Methodenbeitrag: Manuelle Annotation

Janina Jacke [1]

1. Christian-Albrechts-Universität zu Kiel

Thema:	Manuelle Annotation	DOI:	10.48694/for-text.3748
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-01-28 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

Unter (digitalem) *manuellem Annotieren* (vgl. Annotation) versteht man die Praxis, in Texten digital Hervorhebungen oder Anmerkungen (vgl. Metadaten) anzubringen. Diese können ganz unterschiedlichen Zwecken dienen – beispielsweise der Strukturierung von Texten, ihrer sprachlichen oder inhaltlichen Beschreibung, ihrer Kontextualisierung oder Interpretation.
Für taxonomiebasiertes Annotieren werden sogenannte *Tagsets* (hierarchisch organisierte Kategoriensysteme) verwendet. Digitale Annotationen werden in Form von *Markup* (vgl. Markup Language) gespeichert. Im Vergleich zum nicht-digitalen Annotieren ergeben sich dadurch zusätzliche Möglichkeiten der Systematisierung und Auswertung der Annotationen. Im Gegensatz zu (teil-)automatisierten Annotationen werden manuelle digitale Annotationen vom Forschenden jeweils einzeln und gezielt im Text angebracht.

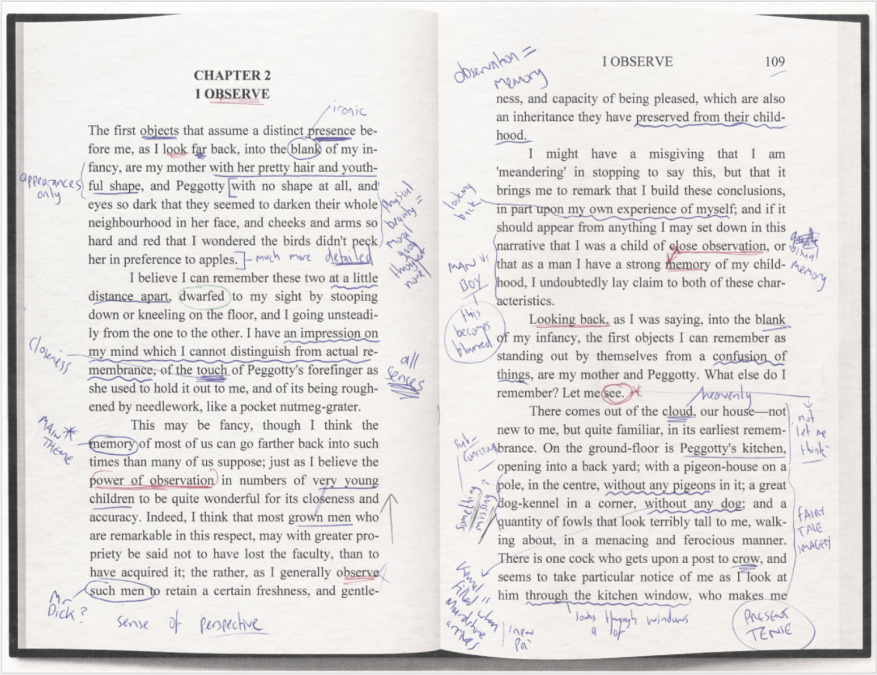
2. Anwendungsbeispiel

Nehmen wir an, Sie möchten die Rolle der Frau im Werk Max Frischs (vgl. Korpus) untersuchen. Sie planen, im ersten Schritt in *Homo faber* die Stellen zu markieren, in denen Frauenfiguren auftreten oder die Haltung des Protagonisten zu Frauen deutlich wird. Zusätzlich zu den Markierungen wollen Sie an den relevanten Stellen auch konkrete Ideen, Interpretationsansätze oder Verweise auf textexterne Informationen (z. B. auf feministische Theorien oder Frischs Biografie) notieren. Wenn Sie diese Vorhaben mit einem geeigneten Annotationsprogramm digital durchführen, können Sie beispielsweise, je nach Bedarf,

- die Annotationen übersichtlich anzeigen lassen,
- sie leicht in thematische Gruppen sortieren,
- digital gestützt eine Annotationstaxonomie (vgl. Tagset) für systematischere Analysen entwickeln oder
- abfragen (vgl. Query), wo, wie oft und in welchen Kombinationen bestimmte annotierte Phänomene auftreten.

3. Literaturwissenschaftliche Tradition

Annotationen gehören schon seit Jahrhunderten zu den textwissenschaftlichen Kernpraktiken (Moulin 2010). Im Gegensatz zum *Kommentar* (vgl. Kommentar), mit dem Annotation konzeptuell einige Überschneidungen aufweist, ist der Begriff der Annotation noch wenig theoretisiert. Textkommentare dienen meist der Erläuterung oder Interpretation literarischer Texte. Sie können entweder selbst Textform annehmen oder den Charakter von Anmerkungen haben, die auf einzelne Textstellen bezogen sind und in Form von Marginalien oder Glossen in den Text geschrieben werden (Oellers 2000, 302; Häfner 2000, 298 und 301). In dieser zweiten Form lassen sich Kommentare auch als „Annotationen“ bezeichnen.
In der literaturwissenschaftlichen Forschungsdebatte werden Kommentare meist im Zusammenhang mit der Editionsphilologie diskutiert. Wie Oellers (2000) deutlich macht, werden dort unterschiedliche Standards für Kommentare in kritischen Ausgaben diskutiert: „Das Spektrum der behandelten Themen reicht vom Plädoyer für unkommentierte Editionen über die Untersuchung einzelner Aspekte (Anordnung, Inhalt, Umfang, Darstellungsweise) bis zu Überlegungen über die interpretatorische Funktion von Kommentaren“ (ebd. 303).
Wir wollen die Praxis der Annotation hier weiter und permissiver verstehen: als Arbeitsschritt, der Forschungsvorhaben ganz unterschiedlicher Natur unterstützen kann. In dieser individuell auszubuchstabierenden Funktion sind Annotationen weder formal noch inhaltlich reglementiert (siehe Abb. 1).



CHAPTER 2
I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentle-

ness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon my own experience of myself; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong memory of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me

Abb. 1: Analoge Annotationen (Kehoe und Gee 2013, 108–109)

Beispielhaft lassen sich drei Varianten von Annotationen nennen, die in der literatur- und sprachwissenschaftlichen Tradition verankert sind. Sie unterscheiden sich sowohl in formaler als auch in funktional-inhaltlicher Hinsicht.

1. Annotationen können in Form von (nicht systematisierten) *Freitext-Kommentaren* am Text angebracht werden. Im Rahmen literaturwissenschaftlicher Forschungsvorhaben sind solche freien Kommentare besonders geeignet, um den Text beispielsweise mit Ideen, Assoziationen und kontextuellem Wissen anzureichern, um auf dieser Basis später unsere *Interpretationshypothesen* zu formulieren (Jannidis, Kohle und Rehbein 2017, 253).
2. Im Rahmen *taxonomiebasierter Annotation* werden einzelne Textstellen bestimmten literatur- oder sprachwissenschaftlichen (Analyse-)Kategorien zugeordnet. Diese Annotationsvariante kann besonders für *Textbeschreibungen* im Rahmen von Form- oder Inhaltsanalyse verwendet werden – zum Beispiel zur Analyse von Metrik, Reimschemata und rhetorischen Mitteln bei der Lyrikanalyse oder bei narratologischen Untersuchungen erzählender Texte.
3. Im Rahmen der *Textauszeichnung* (vgl. Markup (Textauszeichung)) dienen Annotationen lediglich dazu, Texte grafisch zu strukturieren bzw. zu organisieren (Jannidis 2003, 601) – dem Text werden gewöhnlich keine Kommentare oder Schlagworte beigefügt, die das Textverständnis befördern sollen. So wird bei der *Edition* mithilfe bestimmter Auszeichnungssprachen (siehe Abschnitt 5: Technische Grundlagen) dafür gesorgt, dass ein Text Überschriften und Absätze enthält oder dass einzelne Worte gefettet oder kursiviert erscheinen. Das Äquivalent hierzu im Rahmen individueller Textforschung sind beispielsweise Unterstreichungen.

Die Übergänge zwischen diesen exemplarischen Annotationsvarianten sind oft fließend und eine Kombination ist problemlos möglich.

4. Diskussion

Gegenüber ‚analogen' Formen der Annotation hat digitales manuelles Annotieren einige Vorteile. Generell bietet eine digitale Umgebung viele Möglichkeiten, sich Annotationen *übersichtlich darstellen* zu lassen (vgl. Textvisualisierung (Horstmann und Stange 2024)). Dies ist insbesondere dann wichtig, wenn sich in einem Text viele und viele verschiedene Arten von Annotationen finden.
Darüber hinaus ist es möglich, Annotationen digital unterstützt und teil-automatisiert zu *untersuchen und auszuwerten*. Dies kann mithilfe von Abfragen, sogenannten Queries (vgl. Query), geschehen (Horstmann 2024b). Auf diese Weise ist es zum Beispiel möglich, unsystematische Annotationen zu systematisieren und die Texterforschung dadurch zielgerichteter und prägnanter werden zu lassen. Jannidis, Kohle und Rehbein (2017) weisen ebenfalls auf diesen Vorteil hin, den sie den Tendenzen zur Standardisierung im Zusammenhang mit Annotationen zurechnen. Ihnen zufolge können Annotationen „der Abstrahierung von Einzelvorkommnissen

und deren Ordnung und Zusammenfassung in übergreifenden Klassen dienen“ (ebd. 254).
Liegen bereits systematische, taxonomiebasierte Annotationen vor, können mittels Queries teil-automatisiert statistische Auswertungen (vgl. Text Mining) vorgenommen werden – es lässt sich beispielsweise herausfinden, wie oft bestimmte Phänomene in einem Text auftreten, an welchen Textstellen eine besondere Häufung auszumachen ist oder wie oft bestimmte Phänomene miteinander korrelieren. Neben der erleichterten Auswertung der Annotationen, die durch digitales Arbeiten ermöglicht wird, bietet digitales Annotieren weitere Vorteile. Einer dieser Vorteile liegt in vereinfachten Möglichkeiten des *Zusammenarbeitens* mit anderen Forschenden (vgl. Kollaboratives literaturwissenschaftliches Annotieren (Jacke 2024)). Wenn für das Annotieren webbasierte (vgl. Webanwendung) Programme genutzt werden,

- können Arbeitsergebnisse oder Zwischenschritte anderen Mitarbeitenden leichter zur Verfügung gestellt werden;
- die Kollaborierenden können vorher festgelegte Annotationsaufgaben für einen gemeinsamen Text oder ein Korpus untereinander aufteilen und den Arbeitsfortschritt der anderen online verfolgen;
- oder es kann genuin kollaborativ gearbeitet werden – d. h. dass mehrere Mitarbeitende an den gleichen Annotationsaufgaben arbeiten und den Prozess gemeinsam vollziehen, indem fremde Annotationen eingesehen, kommentiert, ergänzt oder verändert werden können. Trotz eines derart verzahnten Arbeitsprozesses ermöglicht es die digitale Arbeitsumgebung, dass der persönliche Beitrag jedes/r einzelnen Mitarbeitenden zuordenbar bleibt.

Weitere Optionen des digitalen Annotierens werden vor allem im Rahmen strukturalistisch orientierter Forschungsparadigmen als Vorteile gesehen. Diese greifen vor allem dann, wenn taxonomiebasiert annotiert wird. Diese Arbeitsweise fördert nicht nur fokussiertes Close Reading, also ein sehr genaues und textnahes Arbeiten, sondern ist zugleich dazu geeignet, die genutzte Taxonomie auf die Probe zu stellen und ggf. die *literaturwissenschaftliche Theoriebildung* voranzutreiben: Wenn die zum Zweck der Textanalyse modellierten Kategorien bzw. Tags sich als nicht adäquat erweisen (z. B. weil sie unvollständig oder zu unspezifisch sind), muss das Kategoriensystem weiterentwickelt werden. Darüber hinaus ist digitales taxonomiebasiertes Annotieren besonders dafür geeignet, die eigenen Annotationsentscheidungen *kritisch zu prüfen*. Denn zum einen wird durch das textnahe Arbeiten und die Visualisierung der Annotationsentscheidungen konsistentes und regelbasiertes Vorgehen gefördert – zum anderen machen es viele Varianten der Kollaboration notwendig, die eigenen Entscheidungen vor anderen Forschenden begründen und rechtfertigen zu können.
Natürlich gibt es auch Spezifika des digitalen Annotierens, die als Nachteile empfunden werden können. Viele Literaturwissenschaftler*innen lesen schöne Literatur lieber als „echtes“ Buch statt an einem Bildschirm. Möchte man das Lesen nicht strikt von der Analyse und Interpretation trennen, werden auch beim Lesen schon erste Annotationen im Buch vorgenommen. Um diese Annotationen digital weiterzuführen und auszuwerten, müssten diese erst in eine digitale Textversion übertragen werden.
Ein weiterer Nachteil des digitalen Arbeitens besteht darin, dass viele Texte noch nicht in sorgfältig digitalisierter Form vorliegen. Wenn dies der Fall ist, muss zunächst selbst digitalisiert werden (Möglichkeiten der Textdigitalisierung (Horstmann 2024a)). Weiterführende Beiträge zur Diskussion der Praxis des digitalen Annotierens finden sich in Bauer und Zirker (2017).

5. Technische Grundlagen

Mit den richtigen Programmen erfordert digitales manuelles Annotieren mittlerweile kaum noch technisches Wissen: Oft sind die Benutzeroberflächen (vgl. GUI) der Annotationsprogramme selbsterklärend und es werden die Arbeitsabläufe des analogen Arbeitens simuliert (CATMA (Schumacher 2024); WebAnno (Schumacher und Bläß 2024)).
Interessant ist es aber vielleicht dennoch zu wissen, in welcher Weise digitale Annotationen vom Computer gespeichert werden und welche Konsequenzen dies beispielsweise für Möglichkeiten der Weiterverwendung Ihrer Annotationen im Rahmen anderer digitaler Umgebungen und Tools hat. Hierfür ist es sinnvoll, dass Annotationen in Form *standardisierter* (vgl. TEI) Auszeichnungssprachen (vgl. Markup Language) gespeichert werden, beispielsweise XML. Auf diese Weise ist die Chance am höchsten, dass die Daten nach dem Exportieren aus einem Annotationsprogramm für andere Programme lesbar sind, mit denen die Daten gegebenenfalls weiterverarbeitet werden sollen.
Eine weitere relevante Unterscheidung im Zusammenhang mit Speicherformaten von Annotationen ist die zwischen *inline* und *standoff Markup*. Im Falle von inline markup werden die Annotationen im Text selbst gespeichert. Dies ist nur dann sinnvoll, wenn es sich um nicht-interpretative, formale Annotationen handelt. Werden Annotationen dagegen als *standoff Markup* gespeichert, werden die Annotationsdaten außerhalb des Textes abgelegt – ihre Position im Text wird über eine Angabe der Zeichenpositionen vermerkt. Diese Markupvariante erlaubt mehrfache, überlappende und diskrepante Annotationen derselben Textstelle – und ist damit optimal für semantische, interpretative Annotationen geeignet (Piez 2010).

Bibliographie

Bauer, Matthias und Angelika Zirker. 2017. Explanatory Annotation in the Context of the Digital Humanities. *International Journal of Humanities and Arts Computing* 11, Nr. 2: 145–152. doi: 10.3366/ijhac.2017.0189,.

Häfner, Ralph. 2000. Kommentar1. In: *Reallexikon der deutschen Literaturwissenschaft*, hg. von Harald Fricke, 2: H-O:298–302. Berlin (u.a.): de Gruyter.

Horstmann, Jan. 2024a. 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.

———. 2024b. Lerneiheit: Analyse und Visualisierung mit CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3752, https://fortext.net/routinen/lerneinheiten/analyse-und-visualisierung-mit-catma.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Jacke, Janina. 2024. Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3749, https://fortext.net/routinen/methoden/kollaboratives-literaturwissenschaftliches-annotieren.

Jannidis, Fotis. 2003. Textauszeichnung. In: *Reallexikon der deutschen Literaturwissenschaft*, hg. von Jan-Dirk Müller, 3: P-Z:601–602. Berlin (u.a.): de Gruyter.

Jannidis, Fotis, Hubertus Kohle und Malte Rehbein. 2017. Manuelle und automatische Annotation. In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 253–267. Stuttgart: Metzler.

Kehoe, Andrew und Matt Gee. 2013. eMargin. A Collaborative Textual Annotation Tool. *Ariadne*, Nr. 71. http://www.ariadne.ac.uk/issue71/kehoe-gee (zugegriffen: 18. Dezember 2017).

Kindt, Tom und Hans-Harald Müller. 2015. Zum Verhältnis von Deskription und Interpretation. Ein Bestimmungsvorschlag und ein Beispiel. In: *Literatur interpretieren. Interdisziplinäre Beiträge zur Theorie und Praxis*, hg. von Jan Borkowski, Stefan Drescher, Felicitas Ferder, und Philipp David Heine, 73–90. Paderborn: Mentis.

Köppe, Tilmann und Tom Kindt. 2014. *Erzähltheorie. Eine Einführung*. Stuttgart: Reclam Verlag.

Meister, Jan Christoph. 2014. Narratology. In: *the living handbook of narratology*, hg. von Peter Hühn, Jan Christoph Meister, John Pier, und Wolf Schmid. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/narratology (zugegriffen: 24. November 2017).

Moulin, Claudine. 2010. Am Rande der Blätter. Gebrauchsspuren, Glossen und Annotationen in Handschriften und Büchern aus kulturhistorischer Perspektive. *Autorenbibliotheken, Quarto. Zeitschrift des Schweizerischen Literaturarchivs* 30/31: 19–26.

Oellers, Norbert. 2000. Kommentar2. In: *Reallexikon der deutschen Literaturwissenschaft*, hg. von Harald Fricke, 2: H-O:302–303. Berlin (u.a.): de Gruyter.

Piez, Wendell. 2010. Towards Hermeneutic Markup. An Architectural Outline. In: *Digital Humanities 2010. Conference Abstracts*, 202–205. London. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html (zugegriffen: 11. Dezember 2017).

Rudek, Christof. 2016. Rhetorische Lyrikanalyse. Formen und Funktionen von Klang- und Bildfiguren. In: *Handbuch Lyrik. Theorie, Analyse, Geschichte*, hg. von Dieter Lamping, 49–58. Stuttgart: Metzler.

Schumacher, Mareike. 2024. Toolbeitrag: CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3761, https://fortext.net/tools/tools/catma.

Schumacher, Mareike und Sandra Bläß. 2024. Toolbeitrag: WebAnno. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3764, https://fortext.net/tools/tools/webanno.

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.

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.

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.

Kommentar Textkommentare dienen meist der Erläuterung oder Interpretation literarischer Texte. Sie können entweder selbst Textform annehmen oder den Charakter von Anmerkungen haben. Treten sie in Form von Marginalien oder Glossen „in den Texten“ geschrieben auf, lassen sich auch Kommentare als Annotationen bezeichnen.

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.

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.

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.

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.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online" genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

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.