Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren

Janina Jacke [1]

1. Christian-Albrechts-Universität zu Kiel

Thema:	Manuelle Annotation	DOI:	10.48694/for-text.3749
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-04-04 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 *kollaborativem literaturwissenschaftlichem Annotieren* (vgl. Annotation) ist eine Praxis kooperativen Arbeitens zu verstehen, bei der sich mehrere Forschende gemeinsam der Annotation literarischer Texte annehmen. Während hierbei unterschiedliche Modi der Kooperation möglich sind, widmet sich der vorliegende Beitrag ausschließlich einer spezifischen Unterform des kollaborativen Annotierens: der gemeinsamen Arbeit an derselben Textgrundlage vor dem Hintergrund derselben Fragestellung. Diese Form der Kollaboration erfordert ein besonderes Maß an Koordination, um zu fruchtbaren Ergebnissen führen zu können. Dazu gehört die Erarbeitung von *Annotationsguidelines* (‚Richtlinien'), die über die genauen Inhalte und Regeln der Annotation Aufschluss geben. Sie enthalten beispielsweise Definitionen der für die Annotation verwendeten *Kategorien* bzw. *Tags* (vgl. Tagset) sowie Hinweise zu deren *Operationalisierung*. Diese dienen als gemeinsame Grundlage für alle Partizipierenden.

2. Anwendungsbeispiel

Angenommen, Sie möchten in einem kleinen, diachron angelegten Textkorpus (vgl. Korpus) exemplarisch untersuchen, wie sich der Einsatz von erlebter Rede in deutschsprachigen Kurzgeschichten vom 18. bis zum 20. Jahrhundert entwickelt hat. Da es häufig nicht trivial ist, erlebte Rede zu erkennen, und es in vielen Fällen eine Frage der Interpretation ist, ob sie vorliegt oder nicht, haben Sie sich dazu entschlossen, Ihre Untersuchung als kollaboratives Annotationsprojekt anzulegen. Sie möchten deswegen zusammen mit zwei Kolleginnen oder Kollegen erlebte Rede in Ihrem Textkorpus annotieren, um Ihre Analysen und Interpretationen abgleichen und diskutieren zu können, bevor Sie – ausgehend von diesen Analysen – Erkenntnisse hinsichtlich der historischen Entwicklung des untersuchten Phänomens formulieren und Erklärungshypothesen aufstellen.

3. Literaturwissenschaftliche Tradition

Zur Rolle des Annotierens in der literaturwissenschaftlichen Tradition siehe Jacke (2024).
Kollaboratives literaturwissenschaftliches Annotieren ist eine Form des kooperativen Arbeitens. Traditionell gelten die Geisteswissenschaften im Allgemeinen – und die Literaturwissenschaft im Besonderen – als akademische Felder, in denen Individualforschung vorherrscht: „[D]ie Identität der geisteswissenschaftlichen Fachgeschichten [wird] in erster Linie über große Individuen und deren Arbeitsleistungen gestiftet" (Röcke 2016, 7). Obwohl sich auch in der literaturwissenschaftlichen Fachgeschichte teilweise kollaborative Arbeitspraxen nachweisen lassen (Stockhorst, Lepper und Hoppe 2016a), muss deren Reichweite als stark eingeschränkt wahrgenommen werden: Kooperatives Arbeiten findet, wenn überhaupt, fast ausschließlich bei der Arbeit an Editionen, Handbüchern, Einführungen, Literaturgeschichten etc. statt (Schönert 1993).
Anders sieht es dagegen bei (einfacher) „eigenständig realisierbaren Arbeitsformen" aus, insbesondere bei Textanalyse und Interpretation (Stockhorst, Lepper und Hoppe 2016b, 9). Hier arbeiten Forschende nahezu immer allein. Betrifft literaturwissenschaftliche Kollaboration doch einmal Textanalyse oder Interpretation – beispielsweise im Zusammenhang mit literaturwissenschaftlichen „Beobachtungsleistungen" in Verbindung mit relevanten „Anschlusskognitionen", also beispielsweise Interpretationen (Klausnitzer 2016, 85) –, so kommt dies fast nur in universitären Lehrveranstaltungen vor. Kollaborative Textanalyse unter arrivierten Literaturwissenschaftler*innen im Rahmen von Forschungsvorhaben ist dagegen in der traditionellen Literaturwissenschaft eine extreme Ausnahme (Lange 2005, 280).

Da kollaborative Analyse- und Interpretationsarbeit an literarischen Texten in der traditionellen literaturwissenschaftlichen Praxis nicht verankert ist, existiert auch keine literaturwissenschaftliche Tradition, an die die Erstellung von Annotationsguidelines für kollaboratives digitales Annotieren direkt anknüpft. Da Annotationsguidelines standardisierte Grundlagen für (gemeinsam) Forschende darstellen und Konzeptdefinitionen sowie Operationalisierungshinweise enthalten, lassen sich hier jedoch gewisse Analogien zu literaturwissenschaftlichen Handbüchern und Einführungen in die Textanalyse ziehen: Handbücher wie das *Reallexikon der deutschen Literaturwissenschaft* (Fricke u. a. 2000–2007) enthalten Definitionen grundlegender literaturwissenschaftlicher Konzepte, die der Forschungsgemeinschaft als gemeinsamer Standard dienen können. Einführungen in die Textanalyse (Pfister 2001; Burdorf 2015; Lahn und Meister 2016) liefern ebenfalls Definitionen literaturwissenschaftlicher Kategorien und enthalten darüber hinaus oft Beispielanalysen, die als Hinweise für eine adäquate Operationalisierung dieser Kategorien dienen.

4. Diskussion

Kollaboratives literaturwissenschaftliches Annotieren in der hier vorgestellten Variante kann zum einen dem Zweck dienen, Annotationen literarischer Texte zu erstellen, die in gewisser Weise intersubjektiv und somit ‚abgesichert' sind. Dadurch, dass mehrere Forschende dieselbe Textgrundlage mit derselben Zielsetzung annotieren, wird ausgeschlossen, dass sich in den Annotationen eine idiosynkratische und womöglich unplausible bzw. ungerechtfertigte Perspektive auf den Text niederschlägt. Dies ist beispielsweise dann von Nutzen, wenn die Annotation des Textes mit dem Ziel durchgeführt wird, die manuell erstellten Annotationen zu automatisieren (vgl. Machine Learning), wie es in vielen Projekten der digitalen Literaturwissenschaft der Fall ist. Damit dies möglich ist, muss eine hohe Übereinstimmung zwischen menschlichen Annotatoren vorliegen.
Aber auch ohne ein solches übergeordnetes Ziel kann es sinnvoll sein, Texte kollaborativ zu annotieren. Denn solch ein Vorgehen legt offen, welche Fragen sich in Bezug auf einen literarischen Text intersubjektiv beantworten lassen und welche Fragen ‚Interpretationssache' sind. Auf diese Weise können zum einen interessante metatheoretische Erkenntnisse über die Interpretationsabhängigkeit literaturwissenschaftlicher Fragestellungen erzielt werden. Zum anderen kann Aufschlussreiches über die untersuchten literarischen Texte herausgefunden werden, zum Beispiel indem sich leicht die Textstellen identifizieren lassen, an denen sich die Mehrdeutigkeit oder Interpretationsoffenheit eines Textes besonders manifestiert.
Kollaboratives Annotieren birgt aber auch einige potenzielle Nachteile bzw. Schwierigkeiten. So handelt es sich beispielsweise um eine kooperative Arbeitsform, die ausgesprochen zeit-, arbeits- und personalintensiv ist. Insbesondere sind in hohem Maße Koordination und Absprachen erfordert. Forschende könnten sich auch dadurch in ihrer Freiheit und Kreativität eingeschränkt fühlen, dass die durch Annotationen festgehaltenen Interpretationen durch das kollaborative Arbeiten stärker auf dem Prüfstand stehen und einer Rechtfertigung bedürfen, die von den Mitarbeitenden als gültig akzeptiert werden muss (Schönert 1993, 399). Insgesamt gilt, dass bei kollaborativem Annotieren in literaturwissenschaftlichen Zusammenhängen einige Besonderheiten zu beachten sind, um zum einen die Vorteile dieser Arbeitsweise voll ausschöpfen zu können und zum anderen dem besonderen Untersuchungsgegenstand *Literatur* gerecht zu werden. Bisher kommt die Praxis des kollaborativen Annotierens deutlich häufiger in linguistischen Kontexten zur Anwendung als in literaturwissenschaftlichen. Daher haben auch die meisten systematischen Untersuchungen und Hinweise zur Erstellung von Annotationsguidelines und zu sinnvollen Arbeitsabläufen beim kollaborativen Annotieren sprachwissenschaftliche Hintergründe (Pustejovsky, Bunt und Zaenen 2017).
Obwohl einige wichtige Erkenntnisse aus diesen Kontexten auf das literaturwissenschaftliche Annotieren übertragen werden können, gibt es auch relevante Unterschiede. Zum einen dient kollaboratives Annotieren in linguistischen Kontexten fast immer der späteren Automatisierung der Annotationen (vgl. Annotation). Zum anderen spielt textuelle Ambiguität in literarischen Texten eine wichtigere Rolle als beispielsweise in Gebrauchstexten – und sie wird als besonderes Qualitätsmerkmal verstanden (Bauer u. a. 2010). Beides führt dazu, dass in linguistischen Zusammenhängen eine Einigkeit zwischen den partizipierenden Annotator*innen oft stärker forciert wird (Wissler u. a. 2014), als aus literaturwissenschaftlicher Perspektive wünschenswert wäre. Ein Vorteil eines solchen Ansatzes besteht allerdings darin, dass er die Anschließbarkeit computerlinguistischer Verfahren garantiert: Beim Training von Machine Learning-Modellen gilt ein so generiertes Korpus von Annotationen als *(+Ground Truth)*, also als besonders verlässliche Datengrundlage.
Die Erstellung von *Annotationsguidelines* (vgl. Annotationsguidelines) ist von diesen Unterschieden noch nicht so stark betroffen wie die Festlegung sinnvoller Arbeitsprozesse bei der Annotation im Team. Bei Annotationsguidelines handelt es sich um verschriftlichte Anweisungen, die bei der Annotation beachtet werden sollen und allen Partizipierenden als gemeinsame Grundlage dienen. Da bei kollaborativem Annotieren häufig taxonomiebasiert annotiert wird – also mithilfe von Kategoriensystemen bzw. sog. Tagsets –, bedeutet dies, dass die Guidelines normalerweise *Definitionen* der zu verwendenden Annotationskategorien bzw. Tags enthalten. (In diesem Zusammenhang kann es manchmal nützlich sein, auch auf literaturwissenschaftliche Theorietexte zu verweisen, an denen sich die in den Guidelines verwendeten Definitionen orientieren, und ggf. auch zu erläutern, warum diese Definitionen anderen verfügbaren Theorieangeboten vorgezogen wurden.)
Ist ein Tagset komplex und hierarchisch gegliedert, bietet es sich an, in die Guidelines Schaubilder zu integrieren,

die Ordnung und Relation der einzelnen Tags zueinander übersichtlich illustrieren. Wenn die Anwendung der verfügbaren Kategorien auf den Text einer bestimmten Reihenfolge folgen soll (beispielsweise weil manche Analysen/Annotationen auf anderen aufbauen), so muss dies auch in den Guidelines vermerkt sein.
Darüber hinaus ist es notwendig, dass die Guidelines detaillierte Hinweise zur *Operationalisierung* der Tags bzw. Kategorien enthalten. Denn zwischen einer abstrakten Definition einer literaturwissenschaftlichen Analysekategorie und der Frage, wann und wie diese Kategorie tatsächlich zur Anwendung kommt, klafft oft eine theoretisch-methodologische Lücke. Hilfreich sind beispielsweise Auskünfte darüber, wie lang die annotierten Passagen für eine Kategorie in der Regel sind (z.B. Wort, Teilsatz, Satz, längere Textpassagen; werden Satzzeichen am Ende mitannotiert?) und welche Indikatoren an der Textoberfläche darauf hindeuten, dass eine Annotationskategorie zur Anwendung kommen muss oder könnte. Des Weiteren sollten Guidelines auch möglichst für jede Annotationskategorie Beispiel-Passagen aus literarischen Texten enthalten, auf die die Kategorie zutrifft (Gius und Jacke 2016).
Eine wichtige Eigenschaft von Annotationsguidelines ist, dass ihre Erarbeitung ein iteratives Verfahren erfordert. Eine erste Version der Guidelines sollte zwar vor dem ersten Annotieren entworfen werden. Die Anwendung der Guidelines im Rahmen der Annotation fördert jedoch in der Regel Inkonsistenzen oder Unklarheiten in den Guidelines zutage, die behoben werden müssen. Danach können die modifizierten Guidelines dann verwendet werden, um die bisherigen Annotationen zu überarbeiten. Diese Schritte müssen in den meisten Fällen mehrfach vollzogen werden. Dieses iterative Verfahren ist auch in linguistischen Forschungsbeiträgen beschrieben worden (Pustejovsky, Bunt und Zaenen 2017, 24) - den *‚MAMA' cycle* (*Model, Annotate, Model, Annotate*) – jedoch sollten insbesondere im Zusammenhang mit dem Ablauf des kollaborativen Annotationsprozesses die zentralen Unterschiede zwischen linguistischer und literaturwissenschaftlicher Annotation beachtet werden. Um die Mehrdeutigkeit literarischer Texte im Prozess der kollaborativen Annotation und Guidelines-Optimierung adäquat zu berücksichtigen, hat sich folgender Annotationsablauf bewährt (Gius und Jacke 2017):

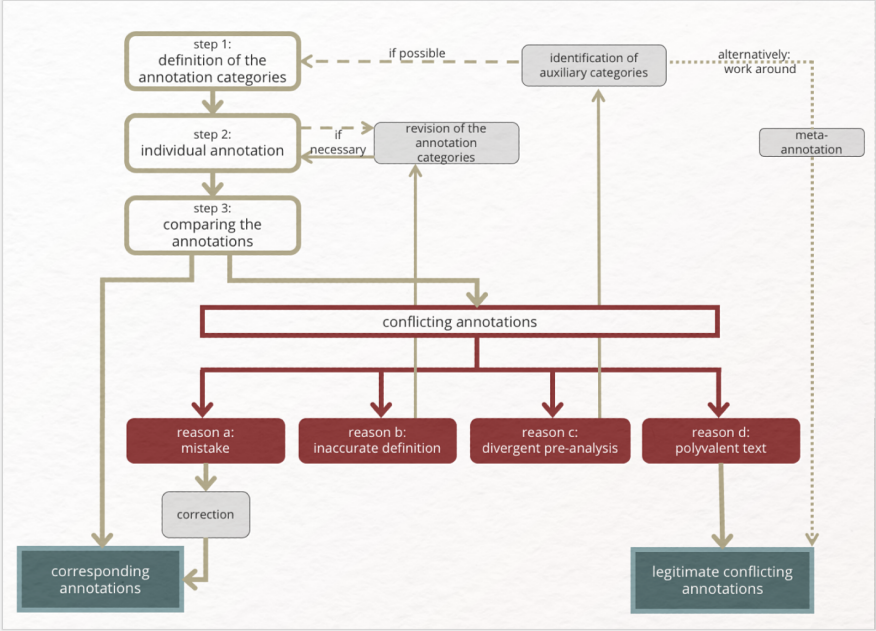

Abb. 1: Arbeitsablauf für kollaboratives literaturwissenschaftliches Annotieren

Im ersten Schritt werden die Guidelines den oben beschriebenen Kriterien folgend verfasst. Im nächsten Schritt wendet jede*r Annotierende diese Guidelines im Rahmen der Annotation desselben Textmaterials in einem individuellen Annotationsdurchgang an. Dieser Schritt hat zum Ziel, dass die Forschenden ihre eigenen Lesarten des Textes entwickeln können, ohne von vornherein von den anderen Mitarbeitenden beeinflusst zu werden. Sollte sich im Rahmen dieses Schritts schon herausstellen, dass die Guidelines überarbeitungsbedürftig sind, erfolgt eine entsprechende Modifikation, nach der dann die individuelle Annotation fortgesetzt werden kann.
Ist diese abgeschlossen, vergleichen die mitarbeitenden Literaturwissenschaftler*innen im dritten Schritt ihre Annotationen und diskutieren die Fälle, in denen unterschiedlich annotiert wurde. Je nachdem, welcher Grund für die fehlende Übereinstimmung herausgestellt wird, werden unterschiedliche Maßnahmen erforderlich: Handelt es sich um einen Verständnisfehler, wird die fragliche Annotation korrigiert. Ist eine Ungenauigkeit in den Guidelines verantwortlich, so müssen diese ein weiteres Mal überarbeitet werden.
Möglicherweise sind aber auch verschiedene theoretische Vorannahmen oder implizit durchgeführte Voranalysen des Textes durch die Annotator*innen der Grund für eine fehlende Übereinstimmung. In diesem Fall ist es entweder möglich, die unterschiedlichen Vorannahmen oder -analysen per Meta-Annotation (s.u.) zu

dokumentieren, um die Annotationsunterschiede nachvollziehbar zu machen und zu legitimieren. Alternativ ist auch eine Einigung hinsichtlich der Vorannahmen oder -analysen möglich, wofür jedoch ggf. zusätzliche Arbeitsschritte notwendig sind. Schließlich können Annotationsunterschiede auch durch eine mehrdeutige oder offene Textstelle zustande kommen. In diesem Fall können die unterschiedlichen Deutungen des Textes ebenfalls mithilfe von Meta-Annotationen vermerkt werden. Diese Art von Annotationsunterschieden ist im literaturwissenschaftlichen Kontext besonders interessant – und die Identifikation der relevanten Textstellen gehört zu den wichtigsten Vorzügen des kollaborativen Annotierens.

5. Technische Grundlagen

Wenn kollaborative Annotation mit geeigneten Annotationstools (bspw. CATMA) durchgeführt wird, benötigen die Forschenden keine technischen Kenntnisse. Folgende Funktionalitäten des verwendeten Tools sind jedoch notwendig bzw. besonders förderlich für kollaboratives Annotieren:

- Das verwendete Tool sollte webbasiert (vgl. Webanwendung) sein oder als Desktopanwendung zumindest mit einer Online-Datenbank synchronisiert werden. So können die Annotierenden leicht von unterschiedlichen Orten aus arbeiten und der spätere Vergleich der Annotationen ist dennoch problemlos möglich.
- Sinnvoll sind zudem Abfrage- (vgl. Query) oder Visualisierungsoptionen (Horstmann und Stange 2024), die den Vergleich von Annotationen vereinfachen.
- Die Möglichkeit von Meta-Annotationen (d.h. bspw. der Einfügung von Kommentaren zu Annotationen) erlaubt es den Annotator*innen, ihre individuellen Annotationsentscheidungen nachvollziehbar zu machen und so den Austausch über mögliche Annotationsunterschiede zu erleichtern.

Während das kollaborative Annotieren selbst keine technischen Kenntnisse erfordert, sieht dies natürlich anders aus, wenn die kollaborativ erstellten Annotationen genutzt werden sollen, um automatische Annotationen zu realisieren. In diesem Fall sind beispielsweise Kenntnisse über Machine-Learning-Verfahren vonnöten.

Externe und weiterführende Links

- CATMA: https://web.archive.org/save/http://catma.de/ (Letzter Zugriff: 03.07.2024)

Bibliographie

Bauer, Matthias, Joachim Knape, Peter Koch und Susanne Winkler. 2010. Dimension der Ambiguität. *Zeitschrift für Literaturwissenschaft und Linguistik* 40, Nr. 158: 7–75. doi: 10.1007/BF03379835,.

Burdorf, Dieter. 2015. *Einführung in die Gedichtanalyse*. 3. aktualisierte und erw. Aufl. Stuttgart: Metzler.

Fricke, Harald, Klaus Weimar, Klaus Grubmüller und Jan-Dirk Müller, Hrsg. 2000–2007. *Reallexikon der deutschen Literaturwissenschaft*. Berlin: de Gruyter.

Gius, Evelyn und Janina Jacke. 2016. Zur Annotation narratologischer Kategorien der Zeit. Guidelines zur Nutzung des CATMA-Tagsets. http://heureclea.de/wp-content/uploads/2016/11/guidelinesV2.pdf.

———. 2017. The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. *International Journal of Humanities and Arts Computing* 11, Nr. 2 (Oktober): 233–254. doi: 10.3366/ijhac.2017.0194, https://www.euppublishing.com/doi/10.3366/ijhac.2017.0194 (zugegriffen: 4. September 2020).

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

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

Klausnitzer, Ralf. 2016. Gemeinsam einsam frei? Beobachter und Beobachtungskollektive an der modernen Universität. In: *Symphilologie. Formen der Kooperation in den Geisteswissenschaften*, hg. von Vinzenz Hoppe, Marcel Lepper, und Stefanie Stockhorst, 73–99. Göttingen: V & R unipress.

Lahn, Silke und Jan Christoph Meister. 2016. *Einführung in die Erzähltextanalyse*. 3., aktualisierte und erweiterte Auflage. Lehrbuch. Stuttgart: Metzler.

Lange, Tanja. 2005. Vernetzte Wissenschaft? Zu Perspektiven computergestützter Kollaboration für Forschung und Lehre in den Geisteswissenschaften. In: *Digitalität und Literalität. Zur Zukunft der Literatur*, hg. von Harro Segeberg und Simone Winko, 271–294. München: Fink.

Pfister, Manfred. 2001. *Das Drama. Theorie und Analyse*. München: Fink.

Pustejovsky, James, Harry Bunt und Annie Zaenen. 2017. Designing Annotation Schemes. From Theory to Model. In: *Handbook of Linguistic Annotation*, hg. von Nancy Ide und James Pustejovsky, 21–72. Dordrecht: Springer.

Röcke, Werner. 2016. Geleitwort. In: *Symphilologie. Formen der Kooperation in den Geisteswissenschaften*, hg. von Stefanie Stockhorst, Marcel Lepper, und Vinzenz Hoppe, 7. Göttingen: V & R unipress.

Schönert, Jörg. 1993. Konstellationen und Perspektiven kooperativer Forschung. In: *Geist, Geld und Wissenschaft*, hg. von Peter J. Brenner, 384–408. Frankfurt am Main: Suhrkamp.

Stockhorst, Stefanie, Marcel Lepper und Vinzent Hoppe, Hrsg. 2016a. *Symphilologie. Formen der Kooperation in den Geisteswissenschaften*. Göttingen: V & R unipress.

———. 2016b. Vom Nutzen und Nachteil der Kooperation für die Philologien. Ein Problemaufriss. In: *Symphilologie. Formen der Kooperation in den Geisteswissenschaften*, hg. von Stefanie Stockhorst, Marcel Lepper, und Vinzenz Hoppe, 9–23. Göttingen: V & R unipress.

Wissler, Lars, Mohammed Almashraee, Dagmar Monett und Adrian Paschke. 2014. The Gold Standard in Corpus Annotation. In: 26. Juni. doi: 10.13140/2.1.4316.3523,.

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.

Annotationsguidelines Annotationsguidelines sind verschriftlichte, projektspezifische Anweisungen, die bei der Annotation beachtet werden sollen und bei kollaborativen Projekten als gemeinsame Grundlage für alle Annotierenden dienen. Taxonomiebasierte Annotationsprojekte enthalten klassischerweise Definitionen der zu verwendenden Annotationskategorien (Tags).

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.

Ground Truth Beim überwachten, maschinellen Lernen bezieht sich der Begriff Ground Truth auf die Genauigkeit der Klassifizierung des Trainingssatzes und wird durch direkte Beobachtung der Daten erhoben. Ein Datensatz, der als Ground Truth bezeichnet wird, ist meist eine manuelle Annotation, die individuell sein darf und Fehler enthalten kann.
Der Goldstandard , im Vergleich, versucht die „Ground Truth“ so genau wie möglich, das heißt ohne Fehler und mit überindividueller Gültigkeit, darzustellen. Ground Truth und Gold Standard werden oft als Synonyme verwendet.

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.

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.

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.

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.

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.