Methodenbeitrag: Korpusbildung

Sandra Bläß

.

Thema:	Korpusbildung	DOI:	10.48694/fortext.3708
Jahrgang:	1	Ausgabe:	2
Erscheinungsdatum:	12-06-2024	Erstveröffentlichung:	2019-12-16 auf fortext.net
Lizenz:			open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition

Ein digitales Korpus ist eine maschinenlesbare Sammlung von Texten, die den Gegenstand Ihrer Untersuchungen im Feld digitaler Literaturwissenschaft bildet. Folglich konzipieren Sie es meist bereits mit einem Ziel oder einer Fragestellung. Je nach Methode oder Disziplin variieren die Textanzahl und nötigen Preprocessing-Maßnahmen (vgl. Preprocessing). Häufig werden Korpora jedoch mit Metadaten angereichert; vor allem in der Korpuslinguistik, wo das quantitative Auswerten (vgl. Distant Reading) von Korpora seinen Ursprung hat, werden Textsammlungen durch ausführliche grammatikalische (vgl. POS) Annotationen (vgl. Annotation) ergänzt.

2. Anwendungsbeispiel

Sie möchten Naturmotive in Dramen des Biedermeier und des Vormärz vergleichen, um zu sehen, wie diese die verschiedenen zeitgenössischen politischen Einstellungen und Lebensweisen verarbeiten. Hierfür planen Sie, digitale Methoden zu Hilfe zu nehmen. Sie haben sich bislang noch nicht festgelegt, welche konkreten Texte Sie untersuchen möchten, und müssen nun ein Korpus zusammenstellen.

3. Literaturwissenschaftliche Tradition

Zwar behandelt Literaturwissenschaft per se Textmaterialien, allerdings ist das Erforschen und damit auch das Ansammeln größerer Mengen von Primärtexten eher aus anderen Bereichen bekannt, beispielsweise aus Bibliotheken. Lauer zufolge liegen die Ursprünge der Digital Humanities in Traditionen der Philologie, in denen es, bevor man interpretierte, noch mehr um das Sammeln, Katalogisieren, Kontextualisieren und Ordnen gegangen sei (Lauer 2013, 101). Eco (1993) hingegen sieht 1977 das Ziel wissenschaftlicher Arbeiten als tiefergehende Deutung und befindet, der Fokus auf ein breites Themenspektrum sei „immer Ausdruck von Hochmut“ (ebd., 17). Je kleiner das Thema, desto geringer fielen die Lücken aus. Konträr hierzu spricht Moretti sich für Fragestellungen aus, die sich auf viele Texte beziehen, da die wissenschaftliche Einengung auf Einzeltexte des Kanons einen Großteil der existierenden Literatur außen vor lasse und damit umso größere blinde Flecken erzeuge (Moretti 2016, 46–50).

Ein weiteres Kriterium dafür, welche Texte in eine Untersuchung einbezogen werden, ist für Eco, ob umliegende Bibliotheken sie vorrätig haben. Um das herauszufinden, mussten zuerst mithilfe zweier Zettelkataloge, einem alphabetisch und einem inhaltlich sortierten, erste Bücher recherchiert werden, um aus deren Literaturlisten nach und nach eine Bibliografie erstellen zu können. Für die darin enthaltenen Bücher musste dann erneut im Zettelkatalog recherchiert werden, ob sie Teil des Bibliotheksbestands waren (Eco 1993, 65–67). Darauf konnte man sie entweder dem Freihandbereich entnehmen oder mit einem händisch ausgefüllten Leihzettel aus dem Magazin bestellen. Nicht vorrätige Bücher konnten überregional angefragt werden, was jedoch mehrere Wochen in Anspruch nehmen konnte (Heidtmann, Fertig und Ulrich 1979, 33–36).

Ein großer Unterschied zwischen der damaligen und heutigen Literaturbeschaffung ist also medialer Art. Hier gab es vor allem in den letzten Jahren einflussreiche Entwicklungen: So war die CD-ROM 1979 eine große Entdeckung für Veröffentlichungen von Textsammlungen, da sie als erster Speicherträger ein ganzes Buch fassen konnte. Das Grimmsche Wörterbuch auf CD-ROM war daher Ende des 20. Jahrhunderts noch eine Sensation, ähnlich die CD-ROM-Beilagen mit Textfaksimiles in der Colli/Montinari-Ausgabe des Nietzsche-Gesamtwerks, die inzwischen online zugänglich sind (Lauer 2013, 108). Aktuell wird die Edition 15 vom Projekt Gutenberg auf USB-Sticks verkauft, die das Gesamtkorpus des Projekts enthalten und damit ca. 10.000 Werke von etwa 2.000 Autor*innen.

4. Diskussion

Ihre Korpusbildung beginnt mit Fragestellungen, Thesen und Überlegungen zu dem Bereich, den Sie untersuchen wollen, um sich Gedanken über das dafür benötigte Untersuchungsmaterial machen zu können: Welche Daten möchten Sie untersuchen? Möchten Sie sich innerhalb einer Zeit, einer Gattung, einem Raum, des Werks bestimmter Autor*innen bewegen oder Vergleiche ziehen? Welche Methoden wenden Sie dafür an? Wie groß muss Ihr Korpus dazu sein (Kreuz 2018, 59)?

Diverse digitale Arbeitsweisen sind bei der wissenschaftlichen Textrecherche und -beschaffung bereits Standard: Zettelkästen wurden größtenteils durch digitale Kataloge (vgl. OPAC) ersetzt, Fernleihen und Vormerkungen können online vorgenommen werden, statt mit Fotokopien wird meist mit Scans gearbeitet und nicht bei jedem Text ist man darauf angewiesen, dass eine lokale Bibliothek ihn in Buchform vorliegen hat, da inzwischen viel digitalisiert und über Bibliotheksnetze abrufbar ist. Digitale Bibliothekskataloge und Textsammlungen (vgl. Korpus) erhöhen die Zugänglichkeit von Texten also bei geringerem Aufwand enorm. Bei der Textauswahl lohnt es auch, auf analoge Bibliografien (Flüh 2024) und Werkausgaben zurückzugreifen, die möglicherweise bereits viele Werke versammeln, die Sie für Ihre Arbeit nutzen können. Einige digitale Angebote wie Datenbanken entwickeln diese Unterstützung weiter, um an Recherchen und Sammlungen anderer anzuschließen (Lauer 2013, 105f.). Lauer stellt fest, dass die postulierte Zugänglichkeit für alle jedoch durch die in der Digitalisierung verstärkte Kommerzialisierung und Lizenzierung von Texten eingeschränkt werde (vgl. ebd., 113).

Viele Texte finden Sie digital in Onlinesammlungen oder als e-Book (vgl. Abbildung 1). Eine Lücke hierbei bilden jedoch Texte zwischen 1945 und etwa 1990, da sie noch nicht gemeinfrei sind, teilweise aber auch nicht gut genug verkäuflich, um jetzt noch als e-Book verlegt zu werden. Einige größere Bibliotheken wie beispielsweise die DNB stellen auch digitale Korpora für Untersuchungen zur Verfügung, die dann aber häufig nur vor Ort nutzbar sind; zudem gibt es einzelne Korpora über Onlineressourcen, beispielsweise beim DTA (Horstmann und Kern 2024). Möglicherweise sind manche der von Ihnen benötigten Texte auch (noch) gar nicht digital verfügbar. Wenn Sie sie trotzdem in Ihre Betrachtungen einbeziehen möchten, müssen Sie entweder eine Digitalisierung (Horstmann 2024b) durchführen oder beauftragen. Eine weitere Alternative wäre, beim Verlag anzufragen, ob dort eine digitale Textversion vorliegt und Sie für eine wissenschaftliche Nutzung Zugriff darauf erhalten können. Sowohl bei selbst erstellten Digitalisaten als auch bei Onlinequellen sollten Sie auf gute Textqualität achten, da beispielsweise Kontaminierung („noise“) von nicht korrigierten OCR-Scans erhebliche Verfälschungen Ihrer Untersuchungsergebnisse bedeuten kann (Eder 2013). Einige Argumente, welche Texte Sie in Ihr Korpus aufnehmen, können sich folglich auf finanzielle oder zeitliche Ressourcen beziehen, die bei Texten, die digital nicht frei (vgl. Open Access) oder gar nicht verfügbar sind, eine Rolle spielen können. Woher Ihre Textdigitalisate stammen, ist ein wichtiger Aspekt der Dokumentation Ihres Vorgehens, weswegen Sie während der Recherche darauf achten sollten, die jeweiligen Textquellen zu notieren.

Abb. 1: Die Digitalisierung von immer mehr Büchern ermöglicht Ihnen eine leichtere Zugänglichkeit von vielen Orten aus sowie verschiedene Formen der digitalen wissenschaftlichen Weiterverwendung.

Häufig wird als Hindernis der digitalen Korpusbildung das Urheberrecht gesehen (*Hinweis*: Die folgenden Erklärungen beziehen sich auf das Urheberrecht in Deutschland, Stand Januar 2020). Texte sind ab siebzig Jahren nach dem Tod der Urheber*innen gemeinfrei, sodass Sie neuere Werke nur eingeschränkt digitalisieren dürfen:

In der Lehre können Sie bis zu 15% eines Werks vervielfältigen und innerhalb des Rahmens der Lehrveranstaltung weitergeben. In der Forschung dürfen Sie bis zu 75% eines Werks vervielfältigen, zudem können Sie es einem Forschungskreis zu 15% zugänglich machen. Beides gilt nur für längere Texte, kürzere von bis zu 25 Seiten können komplett verwendet werden. Diese Einschränkungen sind nicht gegeben, wenn Sie (und, in der Lehre, Ihre Studierenden) mit analogen Texten arbeiten. 2018 gab es in Form des Urheberrechts-Wissensgesellschafts-Gesetzes Anpassungen zugunsten wissenschaftlicher Nutzung: Für Text Mining sowie Data Mining ist es erlaubt, größere Teile des Textes zu vervielfältigen, um ihn dafür verwenden zu können. Problematisch ist jedoch noch immer die weitere Nutzbarmachung von erstellten Korpora, denn für Forschungs- und Lehrprojekte wird vorausgesetzt, dass Sie das Korpus nach Beendigung der Tätigkeit löschen oder bei einer Archivinstitution ablegen lassen, was die Wiederverwendung von erstellten Korpora durch andere erschwert. Weiterhin können Sie sich bei den Rechteinhaber*innen erkundigen, ob Sie die Texte auch über die genannten Regelungen hinaus nutzen dürfen. Die Deutsche Forschungsgemeinschaft (2013) empfiehlt allerdings, solche Anfragen möglichst früh durchzuführen, da schwer abzusehen ist, wie lang solche Prozesse dauern werden, und damit Sie potentielle Lizenzgebühren in Ihre Projektplanung einbeziehen können.

Schöch (2017, 223) erklärt, dass digitale Literaturwissenschaften vorwiegend mit Korpora im Sinne größerer Datensammlungen arbeiten als mit nur wenigen Einzeltexten. In dieser Hinsicht kann die Vorbereitung eines Korpus für eine digitale Weiterverarbeitung mehr Zeit in Anspruch nehmen, als wenn Sie zum Beispiel nur zwei Romane miteinander vergleichen würden: Schließlich wählen Sie dann mehr Texte aus (für weitere Informationen zur Zusammenstellung Ihres Korpus siehe Punkt 5), recherchieren sie, digitalisieren sie möglicherweise selbst und lassen sie ein Preprocessing durchlaufen. Andererseits eröffnen sich weitere Untersuchungsperspektiven, -gegenstände und -methoden, wenn Sie mit digitalisierten Texten arbeiten. Darunter fallen auch Techniken, die analog bereits gängig sind, in Ihrer digitalen Variante aber viele Vorteile haben, beispielsweise manuelle Annotation (Jacke 2024) oder das Durchsuchen (vgl. Query) Ihrer Notizen.

5. Technische Grundlagen

Grundsätzlich sind die Kriterien, nach denen ein linguistisches Korpus zusammengestellt wird, genauer ausdifferenziert als die für ein literaturwissenschaftliches Korpus. Da es sich aber bei beiden um Sprachkorpora handelt, schließt Schahparonjan, dass die in der Linguistik angewandten Maßstäbe auch hier sinnvoll genutzt (vgl. Domäneadaption) werden können: Repräsentativität, Ausgewogenheit, Vergleichbarkeit und Größe. Ihr zufolge ist ein Korpus repräsentativ, wenn es stellvertretend für den gesamten Sprachbereich gelten kann, den man mit ihm untersucht, und ausgewogen, wenn es Subphänomene relational angemessen einbezieht (Schahparonjan 2012, 131f.), also beispielsweise verschiedene Genres so stark im Korpus vertreten sind, wie sie in einem untersuchten Zeitraum vorgekommen sind. Als Maßstab dafür, welche Texte, Gattungen und Autor*innen zu einer bestimmten Zeit als repräsentativ gesehen werden können, wird in vielen Studien geprüft, was dann bei Leser*innen und Kritiker*innen jeweils gerade populär war. Das Korpus ist vergleichbar, wenn es mit anderen Korpora mit einem ähnlichen Thema in Beziehung gesetzt werden kann, sich also beispielsweise zwei Korpora zu romantischen Erzählungen für ähnliche Fragestellungen verwenden lassen. Zur Größe gibt es keine generelle Faustregel, wie viele Texte und/oder Wörter ein Korpus fassen muss. Es braucht jedoch eine gewisse Größe, um statistische Aussagekraft zu haben und seltenere Merkmale überhaupt messen zu können (Schahparonjan 2012, 133). So zeigt beispielsweise Eder (2010), dass Korpora mit Einzeltexten, die jeweils weniger als 2500 Wörter umfassen, keine belastbaren Ergebnisse in stilometrischen Projekten (Horstmann 2024a) liefern, da diese vor allem auf statistischer Auswertung beruhen.

Je größer und diverser ein Korpus gestaltet ist, desto repräsentativer ist es tendenziell, zugleich funktioniert Stilometrie Rißler-Pipka zufolge am besten, wenn der Erscheinungszeitraum, die Gattung und der Umfang der Einzeltexte einheitlich ist (Rißler-Pipka 2018). Mit dieser Art der Korpuskonzeption umgeht sie das Problem, dass Texte verschiedener Gattungen in ihrem Vokabular teilweise unterschiedlich beschaffen sind und dass Wörter im Lauf der Zeit verschiedene Schreibweisen haben, weswegen sie bei Wortzählungen Daten verfälschen können. Letzteres lässt sich allerdings auch durch Normalisierungen oder Stoppwörter (vgl. Stoppwortliste) umgehen, die dokumentiert werden müssen (Aichner 2015). Manche Ressourcen bieten auch gesondert normalisierte Textdateien an. Modrow (2016, 177f.) wiederum entscheidet sich in ihrer Studie zur digitalen manuellen Annotation gegen ein repräsentatives Korpus und für ein kleineres, um ein komplexes Tagset anzuwenden, statt nur statistische Aussagen treffen zu können. Die Deutsche Forschungsgemeinschaft (2019) erklärt, dass Größe, Datenqualität und die Tiefe der Erschließung gegeneinander abgewogen werden müssen und in jedem Fall eine detaillierte Begründung zur Korpusbildung von zentraler Bedeutung ist, vor allem wenn es für weitere Forschung über das eigene Projekt hinaus nutzbar gemacht werden soll. Wie Sie Ihr Korpus letztlich konzipieren, hängt folglich stark davon ab, welche Methode(n) Sie darauf anwenden wollen: Während statistisch arbeitende und Machine Learning basierte (vgl. Machine Learning) Techniken bei größeren Korpora aussagekräftigere Ergebnisse erzielen, lassen sich manuell und nah am Text arbeitende Methoden besser mit kleineren Korpora umsetzen.

Gängige Textformate für viele digitale Tools sind das Reintextformat (vgl. Reintext-Version) TXT oder XML. Insbesondere XML wird häufig empfohlen, da darin Text und Metadaten voneinander getrennt werden, die durch den TEI-XML-Standard (vgl. TEI) für andere leichter interpretier- sowie weiterverwendbar sind, und weil es sich gut archivieren und in andere Formate, u. a. HTML, konvertieren lässt (siehe Schöch (2017, 227) zur Umwandlung von Texten in XML sowie deren Annotation Percillier (2017)). Darüber hinaus sollten Sie die Texte in UTF-8 (vgl. Unicode/UTF-8) codieren, da darin auch Zeichen fremdsprachiger Texte oder Textanteile korrekt angezeigt werden, anstatt Sie in Fragezeichen, Kästchen oder Sonderzeichen umzuwandeln (Schahparonjan 2012, 140). Diese Prozesse können Sie häufig in gängigen Textverarbeitungsprogrammen mit Benutzeroberflächen (vgl. GUI) umsetzen, sodass Sie sich zur Erstellung digitaler Korpora nicht unbedingt komplexes technisches Vorwissen aneignen müssen. Wenn Ihre Texte aus verschiedenen Quellen stammen, liegen Sie möglicherweise in verschiedenen Formaten und mit unterschiedlichen Metadaten vor. Idealerweise vereinheitlichen Sie Ihre Daten in diesem Fall (Schöch 2017, 227f.).

Weitere sinnvolle Bearbeitungsmöglichkeiten Ihrer Daten liegen im Löschen oder Markieren von Zeilen, die nicht zum eigentlichen Text gehören, in der Anpassung von nicht-druckbaren Zeichen und Sonderzeichen sowie im Zusammenfügen von in der Silbentrennung aufgespaltenen Wörtern (Schahparonjan 2012, 140f.). Darüber hinaus gibt es Möglichkeiten zur Textverarbeitung und -auswertung, die vermehrt in der Linguistik verwendet werden, sich aber auch für einige literaturwissenschaftliche Anwendungsbereiche nutzen lassen, u. a. Lemmatisierung (vgl. Lemmatisieren), der TreeTagger, das NLTK, Wordsmith, AntConc oder das LDA-Toolkit.

Externe und weiterführende Links

- Projekt Gutenberg: https://web.archive.org/save/https://www.projekt-gutenberg.org/ (Letzter Zugriff: 04.06.2024)
- Urheberrechts-Wissensgesellschafts-Gesetzes: https://web.archive.org/save/https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/BGBl/Bgbl_UrhDaG.pdf?__blob=publicationFile&v=3 (Letzter Zugriff: 04.06.2024)
- LDA-Toolkit: https://web.archive.org/save/https://diskurslinguistik.net/forschung/software/lda-toolkit/ (Letzter Zugriff: 04.06.2024)
- TreeTagger: https://web.archive.org/save/https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Letzter Zugriff: 04.06.2024)
- DNB: https://web.archive.org/save/https://www.dnb.de/DE/Home/home_node.html (Letzter Zugriff: 04.06.2024)
- AntConc: https://web.archive.org/save/https://www.laurenceanthony.net/software/antconc/ (Letzter Zugriff: 04.06.2024)
- NLTK: https://web.archive.org/save/https://www.nltk.org/ (Letzter Zugriff: 04.06.2024)
- Wordsmith: https://web.archive.org/save/https://www.lexically.net/wordsmith/ (Letzter Zugriff: 04.06.2024)

Bibliographie

Aichner, Christof. 2015. Die Korrespondenz von Leo von Thun-Hohenstein: Eine Dokumentation. *thun*. https://thun-korrespondenz.acdh.oeaw.ac.at/pages/index.html (zugegriffen: 7. Januar 2020).

Bundesministerium für Bildung und Forschung. 2019. Was ist in Lehre und Forschung gesetzlich erlaubt? *Urheberrecht in der Wissensschaft*. https://www.bildung-forschung.digital/de/was-ist-in-lehre-und-forschung-gesetzlich-erlaubt-2652.html (zugegriffen: 7. Januar 2020).

Deutsche Forschungsgemeinschaft. 2013. *Handreichung: Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf (zugegriffen: 9. Januar 2020).

———. 2019. *Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (zugegriffen: 7. Januar 2020).

Eco, Umberto. 1993. *Wie man eine wissenschaftliche Abschlußarbeit schreibt. Doktor-, Diplom- und Magisterarbeit in den Geistes- und Sozialwissenschaften*. Heidelberg: Müller.

Eder, M. 2013. Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing* 28, Nr. 4 (1. Dezember): 603–614. doi: 10.1093/llc/fqt039, https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqt039 (zugegriffen: 5. Dezember 2019).

Eder, Maciej. 2010. Does Size Matter? Authorship Attribution, Small Samples, Big Problem. In: *Digital Humanities 2010. Conference Abstracts*. doi: 10.1093/llc/fqt066, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.385.2063&rep=rep1&type=pdf (zugegriffen: 7. Januar 2020).

Flüh, Marie. 2024. Methodenbeitrag: Digitales Bibliografieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3786, https://fortext.net/routinen/methoden/digitales-bibliografieren.

Heidtmann, Frank, Eymar Fertig und Paul S. Ulrich. 1979. *Wie finde ich Literatur zur deutschen Literatur*. Berlin: Berlin Verlag.

Hirschmann, Hagen. 2019. *Korpuslinguistik. Eine Einführung*. Stuttgart: Metzler.

Horstmann, Jan. 2024a. Methodenbeitrag: Stilometrie. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 1. Stilometrie (26. Februar). doi: 10.48694/fortext.3769, https://fortext.net/routinen/methoden/stilometrie.

———. 2024b. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

Horstmann, Jan und Alexandra Kern. 2024. Ressourcenbeitrag: Deutsches Textarchiv (DTA). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

Ivanovic, Christine. 2017. Die Vernetzung des Textes: Im Möglichkeitsraum digitaler Literaturanalyse. *Zeitschrift für digitale Geisteswissenschaften*. doi: 10.17175/2016_010,.

Jacke, Janina. 2024. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Kreuz, Christian D. 2018. *Das Konzept „Schuld" im Ersten Weltkrieg und in der Weimarer Republik. Linguistische Untersuchungen zu einem brisanten Thema*. Bremen: Hempen.

Lauer, Gerhard. 2013. Die digitale Vermessung der Kultur. Geisteswissenschaften als Digital Humanities. In: *Big Data. Das neue Versprechen der Allwissenheit*, hg. von Heinrich Geiselberger und Tobias Moorstedt, 99–116. Berlin: Suhrkamp.

Lautenschläger, Sina. 2016. *Geschlechtsspezifische Körper- und Rollenbilder*. Berlin, Boston: de Gruyter.

Lemnitzer, Lothar und Heike Zinsmeister. 2015. *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Modrow, Lena. 2016. *Wie Songs erzählen. Eine computergestützte, intermediale Analyse der Narrativität*. Frankfurt am Main: Peter Lang.

Moretti, Franco. 2016. *Distant Reading*. Konstanz: Konstanz University Press.

Percillier, Michael. 2017. Creating and Analyzing Literary Corpora. In: *Data Analytics in Digital Humanities*, hg. von Shalin Hai-Jew, 91–118. Multimedia Systems and Applications. Cham: Springer.

Rißler-Pipka, Nanette. 2018. Die Digitalisierung des goldenen Zeitalters - Editionsproblematik und stilometrische Autorschaftsattribution am Beispiel des Quijote. *Zeitschrift für digitale Geisteswissenschaften* 4, Nr. 3. doi: 10.17175/2018_004,.

Schahparonjan, Anna. 2012. *Stilistische Untersuchungen an den Werken von Lion Feuchtwanger und Thomas Mann. Eine korpuslinguistische Studie*. Hamburg: Kovač.

Schöch, Christof. 2017. Aufbau von Datensammlungen. In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 223–233. Stuttgart: Metzler.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd" + „space", geben „Terminal" ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R", geben „cmd.exe" ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values* . Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv` , sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch

Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

Domäneadaption Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

OPAC OPAC steht für *Online Public Access Catalogue* und bezeichnet online zugängliche Bibliothekskataloge.

Open Access Open Access bezeichnet den freien Zugang zu wissenschaftlicher Literatur und anderen Materialien im Internet.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

POS PoS steht für *Part of Speech* , oder „Wortart" auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt" oder „vorbereitet" werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Query *Query* bedeutet „Abfrage" oder „Frage" und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

Type/Token Das Begriffspaar „Type/Token" wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär." beinhaltet beispielsweise fünf Worttoken („Ein", „Bär", „ist", „ein", „Bär") und drei Types, nämlich: „ein", „Bär", „ist". Allerdings könnten auch vier Types, „Ein", „ein", „Bär" und „ist", als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Unicode/UTF-8 Unicode ist ein internationaler Standard, der für jedes Schriftzeichen oder Textelement einen digitalen Code festlegt. Dabei ist UTF-8 die am weitesten verbreitete Kodierung für Unicode-Zeichen. UTF-8 ist die international standardisierte Kodierungsform elektronischer Zeichen und kann von den meisten Digital-Humanities-Tools verarbeitet werden.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.