Nachnutzbar, aber nicht reproduzierbar. Abgeleitete Textformate für die Arbeit mit urheberrechtsbewehrten Texten

Philippe Genêt [1], Lukas Weimer [2]

[1] Deutsche Nationalbibliothek, Frankfurt am Main
[2] Niedersächsische Staats- und Universitätsbibliothek Göttingen

Der vorliegende Bericht behandelt die Herausforderungen und Lösungsansätze im Umgang mit urheberrechtlich geschützten Texten in der wissenschaftlichen Forschung. Trotz der Einführung der Text und Data Mining-Schranke (§60d UrhG) im Jahr 2018 bleiben zahlreiche Einschränkungen bestehen, die die freie Nutzung dieser Ressourcen durch Forschende verhindern. Bibliotheken, die den gesetzlichen Auftrag haben, ihre Sammlungen der Allgemeinheit zugänglich zu machen, engagieren sich daher u. a. in Text+, einem Konsortium der Nationalen Forschungsdateninfrastruktur.

Text+ zielt darauf ab, sprach- und textbasierte Forschungsdaten dauerhaft zu bewahren und für die Forschung verfügbar zu machen. Ein zentraler Ansatz dabei ist die Nutzung abgeleiteter Textformate (ATFs), die mittels Methoden des Text- und Data Minings erstellt werden. Mit ATFs können Forschungsfragen beantworten werden, ohne die Rechte der Urheberrechtsinhaber zu beeinträchtigen. Die Herstellung von ATFs erfolgt durch Anreicherungen wie Part-of-Speech-Tagging und Lemmatisierung und durch anschließende gezielte Informationsreduktion, die eine Rekonstruktion des Ausgangstextes verhindert.

In Text+ arbeiten verschiedene Universitäten und außeruniversitäre Einrichtungen an der Standardisierung und rechtlichen Klärung von ATFs. Zudem werden Tools und Textkorpora bereitgestellt, die nur in Form von ATFs veröffentlicht werden dürfen. Das Projekt fördert auch die Forschung zur Eignung und Rekonstruierbarkeit von ATFs und betreibt Disseminationsmaßnahmen, um das Thema einer breiteren Öffentlichkeit zugänglich zu machen. Die DNB untersucht zudem im Projekt CORAL die Möglichkeit, Sprachmodelle mit ATFs zu trainieren, ohne Originaltexte reproduzieren zu können. Durch diese Aktivitäten möchten die beteiligten Bibliotheken ihre Relevanz für die Forschungsgemeinschaft unterstreichen und sich weiter als wertvolle Partner der Wissenschaft etablieren.

Publiziert in: Vincent Heuveline, Philipp Kling, Florian Heuschkel, Sophie G. Habinger und Cora F. Krömer (Hrsg.): E-Science-Tage 2025. Research Data Management: Challenges in a Changing World. Heidelberg: heiBOOKS, 2025. DOI: https://doi.org/10.11588/heibooks.1652.c23936 (CC BY-SA 4.0).

Keywords: Abgeleitete Textformate, Urheberrecht, NFDI, Text+

1 Herausforderung Urheberrecht

Forschung mit urheberrechtsbewehrten Texten unterliegt zahlreichen Einschränkungen. Daran ändert auch die Einführung der sogenannten Text und Data Mining-Schranke (§60d UrhG[1]) im Jahre 2018 nur wenig (Iacino, Kamocki und Leinen 2023). Zwar erlaubt sie die Analyse urheberrechtlich geschützten Materials mit maschinellen Methoden, wie sie zum Beispiel die Computational Literary Studies bzw. die Digital Humanities anwenden, zu wissenschaftlichen Zwecken grundsätzlich – von einer freien Nutzung dieser Ressourcen durch Forschende kann jedoch keine Rede sein. Vielmehr sind Zusammenstellung, öffentliche Zugänglichmachung und Archivierung von Forschungsdaten weiterhin an zahlreiche Bedingungen geknüpft. Der lange Wirksamkeitszeitraum des Urheberrechts – es greift bis 70 Jahre nach dem Tod der Autorin bzw. des Autors – bewirkt, dass sich etwa die digitale Literaturwissenschaft sehr intensiv mit Werken aus dem 19. Jahrhundert beschäftigt, kaum aber mit neueren Texten.

Bibliotheken haben ein genuines Interesse und häufig auch den gesetzlichen Auftrag[2], ihre Sammlungen für die Allgemeinheit nutzbar zu machen. Daraus erwächst der Wunsch, der Wissenschaft Zugang auch zu Werken zu ermöglichen, die dem Urheberrecht unterstehen. Ein Ausdruck dieses Wunsches ist das Engagement mehrerer Bibliotheken in Text+[3], einem der 26 Fach- und Methodenkonsortien der Nationalen Forschungsdateninfrastruktur (NFDI[4]).

Die Mission von Text+ ist es, sprach- und textbasierte Forschungsdaten dauerhaft zu bewahren und für die Forschung verfügbar zu machen. Text+ nutzt dazu ein Netzwerk aus Daten- und Kompetenzzentren, die ihre Ressourcen und Erfahrung zur Verfügung stellen. Die Daten und Dienste können zentral abgerufen und verarbeitet werden, wie beispielsweise über die Text+ Registry[5] oder die Federated Content Search[6]. Viele der zur Verfügung gestellten Ressourcen können frei genutzt werden, einige jedoch sind aufgrund von rechtlichen Beschränkungen nur begrenzt zugänglich. Sehr umfangreiche und überwiegend urheberrechtlich geschützte Bestände bringen hier insbesondere die Deutsche Nationalbibliothek (DNB) und die Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB Göttingen) ein.

1 https://www.gesetze-im-internet.de/urhg/___60d.html; *Besucht am 26. März 2025.*
2 Vgl. zum Beispiel das Gesetz über die Deutsche Nationalbibliothek, insbesondere §2 Abs. 1, https://www.gesetze-im-internet.de/dnbg/BJNR133800006.html; *Besucht am 26. März 2025.*
3 https://text-plus.org/; *Besucht am 26. März 2025.*
4 https://www.nfdi.de/; *Besucht am 26. März 2025.*
5 https://registry.text-plus.org/; *Besucht am 26. März 2025.*
6 https://fcs.text-plus.org/; *Besucht am 26. März 2025.*

2 Die Antwort: abgeleitete Textformate (ATF)

Mit dem Ziel, auch solche Werke für die Wissenschaft nutzbar zu machen, rücken in der gemeinsamen Arbeit in Text+ abgeleitete Textformate (ATF, siehe Schöch u. a. 2020) in den Fokus. Sie können – unter Verwendung von Methoden des Text- und Data Minings – so hergestellt werden, dass einerseits das Ergebnis noch die Beantwortung mindestens einer Forschungsfrage ermöglicht, der verbleibende Rest andererseits aber die Rechte der Urheberrechtsinhaber nicht mehr beeinträchtigt. Voraussetzung dafür ist u. a., dass keine urheberrechtlich geschützten Elemente übernommen werden und dass keine triviale Möglichkeit zur Rekonstruktion des Ausgangstextes besteht. Solche ATFs können daher frei veröffentlicht werden. Bei der Erzeugung von mehr als einem ATF von einem Dokument bzw. einem Korpus ist darauf zu achten, dass auch durch deren Kombination keine Rekonstruktion möglich ist. Nicht alle ATFs sind automatisch rechtefrei. Ist eine Rekonstruktion des Originaltextes ohne größeren Aufwand möglich, fällt das ATF weiterhin unter das Urheberrecht (vgl. Iacino u. a. 2024).

2.1 Herstellung von ATFs

ATFs werden auf der Basis des Ausgangstextes durch die Anwendung einer Reihe von Veränderungen erstellt. Hierzu gehören in einem ersten Schritt meist Anreicherungen wie Part-of-Speech-Tagging (POS), Lemmatisierung und/oder Verlinkungen von Named Entities zu Normdaten in Form von Annotationen. Ebenso sind statistische Analysen auf dem Ausgangstext als Vorverarbeitungsschritte sinnvoll, diese sind ebenfalls als Annotationen zu sehen. Danach erfolgt eine gezielte Informationsreduktion, die einerseits auf einer Reihe von Veränderungen basiert, die typischerweise automatisiert ablaufen, und ganz wesentlich auf der Entscheidung, welche Granularität diesen Operationen zu Grunde liegen.

Vier Operationen stehen für die Informationsreduktion zur Verfügung: Löschen, Behalten, Ersetzen und Vertauschen. Diese können auf verschiedenen Granularitätsebenen greifen (z. B. auf Token-, Satz- oder Absatzebene) und in Bezug auf unterschiedliche Größen (z. B. pro Dokument, pro Werk oder pro Korpus). Granularität und Bezugsgrößen sind für die Wiederherstellbarkeit des Ausgangstextes entscheidend. Ist beispielsweise die Wortreihenfolge in einem kurzen Satz randomisiert, ist die Reproduktion des ursprünglichen Wortlauts recht einfach. Ist die Reihenfolge der Wörter hingegen auf einer ganzen Seite oder gar in einem ganzen Dokument vertauscht, ist es – auch mit maschinellen Mitteln – in der Regel unmöglich, den exakten Originalwortlaut zu rekonstruieren.

Unterschieden werden vor allem tokenbasierte und vektorbasierte ATFs. Bei Letzteren werden einzelne Wörter durch hochdimensionale Vektoren mit numerischen Einträgen ersetzt. So entstehen zum Beispiel Word Embeddings, die unter anderem beim Training

von Sprachmodellen zum Einsatz kommen. Im Folgenden wenden wir uns jedoch lediglich tokenbasierten ATFs zu.

2.2 Beispiele für ATFs

Tabelle 1: Beispiel für bag of words, hier: Vergleich der Häufigkeit von 15 Tokens in drei Romanen von Charles Dickens.[7]

	Dickens_Barnaby	**Dickens_Cricket**	**Dickens_Dombey**
,	26044	3192	37907
the	12027	1443	15884
‘	10303	41	15587
and	10267	1173	13548
.	9629	1351	13442
to	6274	761	9314
of	6378	698	9147
a	5470	689	7125
in	4396	526	6164
his	4151	337	5067
I	3313	596	4841
that	3179	365	5038
he	3242	273	3896
was	2888	392	4088
it	2497	442	3836

Ein prominentes, vielseitig einsetzbares und darum häufig verwendetes ATF ist die einfache Dokument-Term-Matrix, auch bag of words genannt. Dabei handelt es sich um eine Liste der im Einzeltext vorkommenden Tokens und deren absoluter Häufigkeit. Sie kann erweitert werden, indem zum Beispiel die Wortform, die Wortart oder das Lemma der Tokens angegeben werden. Diese einfache Frequenzliste dient zahlreichen analytischen Zwecken, etwa der Klassifikation oder dem Clustering von Texten, und reicht sogar für die Autorschaftsattribution aus.

Ein ebenfalls sehr bekanntes Beispiel für ATFs sind n-Gramme, seit sie von Google LLC im großen Stil veröffentlicht wurden[8]. Ein n-Gramm ist eine Folge von n aufeinanderfolgenden Tokens, die üblicherweise als alphabetische oder nach Häufigkeit sortierte Listen dargestellt werden. Als Sonderform gilt die Ranking-Liste, bei der statt der Frequenz

7 https://github.com/dh-trier/tmr/blob/gh-pages/target/en/three/frq-lower.csv; *Besucht am 26. März 2025.*

8 https://storage.googleapis.com/books/ngrams/books/datasetsv3.html; *Besucht am 26. März 2025.*

nur der Rang in der nach Frequenz sortierten Liste angegeben wird. Mit diesem Format kann zum Beispiel Stilometrie betrieben oder der grammatische Wandel über Zeiträume hinweg betrachtet werden. Zudem werden sie in der automatischen Spracherkennung verwendet.

Tabelle 2: Beispiel für n-Gramme, hier: Trigramme aus einem Korpus von fünf Romanen Theodor Fontanes.[9]

n-Gramm	**Häufigkeit**
gott sei dank	43
ja gnädigste frau	17
auch heute wieder	13
doch auch wieder	11
ist doch auch	11
ist immer so	10
gnädigste frau ist	10
war so war	10
nein gnädigste frau	10
wird ja wohl	9
ist doch immer	9
das ist recht	9

Ein drittes Verfahren operiert mit der selektiven Maskierung oder Löschung von Tokens. Löscht man beispielsweise den Sprechtext aus Dramen, so ist das Ergebnis nicht mehr urheberrechtlich geschützt, dient aber noch verschiedenen wissenschaftlichen Zwecken wie etwa der Netzwerkanalyse.

3 Aktivitäten in Text+ und darüber hinaus

In Text+ arbeiten die SUB Göttingen und die DNB gemeinsam mit Partner:innen aus der Forschung – insbesondere von der Universität Trier und dem Leibniz-Institut für Deutsche Sprache – an verschiedenen Aspekten der abgeleiteten Textformate. Mit der Entwicklung eines DIN-Standards soll die Herstellung und der Gebrauch von ATFs vereinheitlicht werden. Daneben wird in Text+ auch die komplexe rechtliche Lage beleuchtet – so erscheint zum Beispiel die Publikation „Legal status of Derived Text Formats“ (Iacino u. a. 2024) in der kommenden Ausgabe 3/2024 der Zeitschrift Recht und Zugang.

9 https://github.com/dh-trier/tmr/blob/gh-pages/target/de/fontane/ngr-wordform-3/allngrs.tsv; *Besucht am 26. März 2025.*

10 https://textgridrep.org/project/TGPR-adf9b705-1533-b0ef-491b-674878350ecb; *Besucht am 26. März 2025.*

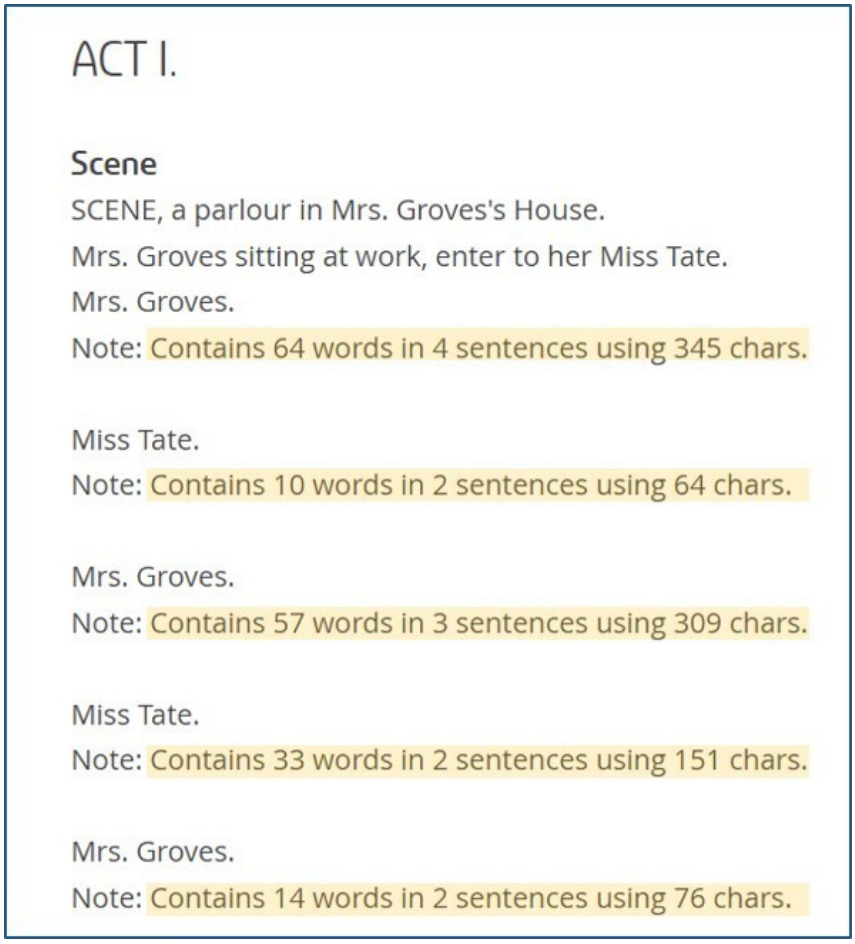

Abbildung 1: Beispiel für Maskierung/Löschung von Tokens, hier: Auszug aus dem American Drama Corpus im TextGrid Repository.[10]

Einige der von Text+ bereitgestellten Tools bieten sich für die Weiterverarbeitung von ATFs an, wie zum Beispiel die Natural Language Processing-Pipeline MONAPipe[11] der SUB Göttingen sowie das von der Universität Tübingen bereitgestellte Annotationstool WebLicht[12]. Gleichzeitig bietet Text+ über seine Datenzentren Textkorpora an, die nur in Form von ATFs veröffentlicht werden dürfen.

Um mehr Wissenschaftler:innen zu motivieren, sich mit dem Thema auseinanderzusetzen, fördert Text+ ebenfalls die einschlägige Forschung, zum Beispiel zur Eignung bestimmter ATFs für einzelne Forschungsfragen (Du und Schöch 2025; Du 2023) oder zur Rekonstruierbarkeit von Ausgangstexten (Kugler u. a. 2023). Für eine breitere Rezeption und Bekanntmachung des Themas dienen schließlich zahlreiche Disseminationsmaßnahmen wie Vorträge, Workshops und Informationsmaterial im Text+ Webportal[13].

11 https://pypi.org/project/monapipe/; *Besucht am 26. März 2025.*
12 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/Main_Page.html; *Besucht am 26. März 2025.*
13 https://text-plus.org/themen-dokumentation/atf/; *Besucht am 26. März 2025.*

Die DNB lotet daneben noch im Forschungsprojekt CORAL[14] die Möglichkeit aus, ob und wie leistungsfähige Sprachmodelle mit ATFs trainiert werden können, die das Wissen aus urheberrechtlich geschützten Quellen in sich tragen, ohne Gefahr zu laufen, Originaltexte reproduzieren zu können.

Mit dem Engagement möchten die in Text+ aktiven Bibliotheken ihre Relevanz für die Forschungscommunitys unterstreichen und ihnen noch bessere Partnerinnen werden.

Danksagungen

Die Autoren danken den Mitgliedern der AG Abgeleitete Textformate in der Task Area Collections in Text+: Florian Barth, José Calvo Tello, Keli Du, Jörg Knappen, Daniel Kurzawe, Peter Leinen, Piroska Lendvai, Christof Schöch, Thorsten Trippel und Andreas Witt.

Beiträge

- Philippe Genêt: Verschriftlichung – Originalentwurf
- Lukas Weimer: Verschriftlichung – Überprüfung & Überarbeitung

Mittelgeber

Die vorliegende Publikation wurde im Rahmen des Konsortiums Text+ im Kontext der Arbeit des Vereins Nationale Forschungsdateninfrastruktur (NFDI) e.V. verfasst. NFDI wird von der Bundesrepublik Deutschland und den 16 Bundesländern finanziert, und das Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projektnummer 460033370. Die Autoren bedanken sich für die Förderung sowie Unterstützung. Ein Dank geht außerdem an alle Einrichtungen und Akteur:innen, die sich für den Verein und dessen Ziele engagieren. Das diesem Bericht teilweise zugrunde liegende Vorhaben CORAL wird mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS24077 A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

14 https://coral-project.github.io/index.html; *Besucht am 26. März 2025.*

Interessenkonflikte

Es liegen keine Interessenkonflikte der Autoren vor.

Literaturverzeichnis

Du, Keli. 2023. „Understanding the impact of three derived text formats on authorship classification with Delta“. In *DHd Open Humanities Open Culture. 9. Tagung des Verbandes.* Zenodo. https://doi.org/10.5281/ZENODO.7715299.

Du, Keli und Christof Schöch. 2025. „Shifting Sentiments? What happens to BERT-based Sentiment Classification when derived text formats are used for fine-tuning“. In *DH-Conference 2024.* To appear.

Iacino, Gianna, Pawel Kamocki, Keli Du, Christof Schöch, Andreas Witt, Philippe Genêt und José Calvo Tello. 2024. „Legal status of Derived Text Formats“. *Recht und Zugang* 3 / 2024.

Iacino, Gianna, Pawel Kamocki und Peter Leinen. 2023. *Assessment of the Impact of the DSM-Directive on Text+.* Technischer Bericht. Zenodo. https://doi.org/10.5281/ZENODO.12759960.

Kugler, Kai, Simon Münker, Johannes Höhmann und Achim Rettinger. 2023. „InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline“. *Journal of Computational Literary Studies* 2 (1). https://doi.org/10.48694/JCLS.3572.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen“. *Zeitschrift für digitale Geisteswissenschaften* 5. https://doi.org/10.17175/2020_006.