Prof. Dr. Christof Schöch, Dr. Frédéric Döhl, Prof. Dr. Achim Rettinger, Prof. Dr. Evelyn Gius, JProf. Dr. Peer Trilcke, Dr. Peter Leinen, Prof. Dr. Fotis Jannidis, Dr. Maria Hinzmann und Jörg Röpke*

Abgeleitete Textformate: Prinzip und Beispiele

I. Einleitung

Der vorliegende Beitrag macht Vorschläge für einen pragmatischen Umgang mit der derzeitigen rechtlichen Situation bei der Nutzung von Methoden des Text und Data Mining (TDM)[1] in den Digital Humanities (DH) und speziell in den Computational Literary Studies (CLS).[2] Vor dem Hintergrund der nun verfügbaren Schranke zugunsten von TDM, aber auch der weiterhin bestehenden Hürden, insbesondere bezüglich des Teilens von Textbeständen, die für das TDM relevant sind, ist es das Anliegen des vorliegenden Beitrags, Perspektiven und Möglichkeiten für die Erstellung, Analyse und Anschlussforschung an solchen Textsammlungen, die auf der Grundlage von urheberrechtlich geschützten Textbeständen entstanden sind, zu eröffnen. Ziel ist es, die offene Publikation und freie Nachnutzbarkeit von *abgeleiteten Textformaten* für die Nachvollziehbarkeit von Analyseergebnissen für Dritte auch außerhalb formaler Qua-

* Christof Schöch ist Inhaber der Professur für Digital Humanities und Ko-Direktor des Trier Center for Digital Humanities. Frédéric Döhl ist Strategiereferent für Digital Humanities der Deutschen Nationalbibliothek. Achim Rettinger ist Inhaber der Professur für Computerlinguistik an der Universität Trier. Evelyn Gius ist Inhaberin der Professur für Digital Philology – Neuere Deutsche Literaturwissenschaft an der TU Darmstadt. Peer Trilcke ist Inhaber der Juniorprofessur deutsche Literatur des 19. Jahrhunderts mit dem Schwerpunkt Theodor Fontane an der Universität Potsdam und Leiter des Theodor-Fontane-Archivs. Peter Leinen ist Leiter des Fachbereichs Informationsinfrastruktur an der Deutschen Nationalbibliothek. Fotis Jannidis ist Inhaber der Professur für Computerphilologie und Neuere Deutsche Literaturgeschichte an der Universität Würzburg. Maria Hinzmann ist Projektkoordinatorin am Trier Center for Digital Humanities. Jörg Röpke ist Leiter der Abteilung Informationstechnologie, Forschungs- und Publikationsdienste an der Universitätsbibliothek Trier. Dieser Beitrag beruht auf dem DFG-Expertenworkshop Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte, der im November 2019 / Januar 2020 an der Universität Trier stattgefunden hat. Der Beitrag ist ein Auszug aus einer umfassenderen Darstellung, vgl. Schöch et al. ZfdG 2020.

1 Vgl. zu TDM u.a. *Hotho/Nürnberger/Paaß*, LDV Forum 2005, 19; *Allahyari et al.*, arXiv:1707.02919 (2017).

2 Vgl. u.a. *Jannidis et al.*, Digital Humanities: eine Einführung, 2017; *Jockers*, Macroanalysis – Digital Methods and Literary History, 2013.

 DOI: 10.5771/2699-1284-2020-2-160

litätssicherungsprozesse und für Anschlussforschung ohne rechtliche Einschränkungen zu ermöglichen.

II. Das Prinzip der abgeleiteten Textformate

Vor dem in den Beiträgen von *Jotzo* (RuZ 2020, 128) und *Grisse* (RuZ 2020, 143) 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; ggfs. 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 (bspw. durch linguistische Annotation) als auch eine Informationsreduktion (bspw. durch Löschung der Wortformen oder Aufhebung der Sequenzinformation) darstellen, in sogenannte abgeleitete Textformate verwandelt. Diese Verarbeitungsroutinen können ggfs. 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. 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.

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

1. die Information über die Wortform des Tokens, also die Abfolge der Zeichen; mit oder ohne Berücksichtigung der Groß-/Kleinschreibung;
2. das Lemma, also die unflektierte Grundform, wie man sie als Wörterbucheintrag finden würde;
3. die Wortart, also die grammatikalische Klasse (Substantiv, Verb, Adjektiv, Pronomen, etc.), gegebenenfalls und sofern relevant auch weitere morpho-syntaktische Informationen (Genus, Numerus, Casus);
4. die Bedeutung, also der semantische Gehalt des Wortes; repräsentiert beispielsweise über Zuordnung eines Wortvektors aus einem Word Embedding Model oder eines Synsets in WordNet;
5. 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;
6. 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;
7. die Häufigkeit des Tokens, wobei die Häufigkeit im Satz, im jeweiligen Segment, im Gesamttext oder in einer Gruppe von Texten gemeint sein kann; zudem kann die Häufigkeit als binäre, absolute oder relative Häufigkeit ausgedrückt werden.

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. Spezifiziert über die jeweils gewählten Parameter (insbesondere: Segmentlänge), ergeben sich eine Vielzahl möglicher Transformationen der Ausgangstexte in verschiedene abgeleitete Textformate. Unterschiedliche Textformate eignen sich dabei je unterschiedlich gut für bestimmte Analyseverfahren.

III. Vorschläge für abgeleitete Textformate

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 bzw. sinnvoll) jeweils Beispiele oder Ausschnitte der entstehenden Dateiformate abgebildet.[3] 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. Fronthaus, Seitenflügel und Kirchhofsmauer bildeten ein einen kleinen Ziergarten umschließendes Hufeisen, an dessen offener Seite man eines Teiches mit Wassersteg und angeketteltem Boot und dicht daneben einer Schaukel gewahr wurde, deren horizontal gelegtes Brett zu Häupten und Füßen an je zwei Stricken hing – die Pfosten der Balkenlage schon etwas schief stehend. Zwischen Teich und Rondell aber und die Schaukel halb versteckend standen ein paar mächtige alte Platanen.“*[4]

1. Token-basierte Textformate

Zunächst gehen wir auf token-basierte 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.

a) 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

3 Vollständige Beispiele für die hier beschriebenen abgeleiteten Textformate sowie der sie erzeugende Programmcode liegen in einem Github-Repository vor, abrufbar unter: https://github.com/dh-trier/tmr, zuletzt abgerufen am 18.10.2020.

4 *Fontane*, Effi Briest, 1894–95, S. 3.

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 bzw. des Gesamtvokabulars) bestimmt wird.

Aufgrund der prinzipiellen Einfachheit des Formats enthält es nur eine kleine Anzahl von Parametern, die es genauer spezifizieren. Hierzu gehören insbesondere die beiden folgenden Parameter:

1. welche Tokenisierung angesetzt wird, d.h. welche Definition von Token für die Segmentierung des Textes in einzelne Tokens, bspw. Wörter, verwendet wird;
2. 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 bspw. als Wortvektor –, angeboten werden.

Rang	**Token (Wortform_Wortart_Lemma)**	fontane_effi-briest
1	,_PUN_,	10307
2	._PUN_.	5204
3	und_KON_und	4087
4	«_PUN_«	1937
5	»_PUN_»	1937
6	die_ART_die	1927
7	ich_PPER_ich	1715
8	sie_PPER_sie	1703
9	das_ART_der/die/das	1618
10	der_ART_der/die/das	1510
11	es_PPER_es	1410
12	nicht_PTKNEG_nicht	1364
13	in_PRP_in	1102
14	so_ADV_so	1020
15	ist_VAFIN_sein	998
16	zu_KONJ_zu	983
...	...	...

Tabelle 1: Ausschnitt aus der Term-Dokument-Matrix für Fontanes Effi Briest. Hier mit Wortform, Lemma und Wortart-Information, absteigend sortiert nach absoluter Häufigkeit.

Dieses abgeleitete Textformat kann folgendermaßen eingeschätzt werden:

- Für einige Analyseverfahren, insbesondere für einfache Varianten der Klassifikation und des Clustering bspw. für Fragen der Autorschaftsattribution, und für einfache Distinktivitätsmaße ist das Format geeignet. Für viele andere Verfahren, darunter für Topic Modeling, Sentiment Analyse, 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.

b) 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 die folgenden: Wie bei den meisten Textformaten sind dies auch hier die Tokenisierung

und die über das Token verfügbare Information. Wesentlich sowohl aus Anwendungs- als auch aus rechtlicher Perspektive ist hier allerdings der 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_, Seiten-
flügel_NN_Seitenflügel breiten_ADJA_breit die_ART_die hin_ADV_hin wäh-
rend_KOUS_während angebauter_ADJA_angebaut der_ART_die nach_APPR_nach
ein_ART_eine Schatten_NN_Schatten auf_APPR_auf einen_ART_eine rechtwink-
lig_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 die-
sen_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>
```

Auszug 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.

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 nur eine minimale 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. Für Topic Modeling ist das Format ebenfalls gut geeignet. Nur bei einer sehr geringen Segmentlänge oder bei einer Segmentierung in Sätze erscheint eine einfache Sentiment Analyse denkbar. Verfahren der Netzwerkanalyse sind denkbar, wären aber auf eine vorgängige, hochwertige Named Entity Recognition und Coreference 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 Sentiment Analyse.

- 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 kleineren Segmentlänge (bspw. <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).

c) 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 (vgl. 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

Auszug 2: Abfolge der Tokens mit Annotation bei selektiver Entfernung der Wortform- und Lemma-Information für den Beginn von Fontanes Effi Briest.

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 die 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 Sentiment Analyse 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 ursprüng-

lichen 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.

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) bzw. grundsätzlich unter Rückgriff auf ein umfangreicheres Korpus ermittelt werden (Wortembeddings).

a) 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 1a)). 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 bzw. sehr viele Einzeltexte hinweg zu berechnen. Die Parameter eines solchen Formats sind (neben der Tokenisierung und Annota-

tion) die N-Gramm-Länge und die Bezugsgröße für die N-Gramm-Häufigkeiten, bspw. jeweils alle Texte einer Textsorte und/oder eines Jahres (Tabelle 2).

Rang	N-Gramm	Häufigkeit
1	gott sei dank	43
2	ja gnädigste frau	17
3	auch heute wieder	13
4	doch auch wieder	11
5	ist doch auch	11
6	ist immer so	10
7	gnädigste frau ist	10
8	war so war	10
9	nein gnädigste frau	9
10	wird ja wohl	9
11	ist doch recht	9
12	doch immer noch	9
...	...	...

Tabelle 2: Häufigkeiten von 3-Grammen über mehrere Texte hinweg, bei einer Mindesthäufigkeit von 5. Beispieldaten auf der Grundlage von fünf Erzähltexten von Theodor Fontane.

Ein solches Format kann folgendermaßen eingeschätzt werden:

- Aus Anwendungsperspektive ist das Einsatzspektrum eines solchen Textformats sicherlich geringer als bei den Einzeltext-basierten 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 bzw. 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 längere N-Gramm-Ketten zusammensetzen, nämlich gerade da, wo häufig verwendete sprachliche Muster verwendet werden; eine entsprechende Metrik müsste also probabilistisch vorgehen.

b) 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, NLP) 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 Sentiment Analyse 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 Analyse oft entscheidende, Informationen verloren. Das Potential dieser Methoden für die Analyse von Textbeständen sowie als urheber-

rechtlich 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 Wortembeddings populär, die mit Verfahren des maschinellen Lernens Parametervektoren aus großen Textsammlungen schätzen.[5] 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 Wortembedding latente Informationen über die Wahrscheinlichkeit des Auftretens von Wörtern im Umfeld. Diese Art von Wortembeddings sind 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 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

5 Vgl. *Mikolov et al.*, Advances in neural information processing systems, 2013, S. 3111.

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.

- 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.

c) Kontextualisierte Embeddings

Der nächste essentielle Schritt zur Verbesserung bestehender NLP-Verfahren wurde durch das Kontextualisieren von Wortembeddings erreicht. Dabei wird Satz für Satz und Wort für Wort erst eine Ersetzung durch Wortembeddings 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.[6] 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.

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

6 Siehe *Devlin et al.*, arXiv 2018.

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 Analysebenchmarks wie Sentimentanalyse, semantische Textähnlichkeit oder Paraphrasierung erreichen weit bessere Ergebnisse als bisherige Verfahren; oft übertreffen diese sogar menschliche Fähigkeiten von Nicht-Experten/innen.[7]

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 bzw. Transformer. Diese erlauben es, die kontextfreien und die kontextsensitiven Vektoren für Texte zu gewinnen, was zahlreiche Anwendungen in allen modernen digitalen textanalytischen Verfahren erlaubt.

3. Fazit

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 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-syntaktische Annotation ist ebenso von Vorteil. Der Parameter der Segmentlänge wird voraussichtlich eine Abwägungsfrage bleiben: Aus der Perspektive der Analyseverfahren sind kleine Segmentlängen grundsätzlich wünschenswert (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.

Nicht verschwiegen werden sollen einige inhärente Nachteile, die mit dem Modell der abgeleiteten Textformate verbunden sind. 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

7 Vgl. GLUE Benchmark, abrufbar unter: https://gluebenchmark.com, zuletzt abgerufen am 18.10.2020.

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 ist. Dennoch überwiegen aus unserer Sicht die Vorteile dieser Strategie gegenüber den Alternativen bzw. 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.