Urheberrecht und Digital Humanities

Prof. Dr. Benjamin Raue/Prof. Dr. Christof Schöch*

Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate

– Versöhnung von Urheberrecht und textbasierter Forschung –

I. Bestandsaufnahme

Ausschließlichkeitsrechte der Urheber können gerade die Analyse großer Datenmengen behindern, die zeitgenössische, urheberrechtlich noch geschützte Texte enthalten. Sowohl die neue Text und Data Mining-Schranke in § 60 d UrhG als auch Lizenzierungsmodelle von Verlagen oder *closed room*-Zugänge von Gedächtniseinrichtungen haben für die wissenschaftliche Praxis entscheidende Nachteile (dazu II.). Deswegen haben wir das Potenzial abgeleiteter Textformate untersucht, die durch Informationsreduktion aus dem Schutzbereich des Urheberrechts gelangen und dennoch eine ausreichende Informationsgrundlage für eine Reihe von digitaler Analysemethoden sein können.

1. Das weitgehend analoge 20. und 21. Jahrhundert

Als Folge des Urheberschutzes wird etwa in den Computational Literary Studies (CLS)[1] eine erhebliche Einschränkung der digital analysierbaren Textbestände beklagt.[2] Für diese gebe es ein sehr kurzes *window of opportunity*: Es öffnet sich um 1800, weil sich frühere Schriften mit den verfügbaren Texterkennungstechniken nur

* *Benjamin Raue* ist Inhaber der Professur für Zivilrecht, Recht der Informationsgesellschaft und des Geistigen Eigentums sowie Direktor des Instituts für Recht und Digitalisierung (IRDT) an der Universität Trier; *Christof Schöch* ist Inhaber der Professur für Digital Humanities und Direktor des Trier Center for Digital Humanities (TCDH). Dieser Beitrag fasst die Ergebnisse der zwei DFG-Expertenworkshops „Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte“ zusammen, die im November 2019 und Januar 2020 an der Universität Trier stattgefunden haben.

1 Die CLS sind ein Teilbereich der Digital Humanities, in dem quantitative Methoden aus Statistik, Informatik und Natural Language Processing für die Bearbeitung literaturwissenschaftlicher Fragestellungen eingesetzt werden.

2 Dazu *Schöch* et al, ZfdG 2020.

 DOI: 10.5771/2699-1284-2020-2-118

unzureichend erschließen lassen. Wegen der urheberrechtlichen Restriktionen enden die verfügbaren Textbestände um das Jahr 1920, weil bei später veröffentlichten Werken die Wahrscheinlichkeit steigt, dass die lange urheberrechtliche Schutzfrist von 70 Jahren p.m.a. noch nicht abgelaufen ist. Die betroffenen Forscherinnen und Forscher empfinden es als starke Beeinträchtigung ihrer Wissenschaftsfreiheit, dass sie ihre Forschungsschwerpunkte nicht primär aufgrund ihres Erkenntnisinteresses bestimmen können, sondern von rechtlichen Faktoren abhängig machen müssen, wenn das Forschungsinteresse nicht mit verfügbaren Textkorpora aus Literatur des 19. Jahrhunderts befriedigt werden kann. Aus literaturwissenschaftlicher Perspektive sind das 20. und 21. Jahrhundert zwar keine dunklen, aber weitgehend analoge Jahrhunderte.

Es ist zudem ein erhebliches Problem, wenn Modelle des Maschinellen Lernens ohne die zum Trainieren eingesetzten Daten publiziert werden müssen. Das gilt insbesondere für komplexe (Sprach-)Modelle der Computerlinguistik. Wer die Trainingsdaten aus urheberrechtlichen Gründen nicht offenlegen darf, schränkt die Reproduzierbarkeit der Ergebnisse und damit deren Anerkennung im internationalen Forschungsdialog ein. Darüber hinaus erschwert das Urheberrecht die Etablierung von Referenzdatensätzen, sogenannter „Benchmarks", die ihre Funktion ebenfalls nur ausüben können, wenn sie international frei zugänglich gemacht werden können.

2. DFG Expertenworkshops „Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte"

Diese Bestandsaufnahme war Anlass für zwei DFG-Expertenworkshops „Strategien für die Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte".[3] Im vorliegenden Beitrag fassen wir die Ergebnisse der beiden Workshops zusammen, die in den Beiträgen von *Jotzo* (RuZ 2020, 128), *Grisse* (RuZ 2020, 143) und *Schöch et al* (RuZ 2020, 160) in diesem Heft weiter vertieft werden. Der Workshop hat das Potenzial abgeleiteter Textformate im Hinblick darauf untersucht, ob mit ihrer Hilfe rechtssicher große Textkorpora aktueller, urheberrechtlich noch geschützter Textbestände öffentlich zugänglich gemacht werden können. Hinter dem Konzept der abgeleiteten Textformate steht die Idee, urheberrechtlich geschützte Texte durch eine gezielte Informationsreduktion so zu transformieren, dass die urheberrechtlich geschützte Textstruktur irreversibel verloren geht. Ein Werkgenuss im eigentlichen Sinne wird dadurch unmöglich. Weil den Urhebern die in ihren Werken enthaltenen Informationen nicht zugewiesen sind,[4] ist deren Verwertungsinteresse durch ein öffentliches Zugänglichmachen der abgeleiteten Textformate nicht weiter betroffen, solange die ursprüngliche Struktur nicht mit verhältnismäßigem Aufwand rekonstruiert werden kann. Die Herausforderung ist dabei, Reduktionsverfahren zu entwickeln, die

3 Tagungsbericht von *Erler*, RuZ 2020, 108, 108 ff.

4 *Schack*, Urheber- und Urhebervertragsrecht, 9. Aufl. 2019, Rn. 196; *Obergfell*, FS Büscher, S. 223, 225; Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60 d Rn. 1; *Raue*, GRUR 2017, 10, 14.

zuverlässig die urheberrechtlich relevanten Elemente entfernen, aber den Korpora so viel Informationen belassen, dass mit ihnen in den digitalen Literaturwissenschaften verbreitete Analyseverfahren weiterhin möglich sind.

II. Lösungsansätze

Für den urheberrechtskonformen Zugang zu großen Textkorpora können vier Lösungsansätze identifiziert werden.[5]

1. Text und Data Mining Schranke

Sowohl der deutsche als auch der europäische Gesetzgeber haben erkannt, welchen Wettbewerbsnachteil das Urheberrecht bei der Analyse großer Textmengen und sonstiger Datenkorpora für die deutsche bzw. europäische Forschungslandschaft haben kann.[6] Deswegen hat der deutsche Gesetzgeber in § 60 d UrhG eine Urheberrechtsschranke für das nicht-kommerzielle Text und Data Mining eingeführt, die das Erstellen, Aufbereiten und Nutzen eines Textkorpus für die wissenschaftliche Forschung erlaubt, soweit die Forscher rechtmäßigen Zugang zu den Schutzgegenständen haben.[7]

Die neue Text und Data Mining-Schranke schafft Rechtssicherheit und Forschungsfreiheit, soweit Forscher eigene und öffentlich zugängliche Textbestände mit algorithmischer Unterstützung erforschen wollen. Wesentlicher Nachteil ist momentan aus Sicht der Wissenschaftler, dass die geltende Fassung des § 60 d III UrhG anordnet, dass Forschende das Korpus und sämtliche Vervielfältigungsstücke des Ursprungsmaterials nach Abschluss der Forschungsarbeiten löschen müssen. Diese Löschungspflicht hat die Text und Data Mining-Schranke in Art. 3 II DSM-RL 2019/790 nun entschärft. Sie erlaubt Forschern nach Ablauf der Umsetzungsfrist im Juni 2021, die Textkorpora ohne zeitliche Beschränkung für die weitere wissenschaftliche Forschung aufzubewahren und zu verwenden.[8]

Die Text und Data Mining-Schranke hat einen weiteren Nachteil. Sie ermöglicht zwar den Aufbau, die Optimierung und die automatisierte Analyse des Textkorpus, nicht aber dessen offene Publikation und Weitergabe. Die Text und Data Mining-Schranke erlaubt lediglich, die Textkorpora in einer Forschungsgruppe bzw. zur Überprüfung der wissenschaftlichen Ergebnisse im Rahmen eines Peer Review-Prozesses zugänglich zu machen, demnächst auch für die spätere Überprüfung wissenschaftlicher Erkenntnisse (Art. 3 II DSM-RL 2019/790). Abgesehen davon dürfen Forscher ihre Datengrundlage der (wissenschaftlichen Fach-)Öffentlichkeit nicht zur Verfügung stellen und ihr so erlauben, die Analyseergebnisse auch außerhalb formaler Qualitätssiche-

5 Zu den Lösungsansätzen 2 und 3 auch *Schöch* et al, ZfdG 2020.
6 Vgl. ErwGr 10 DSM-RL 2019/790.
7 Zu § 60 d UrhG etwa: *Raue*, CR 2017, 656; *Spindler*, ZGE 2018, 273, 277 ff.; *Specht*, OdW 2018, 285; *Obergfell*, FS Büscher, S. 223, 228 ff.
8 *Spindler*, CR 2019, 277, 279 f.; *Raue*, ZUM 2019, 684, 688.

rungsprozesse transparent nachzuvollziehen. Auch die Nachnutzung teilweise sehr aufwendig aufbereiteter Textkorpora für die Anschlussforschung Dritter wird dadurch verhindert. Beide Aspekte sind jedoch für eine transparente und nachhaltige Forschung sowie den Zugang zu großen Referenzkorpora von erheblicher Bedeutung.

Darüber hinaus berechtigt die Schranke nur zum Minen von Texten, zu denen die Forschenden rechtmäßigen Zugang haben.[9] Deswegen müssen sie in analoger Form vorliegende Textbestände selbst digitalisieren[10] und ggf. für die weitere Verwendung aufbereiten. Das erschwert und verteuert das Zusammenstellen sehr großer Textkorpora oder die Etablierung moderner Referenzkorpora, mit denen die Leistungsfähigkeit neuer Analysetools bzw. -ansätze demonstriert werden kann.

Der Gesetzgeber steht aber vor dem Dilemma, dass er ohne eine Beschränkung der öffentlichen Zugänglichmachung kaum das berechtigte und in internationalen Konventionen[11] abgesicherte Interesse der Rechteinhaber sicherstellen kann, durch die Text und Data Mining-Schranke die weitere Verwertung ihrer Inhalte nicht unmöglich zu machen.

2. Lizenzierungslösungen

Soweit kein eigener Zugang zu (größeren) Textkorpora besteht, stellen etwa große Verlage Zugang zu ihren Textkorpora über Schnittstellen (API) zur Verfügung. Allerdings sind diese Textbestände im Regelfall auf das Repertoire des jeweiligen Verlags beschränkt; auch gewähren nicht alle Verlage Zugang. Um ihre weiteren Verwertungsinteressen nicht zu gefährden, untersagen Verlage im Regelfall die Weitergabe der Textkorpora.

Teilweise beschränken die Datenbankinhaber weitergehend sogar den direkten Zugriff auf die Textbestände. Den Forschern ist es in diesen Fällen unmöglich, die Texte über die API herunterzuladen und die Qualität des Forschungskorpus durch Annotationen, Normalisierung oder durch die Ergänzung mit anderen Datensätzen zu verbessern. Zudem können die Forschenden typischerweise auch nicht ihre selbstentwickelten Analyseprogramme verwenden, sondern sind auf die Analysetools des jeweiligen Datenbankanbieters beschränkt.

9 RegE, BT-Drs. 18/12329, S. 41; *Raue*, CR 2017, 656, 658; Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60 d Rn. 4.

10 Wozu sie § 60 d UrhG berechtigt, RegE, BT-Drs. 18/12329, S. 41; *Raue*, CR 2017, 656, 659; Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60 d Rn. 4. Sie dürfen dafür auch Dritte einsetzen, vgl.: Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60 d Rn. 5.

11 Etwa durch den 3-Stufen-Test in Art. 13 TRIPs und Art. 10 I WCT, wonach Beschränkungen und Ausnahmen von ausschließlichen Rechten auf bestimmte Sonderfälle beschränkt werden müssen und diese weder die normale Auswertung des Werkes beeinträchtigen noch die berechtigten Interessen des Rechtsinhabers unzumutbar verletzen dürfen.

3. Closed Room-Zugang (§§ 60 e IV, 60 f I UrhG)

Beim sogenannten *closed room*-Zugang machen insbesondere öffentlich zugängliche Bibliotheken und Archive, gestützt auf §§ 60 e IV, 60 f I UrhG, ihre Bestände auf Workstations in den Räumen der Institution zugänglich. Den Zugriff von außerhalb, auch mittels VPN-Clients, wollte der Gesetzgeber nicht freistellen.[12] Zwar wird hier den Forschern typischerweise ermöglicht, ihre eigenen Analysetools zu verwenden. Erheblicher Nachteil ist neben der Ortsgebundenheit und der fehlenden Möglichkeit, diesen Datensatz durch die Kombination mit anderen Datensätzen zu erweitern, dass die Forschenden das Textkorpus weder für die eigene Nachnutzung kopieren noch Dritten für die wissenschaftliche Nachprüfung zur Verfügung stellen dürfen.

4. Abgeleitete Textformate

Dementsprechend lösen die Lizenzierungs- und *closed room*-Lösungen im Vergleich zu der Freistellung durch die Text und Data Mining-Schranke lediglich das Problem des rechtmäßigen Zugangs, nicht aber die nur sehr eingeschränkte Möglichkeit, Dritten die verwendeten Textkorpora zur Verfügung zu stellen. Diese Schwächen können abgeleitete Datenformate lösen, bei denen durch eine Informationsreduktion der urheberrechtliche Schutz der gespeicherten Texte entfällt. Der große und entscheidende Vorteil dieser Lösung ist, dass auf diese Weise sehr große Textkorpora öffentlich zugänglich gemacht und weiterverwendet werden können, ohne dass die berechtigten Interessen der Urheber dadurch berührt werden. Der auf der Hand liegende Nachteil dieser Lösung liegt in ihrer Konstruktion begründet: Durch die Informationsreduktion können mit diesen abgeleiteten Textformaten nicht alle Analyseverfahren durchgeführt werden. Zudem kann in einzelnen Fällen die Qualität der Ergebnisse der dennoch möglichen Analyseverfahren sinken.[13] Weil diese Lösung aber als einzige eine Veröffentlichung der verwendeten Textkorpora und eine Zusammenführung verschiedener Textkorpora ermöglicht, soll sie nun näher dargestellt werden.

III. Anforderungen an abgeleitete Textformate aus Sicht der Rechtswissenschaft und der Anwendungswissenschaften

1. Urheberrechtliche Vorgaben und Lösungsstrategien

a) Schutzfähigkeit

Aus urheberrechtlicher Sicht muss zunächst definiert werden, was die urheberrechtliche Schutzfähigkeit von Texten ausmacht.[14] Bei Sachtexten sind die in ihnen enthalte-

12 BT-Drs. 16/1828, S. 26; Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60e Rn. 17.
13 Hierzu besteht weiterer Forschungsbedarf, dazu: *Schöch* et al, ZfdG 2020.
14 Dazu ausführlich *Jotzo*, RuZ 2020, 128, 131 ff.

nen Informationen urheberrechtlich nicht schutzfähig.[15] Die nach § 2 II UrhG erforderliche Individualität können daher allein ihre Struktur und Gedankenführung sowie individuelle Formulierungen erreichen. Gerade bei Gebrauchstexten werden hieran relativ hohe Anforderungen gestellt, so dass kleinere Textausschnitte typischerweise urheberrechtlich schutzunfähig sind.[16] Literarische Texte weisen dagegen mit hoher Wahrscheinlichkeit eigenschöpferische Inhalte und Formulierungen auf. Zudem können fiktionale Inhalte, insbesondere Erzählstränge, Charakteristika der Figuren und deren Beziehungsgefüge urheberrechtlich geschützt sein.[17]

b) Reduktion urheberschutzbegründender Merkmale

Dementsprechend müssen aus abgeleiteten Textformaten die schutzbegründenden individuellen Formulierungen, Strukturen und fiktionalen Elemente entfernt werden.[18]

aa) Nichterkennbarkeit schutzfähiger Elemente

Als maßgebliches Kriterium dafür wird man die Erkennbarkeit der schutzbegründenden Elemente heranziehen müssen.[19] Vervielfältigungen und das öffentliche Zugänglichmachen eines Textkorpus greifen danach nicht mehr in das Urheberrecht der Ursprungstexte ein, wenn das Korpus keine schutzfähigen Teile enthält bzw. diese nicht mehr erkennbar sind.[20]

Problematisch ist, dass kleinere Textausschnitte nur *typischerweise* urheberrechtlich schutzunfähig sind.[21] Wird ein großes Textkorpus automatisiert aus einer Vielzahl von kurzen Textausschnitten zusammengestellt, kann daher nicht für alle Textarten mit Sicherheit ausgeschlossen werden, dass das Korpus vereinzelt schutzfähige Textausschnitte enthält.[22] Diese Problematik lässt sich aus unserer Sicht aber mit der Schranke des § 57 UrhG lösen. Vereinzelte urheberrechtlich schutzfähige Elemente sind in einem sehr großen Textkorpus als „unwesentliches Beiwerk neben dem eigentlichen Gegenstand der Vervielfältigung, Verbreitung oder öffentlichen Wiedergabe anzusehen". § 57

15 *Schack*, Urheber- und Urhebervertragsrecht, 9. Aufl. 2019, Rn. 196; *Obergfell*, FS Büscher, S. 223, 225; Dreier/Schulze-*Dreier*, UrhG, 6. Aufl. 2018, § 60 d Rn. 1; *Raue*, GRUR 2017, 10, 14.

16 Dazu *Jotzo*, RuZ 2020, 128, 136 f. mwN.

17 *Jotzo*, RuZ 2020, 128, 135 mwN.

18 Dazu ausführlich *Grisse*, RuZ 2020, 143, 146 ff.

19 *Grisse*, RuZ 2020, 143, 146 ff. in Verallgemeinerung von EuGH, Urt. v. 29.07.2019 – C-476/17, ECLI:EU:C:2019:624 = GRUR 2019, 929 Tz. 26 ff., 39 – *Pelham*.

20 *Grisse*, RuZ 2020, 143, 146, 149 f.

21 Der EuGH hat entschieden, dass Textausschnitte mit elf Wörtern schutzfähig sein *können*: EuGH, Urt. v. 16.07.2009 – C-5/08, ECLI:EU:C:2009:465 = GRUR 2009, 1041 Tz. 48 – *Infopaq I*. Vgl. hierzu das Beispiel des OLG München, Urt. v. 17.09.2009 – 29 U 3271/09, ZUM 2009, 970 f. – *Typisch München!*, wonach die Formulierung „Vom Ernst des Lebens halb verschont / Ist der schon, der in München wohnt" urheberschutzfähig ist. Dazu *Jotzo*, RuZ 2020, 128, 136.

22 *Jotzo*, RuZ 2020, 128, 137; *Grisse*, RuZ 2020, 143, 148.

UrhG setzt Art. 5 Abs. 3 lit. i InfoSoc-RL um, wonach eine „beiläufige Einbeziehung eines Werks oder sonstigen Schutzgegenstands in anderes Material" freigestellt wird. Diese Voraussetzungen erfüllen vereinzelt schutzfähige Textsequenzen, weil sie in das Textkorpus nicht wegen ihrer einprägsamen, individuellen Formulierung einbezogen worden sind, sondern allein wegen der darin enthaltenen Sequenzinformationen, die urheberrechtlich gerade nicht schutzfähig sind. Die Textteile werden also nicht wegen, sondern trotz ihres schutzfähigen Gehalts aus unvermeidbaren technischen Gründen[23] in das Textkorpus einbezogen. Nach dem BGH liegt die Voraussetzung der „beiläufigen Einbeziehung" immer dann vor, wenn einem durchschnittlichen Betrachter ein urheberschutzfähiger Bestandteil in dem eigentlichen Gegenstand, also hier dem Gesamttextkorpus, nicht auffällt.[24] Ein großes Textkorpus wird von einem durchschnittlichen Betrachter oder durchschnittlichen Nutzer aber gerade nicht lesend, sondern nur in Form einer statistischen Analyse betrachtet. Dabei wird ein vereinzeltes schutzfähiges Textelement gerade nicht wahrgenommen, sondern lediglich bei der statistischen Auswertung berücksichtigt.

bb) Nicht-Rekonstruierbarkeit

Neben der Nichterkennbarkeit wird man als zweite Anforderung von einem urheberrechtsfreien, abgeleiteten Textformat verlangen müssen, dass die ursprünglichen Texte nicht aufgrund von Positionsangaben der Textsequenzen oder sonstiger Sequenzinformationen mit verhältnismäßigem Aufwand rekonstruierbar sind.[25] Denn wenn die ursprünglichen, urheberschutzfähigen Texte von Nutzern rekonstruiert werden könnten, wären die abgeleiteten Textformate geeignet, Verwertungsinteressen der Urheber zu beeinträchtigen.

c) Rechtmäßigkeit der Umwandlungshandlung

Um den Ursprungstext in ein abgeleitetes Textformat zu transformieren, muss dieser im Arbeitsspeicher eines Computers abgespeichert werden. Eine solche Vervielfältigung fällt in den Schutzbereich des Vervielfältigungsrechts nach § 16 UrhG. Daher muss sich der Umwandelnde für die Transformation auf eine urheberrechtliche Schranke berufen können.[26] Grundvoraussetzung für alle Schranken ist im Regelfall, dass der Umwandelnde rechtmäßigen Zugang zu den Ursprungstexten hat.[27] In diesem Fall kann die Umwandlung des Ursprungstextes in ein urheberrechtsfreies Format auf

23 Vgl. dazu EuGH, Urt. v. 18.10.2012 – C-173/11, ECLI:EU:C:2012:642 Tz. 36 – *Football Dataco/Sportradar;* EuGH, Urt. v. 12.07.2011 – C-324/09, ECLI:EU:C:2011:474 Tz. 64 – *L'Oréal/eBay;* BGH, Urt. v. 09.11.2017 – I ZR 134/16 = GRUR 2018, 417 Tz. 46 ff. – *Resistograph* (zur Begrenzung des internationalen Anwendungsbereichs des Urheber- bzw. Markenrechts).

24 BGH, Urt. v. 17.11.2014 – I ZR 177/13 = GRUR 2015, 667 Tz. 27 – *Möbelkatalog*.

25 *Grisse*, RuZ 2020, 143, 150 f.

26 Dazu ausführlich *Grisse*, RuZ 2020, 143, 154 f.

27 *Grisse*, RuZ 2020, 143, 154.

§ 44 a UrhG gestützt werden, wenn der Ursprungstext nur vorübergehend im Arbeitsspeicher vervielfältigt und danach automatisch aus diesem wieder gelöscht wird.[28] Weil das ursprüngliche Werk in diesem Fall im abgewandelten Textformat nicht mehr erkennbar ist, sind urheberpersönlichkeitsrechtliche Interessen nicht betroffen.[29]

2. Identifikation geeigneter Textformate

a) Verfahren

Vor diesem Hintergrund wurden im Workshop verbreitete Analyseverfahren der Digital Humanities daraufhin untersucht, ob sie auch auf Textkorpora angewendet werden können, deren Informationsgehalt erheblich reduziert wurde, um die urheberrechtlichen Vorgaben zu erfüllen. Aus Sicht der Anwendungswissenschaften lassen sich zwei große Gruppen von Analyseverfahren unterscheiden.[30] Der ersten Gruppe zuzuordnen sind Analyseverfahren, die auf eine präzise Sequenzinformation der einzelnen Wortformen im Textverlauf angewiesen sind (insb. die meisten Verfahren aus der Sentiment Analyse, einige Verfahren der Netzwerkanalyse, Verfahren des Text Re-Use sowie das Erstellen von Sprachmodellen).[31] Zur zweiten Gruppe gehören Analyseverfahren, die eine solche präzise Sequenzinformation nicht erfordern. Diese Verfahren können auf Worthäufigkeiten operieren, profitieren aber stark davon, wenn sich diese Häufigkeiten auf kleinere Segmente innerhalb der Texte beziehen sowie wenn die Information über die Reihenfolge der Segmente im Gesamttext erhalten bleibt (dazu gehören Verfahren der Autorschaftsattribution, die Extraktion distinktiver Merkmale und das Topic Modeling sowie ggf. einige Verfahren der Netzwerkanalyse).

b) Geeignete abgeleitete Textformate

Vereinfachend lassen sich Texte als eine Zeichenfolge auffassen, die man in verschiedene Teilabschnitte gliedern kann: Token (bspw. einzelne Wörter), Sätze, Segmente und Gesamttext. Die Token wiederum können mit explizierenden Informationen angereichert werden (Wortform, Lemma, Wortart, semantischer Gehalt, Relationen, Sequenzinformation, Häufigkeit). Abgeleitete Textformate entstehen dadurch, dass eine gezielte Kombination von Informationsanreicherung und Informationsreduktion gewählt wird.[32] Es lassen sich zwei Grundansätze unterscheiden:

28 *Grisse*, RuZ 2020, 143, 156 ff. sowie zur Anwendbarkeit von §§ 60 c und 60 d UrhG.
29 *Grisse*, RuZ 2020, 143, 158 f.
30 Dazu ausführlich *Schöch* et al, RuZ 2020, 160, 163, 169.
31 Die einzelnen Verfahren werden vorgestellt von: *Schöch* et al, ZfdG 2020 (im Erscheinen).
32 Dazu ausführlich *Schöch* et al, RuZ 2020, 160, 161.

aa) Token-basierte Textformate

Token-basierte, abgeleitete Textformate gehen von den einzelnen Wörtern als Grundeinheit aus und basieren darüber hinaus in der Regel auf dem vollständigen Einzeltext als Analyseeinheit. Im einfachsten Fall enthält ein solches Format lediglich die Information über die Häufigkeit der verschiedenen Wortformen in jedem einzelnen Text als Ganzes, was einer Entfernung jeglicher Sequenzinformation entspricht. Solche Formate sind im Regelfall urheberrechtlich unbedenklich.[33]

Etwas reichhaltigere Textformate beruhen auf der beschriebenen Anreicherung der Tokens um sprachlich relevante Informationen (wie Lemma, Wortart, Syntax, Semantik); auf der Segmentierung der Texte in kleinere Einheiten, sodass die Häufigkeitsinformationen sich feiner aufgliedern; und auf der gezielten Entfernung von Informationen wie der Wortform oder des Lemmas unter Beibehaltung bspw. der Wortart-Information. Je nach spezifischem Textformat und den verwendeten Parametern ergeben sich hier unterschiedliche Einschätzungen bezüglich der Nützlichkeit für die Forschung in den Digital Humanities einerseits, bezüglich der urheberrechtlichen Beurteilung andererseits. Aus rechtlicher Perspektive kann das Löschen hinreichend vieler Stopp- oder Funktionswörter (zB Artikeln, Konjunktionen, Pronomen, Präpositionen, Modal- und Hilfsverben)[34], ggf. kombiniert mit der Entfernung von Eigennamen ein gangbarer Weg sein.[35] Vielversprechender – sowohl aus Sicht der Anwender als auch des Urheberrechts – sind aber Formate mit gestörten Sequenzinformationen.[36] Hierzu werden die Wörter eines Textsegments, bildlich gesprochen, durcheinandergewirbelt und so die urheberrechtlich geschützte Struktur und ggf. ein fiktionaler Inhalt unkenntlich gemacht. Bei einer ausreichend großen Segmentlänge von etwa 50 bis 100 Wörtern ist der ursprüngliche Text kaum noch rekonstruierbar, aber für eine Reihe von Analysemethoden nutzbar.

bb) Textformate auf (Teil-)Korpusebene

Eine andere Gruppe von Textformaten verwendet nicht die vollständigen Einzeltexte als Bezugsgröße, sondern größere Bestände von Texten. Diese werden aber so in Teilbestände gegliedert (bspw. nach Jahr und/oder Textsorte), dass dennoch analytisch gestützte Aussagen über die stilistischen Eigenschaften oder inhaltlichen Muster der Textgruppen als Ganzes gemacht werden können. Hierzu gehören Textformate, die statt der Häufigkeit einzelner Tokens die Häufigkeiten von Wortsequenzen einer voreingestellten Länge von N-Wörtern (sog. N-Gramme)[37] verfügbar machen: auf Einzeltext-Ebene wäre hier die Rekonstruierbarkeit trivial; wenn die N-Gramm-Häufigkeiten sich aber auf eine größere Textmenge beziehen und seltene N-Gramme gelöscht

33 *Grisse*, RuZ 2020, 143, 152.
34 Hierzu *Schöch et al*, RuZ 2020, 160, 167.
35 *Grisse*, RuZ 2020, 143, 152.
36 Dazu aus Anwendersicht: *Schöch et al*, RuZ 2020, 160, 165 ff. und aus urheberrechtlicher Perspektive *Grisse*, RuZ 2020, 143, 153.
37 Dazu mit Beispielen: *Schöch et al*, RuZ 2020, 160, 169 ff.

werden, ist das vermutlich nicht mehr der Fall. Soweit die Ursprungstexte nicht rekonstruierbar sind, sind solche Formate urheberrechtlich unbedenklich, solange keine zu großen N-Gramme gewählt werden.[38] 5-Gramme etwa, also Textsegmente mit der Länge von 5 Wörtern, bilden bei den meisten Textgattungen keine urheberschutzfähigen Textteile. Sollten sehr ungewöhnliche Formulierungen ausnahmsweise doch schutzfähig sein, so würden diese wegen ihrer Seltenheit gelöscht werden. Sollte auch dieser Filter versagen, etwa weil das Fragment oft zitiert wird, darf es wegen § 57 UrhG trotzdem in große Textkorpora aufgenommen werden.[39]

In diese Gruppe gehören zudem vektorbasierte Textformate, die auf der Repräsentation der Wörter in Form von Vektoren beruhen, die deren syntaktische und semantische Eigenschaften kodieren und damit für Analysen nutzbar machen.[40] Diese werden vor allem für die computergestützte Verarbeitung natürlicher Sprache verwendet (Natural Language Processing, NLP). Urheberrechtlich sind sie grundsätzlich unbedenklich, weil die Texte mangels Textgestalt für einen Menschen nicht mehr lesbar sind. Es muss aber sichergestellt sein, dass die ursprüngliche Textform nicht rekonstruiert werden kann.

IV. Fazit

Intelligent konstruierte, abgeleitete Textformate ermöglichen, große Korpora urheberrechtlich geschützter Texte zu erforschen und frei zugänglich zu machen. Damit gleichen sie die Schwächen anderer Lösungen aus, mit denen die Interessen von Urhebern und Forschern in Einklang gebracht werden. Die abgeleiteten Textformate sind allerdings nur deswegen urheberrechtsfrei, weil bei ihnen Wortgruppen, die Textstruktur bzw. andere Informationen gezielt entfernt wurden. Deswegen sind mit ihnen viele, aber nicht alle wissenschaftliche Analysemethoden möglich. Als strategische Ergänzung vorhandener Lösungsansätze können sie das literaturwissenschaftliche 20. und 21. Jahrhundert trotzdem digital aufhellen.

38 *Grisse*, RuZ 2020, 143, 154.
39 Dazu oben III.1.b)aa).
40 Dazu *Schöch et al*, RuZ 2020,160, 171 ff.