklärt

All dies sind nur Beispiele für nach wie vor bestehende Hürden und Unsicherheiten. Und hier sind nur Urheber- und Leistungsschutzrecht angesprochen, das Datenschutzrecht kommt erst noch,[19] ist aber ebenfalls an vielen Stellen zentral für die Arbeit mit Gegenwartskultur. So kann Forschung, die prinzipiell auf verfügbaren Vorgängerarbeiten aufbaut und so fortschreitet, nicht effektiv und effizient funktionieren.

Dieses Recht wird nun durch die Europäische Richtlinie 2019/790 zum Urheberrecht im digitalen Binnenmarkt (DSM-RL) von 2019 nochmals geändert werden, die derzeit bis Juni 2021 in nationales Recht umgesetzt wird.[20] Die Umsetzung der DSM-RL wird das deutsche TDM-Recht in signifikanten Punkten modifizieren und dabei zumindest einige der vorgenannten Hürden und Unsicherheiten adressieren. So gilt insbesondere:

- Die Befristung des UrhWissG und damit des TDM auf lediglich 5 Jahre wird überflüssig
- Die Pflichten zur Vergütung und Quellenangabe entfallen
- Das TDM selbst an Einzelwerken wird zulässig
- Die Option zur Beschränkung der Nutzung online verfügbarer Inhalte wird aufgehoben
- Und insbesondere werden die Korpora nach Projektende für Anschlussforschung nachnutzbar

Gerade Letzteres ist ein großer, potentiell folgenreicher Schritt. Doch so begrüßenswert all dies aus Sicht der Forschung ist, die Korpora können auch nach Inkrafttreten des neuen Rechts voraussichtlich weiterhin nicht allgemein öffentlich zugänglich gemacht werden. Dies gilt selbst, wenn dieser Zugang auf die dann vom neuen Recht privilegierten Akteursgruppen (Forschungsorganisationen, Einrichtungen des Kulturerbes und Einzelforschende) und ausschließlich zum Zwecke nichtkommerzieller, wissenschaftlicher Forschung durch diese beschränkt würde. Sie sind stattdessen unter angemessenen Sicherheitsvorkehrungen nicht öffentlich aufzubewahren. Und es ist zwar wenig konturiert im anstehenden neuen Recht, aber die nun sogar ausdrücklich vorgesehene Voraussetzung des rechtmäßigen Zugangs zum Ursprungsmaterial lässt erwarten, dass die Korpora jedenfalls (jenseits gemeinsamer Projektarbeit mit den Zugangsberechtigten an den Korpora) nicht von Dritten ohne deren eigenen Zugang genutzt werden dürfen. Sonst könnte diese zentrale Bedingung im Interessenausgleich zwischen Rechteinhaber*innen und Forschenden leicht umgangen werden. Die Korpora werden demnach bedingt nachnutzbar, aber nicht publizierbar.

[17] Vgl. § 95b Abs. 1 S. 1 Nr. 11 UrhG.
[18] Vgl. § 95b Abs. 3 UrhG.
[19] Vgl. Art. 5, 6, 89 DSGVO.
[20] Vgl. zu DSM-RL und DH/TDM Dreier 2019, passim; Ducato / Strowel 2019, passim; Flechsig 2019, passim; Geiger et al. 2019, passim; Raue 2019, passim; Raue 2020, passim; Schaper / Verweyen 2019, passim; Spindler 2019, passim; Steinbrecher 2019, passim; Stieper 2019, passim.

3. Das Prinzip der abgeleiteten Textformate

Vor dem skizzierten rechtlichen Hintergrund ist die Grundidee der abgeleiteten Textformate im Kern folgende: Es wird von Beständen urheberrechtlich geschützter Volltexte ausgegangen (die Ausgangstexte, im Urheberrecht als ›Ursprungsmaterial‹ bezeichnet; gegebenenfalls auch bereits als Korpus vorliegend), zu denen eine Institution legalen Zugang hat. Diese Textbestände werden durch die Anwendung von Verarbeitungsroutinen, die im Wesentlichen sowohl eine gezielte Informationsanreicherung (beispielsweise durch linguistische Annotation) als auch eine Informationsreduktion (beispielsweise durch Löschung der Wortformen oder Aufhebung der Sequenzinformation) darstellen, in sogenannte abgeleitete Textformate verwandelt.[21] Diese Verarbeitungsroutinen können gegebenenfalls in Verbindung mit einem konkreten Forschungsvorhaben der eigenen Institution oder Dritter angewendet werden. Das einfachste Beispiel für ein solches abgeleitetes Textformat wäre eine Tabelle, die für einen Textbestand die Häufigkeiten jedes Wortes in jedem Text festhält.

Solche abgeleiteten Textformate können für einen unstrukturierten Gesamtbestand, einen größeren Teilbestand oder aber für einen gezielt zusammengestellten, für die Bearbeitung einer bestimmten Forschungsfrage geeigneten Teilbestand von Texten erstellt werden.[22] Die abgeleiteten Textformate sind dabei so gestaltet, dass die Texte in der dann vorliegenden Form einerseits nicht mehr in den Geltungsbereich des Urheberrechts fallen, andererseits dennoch die Anwendung möglichst vielfältiger quantitativer Analysen der Texte erlauben. Zwar enthalten die abgeleiteten Textformate abhängig von der jeweiligen Umsetzung verglichen mit den Ausgangstexten manche Informationen nicht mehr, sodass nicht alle relevanten Analyseverfahren uneingeschränkt umsetzbar sind (siehe hierzu ausführlich Abschnitt 5). Dafür können solche Datenbestände ohne Einschränkungen gespeichert, in der Forschung genutzt, veröffentlicht und von Dritten nachgenutzt werden. Über die Nachnutzung und Veröffentlichung hinaus ist – bei rechtmäßigem Zugang zum Ursprungsmaterial – zudem auch die Erstellung der abgeleiteten Textformate erlaubnisfrei möglich.

Die Idee der abgeleiteten Textformate ist nicht neu, vielmehr gibt es bereits mehrere Beispiele für die erfolgreiche Umsetzung dieses Prinzips. Zu den prominentesten Beispielen für den Einsatz abgeleiteter Textformate zählen das *Google Ngram Dataset* sowie das *HTRC Extracted Features Dataset* der Hathi Trust Digital Library. Das Google Ngram Dataset basiert auf einem Korpus von rund 8 Millionen Büchern in mehreren Sprachen, das auch eine starke diachrone Dimension aufweist. Der Datensatz enthält für jedes Wort (und für jedes N-Gramm der Länge 2–5) in jedem Jahr die Angabe darüber, wie häufig das Wort vorkommt und in wie vielen verschiedenen Dokumenten es vorkommt.[23] Seit der Version 2 von 2012 sind die Texte auch nach Wortarten annotiert. Das HTRC Extracted Features Dataset enthält für jede einzelne Seite im umfangreichen Dokumentenbestand der Hathi Trust Digital Library das Inventar an Tokens mit Angabe der Wortart und der Häufigkeit auf der Seite.[24] Derzeit enthält der Datensatz diese Informationen für rund 15 Millionen Dokumente oder fast 6 Milliarden Seiten. Knapp zwei Drittel des Datenbestands sind dabei urheberrechtlich geschützt, der Rest ist gemeinfrei. Der Datensatz kann ganz oder in Teilen heruntergeladen oder mit dem *HTRC Feature Reader* ausgelesen werden. Die Dokumente sind nach Seiten segmentiert und mit detaillierten bibliografischen Metadaten angereichert.

Weitere Beispiele für den Einsatz abgeleiteter Textformate sind der *CrossAsia N-Gram Service* der Staatsbibliothek zu Berlin, die Datensätze im Paket *stylo für R*, die *Leipzig Corpora Collection* und das *Open Super-large Crawled ALMAnaCH coRpus* (OSCAR). Im Fall des CrossAsia N-Gram Service sind die lizenzierten Volltextbestände registrierten Nutzer*innen vorbehalten, frei verfügbar sind aber die davon abgeleiteten Häufigkeitsinformationen, die für verschiedene N-Gramm-Größen angeboten werden. Für jedes Dokument und jede N-Gramm-Größe liegt eine separate Datei mit den Häufigkeitsinformationen vor sowie eine separate Metadatentabelle. Mit insgesamt rund 13.000 Dokumenten handelt es sich um einen substantiellen Datenbestand historischer Texte. Im Fall von stylo (einem Werkzeug für stilometrische Analysen in der Statistikumgebung R), werden mehrere Sammlungen urheberrechtlich geschützter Texte in Form einer einfachen *Term-Dokument-Matrix* bei der Installation mitgeliefert und stehen direkt für Beispielanalysen zur Verfügung. Es handelt sich um vergleichsweise kleine Datensätze: Der Galbraith-Datensatz enthält beispielsweise die Häufigkeiten der 3.000 häufigsten Wörter für 28 Romane, ohne weitere Segmentierung.[25] Die Leipzig Corpora Collection bietet für rund 250 Sprachen teils sehr umfangreiche Korpora an, die auf überwiegend urheberrechtlich geschützten Texten aus dem Internet basieren, von denen ein zufälliges Sample einzelner Sätze verfügbar gemacht wird.[26] OSCAR ist ein äußerst umfangreiches Korpus, in dem 166 Sprachen vertreten sind, wobei hier pro Sprache die Reihenfolge der Zeilen gegenüber den Ausgangsdokumenten randomisiert wurde.[27]

---

[21] Das Verfahren hat eine gewisse Beziehung zum Prinzip der *differential privacy* bei sozialwissenschaftlichen Datenerhebungen. Dort wird durch gezielte Randomisierung von Antworten die Privatsphäre der einzelnen Teilnehmenden geschützt, ohne dass die Verlässlichkeit von Schlussfolgerungen auf der Ebene aller Befragten eingeschränkt wird. Vgl. u. a. Dwork / Roth 2014, 5–27.
[22] Erst wenn ein Datenbestand für eine Analyse zusammengestellt und gegebenenfalls systematisiert und aufbereitet wurde, spricht das Urheberrecht von einem ›Korpus‹.
[23] Vgl. Lin et al. 2012, passim.
[24] Vgl. Bhattacharyya et al. 2015, passim; Jett et al. 2020, passim.
[25] Vgl. Eder et al. 2016, passim.
[26] Vgl. Goldhahn et al. 2012, passim.

Diese Beispiele zeigen, dass die Vorteile der Idee abgeleiteter Textformate durchaus bereits erkannt worden sind. Es wird aber auch deutlich, dass erstens die Umsetzung bisher nur wenig programmatisch erfolgt, denn es gibt kaum Forschungsliteratur, die sich spezifisch diesem Thema widmet, und dass es zweitens bisher kaum Bemühungen um eine Standardisierung von Formaten und Strategien über Einzelprojekte oder einzelne Institutionen hinweg gibt.

Wie kann man sich dem Konzept der abgeleiteten Textformate also grundsätzlicher nähern? Zunächst ist zu konstatieren, dass ein Text nur scheinbar aus einer schlichten Abfolge von Wortformen oder gar Zeichen besteht. Denn für die verstehende Lektüre eines Textes ist die Kenntnis nicht nur der Wortformen und ihrer genauen Reihenfolge notwendig, sondern auch die Kenntnis der Bedeutung und grammatikalischen Funktion der Wortformen im Satz sowie der semantischen und syntaktischen Beziehungen zwischen den Wortformen. Hinzu kommen noch Kontext und Pragmatik des Textes über die Satzgrenzen hinaus.

Um eine systematische Modellierung der abgeleiteten Textformate vorzunehmen, wird hier abstrahierend davon ausgegangen, dass ein Text lediglich in die folgenden Teile zu gliedern ist:

- Token (vereinfacht gesagt: ein einzelnes Wort)
- Satz
- Segment (Abschnitte fester, aber willkürlicher Länge)
- Gesamttext[28]

Ausgehend von dieser abstrakten Modellierung kann man damit zusammenfassend von den folgenden Teilinformationen ausgehen, die sich jeweils auf ein Token im Text beziehen:

- Die Information über die Wortform des Tokens, also die Abfolge der Zeichen; mit oder ohne Berücksichtigung der Groß-/Kleinschreibung
- Das Lemma, also die unflektierte Grundform, wie man sie als Wörterbucheintrag finden würde
- Die Wortart, also die grammatikalische Klasse (Substantiv, Verb, Adjektiv, Pronomen, etc.), gegebenenfalls und sofern relevant auch weitere morpho-syntaktische Informationen (Genus, Numerus, Casus)
- Die Bedeutung, also der semantische Gehalt des Wortes; repräsentiert beispielsweise über Zuordnung eines Wortvektors aus einem *Word Embedding Model* oder eines Synsets in WordNet
- Die Relationen, also insbesondere die syntaktische Rolle des Tokens und seine Beziehung zu anderen Tokens im Satz, wie etwa die Bestimmung als Prädikat oder die Auflösung der Referenz eines Pronomens
- Die Sequenzinformation, also die syntagmatische Position des Tokens relativ zu anderen Tokens im Satz; die Position des Satzes relativ zu anderen Sätzen im Segment; und die Position des Segments im Text
- Die Häufigkeit des Tokens, wobei die Häufigkeit im Satz, im jeweiligen Segment oder im Gesamttext gemeint sein kann; zudem kann sie als binäre, absolute oder relative Häufigkeit ausgedrückt werden

Die meisten der hier beschriebenen abgeleiteten Textformate beruhen grundsätzlich auf einem tokenisierten und linguistisch annotierten Text. Ein solches annotiertes Textformat erlaubt die grammatikalische Disambiguierung bestimmter Wörter, die Zusammenfassung mehrerer unterschiedlicher Wortformen auf der Ebene ihres gemeinsamen Lemmas oder ihrer gemeinsamen Wortart und die Suche und Filterung auf Wortart-Ebene. Weitere Annotationsebenen, wie bezüglich der syntaktischen Funktion im Satz oder als semantische und morpho-syntaktische Beschreibung von Wörtern durch Wort-Vektoren, sind ebenso denkbar, aber deutlich aufwändiger und daher weniger verbreitet.

Abgeleitete Textformate können auf der Grundlage des skizzierten Verständnisses von Text solchermaßen definiert werden, dass man jeweils beschreibt, welche der genannten, unterschiedlichen Teilinformationen durch Annotation expliziert werden, welche vereinfacht oder entfernt werden, und welche erhalten bleiben. Außerdem ergeben sich aus diesem Verständnis der Textformate eine Reihe von Parametern eines Textformats. Diese Parameter können sich insbesondere auf die folgenden Punkte beziehen:

- Welche Strategie der Tokenisierung wurde angewandt, wobei insbesondere der Umgang mit zusammengesetzten Begriffen geklärt sein muss. Besteht ›New York Times‹ aus einem, zwei oder drei Tokens?
- Welche Informationen werden für jedes Token bereitgestellt: die Wortform, das Lemma, die Wortart, die syntaktische Rolle, eine Repräsentation der Semantik, oder eine Kombination dieser Informationen?
- Wird der Text in einzelne Tokens gegliedert oder bilden Folgen mehrerer Tokens (sogenannte N-Gramme) die Darstellungseinheit?
- In welchem Umfang bleibt die Sequenzinformation erhalten?

[27] Vgl. Suárez et al. 2019, passim.
[28] Die Frage der texteigenen Untergliederung in Kapitel und Absätze (in Erzähltexten und Essays), Szenen, Akte und Reden (in Dramen) oder in Strophen (in der Lyrik) bleibt hier zunächst unberücksichtigt, wäre aber grundsätzlich in das Modell integrierbar.

- Gibt es eine Aufteilung jedes Gesamtdokumentes in Segmente und/oder Sätze, und wenn ja, welche Länge (in Worten oder Sätzen) haben die Segmente?
- Schließlich: Handelt es sich um ein Textformat, das grundsätzlich das einzelne Dokument innerhalb eines Korpus als Einheit beibehält, oder werden die Dokumentgrenzen aufgelöst?

Durch die Festlegung der jeweils explizierten, reduzierten oder entfernten und erhaltenen Informationen, spezifiziert über die jeweils gewählten Parameter, ergeben sich eine Vielzahl möglicher Transformationen der Ausgangstexte in verschiedene abgeleitete Textformate. Unterschiedliche Textformate eignen sich dabei je unterschiedlich gut für bestimmte Analyseverfahren. Der folgende Abschnitt nimmt dieses Verhältnis genauer in den Blick.

4. Verbreitete Analyseverfahren der Digital Humanities und ihr Informationsbedarf

Aus der Perspektive der Anwendung von TDM ist es wünschenswert, dass die abgeleiteten Formate einen möglichst geringen Informationsverlust gegenüber den Ausgangstexten erfahren. Dies soll es ermöglichen, dass auf Grundlage der abgeleiteten Formate möglichst weitreichende und vielfältige Analysemethoden eingesetzt werden können. Dabei ist je nach Analyseverfahren auch entscheidend, welche Arten von Informationen jeweils beibehalten oder reduziert wurden. Im vorliegenden Abschnitt werden daher mehrere in den DH übliche Verfahren in ihren Grundprinzipien knapp erläutert und speziell mit Blick auf ihre jeweiligen Anforderungen an den Informationsgehalt der abgeleiteten Textformate hin charakterisiert. Auf dieser Grundlage können dann verschiedene Textformate auf ihr Einsatzspektrum hin geprüft werden. Da unterschiedliche Analyseverfahren unterschiedliche Anteile der gesamten Textinformation nutzen, ist es wahrscheinlich, dass mehrere unterschiedliche Textformate notwendig sind, um eine breite Menge an Analyseverfahren zu unterstützen. Dies können jedoch auch nicht beliebig viele Formate sein, da sonst der Kuratierungsaufwand massiv ansteigt und zudem die Möglichkeit besteht, dass Ausgangstexte aus der Kombination der Formate rekonstruiert werden können.

Die hier diskutierten Analyseverfahren sind in den DH und insbesondere in den Computational Literary Studies (CLS) einschlägig, wobei die Darstellung keinesfalls Anspruch auf Vollständigkeit erhebt, sondern exemplarisch angelegt ist. Die folgenden Verfahren werden diskutiert:

- Die Klassifikation und Clustering von Texten u. a. für die *Autorschaftsattribution*
- Die Extraktion distinktiver Merkmale
- Die semantische Analyse mit *Topic Modeling*
- Die Analyse von Polarität mit *Sentimentanalyse*
- Der Blick auf Figurenbeziehungen mit der *Netzwerkanalyse*
- Die Analyse von Beziehungen zwischen Texten beispielsweise beim *Text Re-Use*
- Sowie allgemein der Einsatz von Sprachmodellen für verschiedenste Aufgaben

4.1 Klassifikation und Clustering von Texten inkl. Autorschaftsattribution

Überwachtes und nicht-überwachtes maschinelles Lernen (also Klassifikation und Clustering) werden im Bereich des Text Mining mit guten Resultaten durchgeführt, um Texte mit bestimmten Inhalten oder anderen semantischen Gemeinsamkeiten in größeren Textsammlungen zu finden. Dies erfolgt regelmäßig auf der Grundlage von Term-Dokument-Matrizen.[29] Eine besondere Anwendungsdomäne, bei der es um die Zuordnung von Texten zu ihren Autor*innen aufgrund von (lexikalischer, stilistischer etc.) Textähnlichkeit geht, ist die stilometrische Autorschaftsattribution.[30] Grundprinzip ist hier die Erfassung des lexikalischen, insbesondere stilistischen Grades der Ähnlichkeit zwischen mehreren Texten. Die Ähnlichkeit kann auf der Ebene ganzer Texte, Teilsegmente oder Sätze ermittelt werden sowie auf der Grundlage der Häufigkeit von Wortformen oder anderen Merkmalen erfolgen. Da diese Verfahren in den allermeisten Fällen ohnehin auf einer Term-Dokument-Matrize operieren und keine Sequenzinformation berücksichtigen, können sie gut auch mit einfachen abgeleiteten Formaten unterstützt werden.

[29] Vgl. Feldman / Sanger 2007, passim.
[30] Vgl. Stamatatos 2009, passim.

4.2 Extraktion distinktiver Merkmale

Ein weiteres Verfahren aus dem Methodeninventar der CLS ist die Extraktion distinktiver Merkmale. Hier geht es um die Identifikation von Wortformen oder anderen Merkmalen, die für einen Text oder eine Textgruppe im Vergleich mit einer anderen Textgruppe charakteristisch sind.[31] Dies erlaubt es beispielsweise, die stilistischen und inhaltlichen Eigenheiten einer Autorin oder eines Autors, einer bestimmten Textsorte oder einer Epoche zu ermitteln und für weitere Analysen zu nutzen. Literaturwissenschaftliche Anwendungsbeispiele gibt es u. a. zu Shakespeare, dem britischen Roman oder dem französischen Drama.[32]

Sehr einfache Varianten dieses Verfahrens erfordern keine Sequenzinformation, sondern beruhen lediglich auf einem Vergleich der relativen Häufigkeit der Merkmale in den beiden Textgruppen. Etwas avanciertere Verfahren vergleichen lediglich die Verteilung der Häufigkeiten in den Texten der beiden Textgruppen und sind daher ebenfalls nicht auf Sequenzinformation angewiesen. Präzisere Verfahren benötigen allerdings Informationen zur Verteilung der Merkmale auch innerhalb der Texte (die sogenannte Dispersion). Diese Verfahren sind auf Sequenzinformation angewiesen, wobei eine geringe Segmentgröße zwar grundsätzlich wünschenswert ist, die Verfahren aber von jeder Segmentierung der vollständigen Texte in kleinere Teile profitieren. Für manche Varianten des Verfahrens ist die Information über die Reihenfolge der Segmente im Text nützlich, für andere ist sie nicht ausschlaggebend.

4.3 Topic Modeling

Topic Modeling ist ein Verfahren aus dem Bereich des unüberwachten Machine Learning mit dem latente, im weitesten Sinne semantische Strukturen in größeren Textsammlungen entdeckt werden können. Dabei werden aufgrund des wiederholten, gemeinsamen Vorkommens von Wörtern beziehungsweise dem wiederholten Vorkommen von Wörtern in ähnlichen Kontexten Gruppen von Wörtern gebildet, zwischen denen eine wie auch immer geartete semantische Beziehung besteht, wobei dann die Verteilung dieser Wortgruppen in der Textsammlung ermittelt werden kann.[33] Die semantische Beziehung der Wörter kann sich bei fiktionalen Texten u. a. auf ein abstraktes Thema (Gerechtigkeit, Fortschritt), ein wiederkehrendes erzählerisches Motiv (Eisenbahnfahrt, Konzertbesuch) oder das Vokabular für die Beschreibung von Handlungsorten (Innenräume, Landschaften) beziehen. Es gibt in den CLS zahlreiche Anwendungsbeispiele für diese Methode, sei es zu Tagebüchern, Romanen, Dramen oder Lyrik.[34]

Topic Modeling profitiert auf jeden Fall von der Verfügbarkeit der Information zu Lemma und Wortart, weil mit dieser Information semantisch eng verwandte Wörter zusammengeführt und Funktionswörter sowie Namen besser herausgefiltert werden können, als wenn lediglich die Häufigkeitsinformation verfügbar ist. Zudem ist für Topic Modeling, zumindest bei umfangreicheren Texten wie Theaterstücken und insbesondere Romanen eine Segmentierung der Texte in kleinere Segmente notwendig, um möglichst präzise und semantisch kohärente Topics zu erhalten. Auch hier gilt allerdings, dass eine geringe Segmentgröße grundsätzlich wünschenswert ist, die Verfahren aber von jeder Segmentierung der vollständigen Texte in kleinere Teile profitieren. Ideal wäre für das Topic Modeling vermutlich eine Segmentbildung, bei der Sätze oder sogar Absätze nicht getrennt werden, aber auch eine Segmentierung in Textabschnitte willkürlicher Länge in Wörtern (beispielsweise Segmente einer Länge von 500 Wörtern) ist nützlich. Information zur genauen Position eines Wortes im Text ist nicht notwendig, da das Verfahren in der Regel ohnehin dem *Bag-of-Words*-Modell folgt. Allerdings ist die Information über die Position eines Segmentes im Gesamttext durchaus nützlich, weil dadurch Muster in der Topic-Prävalenz nach Position im Text (beispielsweise Textanfang vs. Textende) möglich werden.

4.4 Netzwerkanalyse

Die Netzwerkanalyse beruht in der Regel auf der Erkennung der Entitäten, die als Knoten des Netzwerks dienen sollen, etwa Figuren, sowie auf einem Kriterium, aus dem sich eine Relation zwischen den Entitäten ergibt, beispielsweise die Interaktion in einem Dialog oder die Erwähnung in einem bestimmten Segment innerhalb des Textes.[35] Die Durchführung von Netzwerkanalysen setzt dabei in der Regel ein hohes Maß an automatisierter oder händischer Vorverarbeitung eines Textes voraus. In Hinblick auf die Extraktion von Entitäten müssen z. B. Figuren mittels Verfahren der *Named Entity Recognition* identifiziert werden; ist eine höhere Präzision angestrebt, müssen zudem pronominale Bezüge mittels Verfahren der *Coreference Resolution* aufgelöst werden. Alternativ können hier – wie etwa bei dramatischen Texten in Gestalt der Sprecherangaben

[31] Vgl. Kilgarriff 2001, passim.
[32] Vgl. Craig / Kinney 2009, passim; Schöch 2018, passim; Hoover 2010, passim.
[33] Vgl. Blei 2011, passim.
[34] Vgl. Blevins 2010, passim; Jockers 2013, passim; Rhody 2012, passim; Schöch 2017b, passim.
[35] Vgl. Jannidis 2017, passim; Trilcke 2013, S. 223–226 sowie S. 236–246.

– vorgegebene Strukturinformationen aus den Texten übernommen oder aber eine händische Auszeichnung von Figuren vorgenommen werden. Die Extraktion von Relationen setzt bei avancierteren Verfahren die computerlinguistische Identifikation von Formen der Rede- und Gedankenwiedergabe, einschließlich Adressant- und Adressatenzuweisung voraus (etwa: Wer spricht mit wem? Wer denkt an wen?). Weniger avancierte Verfahren greifen auf vorgegebene Segmentierungen in den Texten zurück, etwa Szenenstrukturen in dramatischen Texten oder Kapitelstrukturen in narrativen Texten (welche Figuren werden in einer Szene/einem Kapitel genannt?). Alternativ werden stärker formale Segmente (etwa Absätze) oder willkürliche Segmentierungen (etwa eine bestimmte Anzahl von Wörtern) als Definitionsgrundlage für die Extraktion von Relationen verwendet.

Obwohl aktuelle Studien in der Netzwerkanalyse literarischer Texte vermehrt komplexe semantische und sprechaktpragmatische Informationen für die Extraktion und Spezifizierung von Relationen verwenden, basiert ein Großteil der bisherigen Forschung vor allem a) auf einer zuverlässigen Identifikation von Entitäten und b) auf einer basalen, recht großzügigen Segmentierung. Auf sehr großen Korpora lassen sich insofern bereits mittels sehr weniger Informationen abstrakte Strukturmodelle vergleichend untersuchen. Eine in diesem Sinne abstrakte Netzwerkanalyse erfordert lediglich die Information darüber, welche *Named Entities* in welchen großzügig gefassten Textsegmenten (von z. B. 500 Wörtern) vorkommen, sowie die Information über die Position der Segmente im Text, kann aber auf alle übrigen semantischen, syntaktischen oder positionalen Informationen verzichten. Als problematisch erweist sich jedoch das Verfahren der Named Entity Recognition, dessen Ergebnisse in der Regel insbesondere in Hinblick auf die Zuordnung unterschiedlicher Named Entities zu Figuren (z. B. Vater und Ehemann als eine Entität) keine befriedigenden Ergebnisse liefern und insofern regelmäßig die Überprüfung und Nachkorrektur anhand des Volltextes voraussetzen.

4.5 Sentimentanalyse

Ziel der Sentimentanalyse ist die Ermittlung der Polarität (positiv, neutral, negativ) eines Textabschnittes, beispielsweise eines Satzes oder Segmentes.[36] In Erweiterung des Paradigmas kann es auch um die Ermittlung der Prävalenz verschiedener Basisemotionen (Freude, Furcht, Trauer, Glück, etc.) in einem Textabschnitt gehen. Neben dem Fokus – Analyse von Sentiment oder von Emotionen – unterscheiden sich die verschiedenen Ansätze darin, ob für die Analyse Wörterbücher, manuelle Annotationen oder Verfahren des Machine Learning genutzt werden.[37]

Eine Herausforderung für die Sentimentanalyse sind Phänomene wie Verneinung, deren Skopus wörterbuchbasierte Ansätze ermitteln und berücksichtigen müssen, um die korrekte Polarität oder Basisemotion zu ermitteln. Diese Phänomene sind in der Regel nur mit einer syntaktischen Analyse, in jedem Fall aber mit einer gewissen lokalen Sequenzinformation, in den Griff zu bekommen. Gleichzeitig sind sie in literarischen Texten vergleichsweise häufig, da diese ein hohes Maß an so genanntem uneigentlichem Sprechen aufweisen, das u. a. Metaphern, Sarkasmus und Ironie umfasst. Im Falle der manuellen Annotationen müssen hingegen die entsprechend annotierten Textabschnitte – also der Text, aus dem Quellenmaterial zusammen mit einer ihm zugewiesenen Annotation eines Sentiments beziehungsweise einer Emotion – zugänglich sein. Die Machine-Learning-Verfahren werden schließlich üblicherweise mindestens auf vollständigen Sätzen, die z. T. ebenfalls annotiert werden, entwickelt und eingesetzt.[38] Einfache abgeleitete Formate ohne lokale Sequenzinformation sind daher für dieses Verfahren vermutlich nicht geeignet.

Für alle Verfahren gilt außerdem, dass für die Weiter- oder Neuentwicklung eines Verfahrens mehr Textinformationen nötig sind, da sowohl die Wörterbucherstellung als auch das maschinelle Lernen zumeist über Annotationen des Textmaterials in Bezug auf seine Sentimentwerte erfolgt. Außerdem benötigen Verfahren der Sentimentanalyse Sequenzinformationen (Position von Satz oder Segment im Gesamttext), sobald sie Analysen über den Textverlauf vornehmen, etwa die Sentiment-Ausprägung eines Themas oder einer Figur im Verlauf des gesamten Textes oder als Kontraste zwischen Anfang und Ende eines Textes.

4.6 Text Re-Use

Bei Verfahren, die unter Text Re-Use zusammengefasst werden, geht es um die Identifikation von identischen oder sehr ähnlichen Passagen in mehreren Texten oder in einer umfangreichen Textsammlung, wie z. B. den Nachweis von Shakespeare-Zitaten in anderen Texten.[39] Dabei kann der Text Re-Use auch in verschiedenen Sprachen und über verschiedene Medien hinweg analysiert werden.[40]

[36] Vgl. Liu 2012, insbesondere S. 5–101.
[37] Für eine Übersicht von Sentimentanalyse-Ansätzen in den CLS vgl. Kim / Klinger 2019, passim.
[38] Vgl. Pang / Lee 2008, passim.
[39] Vgl. Hohl-Trillini / Quassdorf 2010, passim.
[40] Vgl. Burghardt et al. 2019, passim.

Aufgrund des zum Teil sehr unterschiedlichen Begriffs von Text Re-Use und durch die unterschiedlichen Textsorten, die in das Untersuchungskorpus integriert werden, unterscheiden sich die Verfahren stark voneinander.[41] In den Grundzügen funktionieren die Verfahren aber so, dass sie nach Ähnlichkeiten in Texten suchen, indem sie diese paarweise vergleichen und Passagen ausfindig machen, die in beiden Texten identisch sind oder aber durch nur wenige Lösch- oder Ergänzungsoperationen ineinander überführt werden können. Das Konzept von Identität oder Ähnlichkeit kann dabei syntaktisch, lexikalisch und/oder semantisch bestimmt sein. Entsprechend reicht die Bandbreite der von Text Re-Use-Anwendungen aufgefundenen Passagen von wörtlichen Zitaten über die Nutzung bestimmter morphosyntaktischer oder syntaktischer Muster bis hin zu freien Paraphrasen oder gar der lexikalisch und syntaktisch kaum sichtbaren Übernahme von Gedankengängen.

Mit Blick auf geeignete abgeleitete Textformate ist allen Zugängen gemeinsam, dass die Sequenzinformation unabdingbar ist. Darüber hinaus basieren Text Re-Use-Zugänge je nach ihrer Modellierung auf erweiterten N-Gramm-Analysen, syntaktischen Analysen oder anderen hier beschriebene TDM-Verfahren wie Topic Modelling und Word Embeddings. Entsprechend sind weitere der oben aufgelisteten Parameter abgeleiteter Textformate einzeln oder kombiniert nötig.

4.7 Sprachmodelle / Word Embeddings

Zahlreiche der Verfahren, die in den Computational Literary Studies verwendet werden, entstammen der Computerlinguistik beziehungsweise dem *Natural Language Processing* (NLP). Seit 2017 hat sich in der Verarbeitung von natürlicher Sprache ein Ansatz als besonders erfolgreich erwiesen: die Erstellung großer Sprachmodelle mit tiefen neuronalen Netzen, die dann in einem letzten Schritt auf die spezifische Aufgabe abgestimmt werden.[42] Diese Ansätze sind so erfolgreich, dass sie alle anderen weitgehend verdrängt haben, da diese Modelle in sehr hohem Maße Informationen über semantische, syntaktische und andere Aspekte von Sprache enthalten. Allerdings braucht man für ihre Erzeugung die ganzen Texte beziehungsweise zumindest vollständige Sätze. Im Gegenzug ist die Information, aus welchem Einzeltext ein Satz nun gerade kommt (oder zumindest die Information, in welcher Reihenfolge die Sätze in einem Text vorkommen), im Grunde unerheblich. Wichtiger ist, dass die Zusammensetzung der gesamten Textsammlung in ihren wesentlichen Parametern (u. a. Anteile der enthaltenen Textsorten, Anteile verschiedener zeitlicher Abschnitte) bekannt ist. Selbst wenn man so vorgehen würde, dass die großen Sprachmodelle von den Gedächtnisinstitutionen trainiert werden (wofür auch spricht, dass dieser Schritt sehr zeit- und rechenintensiv ist), so würde der letzte Schritt, das Abstimmen auf das spezifische Problem beziehungsweise die spezifische Anwendungsdomäne, immer noch voraussetzen, dass man zumindest eine größere Menge ganzer Sätze für diesen Schritt zur Verfügung hat.

4.8 Zwischenfazit Analyseverfahren

In der Summe zeigt sich hier, dass man wohl zwei große Gruppen von Analyseverfahren unterscheiden kann, wenn man ihre Anforderungen an den Informationsgehalt der abgeleiteten Textformate zu Grunde legt: einerseits diejenigen Formate, die auf eine präzise, insbesondere auch lokale, Sequenzinformation der einzelnen Tokens im Textverlauf angewiesen sind (dazu gehören wohl die meisten Verfahren aus der Sentimentanalyse, einige Verfahren der Netzwerkanalyse, sicherlich fast alle Verfahren des Text Re-Use sowie das Erstellen von Sprachmodellen); und andererseits diejenigen Formate, die diese Art von Sequenzinformation nicht erfordern, aber von einer nicht zu großen Segmentlänge sowie der Information über die Reihenfolge der Segmente im Gesamttext profitieren (dazu gehören sicherlich die Autorschaftsattribution, die Extraktion distinktiver Merkmale und das Topic Modeling, gegebenenfalls auch einige Verfahren der Netzwerkanalyse).

Über die Nützlichkeit für bestimmte Analyseverfahren in den Digital Humanities hinaus müssen abgeleitete Textformate weitere Kriterien erfüllen, damit ihre Erstellung und Publikation in der Praxis tatsächlich umgesetzt werden kann. Diese kommen insbesondere aus den Rechtswissenschaften und betreffen die urheberrechtliche Unbedenklichkeit der abgeleiteten Formate. Wann die abgeleiteten Textformate so gestaltet sind, dass die verbliebenen Informationen klar nicht mehr urheberrechtlich relevant sind, hängt von mehreren Faktoren ab, die von Karina Grisse und Florian Jotzo genauer diskutiert werden,[43] zu denen aber sicherlich Folgendes gehört:

- Es handelt sich weder um Vervielfältigungen noch um Bearbeitungen der Primärtexte im Sinne des Urheberrechtsgesetzes
- Die Menge an zusammenhängendem Text liegt unter einer bestimmten Schwelle
- Der Werkgenuss (im Sinne der ›normalen‹ Lektüre durch einen Menschen) ist ausgeschlossen
- Der Grad der Wiedererkennbarkeit des Textes, insbesondere seiner individuellen ästhetischen Qualitäten, ist für Normalbürger*innen gering

[41] Für eine Übersicht, vgl. Büchler et al. 2014, passim.
[42] Beispielsweise BERT, vgl. Devlin et al. 2018, passim.
[43] Vgl. Grisse 2020, passim und Jotzo 2020, passim.

- Die Rekonstruktion des Textes (oder auch kleinerer, aber urheberrechtlich relevanter Passagen des Textes) ist nicht trivial und/oder mit Unsicherheiten behaftet

Aus der Perspektive der Gedächtnisinstitutionen (also Bibliotheken, Archive, Museen, die digitale, urheberrechtlich geschützte Textbestände vorhalten) und damit aus Anbietersicht, sind abgeleitete Formate zudem dann geeignet, wenn sie leicht zu erstellen, zu speichern und vorzuhalten sind. Dies kann u. a. bedeuten, dass sie für jedes Textdokument unabhängig von anderen Dokumenten erstellbar sein sollten, damit ein inkrementeller Bestandsaufbau abgeleiteter Formate möglich ist. Dies ist bei den token-basierten Formaten der Fall, nicht aber insbesondere bei den auf *Wort-Embeddings* basierenden Formaten, die zur Erstellung umfangreiche Textbestände als Gesamtheit erfordern.[44]

Aus Anwendungsperspektive ist sicherlich unabhängig von der gewählten Analysemethode und dem spezifischen Format wünschenswert, dass abgeleitete Textformate in einem einfach zu verarbeitenden Datenformat vorliegen, was insbesondere bei einfach strukturierten und weit verbreiteten, standardisierten Formaten wie XML, JSON oder CSV der Fall sein dürfte. Außerdem sollten abgeleitete Textformate mit reichhaltigen, auch fachwissenschaftlich relevanten Metadaten, publiziert werden, sodass die Dokumentation der Provenienz der Texte und ihre fachwissenschaftliche Einordnung gewährleistet sind.

Ob diese Kriterien bei den hier vorgeschlagenen Textformaten jeweils gegeben sind, wird im folgenden Abschnitt kurz zusammengefasst dargestellt.

5. Vorschläge für abgeleitete Formate

In den folgenden Abschnitten wird eine Auswahl konkreter, abgeleiteter Textformate beschrieben und diskutiert. Diese Auswahl beruht auf einer vorgängigen Beurteilung einer größeren Anzahl von Formaten und beinhaltet nur solche Formate, die grundsätzlich aus den oben genannten Perspektiven zumindest vielversprechend erscheinen.

Zur Veranschaulichung der Textformate werden im Text (sofern möglich beziehungsweise sinnvoll) jeweils Beispiele oder Ausschnitte der entstehenden Dateiformate abgebildet.[45] Die verwendeten Texte sind gemeinfrei, sodass die Ausgangstexte im Sinne der Transparenz der Transformationsverfahren mit publiziert werden können. Zur Illustration der Formate soll das erste Kapitel aus dem Roman Effi Briest (1894–95) von Theodor Fontane dienen; um einen direkten Vergleich mit dem Ausgangstext und eine Einschätzung bezüglich Werkgenuss und Wiedererkennbarkeit zu ermöglichen, sei der Anfang des Kapitels hier zitiert:

> »In Front des schon seit Kurfürst Georg Wilhelm von der Familie von Briest bewohnten Herrenhauses zu Hohen-Cremmen fiel heller Sonnenschein auf die mittagsstille Dorfstraße, während nach der Park- und Gartenseite hin ein rechtwinklig angebauter Seitenflügel einen breiten Schatten erst auf einen weiß und grün quadrierten Fliesengang und dann über diesen hinaus auf ein großes, in seiner Mitte mit einer Sonnenuhr und an seinem Rande mit Canna indica und Rhabarberstauden besetztes Rondell warf. Einige zwanzig Schritte weiter, in Richtung und Lage genau dem Seitenflügel entsprechend, lief eine ganz in kleinblättrigem Efeu stehende, nur an einer Stelle von einer kleinen weißgestrichenen Eisentür unterbrochene Kirchhofsmauer, hinter der der Hohen-Cremmener Schindelturm mit seinem blitzenden, weil neuerdings erst wieder vergoldeten Wetterhahn aufragte.«[46]

5.1 Tokenbasierte Textformate

Zunächst gehen wir auf tokenbasierte Textformate ein, die zugleich solche Formate sind, bei denen die Grundeinheit der Erstellung und Publikation in der Regel einzelne, vollständige Texte sind. Dies gilt für die anschließend vorgestellten Textformate, die auf N-Grammen oder Vektoren beruhen, nicht in gleicher Weise.

[44] Karina Grisse (Grisse 2020) geht in ihrem Beitrag ausführlicher auf die Frage der rechtlichen Einschätzung verschiedener Textformate ein.
[45] Weitere Beispiele für die hier beschriebenen abgeleiteten Textformate sowie der sie erzeugende Programmcode liegen in einem Github-Repository vor.
[46] Fontane 2012, S. 7 vor.

5.1.1 Einfache Term-Dokument-Matrix

Das erste, sehr einfache abgeleitete Textformat ist die einfache Term-Dokument-Matrix. Sie besteht für jeden Einzeltext aus einer Liste der vorkommenden Tokens und ihrer absoluten Häufigkeit im Ausgangstext. Dabei kann zunächst für jeden Ausgangstext eine Datei erhalten bleiben (Tabelle 1). In der Praxis kann durch Zusammenführen mehrerer solcher Häufigkeitslisten eine ganze Textsammlung in Form einer Term-Dokument-Matrix repräsentiert werden, deren Größe von der Anzahl der enthaltenen Texte und der Anzahl der Types (d. h. der unterschiedlichen Wörter beziehungsweise des Gesamtvokabulars) bestimmt wird.

Aufgrund der prinzipiellen Einfachheit des Formats enthält es nur eine kleine Anzahl von Parametern, die es genauer spezifizieren:

- Welche Tokenisierung angesetzt wird, d. h. welche Definition von Token für die Segmentierung des Textes in einzelne Tokens, beispielsweise Wörter, verwendet wird
- Welche Informationen auf Token-Ebene jeweils mitgeführt werden (und dafür auch erhoben werden müssen), d. h. ob lediglich die Wortform des Tokens, oder aber weitere Informationen über das Token – wie beispielsweise das Lemma, die Wortart, morphologische Information, die syntaktische Rolle im Satz oder eine Repräsentation der Wortbedeutung, beispielsweise als Wortvektor (siehe hierzu auch Abschnitt 5.2.2) –, angeboten werden

Rang	Token (Wortform_Wortart_Lemma)
1	,_PUN_,
2	._PUN_.
3	und_KON_und
4	«_PUN_«
5	»_PUN_»
6	die_ART_die
7	ich_PPER_ich
8	sie_PPER_sie
9	das_ART_der/die/das
10	der_ART_der/die/das
11	es_PPER_es
12	nicht_PTKNEG_nicht
13	in_PRP_in
14	so_ADV_so
15	ist_VAFIN_sein
16	zu_KONJ_zu
…	…

Tab. 1: Ausschnitt aus der Term-Dokument-Matrix für Fontanes Effi Briest. Hier mit Wortform, Lemma und Wortart-Information, absteigend sortiert nach absoluter Häufigkeit. [Schöch et al. 2020]

Dieses abgeleitete Textformat kann folgendermaßen eingeschätzt werden:

- Für einige Analyseverfahren, insbesondere für einfache Varianten der Klassifikation und des Clustering beispielsweise für Fragen der Autorschaftsattribution, und für einfache Distinktivitätsmaße ist das Format geeignet. Für viele andere Verfahren, darunter für Topic Modeling, Sentimentanalyse, Netzwerkanalyse oder Text Re-Use ist dieses Textformat hingegen nicht ausreichend informationsreich: Insbesondere die vollständige Abwesenheit von Sequenzinformation auf allen Ebenen führt dazu, dass keine Verfahren eingesetzt werden können, die die (im Falle der Belletristik oft sehr umfangreichen) Texte nicht nur als Ganzes betrachten
- Aus technischer Sicht ist sicherlich ein Vorteil dieses Textformats, dass es mit relativ trivialen Mitteln erstellt werden kann. Wie einfach das ist, hängt allerdings insbesondere vom oben genannten, zweiten Parameter ab. Denn die dafür jeweils notwendige linguistische Annotation ist nicht in allen Fällen trivial und in so gut wie keinem Fall gibt es nur eine einzige,

standardisierte Vorgehensweise. Aus Anwendersicht ist zudem die einfache Nutzbarkeit eines solchen Formats ein Vorteil. Viele relevante Tools (u. a. Excel, Calc, R und Python) können eine solche Repräsentation in Form einer CSV-Datei direkt importieren und weiter verarbeiten
- Aus rechtlicher Sicht ist die einfache Term-Dokument-Matrix ein ganz klar unbedenkliches Format. Eine Rekonstruktion des Ausgangstextes ist ebenso klar ausgeschlossen wie der Werkgenuss durch die Leser*innen oder auch nur die intuitive Wiedererkennbarkeit des Ausgangstextes. Dass die stilometrische Autorschaftsattribution in der Lage ist, das individuelle stilistische Profil eines Autors aus einer solchen Matrix abzuleiten, bedeutet nicht, dass die individuellen Eigenschaften des Autors ohne technische Unterstützung erkennbar wären
- Aus Anbietersicht schließlich ist das Format ebenfalls vergleichsweise unproblematisch, da es einfach erstellt werden kann, nach und nach Texte transformiert werden können und keine besonders umfangreichen Datenbestände entstehen

In der Summe kann die Term-Dokument-Matrix demnach als rechtlich unbedenkliches, technisch eher unproblematisches, in der Anwendung aber eingeschränkt nützliches Format beschrieben werden. Es stellt damit in gewisser Weise die *Baseline* der abgeleiteten Formate dar.

5.1.2 Segmentweise Aufhebung der Sequenzinformation

Die Grundidee dieses abgeleiteten Formats ist es, die Reihenfolge der Wörter im Textverlauf durcheinanderzuwirbeln. Entscheidend ist hier allerdings, dass dies nicht für einen Einzeltext als Ganzes vorgenommen wird (dann wäre das Format bezüglich des Informationsgehalts mit der einfachen Term-Dokument-Matrix identisch), sondern jeweils nur innerhalb kleinerer Segmente, wobei die ursprüngliche Reihenfolge dieser Segmente im Text aber beibehalten wird (Auszug 1). Es erfolgt also eine selektive Reduktion der Sequenzinformation. Die wesentlichen Parameter dieses Textformats sind wie bei den meisten Textformaten die Tokenisierung und die über das Token verfügbare Information. Wesentlich sowohl aus Anwendungs- als auch aus rechtlicher Perspektive ist hier allerdings der zusätzliche Parameter der Länge der Segmente in Tokens.

```
von_APPR_von Hohen-Cremmen_NN_Hohen-Cremmen Georg_NE_Georg zu_APPR_zu heller_ADJA_hell
des_ART_die fiel_VVFIN_fallen schon_ADV_schon bewohnten_ADJA_bewohnt In_APPR_in der_ART_die
<SEG> Mittagsstille_ADJA_Mittagsstille Gartenseite_NN_Gartenseite und_KON_und erst_ADV_erst
Park-_TRUNC_Park- Dorfstraße_NN_Dorfstraße ,_PUN_, Seitenflügel_NN_Seitenflügel
breiten_ADJA_breit die_ART_die hin_ADV_hin während_KOUS_während angebauter_ADJA_angebaut
der_ART_die nach_APPR_nach ein_ART_eine Schatten_NN_Schatten auf_APPR_auf einen_ART_eine
rechtwinklig_ADJD_rechtwinklig <SEG> großes_ADJA_groß ,_PUN_, auf_APPR_auf mit_APPR_mit
in_APPR_in ein_ART_eine weiß_ADJD_weiß und_KON_und über_APPR_über quadrierten_ADJA_quadrierten
und_KON_und diesen_PDAT_dies auf_APPR_auf Mitte_NN_Mitte seiner_PPOSAT_sein dann_ADV_dann
Fliesengang_NN_Fliesengang hinaus_ADV_hinaus einen_ART_eine grün_ADJD_grün <SEG>
```

Ausz. 1: Ausschnitt aus der Liste der Tokens mit Annotation bei segmentweiser Aufhebung der Sequenzinformation für den Beginn von Fontanes Effi Briest. Hier auf Unigramm-Basis und mit Wortform, Lemma und Wortart-Information sowie einer Segmentlänge von 20 Tokens. Man beachte die Markierung der Segmentgrenzen mit <SEG> nach jeweils 20 Tokens. [Schöch et al. 2020]

Bei diesem abgeleiteten Textformat gibt es keine Abhängigkeit zwischen den Texten in einer Textsammlung, sodass die Texte frei rekombiniert werden können. Die Größe der Segmente hat keine signifikante Auswirkung auf die Größe der resultierenden Dateien, weil nur die Reihenfolge der Merkmale verändert wird. Der Eingriff in die segmentübergreifende Textstruktur ist minimal, das lesende Verständnis des Textes erscheint aber schon bei sehr kleinen Segmentgrößen so gut wie ausgeschlossen.

Dieses abgeleitete Textformat kann folgendermaßen eingeschätzt werden:

- Aus der Anwendungsperspektive erscheint dieses Textformat für eine vergleichsweise große Anzahl von Analysemethoden nützlich, vorausgesetzt, die Segmentlänge wird nicht zu groß angesetzt (<50 Tokens wären sicherlich in einigen Szenarien ausreichend klein): für einfache stilometrische Verfahren auf jeden Fall, zudem auch für avanciertere Verfahren und das Ermitteln distinktiver Merkmale, wofür ein segmentierter Text erforderlich ist, um zu sampeln oder die Dispersion der Merkmale zu berücksichtigen.[47] Für Topic Modeling ist das Format ebenfalls gut geeignet. Nur bei einer sehr geringen Segmentlänge oder bei einer Segmentierung in Sätze erscheint eine einfache Sentimentanalyse denkbar. Verfahren der Netzwerkanalyse sind denkbar, wären aber auf eine vorgängige, hochwertige Named Entity Recognition und Coreference

[47] Nicht geeignet ist das Textformat für einen Spezialfall der stilometrischen Autorschaftsattribution, das sogenannte *rolling Delta*.

Resolution angewiesen. Mit diesem Format nicht durchführbar erscheinen avanciertere Verfahren des Text Re-Use, die stark auf einer feingranularen Sequenzinformation beruhen, oder nicht-triviale Verfahren aus dem Bereich der Sentimentanalyse
- Aus technisch-informatischer Perspektive ist dieses Format unproblematisch, weil es einfach zu erstellen ist und keine besonderen Anforderungen an Speicherkapazitäten oder Datenstruktur erfordert. Es kann eine Datei pro Gesamttext erstellt werden, wodurch ein progressiver Bestandsaufbau ermöglicht wird; zudem erlaubt dies die einfache, nachträgliche Kombination von Texten zu einem je nach Forschungsfrage zusammengestellten Korpus
- Aus rechtlicher Perspektive ist eine Rekonstruktion des Ausgangstextes mit einer so hohen Zuverlässigkeit, dass der Ausgangstext tatsächlich gelesen und verstanden werden könnte, schon bei einer Segmentlänge von >50 aufgrund der exponentiell steigenden Anzahl der möglichen Kombinationen kaum noch denkbar. Mit höherer Segmentlänge sinkt die Rekonstruierbarkeit weiter ab. Bei einer geringen Segmentlänge (beispielsweise <10 Tokens) oder bei einer satzweisen Segmentierung steigt sie hingegen; dann wäre eine Rekonstruierbarkeit des Ursprungstextes in einzelnen Fällen (also nicht für den Gesamttext, aber doch für mehrere längere Abschnitte des Textes) denkbar. Die genauen Verhältnisse wären allerdings erst empirisch nachzuweisen.

Damit handelt es sich hier um ein sehr empfehlenswertes Format, das bei entsprechend geeigneter Wahl des Parameters Segmentlänge (im Bereich von um die 50 Tokens) sowohl aus der Anwendungsperspektive für eine Reihe von Verfahren nützlich ist als auch aus rechtlicher Perspektive als unbedenklich eingeschätzt werden kann. Zu beachten ist zudem, dass das erste abgeleitete Textformat, die einfache Term-Dokument-Matrix, aus diesem Format ebenfalls generiert werden kann (nicht aber umgekehrt).

5.1.3 Selektiv reduzierte Information über einzelne Tokens

Die Grundidee dieses Formats ist es, die vollständige Sequenzinformation im Text beizubehalten, um bestimmte Verfahren zu ermöglichen, die auf diese Information angewiesen sind, dabei aber so viel Information über die einzelnen Tokens zu entfernen, dass dennoch von einer urheberrechtlichen Unbedenklichkeit ausgegangen werden kann. Zahlreiche Varianten sind denkbar, aber eine aus Anwendungssicht nützliche Implementierung dieses Textformats könnte folgendermaßen gestaltet sein: Ausgangspunkt wäre erneut ein tokenisierter und annotierter Text, sodass für jedes Token mindestens Wortform, Lemma und Wortart verfügbar sind. Dann wird beim Erstellen des Textformats aber beispielsweise für alle Funktionswörter (also u. a. Präpositionen, Pronomina und Artikel) die Information über die Wortform und das Lemma entfernt und lediglich die Information über die Wortart beibehalten (Auszug 2). Dadurch bleibt die Sequenzinformation vollständig erhalten, nicht nur in Bezug auf die Abfolge der Inhaltswörter, sondern auch in Bezug auf den exakten Abstand der Wörter zueinander im Ausgangstext. Wichtigster Parameter dieses Textformats ist sicherlich, für welche Wortarten die Information über Wortform und Lemma entfernt wird und für welche nicht.

```
APPR Front_NN_Front ART schon_ADV_schon APPR Kurfürst_NN_Kurfürst Georg_NE_Georg
Wilhelm_NE_Wilhelm APPR ART Familie_NN_Familie APPR Briest_NN_Briest bewohnten_ADJA_bewohnt
Herrenhauses_NN_Herrenhaus APPR Hohen-Cremmen_NN_Hohen-Cremmen fiel_VVFIN_fallen
heller_ADJA_hell Sonnenschein_NN_Sonnenschein APPR ART Mittagsstille_NN_Mittagsstille
Dorfstraße_NN_Dorfstraße PUN KOUS APPR ART TRUNC KON Gartenseite_NN_Gartenseite hin_ADV_hin
ART rechtwinklig_ADJD_rechtwinklig angebauter_ADJA_angebaut Seitenflügel_NN_Seitenflügel
ART breiten_ADJA_breit Schatten_NN_Schatten erst_ADV_erst APPR ART weiß_ADJD_weiß KON
grün_ADJD_grün quadrierten_ADJA_quadrierten Fliesengang_NN_Fliesengang KON dann_ADV_dann
APPR PDAT hinaus_ADV_hinaus APPR ART großes_ADJA_groß PUN APPR PPOSAT Mitte_NN_Mitte APPR ART
Sonnenuhr_NN_Sonnenuhr KON APPR PPOSAT Rande_NN_Rand APPR Canna_NN_Canna indica_NE_indica KON
Rhabarberstauden_NN_Rhabarberstaude besetztes_ADJA_besetzt Rondell_NN_Rondell warf_VVFIN_werfen
PUN
```

Ausz. 2: Abfolge der Tokens mit Annotation bei selektiver Entfernung der Wortform- und Lemma-Information für den Beginn von Fontanes Effi Briest. [Schöch et al. 2020]

Dieses abgeleitete Textformat kann folgendermaßen eingeschätzt werden:

- Dieses Textformat ist (in der beschriebenen Form) für die stilometrische Autorschaftsattribution kaum geeignet, weil für die Stilometrie gerade die feinen Unterschiede in den Häufigkeiten der einzelnen Funktionswörter entscheidend sind. Topic Modeling würde durch ein solches Format aber gut unterstützt, da hier meist ohnehin die Funktionswörter entfernt werden. Für die Ermittlung distinktiver Merkmale wäre das Verfahren nur geeignet, wenn es um die Ermittlung distinktiver Inhaltswörter oder distinktiver Wortarten geht. Für die Netzwerkanalyse ist auch dieses Format nur eingeschränkt nützlich,

da zwar die Eigennamen von Personen und ihr Abstand im Text ersichtlich bleiben könnten, Informationen wie Koreferenz jedoch nicht rekonstruierbar sind. Avanciertere Verfahren der Netzwerkanalyse, die etwa die Zuordnung von Rede- oder Gedankenwiedergabe für die Extraktion und Spezifizierung von Relationen verwenden, sind nicht möglich
- Für Verfahren wie den Text Re-Use hat das Format großes Anwendungspotential, denn Text Re-Use operiert ohnehin häufig mit N-Grammen, die auf den Lemmata der Inhaltswörter reduziert sind, um den *noise* zu reduzieren, der von kleineren stilistischen Varianzen produziert wird, und/oder auf die Inhaltswörter fokussiert ist. Einzig für die Sentimentanalyse wird auch dieses Format wenig gewinnbringend sein, weil vermutlich zu wenig syntaktische Information für die Berücksichtigung von Verneinungen u. ä. erhalten bleibt. Dies wäre allerdings je nach Parameter des Formats auch empirisch zu prüfen
- Aus urheberrechtlicher Sicht erscheint hier problematisch, dass die Wiedererkennbarkeit des Textes aufgrund der Substantive und Eigennamen, die in der ursprünglichen Reihenfolge erhalten bleiben, vergleichsweise hoch ist, auch wenn von einem Werkgenuss wohl nicht die Rede sein kann. Dieser Effekt könnte durch das zusätzliche Entfernen der Eigennamen deutlich reduziert werden. Die Rekonstruierbarkeit erscheint für den korrekten Gesamttext kaum möglich, für kleinere Werkteile aber eventuell denkbar

5.2 Textformate auf Korpus- oder Subkorpusebene

Die im vorigen Abschnitt verhandelten Textformate zeichnen sich alle dadurch aus, dass sie für jeden Einzeltext für sich genommen generiert werden können. Dies ist bei den folgenden Formaten anders, die den Einzeltext überschreiten können (bei den N-Grammen) beziehungsweise grundsätzlich unter Rückgriff auf ein umfangreicheres Korpus ermittelt werden (Wort-Embeddings).

5.2.1 N-Gramme

N-Gramme sind Sequenzen von mehreren aufeinander folgenden Tokens, ohne dass diese einer lexikalischen Einheit oder einer *Multi-Word Expression* entsprechen müssen. Im einfachsten Falle werden bei einem abgeleiteten Textformat, das auf N-Grammen beruht, die Häufigkeiten der in einem Text enthaltenen N-Gramme erhoben, ähnlich wie bei der einfachen Term-Dokument-Matrix (Abschnitt 5.1.1). Weil sie lokale Sequenzinformation beinhalten, sind N-Gramme als Hinweise auf Phänomene wie Kollokationen, Phraseme und andere lexikalisch-stilistische Muster für viele Analyseverfahren relevant. Aus diesem Grund wäre dieses Format besser als die bisher vorgestellten Formate für Text Re-Use geeignet.

Solange die Einheit des jeweiligen Einzeltextes nicht aufgelöst wird, dürfte allerdings aus urheberrechtlicher Perspektive selbst eine einfache Aufstellung der Häufigkeiten von N-Grammen der Größe 2–5 problematisch sein, weil durch die schindelartige Überlagerung mehrerer N-Gramme längere Textsequenzen rekonstruiert werden könnten. Dies gilt selbst dann als problematisch, wenn nicht der vollständige Text rekonstruiert werden kann, sondern nur eine größere Menge von Fragmenten. Wenn die N-Gramm-Häufigkeiten sich auf kleinere Segmente innerhalb eines Textes beziehen, potenziert sich das Problem noch, weil die Rekonstruierbarkeit erleichtert wird. Im Falle des Formats, das auf der selektiv reduzierten Information über einzelne Tokens beruht (Abschnitt 5.1.3), sind allerdings verschiedenste N-Gramme indirekt enthalten, denn aus der ja vollständig vorhandenen, wenn auch nur lückenhaft mit Wortformen versehenen Tokensequenz lassen sich beliebig lange (allerdings wiederum nur teilweise mit Wortformen versehene) N-Gramme bilden.

Es ist allerdings auch möglich, sich vom Einzeltext als Bezugsgröße zu lösen und die N-Gramm-Häufigkeiten über mehrere beziehungsweise sehr viele Einzeltexte hinweg zu berechnen. Die Parameter eines solchen Formats sind (neben der Tokenisierung und Annotation) die N-Gramm-Länge und die Bezugsgröße für die N-Gramm-Häufigkeiten, beispielsweise jeweils alle Texte einer Textsorte und/oder eines Jahres (Tabelle 2).

Rang	N-Gramm
1	gott sei dank
2	ja gnädigste frau
3	auch heute wieder
4	doch auch wieder
5	ist doch auch
6	ist immer so
7	gnädigste frau ist

8	war so war
9	nein gnädigste frau
10	wird ja wohl
11	ist doch recht
12	doch immer noch
...	...

Tab. 2: Häufigkeiten von 3-Grammen über mehrere Texte hinweg, bei einer Mindesthäufigkeit von 5. Beispieldaten auf der Grundlage von fünf Erzähltexten von Theodor Fontane. [Schöch et al. 2020]

Ein solches Format kann folgendermaßen eingeschätzt werden:

- Aus Anwendungsperspektive ist das Einsatzspektrum eines solchen Textformats sicherlich geringer als bei den einzeltextbasierten N-Grammen. Immerhin sind solche Formate aber für bestimmte Fragestellungen und Anwendungen, die sich nicht auf den Einzeltext beziehen, immer noch informativ genug, da N-Gramme auch Informationen zum Sprachgebrauch enthalten. Solche Korpora wären bereits nützlich, um Einsichten in die sprachlichen Regeln bestimmter Felder zu gewinnen, z. B. welche Worte mit einer gewissen Wahrscheinlichkeit auf andere Worte folgen. Sie würden aber auch die Entwicklung und Verbesserung ganz praktischer Anwendungen, z. B. die Verbesserung von themenspezifischer Spracherkennung, unterstützen können. Sind die zugrundeliegenden Teilkorpora ausreichend spezifisch, ist auch die Extraktion distinktiver N-Gramme im Vergleich mehrerer Teilkorpora möglich
- Die Rekonstruierbarkeit dürfte im Gegenzug deutlich eingeschränkt sein. Wenn nun Bibliotheken oder Archive sehr große Bestände etwa als 5-Gramme anbieten und dabei (wie Google) die N-Gramme des Korpus zählen, die in allen Büchern eines Jahres vorkommen, ist es sehr viel schwieriger, wenn nicht unmöglich, einen bestimmten Text oder auch nur längere Passagen beliebiger Texte aus den N-Grammen wiederherzustellen. Dies gilt insbesondere dann, wenn man dem Modell Googles auch in dem Punkt folgt, dass alle N-Gramme, die im Gesamtkorpus eine bestimmte Mindesthäufigkeit nicht haben, auch nicht im Format enthalten sind
- Eine Herausforderung aus Anbietersicht stellt hierbei die Frage dar, welche Aggregation von Einzeltexten innerhalb eines Gesamtkorpus (also z. B. alle digitalen Texte einer Bibliothek) für die Forschung relevant sind und an welchem Punkt eine rechtlich relevante Grenze überschritten wird: Neben der chronologischen Ordnung (jeweils die N-Gramm-Häufigkeiten aller Texte aus einem Jahr), die für Begriffs- und Ideengeschichte, aber auch Sprachgeschichte und andere historische Interessen brauchbar ist, könnte man sich auch andere Aggregationen vorstellen, die eher an Themen beziehungsweise Sachgruppen oder Textsorten orientiert sind (z. B. alle medizinischen oder auf die Wirtschaft bezogenen Texte). Es stellt sich dabei die Frage, wie klein die Gruppe sein kann und ob man sich eine Metrik vorstellen kann, die es einer Bibliothek leicht macht, zu entscheiden, ab welchem Punkt der Herstellbarkeit von längeren N-Gramm-Ketten die Bibliothek davon Abstand nehmen sollte. Löscht man die N-Gramme, die seltener vorkommen als ein bestimmter Schwellenwert besagt, dann kann man nicht alle, aber immer noch manche längeren N-Gramm-Ketten zusammensetzen, nämlich gerade da, wo häufig verwendete sprachliche Muster genutzt werden; eine entsprechende Metrik müsste also probabilistisch vorgehen

5.2.2 Wort-Embeddings

Neben den tokenbasierten Textformaten und den N-Grammen spielen auch vektorbasierte Formate eine zunehmend wichtige Rolle. Die technische Entwicklung im Bereich der computergestützten Verarbeitung natürlicher Sprache (Natural Language Processing) hat aufgrund von vektorbasierten Textformaten seit etwa 2013 enorme Fortschritte zu verzeichnen. Obwohl die Ziele in der NLP-Forschung – hier geht es primär um die Mensch-Maschine-Interaktion – von den zuvor genannten Analyseverfahren der DH teilweise abweichen, werden die entwickelten Verfahren später oft für die DH angepasst oder weiterentwickelt. Wie bei Topic Modeling und Sentimentanalyse ist davon auszugehen, dass viele der vektorbasierten NLP-Verfahren, die derzeit noch wenig in den DH Anwendung gefunden haben, in Zukunft auch dort vermehrt eine Rolle spielen werden.

Die Grundidee vektorbasierter Textformate ist, Sprache nicht als symbolisches Zeichensystem zu betrachten, sondern Wörter und größere Einheiten wie Sätze oder Dokumente in einem algebraischen Vektorraum abzubilden. Man erhält so eine Informationsanreicherung der Wörter über das reine Symbol hinaus, da auch semantische und syntaktische Informationen im zum Wort gehörenden Vektor repräsentiert werden. Allerdings haben vektorbasierte Textformate auch einen offensichtlichen Nachteil, der gerade in den DH entscheidend sein kann: Durch die Umwandlung von Text in Vektoren gehen explizite, für qualitative Analysen oft entscheidende, Informationen verloren. Das Potential dieser Methoden für die Analyse von Textbeständen sowie als urheberrechtlich unbedenkliches Textformat ist aber naheliegend und soll im Folgenden skizziert werden.

Die erste Generation der vektorbasierten Textformate basiert auf dem Zählen des Auftretens von Wörtern im direkten Umfeld eines Wortes. Ein Wort wird somit als die Häufigkeit der anderen Wörter im Korpus repräsentiert und hat damit Ähnlichkeit zu den schon erwähnten Term-Dokument-Matrizen. Mit dem *Word2vec-Verfahren* wurden ab 2013 die Wort-Embeddings populär, die mit Verfahren des maschinellen Lernens Parametervektoren aus großen Textsammlungen schätzen.[48] Jedes Token im Vokabular wird demnach als ein Vektor von reellen Zahlen (üblicherweise wenige Hunderte) dargestellt und nicht mehr als Vektor von natürlichen Zahlen mit der Länge der Anzahl der Types im Korpus (mehrere Tausend). Allerdings haben nun die Werte eines *Word Embedding Vectors* keine explizit interpretierbare Bedeutung mehr. Wo bei den zählbasierten Wortvektoren jeder Eintrag die Häufigkeit des Auftretens eines Wortes im Umfeld repräsentierte, enthält ein Wort-Embedding latente Informationen über die Wahrscheinlichkeit des Auftretens von Wörtern im Umfeld. Diese Art von Wort-Embeddings ist somit komplementär zu den bereits erwähnten abgeleiteten Textformaten zu verstehen. Jeder unterschiedlichen Wortform, alternativ auch jedem unterschiedlichen Lemma, im Korpus wird exakt ein eindeutiger Wortvektor zugeordnet. Dieser repräsentiert die distributionale Semantik dieses Wortes in Bezug auf das gesamte Korpus. Es besteht somit eine global eindeutige Beziehung zwischen Vektor und Token.

In diesem Zusammenhang ergeben sich eine Reihe von denkbaren Szenarien, je nachdem, welche Art von Informationen im Rahmen eines abgeleiteten Textformats angeboten werden.

- Erstens könnte man alle Wortformen in den Ausgangstexten durch ihre Vektoren ersetzen und auch sämtliche Sequenzinformation beibehalten, allerdings um den Preis, dass jegliche Interpretierbarkeit des Textes unmöglich wird. Da bei einem solchen Format dennoch jede Wortform durch einen eindeutigen Vektor repräsentiert ist, kann beispielsweise stilometrische Autorschaftsattribution damit weiterhin bewerkstelligt werden, mit der Einschränkung allerdings, dass die Merkmale nicht interpretierbar sind, weil die jeweils dazugehörige Wortform nicht vorliegt. Aus demselben Grunde wäre ein Verfahren wie Topic Modeling mit einem solchen Textformat zwar technisch möglich, aber wenig aufschlussreich. Urheberrechtlich dürfte das Format völlig unbedenklich sein, insbesondere wenn das Vokabular der so repräsentierten Texte nicht bekannt ist
- Zweitens könnte man das Word Embedding Model als solches publizieren, also die Gesamtheit des Vokabulars einer Textsammlung mit ihren jeweiligen Wortvektoren. Dies ist urheberrechtlich ebenfalls unproblematisch, weil es keinerlei Bezug zu bestimmten Einzeltexten gibt. Allerdings handelt es sich hier dann in erster Linie um eine Ressource zur syntaktisch-semantischen Annotation von Texten, die auf einen geeigneten Textbestand im Sinne einer weiteren Annotationsschicht neben Lemmata und Wortarten angewandt werden könnte. Erstellt man solche Modelle (ähnlich wie für die N-Gramme vorgeschlagen) für verschiedene Subkorpora, kann der Vergleich der Modelle Einblicke in die Sprachentwicklung oder in die konzeptuelle Struktur bestimmter Textsorten bieten[49]
- Schließlich könnte man die oben beschriebenen Textformate über die Annotation nach Lemma und Wortart hinaus mit einer solchen syntaktisch-semantischen Annotationsschicht ausstatten. Eine gewisse Passung zwischen Word Embedding Model und zu annotierenden Texten ist dafür allerdings Voraussetzung. Urheberrechtlich würde dies keinen entscheidenden Unterschied in der Beurteilung des jeweils in Frage stehenden tokenbasierten Textformats bedeuten; vorteilhaft wäre dies aber für verschiedenste Analyseverfahren, die so die Information über die semantischen und syntaktischen Ähnlichkeiten oder Unterschiede der Tokens nutzen könnten

5.2.3 Kontextualisierte Embeddings

Der nächste essentielle Schritt zur Verbesserung bestehender NLP-Verfahren wurde durch das Kontextualisieren von Wort-Embeddings erreicht. Dabei wird Satz für Satz und Wort für Wort erst eine Ersetzung durch Wort-Embeddings durchgeführt, die danach jeweils individuell transformiert werden in Abhängigkeit der Worte die davor und danach in dem konkreten Satz auftreten. Die ersten Verfahren, die erfolgreich dafür eingesetzt wurden, sind rekurrente neuronale Netze, konkret die *Long-Short-Term-Memories*, und seit 2017 transformerbasierte Modelle, speziell das BERT-Modell.[50] Das Grundprinzip besteht darin, statt eines global statischen Vektors pro Type im Korpus einen individuellen Vektor für jedes Token in jedem bestimmten Satzkontext zu generieren. Wäre zuvor in zwei unterschiedlichen Sätzen, die beide ein Wort gemeinsam haben, dieses Wort durch denselben Vektor repräsentiert worden, ist bei kontextualisierten Embeddings jeder Wortvektor unterschiedlich, weil die umgebenden Wörter im Satz unterschiedlich sind. Damit hat jedes im Korpus auftretende Token prinzipiell eine individuelle Vektorrepräsentation und der Rückschluss von Vektor auf Wort ist nicht mehr trivial möglich.

[48] Vgl. Mikolov et al. 2013, passim.
[49] Siehe z. B. Hamilton et al. 2016, passim.
[50] Siehe respektive Devlin et al. 2018, passim; Hochreiter / Schmidhuber 1997, passim und Vaswani et al. 2017, passim.

Die kontextualisierten Embeddings haben damit zwei entscheidende Vorteile gegenüber den bisher dargelegten Textformaten:

- Die Rekonstruktion des ursprünglichen Textes, in dem alle Tokens durch ein kontextualisiertes Embedding ersetzt wurden, ist vermutlich nicht möglich, wenn das Mapping nicht für jedes einzelne Token explizit mit vorliegt. Um eine belastbare Aussage hierzu zu treffen, muss noch theoretische und empirische Forschung betrieben werden. Es lässt sich allerdings vermuten, dass Sätze ab einer gewissen Wortlänge (ca. >3) nicht rekonstruierbar sind. Dies gilt, ohne die Ursprungstexte in irgendeiner Art und Weise zu vereinfachen, also mit vollständigem Erhalt der Wortreihenfolge und Interpunktion. Dies wiederum hat erhebliches Potential für die Forschung auf Satzebene, u. a. zur Satzähnlichkeit und damit auch für Text Re-Use
- Die in einem kontextualisierten Embedding beinhaltete syntaktische und semantische Information ist der anderer Repräsentationsformate deutlich überlegen. Die mit solchen Verfahren gewonnen Ergebnisse auf Analyse-Benchmarks wie Sentimentanalyse, semantische Textähnlichkeit oder Paraphrasierung erreichen weit bessere Ergebnisse als bisherige Verfahren; oft übertreffen diese sogar menschliche Fähigkeiten von Nicht-Expert*innen[51]

Vor diesem Hintergrund sind kontextualisierte Embeddings ein vielversprechender Kandidat für informationsreiche abgeleitete Textformate, die urheberrechtlich unbedenklich sind. Die Einschränkungen für die qualitative Forschung durch den Verzicht auf eine explizite Interpretierbarkeit der Worte bleiben aber auch hier bestehen. Damit befindet man sich mitten in der aktuellen Debatte über die Erklärbarkeit und Verlässlichkeit moderner Verfahren der künstlichen Intelligenz. Ebenso relevant wären kontextualisierte Embedding-Modelle beziehungsweise Transformer. Diese erlauben es, die kontextfreien und die kontextsensitiven Vektoren für Texte zu gewinnen, was zahlreiche Anwendungen in allen modernen digitalen textanalytischen Verfahren erlaubt.

5.3 Zwischenfazit zu den abgeleiteten Textformaten

Die Übersicht über einige denkbare abgeleitete Textformate zeigt, dass es durchaus mehrere vielversprechende Formate gibt, die in der Forschung nutzbringend eingesetzt werden können und die auch aus rechtlicher Sicht umsetzbar erscheinen. Eine kompakte Übersicht bietet Abbildung 1.[52]

abgeleitetes Textformat	Nützlichkeit für die Forschung							rechtliche Beurteilung		
	Stilometrie	Distinktivität	Topic Modeling	Sentiment Analyse	Netzwerkanalyse	Text Re-Use	Sprachmodelle	Schutz vor Wiedererkennbarkeit	Schutz vor Rekonstruierbarkeit	Unmöglichkeit des Werkgenusses
5.1.1. Einfache Term-Dokument-Matrix	+	+	o	–	–	–	–	+	+	+
5.1.2. Segmentweise aufgehobene Sequenzinf.	+	+	+	o	o	–	o	o	o	+
5.1.3. Selektiv reduzierte Information über Tokens	–	o	+	o	o	+	o	o	o	o
5.2.1. N-Gramme auf Teilkorpus-Ebene	–	o	–	–	–	–	+	o	o	+
5.2.2. einfache Wortembeddings	+	–	–	–	–	–	+	+	+	+
5.2.3. kontextualisierte Embeddings	o	o	o	o	o	o	+	+	+	+

Abb. 1: Übersicht über die abgeleiteten Textformate, ihre Nützlichkeit in der Forschung und ihre urheberrechtliche Einschätzung. [Hinzmann / Schöch 2020]

Eine linguistische Annotation zumindest mit der Information über das Lemma und die Wortart erscheint immer wünschenswert und aus rechtlicher Sicht unproblematisch. Die Nutzung eines geeigneten Word Embedding Models im Sinne einer semantisch-syntaktischen Annotation ist ebenso von Vorteil. Der Parameter der Segmentlänge wird voraussichtlich eine Abwägungsfrage bleiben: Aus der Perspektive der Analyseverfahren sind kleine Segmentlängen grundsätzlich besser (nicht zuletzt, weil eine Aggregation auf größere Segmente immer möglich, eine Aufteilung in kleinere Segmente hingegen im Nachhinein nicht möglich ist). Aus rechtlicher Sicht steigt aber in der Regel die Sicherheit, mit der ein Format urheberrechtlich irrelevant ist, mit größerer Segmentlänge an.

[51] Vgl. GLUE Benchmark und Wang et al. 2019, passim.

[52] Diese Übersicht kann lediglich der Orientierung dienen. Die exakte Einschätzung für jedes Format und jedes Verfahren bzw. für verschiedene urheberrechtliche Aspekte hängt auch von den jeweiligen Parametern des Formats ab und muss dem Text entnommen werden.

Nicht alle denkbaren abgeleiteten Textformate sind hier diskutiert worden. Einige sind rechtlich klar bedenklich, andere für die Forschung wenig nützlich. Urheberrechtlich problematisch sind insbesondere Formate, die längere Folgen von Worten beinhalten, also Formate, bei denen die lokale Sequenzinformation nur teilweise aufgehoben ist. Darunter fallen recht klar Formate, die aus vollständigen Sätzen in durcheinandergewirbelter Form bestehen oder aus einem Sample vollständiger Sätze. Beide Formate sind aus der Sicht der Analyseverfahren, die lokale Sequenzinformation erfordern, wie beispielsweise Sentimentanalyse oder Text Re-Use, besonders attraktiv, auch wenn es in den Digital Humanities häufig erforderlich erscheint, dass Analysen auf dem vollständigen relevanten Sprachmaterial beruhen; sie sind aber aus rechtlicher Hinsicht eher problematisch.

Nicht verschwiegen werden sollen einige inhärente Nachteile, die mit dem Modell der abgeleiteten Textformate verbunden sind und die ebenfalls in der Übersicht deutlich werden. Für diese sind sicherlich geeignete Minimierungsstrategien zu entwickeln. Dazu gehört erstens die nicht vollständige Nachvollziehbarkeit der Forschung, weil Analyseprozesse nicht vom Ursprungsmaterial aus nachvollzogen werden können, sondern nur vom verwendeten, abgeleiteten Textformat aus. Dies ist einer der Gründe, warum die Erstellung der abgeleiteten Textformate ein standardisierter und zertifizierter Prozess sein sollte, der die notwendige Vertrauenswürdigkeit der Formate garantiert. Ein weiterer Nachteil ist zweifellos, dass zwar einige, aber eben nicht alle relevanten Analyseverfahren auch mit einem abgeleiteten Textformat umgesetzt werden können. Schließlich ist in Rechnung zu stellen, dass die Erstellung von standardisierten, zertifizierten Beständen an Texten in abgeleiteten Formaten für die Anbietenden mit einem erheblichen Aufwand verbunden sind. Dennoch überwiegen aus unserer Sicht die Vorteile dieser Strategie gegenüber den (eingangs genannten) Alternativen beziehungsweise ist diese Strategie in jedem Fall eine wichtige, komplementäre Maßnahme neben den alternativen Ansätzen. Dies gilt nicht zuletzt auch, weil mit solchen Textbeständen besser als bisher demonstriert werden könnte, welches Potential in der Analyse urheberrechtlich geschützter Textbestände in den DH liegt. Im Kontext eines gesellschaftlichen und rechtlichen Interessenausgleichs zwischen Rechteinhabenden und Anwender*innen von TDM kann dies ein wichtiges Argument sein.

6. Fazit: Eine Forschungsagenda für abgeleitete Textformate

Abschließend soll hier ein Fazit insbesondere zu der Frage formuliert werden, welche Aufgaben und nächsten Schritte auf verschiedene relevante Akteure zukommen, wenn es gelingen soll, die abgeleiteten Textformate in der Praxis der DH zu verankern. Hier wird insbesondere auf die Rolle der Bibliotheken und Archive, der Informatik und der TDM-Anwender*innen in den DH eingegangen.

6.1 Bibliotheken und Archive

Bibliotheken und Archiven kommt eine zentrale Rolle bei der Etablierung abgeleiteter Textformate zu, weil sie rechtmäßigen Zugang zu umfangreichen urheberrechtlich geschützten Textbeständen haben. Schon mit dem UrhWissG wurden sie im UrhG als Institutionen benannt,[53] die urheberrechtlich geschützte Korpora, die von Dritten angefertigt wurden, aufbewahren und selektiv, nämlich zur Überprüfung wissenschaftlicher Qualität, verfügbar machen dürfen und sollen. Die Umsetzung des 2019 verabschiedeten Art. 3 DSM-RL bis Juni 2021 in nationales Recht[54] stärkt die Rolle dieser Einrichtungen im Kontext des TDM nochmals enorm und dies in mehrfacher Hinsicht: Bibliotheken und Archive werden nun ausdrücklich und autonom als Akteure des TDM adressiert und für wissenschaftliche Zwecke privilegiert; zugleich wird das Nachnutzungsverbot für die Korpora für Anschlussforschung aufgehoben werden.[55] Die Frage der wissenschaftlichen Anschlussforschung mittels abgeleiteter Formate tritt also absehbar als eine von dann zwei Säulen neben die Frage der wissenschaftlichen Anschlussforschung an Korpora.[56] Das lässt eine kohärente, in sich stimmige Entwicklungsstrategie der Bestände von der Warte des TDM aus in den Bereich des Möglichen rücken. Und beides ist verbunden mit einem nun expliziten Auftrag des Gesetzgebers an die Bibliotheken und Archive, sich jedenfalls insoweit mit dem Bereich TDM aktiv auseinanderzusetzen. Denn neben Forschungsorganisationen werden insofern nur Kulturerbeeinrichtungen privilegiert, anders als etwa andere Behörden, Bildungsträger, Journalisten oder Privatwirtschaft. Das ist eine Chance für Bibliotheken und Archive, aber zugleich auch eine Verantwortung.

[53] Vgl. §§ 60d Abs. 3 S. 2 iVm 60e, 60f UrhG.
[54] Nach derzeitigem Stand voraussichtlich in §§ 44b, 60d UrhG-E (Entwurfsfassung).
[55] Vgl. Döhl 2020a, passim.
[56] Vgl. Döhl 2020b, passim.

Es bleibt in diesem Kontext im anstehenden neuen Recht bei der nachvollziehbaren, derzeit impliziten, dann aber auch ausdrücklich im Gesetz stehenden Einschränkung, dass die Nutzung der Korpora einen rechtmäßigen Zugang voraussetzt und in bestimmten Fällen sogar nur in den Räumen der privilegierten Einrichtungen gewährt werden darf. Bibliotheken und Archive werden künftig im Kontext von TDM eine aktivere Rolle in Forschungsprozessen einnehmen, wodurch sich die Funktionen der Einrichtungen selbst langfristig verändern werden.

Bibliotheken und Archive sehen es stets und unverändert als ihre genuine Aufgabe an, die Bestände in einer jeweils zeitgemäßen Form zur Nutzung zur Verfügung zu stellen. Die Bibliotheken und Archive können demnach durch die neuen rechtlichen Rahmenbedingungen und sobald geeignete Spezifikationen für grundsätzlich nützliche abgeleitete Textformate vorliegen, in die Rolle der Anbieter von entsprechenden Datensätzen für die Forschung treten. Abgeleitete Formate haben das Potential, sich hier zu einem wichtigen Arbeits- und Angebotsbereich im Kerngeschäft von Bibliotheken und Archiven zu entwickeln.

Abgeleitete Textformate fördern die Perspektive auf eine globale, virtuelle Sammlung für die Forschung, die als Linked Open Data ausgeformt werden kann. Auch daraus ergeben sich weitreichende Konsequenzen für die Rolle, die Funktionen und das institutionelle Selbstverständnis dieser Einrichtungen.[57] Neben der Erstellung und Bereitstellung der abgeleiteten Textformate selbst wird es auch notwendig sein, für die Forschenden in den DH relevante Metadaten über die Texte bereitzustellen oder Wege zu entwickeln, wie diese erhoben werden können. Denn für die Forschung sind oft andere, forschungsnähere oder fachnähere Metadaten notwendig als diejenigen, die üblicherweise aus bibliothekarischer Perspektive erhoben würden.

Eine besondere Herausforderung bedeutet die Dokumentation einzelner Analysen unter wissenschaftlichen Anforderungen. In einem Bestand abgeleiteter Daten, der sich durch diverse technische und fachliche Einflüsse in einem fortwährenden Optimierungsprozess[58] befindet, muss eine besondere Herausforderung auf die Referenzierbarkeit einzelner Datensätze über persistente Identifikatoren sowie deren Echtheitsnachweis und Integrität zukommen. Grundsätzlich bestehen hohe Anforderungen an die Dokumentation, Transparenz und Verlässlichkeit des Prozesses der Erstellung von abgeleiteten Textformaten. Denn die Forschenden, die ein abgeleitetes Textformat für ihre Analysen einsetzen möchten, müssen sich darauf verlassen können, dass die abgeleiteten Textformate wirklich exakt im jeweils dokumentierten Verhältnis zu den Ausgangstexten stehen. Eine direkte Einsicht durch die Nutzenden ist bei den urheberrechtlich geschützten Ausgangstexten im Regelfall ja gerade nicht möglich.

Daraus ergibt sich die Notwendigkeit, abgeleitete Textformate durch detaillierte Spezifikationen zu beschreiben, deren Einhaltung entsprechend geprüft werden kann. Dies schließt die jeweils gewählten Konzepte, Verfahren und Werkzeuge für Tokenisierung, Lemmatisierung, Wortartauszeichnung (einschließlich des jeweils verwendeten Sprachmodells und des Tagsets) mit ein, die allerdings sprachspezifisch beziehungsweise auch sprachstufenspezifisch sind. Diese Spezifikationen sollten von der Anbieterseite gemeinsam mit den TDM-Anwender*innen im Sinne eines Community-Standards entwickelt werden. Der Programmiercode, der für den Verarbeitungsprozess von den Ausgangstexten zu den abgeleiteten Textformaten eingesetzt wird, muss darüber hinaus öffentlich und frei zur Verfügung stehen, damit die Nachvollziehbarkeit und Nachnutzbarkeit des Codes garantiert ist. Ideal wäre es, wenn die Pipeline zur Erstellung der Textformate institutionenübergreifend entwickelt, gepflegt und geprüft würde, sodass es auch hier zu einer Standardisierung kommt. Dann könnte auch eine Dokumentation gemeinsam entwickelt werden, die die Verarbeitungsschritte in Prosa erläutert und anhand urheberrechtlich unbedenklicher Textbestände demonstriert. Eine solche Standardisierung der Prozesse und Formate wird es dann auch erlauben, Bestände abgeleiteter Textformate aus unterschiedlichen anbietenden Institutionen für bestimmte Forschungsvorhaben zu kombinieren.

Aus den genannten Punkten ergeben sich sicherlich auch infrastrukturelle und institutionelle Herausforderungen mit Bezug auf die Personalentwicklung (Data Curation, Rechts- und IT-Expertise), wie sie derzeit vermutlich nur bei größeren Gedächtnisinstitutionen vorhanden sind. Auch bestimmte Rechte und Pflichten der beteiligten Akteure werden in diesem Kontext von Relevanz sein, insbesondere betrifft dies die Service anbietenden Bibliotheken und Archive und die TDM praktizierenden Nutzenden. Hierbei gilt es, unter Berücksichtigung von fachspezifischen, bibliothekarischen, wirtschaftlichen und juristischen Belangen, für alle beteiligten Akteure transparente Rahmenbedingungen zu schaffen. Diese können in einer Data Governance münden und darüber hinaus können inhaltliche Entwicklungen durch Policies gesteuert werden.

[57] Vgl. DBV Sektion 4 2018, passim.
[58] Unter Optimierungsprozess ist beispielsweise die Erweiterung des Datenbestandes, die Optimierung von Algorithmen oder die qualitative Anhebung der Daten durch verbesserte Software für die optische Zeichenerkennung zu verstehen.

6.2 Informatik und Computerlinguistik

Im Zusammenhang mit den abgeleiteten Textformaten ergeben sich auch für die Informatik und Computerlinguistik neue Tätigkeitsfelder. So kann die Computerlinguistik sicherlich dabei mitwirken, standardisierte und zukunftsfähige Textformate sowie insbesondere standardisierte und zertifizierbare Pipelines für das Erstellen von abgeleiteten Textformaten zu entwickeln. Darüber hinaus könnten diese Fächer in den Bereichen Benchmarking und Rekonstruierbarkeit tätig werden.

Die empirische Überprüfung der Wahrscheinlichkeit beziehungsweise des Grades, mit der ein Ausgangstext auf der Grundlage eines abgeleiteten Textformats rekonstruiert werden kann, wird hier sicherlich eine wichtige Aufgabe der Informatik sein. Welcher Anteil eines Textes kann rekonstruiert werden? Wie lang sind die Fragmente, die auf diese Weise entstehen? Mit welcher Wahrscheinlichkeit sind die rekonstruierten Fragmente auch tatsächlich korrekt? Neben solchen Fragen ist hier insbesondere relevant, dass eine Rekonstruktion auch nicht durch die Kombination von Informationen aus mehreren unterschiedlichen abgeleiteten Textformaten möglich sein sollte. Für die Beantwortung dieser Fragen sind gemeinfreie Textbestände, bei denen Ausgangstexte und abgeleitete Formate gemeinsam vorliegen, einsetzbar beziehungsweise erforderlich.

In Kooperation von Informatik oder Computerlinguistik mit den Anwender*innen in den DH sollten Benchmarking-Analysen und vergleichende Untersuchungen durchgeführt werden. Solche Analysen können auf der Grundlage von Datenbeständen, die Ausgangstexte und abgeleitete Textformate gleichermaßen beinhalten, überprüfen, wie groß die Unterschiede in der Performance bestimmter Analysemethoden sind, wenn man sie auf den Ausgangstexten einerseits, verschiedenen abgeleiteten Textformaten andererseits, einsetzt. Derartige Analysen wiederum sind schon jetzt von großem Interesse für die Definition und Spezifikation der abgeleiteten Textformate und können auch auf urheberrechtlich unbedenklichen (beispielsweise gemeinfreien) Beständen durchgeführt werden, bevor eine verlässliche rechtswissenschaftliche Einschätzung eines bestimmten Textformats vorliegt.

6.3 Digital Humanities

Eine wesentliche Aufgabe der DH als die am unmittelbarsten betroffenen Stakeholder wird es sein, eine enge Abstimmung zwischen den verschiedenen Akteuren und Communities sicherzustellen. Denn nur, wenn ein mit allen relevanten Akteuren (insbesondere aus DH, Rechtswissenschaften, Gedächtnisinstitutionen und Informatik) abgestimmtes Inventar von abgeleiteten Textformaten entwickelt wird, können auch entsprechend standardisierte und zertifizierte Pipelines für das Erstellen der Textformate entwickelt und angeboten werden. Diese tragen wiederum entscheidend zur Verlässlichkeit und Nützlichkeit der in dieser Form angebotenen Textbestände für die Forschung in den DH bei.

Mit dem vorliegenden Beitrag haben wir das Ziel verfolgt, erste Schritte zu einer solchen Konsensbildung innerhalb der Fachcommunity zu gehen. Anliegen war uns demnach, die Grundidee der abgeleiteten Textformate und die damit einhergehenden Chancen und Herausforderungen insbesondere in der Community der DH sowie der Bibliotheken und Archive vorzustellen und eine Diskussion über sie anzuregen. Wir hoffen, dass unser Beitrag zeigen konnte, dass das Erstellen, Veröffentlichen und Nutzen solcher abgeleiteten Textformate grundsätzlich möglich und wünschenswert ist. Darüber hinaus hoffen wir, dass es gelungen ist, der Realisierung solcher Datenbestände näher zu kommen und die nächsten notwendigen Schritte unterschiedlicher Akteure herauszuarbeiten.

Bibliographische Angaben

A brief survey of text mining: Classification, clustering and extraction techniques. Hg. von Mehdi Allahyari / Seyedamin Pouriyeh / Mehdi Assefi / Saied Safaei / Elizabeth D. Trippe / Juan Bernardo Gutierrez / Krys Kochut. In: Computing Research Repository (CoRR) arXiv.org. Version 1 arXiv:1707.02919 vom 10.07.2017. [online]

Katharina Anton: § 60d UrhG. In: Recht der elektronischen Medien: Kommentar. Hg. von Gerald Spindler / Fabian Schuster et al. 4. Auflage. München 2019. [Nachweis im GBV]

Christian Berger: Urheberrecht in der Wissensgesellschaft. In: Gewerblicher Rechtsschutz und Urheberrecht 119 (2017), H. 10, S. 953–964. [Nachweis im GBV]

Sayan Bhattacharyya / Peter Organisciak / John Stephen Downie: A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features. In: Interdisciplinary Science Reviews 40 (2015), H. 1, S. 61–77. DOI: 10.1179/0308018814Z.000000000105 [Nachweis im GBV]

David Meir Blei: Introduction to probabilistic topic models. In: Communications of the ACM 55 (2011), H. 4, S. 77–84. [Nachweis im GBV]

Cameron Blevins: Topic Modeling Martha Ballard's Diary. In: Historying.org. Blogbeitrag vom 01.04.2010. [online]

Winfried Bullinger: § 60d UrhG. In: Praxiskommentar Urheberrecht: UrhG, VGG, InsO, UKlaG, KUG, EVtr, InfoSoc-RL. Hg. von Artur-Axel Wandtke / Winfried Bullinger. 5., neu bearbeitete und erweiterte Auflage. München 2019. [Nachweis im GBV]

Manuel Burghardt / Selina Meyer / Stephanie Schmidtbauer / Johannes Molz: ›The Bard meets the Doctor‹ – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who. In: DHd 2019 Digital Humanities: multimedial und multimodal. Konferenzabstracts. Hg. von Patrick Sahle 2019, S. 222–225. DOI: 10.5281/zenodo.2596094

Marco Büchler / Philip Robert Burns / Greta Franzini / Emily Franzini / Martin Müller: Towards a Historical Text Re-use Detection. In: Text Mining: From Ontology Learning to Automated Text Processing Applications. Hg. von Chris Biemann / Alexander Mehler. Heidelberg u. a. 2014, S. 221–238. (= Theory and Applications of Natural Language Processing) [Nachweis im GBV]

Hugh Craig / Arthur Frederick Kinney: Shakespeare, Computers, and the Mystery of Authorship. Cambridge u. a. 2009. [Nachweis im GBV]

Wissenschaftliche Bibliotheken 2025. Hg. von Deutscher Bibliotheksverband e.V. Sektion 4 »Wissenschaftliche Universalbibliotheken«. München 2018. PDF. [online]

Jacob Devlin: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Computing Research Repository (CoRR) in arXiv.org. Version 1 arXiv:1810.04805 vom 11.10.2018). [online]

Frédéric Döhl (2020a): Digital Turn – Gedächtnisinstitutionen und Digital Humanities. Zwischenbericht einer Workshop-Reihe der Deutschen Nationalbibliothek. In: Zeitschrift für Bibliothekswesen und Bibliographie 67 (2020), H. 3–4, S. 213–230. [Nachweis im GBV]

Frédéric Döhl (2020b): Game Changer für Gedächtnisinstitutionen und Digital Humanities? Herausforderungen des neuen Rechts auf wissenschaftliche Nachnutzung von Korpora bei Text und Data Mining, Art. 3 DSM-RL und §§ 44b, 60d UrhG-E. In: Recht und Zugang 1 (2020), H. 2. (im Erscheinen)[Nachweis im GBV]

Thomas Dreier: § 60d UrhG. In: Urheberrechtsgesetz: UrhG. Verwertungsgesellschaftengesetz, Kunsturhebergesetz: Kommentar. Hg. von Thomas Dreier / Gernot Schulze. 6. Auflage. München 2018. [Nachweis im GBV]

Thomas Dreier: Die Schlacht ist geschlagen – ein Überblick. Zum Ergebnis des Copyright Package der EU-Kommission. In: Gewerblicher Rechtsschutz und Urheberrecht 121 (2019), H. 8, S. 771–779. [Nachweis im GBV]

Rossana Ducato / Alain Strowel: Limitations to Text and Data Mining and Consumer Empowerment. In: International Review of Intellectual Property and Competition Law 50 (2019), H. 6, S. 649–684. [Nachweis im GBV]

Katharina de la Durantaye: Neues Urheberrecht für Bildung und Wissenschaft – eine kritische Würdigung des Gesetzentwurfs. In: Gewerblicher Rechtsschutz und Urheberrecht 119 (2017), H. 6, S. 558–567. [Nachweis im GBV]

Cynthia Dwork / Aaron Roth: The Algorithmic Foundations of Differential Privacy. In: Foundations and Trends in Theoretical Computer Science 9 (2014), H. 3–4, S. 211–407. [Nachweis im GBV]

Maciej Eder / Mike Kestemont / Jan Rybicki: Stylometry with R: A package for computational text analysis. In: The R Journal 8 (2016), H. 1, S. 107–121. [online]

Ronen Feldman / James Sanger: The Text Mining Handbook. Advanced approaches in analyzing unstructured data. Cambridge u. a. 2007. [Nachweis im GBV]

Norbert P. Flechsig: Europäisches Urheberrecht in der Digitalität. In: Jur-PC. Internet-Zeitschrift für Rechtsinformatik und Informationsrecht 145 (2019), Abs. 1–320. DOI: 10.7328/jurpcb20193411146

Theodor Fontane: Effi Briest. In: TextGrid Repository. Digitale Bibliothek. Göttingen 2012. Handle: 11858/00-1734-0000-0002-AF56-2.

Christophe Geiger / Giancarlo Frosio / Oleksandr Bulayenko: Text and Data Mining Articles 3 and 4 of the Directive 2019/790/EU. In: Propiedad intelectual y mercado único digital europeo. Hg. von Concepción Saiz García / Raquel Evangelio Llorca. Valencia 2019. (= Centre for International Intellectual Property Studies Research Paper, 8) DOI: 10.2139/ssrn.3470653

Dirk Goldhahn / Thomas Eckart / Uwe Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation. Hg. von Calzolari Nicoletta. (LREC'12, Istanbul, 21.–27.05.2012) Paris 2012. [online]

Karina Grisse: Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte – Rechtliche Bedingungen und Möglichkeiten. In: Recht und Zugang 1 (2020), H. 2. (im Erscheinen)[Nachweis im GBV]

Stefanie Hagemeier: § 60d UrhG. In: BeckOk Urheberrecht. Hg. von Hartwig Ahlberg / Horst-Peter Götting. 28. Edition. München 2020. [Nachweis im GBV]

William Leif Hamilton / Jure Leskovec / Dan Jurafsky: Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Hg. von Association for Computational Linguistics. 2 Bände. (ACL: 54, Berlin, 07.–12.08.2016) Stroudsburg, PA 2016. Bd. 1: Long Papers, S. 1489–1501. DOI: 10.18653/v1/P16-1141

Sepp Hochreiter / Jürgen Schmidhuber: Long short-term memory. In: Neural computation 9 (1997), H. 8, S. 1735–1780. [Nachweis im GBV]

Regula Hohl-Trillini / Sixta Quassdorf: A ›key to all quotations‹? A corpus-based parameter model of intertextuality. In: Literary and Linguistic Computing 25 (2010), H. 3, S. 269–286. [Nachweis im GBV]

David L. Hoover: Teasing out Authorship and Style with t-tests and Zeta 2010. In: Digital Humanities 2010. Conference abstracts. (DH 2010, London, 07–10.07.2010) London 2010. [online] [Nachweis im GBV]

Andreas Hotho / Andreas Nürnberger / Gerhard Paaß: A brief survey of text mining. In: LDV Forum – GLDV Journal for Computational Linguistics and Language Technology 20 (2005), H. 1, S. 19–62. PDF. [online] [Nachweis im GBV]

Fotis Jannidis: Netzwerke. In: Digital Humanities: Eine Einführung. Hg. von Fotis Jannidis / Hubertus Kohle / Malte Rehbein. Stuttgart 2017, S. 147–161. [Nachweis im GBV]

Digital Humanities: eine Einführung. Hg. von Fotis Jannidis / Hubertus Kohle / Malte Rehbein. Stuttgart 2017. [Nachweis im GBV]

Jacob Jett / Boris Capitanu / Deren Kudeki / Timothy Cole / Yuerong Hu / Peter Organisciak / Ted Underwood / Eleanor Dickson Koehl / Ryan Dubnicek / John Stephen Downie: The HathiTrust Research Center Extracted Feature Dataset (2.0). In: wiki.htrc.illinois.edu. Hg. von HathiTrust Research Center. Blogbeitrag v.2.0 vom 16.06.2020. DOI: 10.13012/R2TE-C227

Matthew Lee Jockers: Macroanalysis – Digital Methods and Literary History. Urbana, IL u. a. 2013. [Nachweis im GBV]

Florian Jotzo: Der Schutz großer Textbestände nach dem UrhG. Stolpersteine bei der Nutzbarmachung fremder Textbestände für die Forschung. In: Recht und Zugang 1 (2020), H. 2. (im Erscheinen)[Nachweis im GBV]

Adam Kilgarriff: Comparing Corpora. In: International Journal of Corpus Linguistics 6 (2001), H. 1, S. 97–133. [Nachweis im GBV]

Evgeny Kim / Roman Klinger: A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. In: Zeitschrift für digitale Geisteswissenschaften 4 (2019). DOI: 10.17175/2019_008.

Yuri Lin / Jean-Baptiste Michel / Erez Aiden Lieberman / Jon Orwant / Will Brockman / Slav Petrov: Syntactic Annotations for the Google Books NGram Corpus. In: System demonstrations. 50th annual meeting of the Association for Computational Linguistics 2012. Hg. von Min Zhang. (ACL 2012: 50, Jeju Island, 08.–14.07.2012) Red Hook, NY, S. 169–174. [online] [Nachweis im GBV]

Bing Liu: Sentiment analysis and opinion mining (Synthesis lectures on human language technologies). San Rafael, CA 2012. (= Synthesis Lectures on Human Language Technologies, 16) DOI: 10.2200/S00416ED1V01Y201204HLT016 [Nachweis im GBV]

Tomas Mikolov / Ilya Sutskever / Kai Chen / Greg S. Corrado / Jeffrey Dean: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 25. 26th Annual Conference on Neural Information Processing Systems. Hg. von Christopher J. C. Burges / Léon Bottou / M. Welling / Zarah Ghahramani / Kilian Q. Weinberger. 4 Bände. (NIPS'13: 26, Lake Tahoe, NV, 03.–06.12.2013) Red Hook 2013. Bd. 2: S. 3111–3119. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems [Nachweis im GBV]

Clemens Neudecker / Konstantin Baierer /Maria Federbusch / Matthias Boenig / Kay Michael Würzner / Volker Hartmann / Elisa Herrmann: OCR-D: An end-to-end open source OCR framework for historical documents. In: Europeana Tech (2019), H. 13. Artikel vom 31.07.2019. [online]

Axel Nordemann: § 60d UrhG. In: Fromm / Nordemann. Urheberrecht. Kommentar zum Urheberrechtsgesetz, Verlagsgesetz, Einigungsvertrag (Urheberrecht), neu: zur EU-Portabilitätsverordnung. Hg. von Axel Nordemann / Jan Bernd Nordemann / Christian Czychowski. 12., erweiterte und überarbeitete Auflage. Stuttgart 2018. [Nachweis im GBV]

Bo Pang / Lillian Lee: Opinion Mining and Sentiment Analysis. Boston, MA u. a. 2008, S. 1–135. (= Foundations and Trends in Information Retrieval, 2, 1–2) [Nachweis im GBV]

Thomas Pflüger / Oliver Hinte: Das Urheber-Wissensgesellschafts-Gesetz aus Sicht von Hochschulen und Bibliotheken. In: Zeitschrift für Urheber- und Medienrecht 62 (2018), H. 3, S. 153–161. [Nachweis im GBV]

Benjamin Raue (2017a): Das Urheberrecht der digitalen Wissen(schaft)sgesellschaft. In: Gewerblicher Rechtsschutz und Urheberrecht 119 (2017), H. 1, S. 11–19. [Nachweis im GBV]

Benjamin Raue (2017b): Text und Data Mining. In: Computer und Recht 33 (2017), H. 10, S. 656–662. [Nachweis im GBV]

Benjamin Raue: Rechtssicherheit für datengestützte Forschung. In: Zeitschrift für Urheber- und Medienrecht 63 (2019), H. 8–9, S. 684–693. [Nachweis im GBV]

Benjamin Raue: Die geplanten Text- und Data Mining-Schranken (§§ 44b und 60d UrhG-E). In: Zeitschrift für Urheber- und Medienrecht 64 (2020), H. 3, S. 172–175. [Nachweis im GBV]

Christian Reul / Dennis Christ / Alexander Hartelt / Nico Balbach / Maximilian Wehner / Uwe Springmann / Christoph Wick / Christine Grundig / Andreas Büttner / Frank Puppe: OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. In: Applied Sciences 9 (2019), H. 22, Nr. 4853. Artikel vom 13.11.2019. [online]

Lisa Marie Rhody: Topic Modeling and Figurative Language. In: Journal of Digital Humanities 2 (2012), H. 1. [online]

Haimo Schack: Das neue UrhWissG – Schranken für Unterricht, Wissenschaft und Institutionen. In: Zeitschrift für Urheber- und Medienrecht 61 (2017), H. 11, S. 802–808. [Nachweis im GBV]

Marthe Schaper / Urs Verweyen: Die Europäische Urheberrechtsrichtlinie (EU) 2019/790. In: Kommunikation und Recht 22 (2019), H. 7/8, S. 433–441. [Nachweis im GBV]

Christof Schöch (2017a): Quantitative Analyse. In: Digital Humanities: eine Einführung. Hg. von Fotis Jannidis / Hubertus Kohle / Malte Rehbein. Stuttgart 2017, S. 280–299. [Nachweis im GBV]

Christof Schöch (2017b): Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 11 (2017), H. 2. [online]

Christof Schöch: Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie. In: Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven. Hg. von Toni Bernhart / Marcus Willand / Sandra Richter / Andrea Albrecht. (Scientia Quantitatis, Hannover, 30.09.–02.102014) Berlin u. a. 2018, S. 77–94. DOI:10.1515/9783110523300 [Nachweis im GBV]

Louisa Specht: Die neue Schrankenregelung für Text und Data Mining und ihre Bedeutung für die Wissenschaft. In: Ordnung der Wissenschaft (2018), H. 4, S. 285–289. PDF. [online] [Nachweis im GBV]

Gerald Spindler: Text und Data Mining – urheber- und datenschutzrechtliche Fragen. In: Gewerblicher Rechtsschutz und Urheberrecht 118 (2016), H. 11, S. 1112–1120. [Nachweis im GBV]

Gerald Spindler: Text und Datamining im neuen Urheberrecht und in der europäischen Diskussion. In: Zeitschrift für Geistiges Eigentum 10 (2018), H. 3, S. 273–300. [Nachweis im GBV]

Gerald Spindler: Die neue Urheberrechts-Richtlinie der EU, insbesondere ›Upload-Filter‹ – Bittersweet? In: Computer und Recht 35 (2019), H. 5, S. 277–291. [Nachweis im GBV]

Efstathios Stamatatos: A survey of modern authorship attribution methods. In: Journal of the Association for Information Science and Technology 60 (2009), H. 3, S. 538–556. [Nachweis im GBV]

Judith Steinbrecher: Die EU-Urheberrechtsrichtlinie aus Sicht der Digitalwirtschaft. Zeit für Augenmaß und faktenbasierte Gesetzgebung. In: Multimedia und Recht 22 (2019), H. 10, S. 639–643. [Nachweis im GBV]

Malte Stieper: Das Verhältnis der verpflichtenden Schranken der DSM-RL zu den optionalen Schranken der InfoSoc-RL. In: Gewerblicher Rechtsschutz und Urheberrecht 122 (2019), H. 1, S. 1–7. [Nachweis im GBV]

Pedro Javier Ortiz Suárez / Benoît Sagot / Laurent Romary: Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In: Proceedings of the Workshop on Challenges in the Management of Large Corpora. Hg. von Piotr Bánski / Adrien Barbaresi / Hanno Biber / Evelyn Breiteneder / Simon Clematide / Marc Kupietz / Harald Lüngen / Caroline Iliadi. Mannheim 2019. DOI: 10.14618/ids-pub-9021

Peer Trilcke: Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. In: Empirie in der Literaturwissenschaft. Hg. von Philip Ajouri / Katja Mellmann / Christoph Rauen. Münster 2013, S. 201–247. [Nachweis im GBV]

Ashish Vaswani / Noam Shazeer / Niki Parmar / Jakob Uszkoreit / Llion Jones / Aidan N. Gomez / Łukasz Kaiser / Illia Polosukhin: Attention is all you need. In: Advances in neural information processing systems 30. 31st Annual Conference on Neural Information Processing Systems. Hg. von Ulrike von Luxburg / Isabelle Guyon / Samy Bengio / Hanna Wallach / Rob Fergus / S.V.N. Vishwanathan / Roman Garnett. 10 Bände. (NIPS'17, Long Beach, CA, 04.–09.12.2017) Red Hook, NY 2017. Bd. 9: S. 6000–6010. DOI: 10.5555/3295222.3295349 [Nachweis im GBV]

Alex Wang / Yada Pruksachatkun / Nikita Nangia / Amanpreet Singh / Julian Michael / Felix Hill / Omer Levy / Samuel R. Bowman: Superglue: A stickier benchmark for general-purpose language understanding systems. In: Advances in neural information processing systems 32. Hg. von Hanna Wallach / Hugo Larochelle / Alina Beygelzimer / Florence d'Alché-Buc / Emily Fox / Roman Garnett. (NeurIPS 2019, Vancouver, 08.–14.12.2018) Red Hook, NY 2019, S. 3261–3275. [online] [Nachweis im GBV]

Gesetzestexte

Gesetz über Urheberrecht und verwandte Schutzgesetze (UrhG). Bundesrepublik Deutschland, vertreten durch die Bundesministerin der Justiz und für Verbraucherschutz vom 28.11.2018. [online]

Referentenentwurf für das Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes. Bundesrepublik Deutschland, vertreten durch die Bundesministerin der Justiz und für Verbraucherschutz vom 13.10.2020. PDF. [online]

Richtlinie (EU) 2019/790 des Europäischen Parlaments und des Rates vom 17. April 2019 über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt und zur Änderung der Richtlinien 96/9/EG und 2001/29/EG (DSM-RL) vom 17.04.2019. [online]

Richtlinie 2001/29/EG des Europäischen Parlaments und des Rates vom 22. Mai 2001 zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft (InfoSoc-RL 2001/29/EG) vom 22.05.2001. PDF. [online]

Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung, DSGVO) vom 27.04.2016. [online]

Abbildungslegenden und -nachweise

Tab. 1: Ausschnitt aus der Term-Dokument-Matrix für Fontanes Effi Briest. Hier mit Wortform, Lemma und Wortart-Information, absteigend sortiert nach absoluter Häufigkeit. [Schöch et al. 2020]

Ausz. 1: Ausschnitt aus der Liste der Tokens mit Annotation bei segmentweiser Aufhebung der Sequenzinformation für den Beginn von Fontanes Effi Briest. Hier auf Unigramm-Basis und mit Wortform, Lemma und Wortart-Information sowie einer Segmentlänge von 20 Tokens. Man beachte die Markierung der Segmentgrenzen mit <SEG> nach jeweils 20 Tokens. [Schöch et al. 2020]

Ausz. 2: Abfolge der Tokens mit Annotation bei selektiver Entfernung der Wortform- und Lemma-Information für den Beginn von Fontanes Effi Briest. [Schöch et al. 2020]

Tab. 2: Häufigkeiten von 3-Grammen über mehrere Texte hinweg, bei einer Mindesthäufigkeit von 5. Beispieldaten auf der Grundlage von fünf Erzähltexten von Theodor Fontane. [Schöch et al. 2020]

Abb. 1: Übersicht über die abgeleiteten Textformate, ihre Nützlichkeit in der Forschung und ihre urheberrechtliche Einschätzung. [Hinzmann / Schöch 2020]