Ressourcenbeitrag: Tagset Narratologie (histoire)

Janina Jacke [1]

1. Christian-Albrechts-Universität zu Kiel

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/for-text.3757
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2020-05-04 auf fortext.net
Lizenz:	CC BY SA		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. Kurzbeschreibung

Das Tagset „Narratologie (histoire)“ ist für die Annotation (Jacke 2024a) narrativer Elemente in Texten geeignet. Es enthält grundlegende Kategorien für die Analyse der Geschichte (*histoire*, d. h. dem Inhalt der Erzählung), konkret für die Figuren- und Handlungsanalyse. Das Tagset erhalten Sie auf Zenodo (forTEXT 2020b).

2. Anwendungsbeispiel

Angenommen, Sie haben mithilfe des Tagsets „Narratologie (discours)“ (forTEXT 2020a; Jacke 2024b) bereits anhand eines kleinen Korpus deutschsprachiger Novellen des 18. Jahrhunderts im Rahmen einer ersten explorativen Studie untersucht, wie sich die Erzählweise in kürzeren narrativen Texten im Laufe des Jahrhunderts verändert hat. Möglicherweise interessiert Sie nun darüber hinaus, ob sich ebenfalls Entwicklungen hinsichtlich der Figurentypen, von denen deutschsprachige Novellen des 18. Jahrhunderts erzählen, feststellen lassen. Sie können hierfür die im Tagset Narratologie (histoire) enthaltenen Annotationskategorien für die Figurenanalyse nutzen und bei Bedarf erweitern. Mithilfe geeigneter Abfragen (vgl. Query) können Sie dann beispielsweise auch untersuchen, ob möglicherweise festgestellte Entwicklungen auf *discours*- und *histoire*-Ebene der Erzählungen miteinander korrelieren.

3. Diskussion

Wie auch im Beitrag zum Tagset „Narratologie (discours)“ (Jacke 2024b) beschrieben, handelt es sich bei der Narratologie um eine geisteswissenschaftliche Disziplin, die vornehmlich Modelle für die Analyse von Erzählungen entwickelt. Hervorgegangen aus dem russischen Formalismus und dem französischen Strukturalismus, stellen vor allem viele der früheren Arbeiten beschreibende Analysekategorien zur Verfügung, deren Anwendung weitgehend ohne Rückgriff auf textexternes Wissen oder Interpretation auskommt.
Während die von Genette (2007) im Rahmen seines strukturalistischen Ansatzes entwickelten Kategorien für die Analyse der Erzählweise immer noch häufig zur Anwendung kommen, finden detaillierte formalistisch-strukturalistische Modelle für die Analyse der Geschichte heute kaum mehr direkte praktische Anwendung. Starken Einfluss auf die theoretische Weiterentwicklung von Analysemodellen auf diesem Gebiet hat allerdings Propp (1975) *Morphologie des Märchens*, in der er eine Taxonomie typischer Figuren und Handlungselemente für die Gattung des russischen Zaubermärchens entwickelt. Viele der heute gebräuchlichen Arbeiten zur Analyse von Figuren und Handlung in erzählender Literatur integrieren neben strukturalistischen Elementen beispielsweise auch rezeptionstheoretische oder kognitionswissenschaftliche Erkenntnisse (Margolin 1983; Margolin 1990). Der Grund hierfür liegt darin, dass Leser*innen für die inhaltliche Analyse einer Erzählung stärker auf textexternes Wissen (u. a. auf psychologisches und Genrewissen) zurückgreifen müssen. Die Figuren- und Handlungsanalyse nähert sich dadurch etwas stärker einer (inhaltsspezifizierenden) Textinterpretation an als die Analyse der Form bzw. der Präsentationsweise (Jacke 2014).
Die im Tagset Narratologie (histoire) verwendeten Kategorien für die Figurenanalyse gehen auf Aspekte der Modelle von Hansen (2000) und von Jannidis (2012), die für die Handlungsanalyse auf de Toro (1986) zurück. Die Operationalisierung der Kategorien in Form eines Tagsets für die manuelle Annotation basiert auf den Arbeiten von Gius (2015) und Modrow (2016).
Ein Teil der im Tagset enthaltenen Kategorien für die Figurenanalyse – insbesondere diejenigen für die Annotation von Figurenreferenz – können sinnvoll mit weiteren Verfahren der digitalen Literaturanalyse kombiniert

werden. So kann beispielsweise für das automatische Auffinden bestimmter Formen der direkten Figurenreferenz Named Entity Recognition (Schumacher 2024c) verwendet werden. Sind alle Figurenreferenzen annotiert worden, können die Beziehungen zwischen Figuren sodann auch durch das Verfahren der Netzwerkanalyse (Schumacher 2024b) unterstützt werden.

4. Tagset

Das Tagset ist auf Zenodo als XML-Datei verfügbar und kann in geeignete Tools (beispielsweise CATMA (Schumacher 2024a)) importiert und dort verwendet werden. Abbildung 1 zeigt die im Tagset enthaltenen Tags in ihrer hierarchischen Struktur sowie die Properties (vgl. Property) und Values.

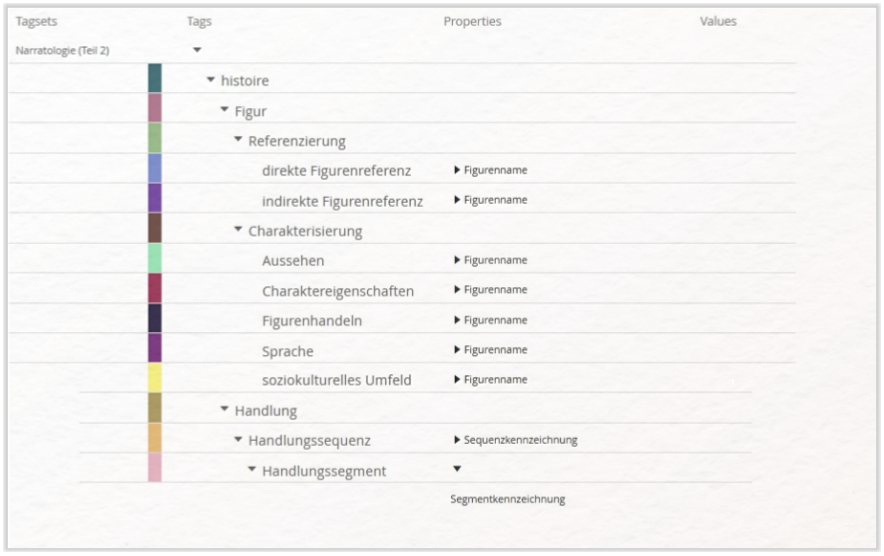

Abb. 1: Tagset Narratologie (histoire)

5. Richtlinien zur Anwendung

Diese Richtlinien enthalten nur spezifische Anwendungshinweise für die Tags, die speziell für die Anwendung bei der Annotation vorgesehen sind. Im Falle der Tags für die Figurenanalyse sind das ausschließlich die Tags auf der untersten Hierarchieebene. Bei der Handlungsanalyse sind sowohl *Handlungssequenz* als auch *Handlungssegment* für die Annotation geeignet, obwohl erstere Kategorie auf einer höheren Hierarchieebene angeordnet ist. Hierarchisch höherliegende Kategorien dienen dagegen vor allem der Systematisierung.
Im Folgenden werden die einzelnen Kategorien kurz definiert – für speziell zur Annotation vorgesehene Tags werden darüber hinaus Hinweise zur Länge der annotierten Passage und zu textuellen Indikatoren angegeben sowie i. d. R. ein Beispiel.

histoire: Die hier versammelten Kategorien dienen der Analyse des Inhalts von Erzählungen. Sofern Bedarf besteht, können den Annotationskategorien jeweils noch weitere Unterkategorien zur genaueren Analyse hinzugefügt werden

5.1 Figur

Figur: Dieses Untertagset ist für die Analyse der Akteure in Erzählungen vorgesehen. Mit den im Folgenden vorgestellten Tags können Textstellen annotiert werden, in denen auf unterschiedliche Weise auf Figuren Bezug genommen wird oder Figuren näher ausgestaltet werden. Um welche Figur es in einer annotierten Passage jeweils geht, kann mithilfe der Property *Figurenname* festgehalten werden. Hier wird bei jeder Annotation als Propertywert der Name der relevanten Figur (oder alternativ eine andere aussagekräftige Kennzeichnung der Figur) eingefügt.

Referenzierung: Diese Kategorien für die Figurenanalyse dienen der Analyse der Bezugnahme auf eine Figur im Text.

- Tag *direkte Figurenreferenz*: Eine direkte Figurenreferenz liegt vor, wenn mithilfe eines Ausdrucks explizit auf eine Figur Bezug genommen wird, beispielsweise mithilfe eines Eigennamens, einer Kennzeichnung oder eines Personalpronomens.

  - Länge der annotierten Passage: i. d. R. ein Wort oder eine Wortfolge.
  - Indikatoren: Eigennamen für Personen, Personalpronomina
  - Beispiele: „*Matteo* wurde krank, die bösartigsten Blattern befielen *ihn*, und *er* mußte viel leiden“ (F. Hebbel: Matteo; Property *Figurenname* für jede Annotation mit dem Wert „Matteo“); „die ungewohnte Pracht, die *das Mädchen* umgab“ (F. Hebbel: Matteo; Property *Figurenname* mit dem Wert „Felicita“)
- Tag *indirekte Figurenreferenz*: Eine direkte Figurenreferenz liegt dann vor, wenn bspw. Anwesenheit oder Handlung einer Figur durch das Erzählte impliziert ist, also wenn beispielsweise eine Handlung (mithilfe einer Passivkonstruktion) wiedergegeben wird oder im Fall nicht-eingeleiteter Figurenrede.
  - Länge der annotierten Passage: i. d. R. Teilsätze bis Absätze
  - Indikatoren: Passivkonstruktionen, autonome Figurenrede
  - Beispiel: „und bin nun zu Hause, und *die Lampe brennt*, und *die Zigarre ist angezündet*“ (A. Schnitzler: Blumen; Property *Figurenname* für beide Annotationen mit dem Wert „Erzähler“)

Charakterisierung: Mithilfe dieser Tags für die Figurenanalyse kann analysiert werden, wie Figuren ausgestaltet sind und wie sie dargestellt werden.

- Tag *Aussehen*: Mit diesem Tag können Passagen annotiert werden, in denen das Äußere einer Figur beschrieben wird, zum Beispiel ihre äußere Erscheinung, ihre Physiognomie oder ihre Kleidung.
  - Länge der annotierten Passage: i. d. R. mehrere Wörter bis mehrere Sätze.
  - Indikatoren: beispielsweise Wörter aus den Wortfeldern *Körper* oder *Kleidung*
  - Beispiele: „Sie war köstlicher geschmückt, als er sie zu irgendeiner Zeit, die höchsten Festtage nicht ausgenommen, erblickt hatte, ein reiches seidenes Kleid umfloß ihre edle Gestalt, und ein goldenes Kreuz, mit roten Edelsteinen besetzt, glänzte an ihrem Halse“ (F. Hebbel: Matteo; Property *Figurenname* mit dem Wert „Felicita“)
- Tag *Figurenhandeln*: Mit diesem Tag werden Textstellen annotiert, in denen Figuren durch ihr Handeln charakterisiert werden, beispielsweise durch Motivation, Absicht oder Art der Handlung. Um festzuhalten, in welcher Hinsicht ihre Sprache sie charakterisiert, können nach Bedarf weitere Unterkategorien oder Properties eingefügt werden.
  - Länge der annotierten Passage: i. d. R. mindestens Teilsätze
  - Indikatoren: z. B. Aktionsverben
  - Beispiele: „Bei dem Krippenmacher war es immer lustig, denn er und sein Sohn, ein flinker lachender Bursche von etwa vierzehn Jahren, waren Musikanten, und *wenn die Monate heranrückten, wo die Krippenmacherei ruhte, da nahmen sie ihre Geigen und spielten an Sonntagen in den kleinen Schenken der Vorstadt zum Tanze auf*“ (A. Christen: Nachbar Krippelmacher; Property *Figurenname* mit dem Wert „Krippelmacher“)
- Tag *Charaktereigenschaften*: Dieser Tag dient der Annotation von Textstellen, in denen Figuren explizit Charaktereigenschaften zugeschrieben werden, beispielsweise durch einen Erzähler, durch andere Figuren oder durch sich selbst.
  - Länge der annotierten Passage: einzelnes Wort bis mehrere Sätze
  - Indikatoren: z. B. Adjektive aus dem Wortfeld *Charakter*
  - Beispiele: „Matteo war ein junger Mann, der, obwohl von niedriger Herkunft, und nicht mit besonderen Talenten ausgestattet, sich durch seine Dienstbeflissenheit und sein stilles, bescheidenes Wesen angenehm zu machen und Vertrauen zu erwecken wußte“ (F. Hebbel: Matteo; Property *Figurenname* für alle Annotationen mit dem Wert „Matteo“)
- Tag *Sprache*: Mit diesem Tag können Passagen annotiert werden, in denen Figuren durch die Sprache, die sie benutzen, oder durch den Inhalt des Gesagten charakterisiert werden. Um festzuhalten, in welcher Hinsicht ihre Sprache sie charakterisiert, können nach Bedarf weitere Unterkategorien oder Properties eingefügt werden.
  - Länge der annotierten Passage: i. d. R. mindestens Teilsätze
  - Indikatoren: Anführungszeichen, Verba dicendi
  - Beispiele: „Jastersky, der blonde Unteroffizier, schreit endlich: »*Also, entweder Sie kommen herunter oder Sie klettern hinauf, Gruber! Sonst melde ich dem Herrn Oberlieutenant…*«“ (R. M. Rilke: Die Turnstunde; Property *Figurenname* mit dem Wert „Unteroffizier Jastersky“)
- Tag *soziokulturelles Umfeld*: Mit diesem Tag können Passagen annotiert werden, die der Charakterisierung einer Figur durch die Einflüsse ihres Umfeldes auf sie dienen. Dazu gehören beispielsweise Hinweise auf Herkunft, Erziehung, Sozialisierung oder die gesellschaftlichen Kreise, in denen eine Figur verkehrt.
  - Länge der annotierten Passage: mindestens Teilsätze
  - Beispiele: „Unter einem Busch, in ärmliche Decken gehüllt, hatten sie ihn gefunden. Kein Mensch wußte, wer seine Eltern waren. Da man ihn nicht verkommen lassen konnte, taufte man ihn und übergab ihn einer Alten, die für seine Pflege etliche Lire von der Gemeinde erhielt“ (M. Janitschek: Poverino; Property *Figurenname* mit dem Wert „Gaetano“)

5.2 Handlung

Handlung: Die hier versammelten Kategorien zur Analyse der *histoire*, also des Inhalts des Erzählten, sind für die Untersuchung der Handlung vorgesehen, also der Zustandsveränderungen bzw. Ereignisse, an denen typischerweise Figuren als Handlungsträger*innen mitwirken.

- Tag *Handlungssequenz*: Eine Handlungssequenz zeichnet sich vornehmlich dadurch aus, dass sie aus miteinander verknüpften Ereignissen besteht, dass ihr bestimmte Figuren zugeordnet sind und sie über eine spezifische zeitliche und räumliche Dimension verfügt. Die Verwendung des Tags *Handlungssequenz* ist vor allem dann sinnvoll, wenn sich in einer Erzählung mehrere Handlungssequenzen finden, die sich beispielsweise abwechseln. Wann immer eine zusammenhängende Passage ausgemacht werden kann, in der von den Ereignissen einer Handlungssequenz berichtet wird, wird diese mit einer *Handlungssequenz*-Annotation versehen. Für jede Annotation wird für die Property _Sequenzkennzeichnun_g ein eindeutiger Name oder eine Nummer vergeben, um die Handlungssequenz zu bezeichnen.
  - Länge der annotierten Passage: mindestens Halbsätze, typischerweise Absätze oder ganze Kapitel
  - Indikatoren: (für den Wechsel zwischen Handlungssequenzen) Worte wie „währenddessen“, „unterdessen“, „in der Zwischenzeit“; ggf. Wechsel der Fokalisierung zwischen Figuren, erkennbar an anderen vorkommenden Figurennamen; zeitlicher/räumlicher Wechsel, erkennbar u. a. an Zeitausdrücken, die größere Zeitsprünge ausdrücken, oder neuen Ortsnamen; insbesondere eine Kombination dieser Indikatoren.
  - Beispiel: „Dieses Anerbieten kam Georgen erwünscht. Ein Wort gab das andere, und mein Schelm von Diener, der an meinen Pariser Streichen großen Anstoß genommen hatte, erzählte Alles, was er von meiner dortigen Lebensart wußte, und pries sich glücklich, daß die Noth mich endlich triebe, eine so angenehme Zuflucht zu wählen. Mir aber verschwieg er aus guten Gründen Alles. Victoire, die mit der Frau desselben Kaufmanns in Paris war, erfuhr bei ihrer Zurückkunft, welche Nachrichten von dem deutschen Bräutigam eingegangen waren […]“ (A. Kähler: Die drei Schwestern; bis „wählen“: Property *Sequenzkennzeichnung* mit Wert „Nebenhandlung Diener und Schwestern“ oder Zahlenwert „2“; von „mir“ bis „alles“: Property *Sequenzbezeichnung* mit Wert „Haupthandlung Bräutigam“ oder Zahlenwert „1“; ab „Victoire“: Property *Sequenzkennzeichnung* mit Wert „Nebenhandlung Diener und Schwestern“ oder Zahlenwert „2“)
- Tag *Handlungssegment*: Der Tag *Handlungssegment* ist ein Untertag zu *Handlungssequenz*. Ein Handlungssegment ist dementsprechend ein Teil einer Handlungssequenz, der einer bestimmten Handlungssequenz eindeutig zugeordnet ist. Mit dem Tag werden Passagen annotiert, die von einem einzelnen Ereignis bzw. der Einzelaktion einer Figur berichten. Besonders fruchtbar kann dieser Tag sein, um Einzelereignisse zu annotieren, denen eine besondere Relevanz für den Verlauf der dazugehörigen Handlungssequenz zugeschrieben wird.
  - Länge der annotierten Passage: mindestens ein Teilsatz, typischerweise einer oder mehrere Sätze
  - Beispiel: „Ein Wort gab das andere, und mein Schelm von Diener, der an meinen Pariser Streichen großen Anstoß genommen hatte, erzählte Alles, was er von meiner dortigen Lebensart wußte, und pries sich glücklich, daß die Noth mich endlich triebe, eine so angenehme Zuflucht zu wählen“ (A. Kähler: Die drei Schwestern; Property *Segmentkennzeichnung* mit Wert „Diener Georg erzählt vom Verhalten des Erzählers in Paris“ oder Zahlenwert)

Externe und weiterführende Links

- CATMA: https://web.archive.org/save/https://catma.de (Letzter Zugriff: 03.07.2024)
- Tagset „Narratologie (discours)“ auf Zenodo: https://web.archive.org/save/https://zenodo.org/records/10519654 (Letzter Zugriff: 03.07.2024)

Bibliographie

de Toro, Alfonso. 1986. *Die Zeitstruktur im Gegenwartsroman: am Beispiel von G. García-Márquez' Cien años de soledad, M. Vargas-Llosas La casa verde & A. Robbe-Grillets La maison de rendez-vous*. Tübingen: Narr.

forTEXT. 2020a. Tagset Narratologie (discours). Zenodo, 27. Januar. doi: 10.5281/zenodo.105196488, https://zenodo.org/records/10519648.

———. 2020b. Tagset Narratologie (histoire). Zenodo, 4. Mai. doi: 10.5281/zenodo.10519654, https://zenodo.org/records/10519654.

Genette, Gérard. 2007. *Discours du récit*. Points Essais 581. Paris: Éd. du Seuil.

Gius, Evelyn. 2015. *Erzählen über Konflikte: Ein Beitrag zur digitalen Narratologie*. Bd. 46. Narratologia. Berlin; Boston: De Gruyter.

Hansen, Per Krogh. 2000. *Karakterens rolle: aspekter af en litterær karakterologi*. Holte: Medusa.

Jacke, Janina. 2014. Is There a Context-Free Way of Understanding Texts? The Case of Structuralist Narratology. *Journal of Literary Theory* 8, Nr. 1 (1. Januar): 118–139. doi: 10.1515/jlt-2014-0005, http://www.degruyter.com/view/j/jlt.2014.8.issue-1/jlt-2014-0005/jlt-2014-0005.xml (zugegriffen: 25. Oktober 2016).

———. 2024a. 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.

———. 2024b. Ressourcenbeitrag: Tagset Narratologie (discours). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3756, https://fortext.net/ressourcen/tagsets/tagset-narratologie-discours.

Jannidis, Fotis. 2012. Character. 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/character (zugegriffen: 8. Mai 2018).

Margolin, Uri. 1983. Characterisation in Narrative: Some Theoretical Prolegomena. *Neophilologus* 67: 1–14. doi: 10.1007/BF01956983,.

———. 1990. Individuals in Narrative Worlds: An Ontological Perspective. *Poetics Today* 11, Nr. 4: 843–871. doi: 10.2307/1773080,.

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

Propp, Vladimir. 1975. *Morphologie des Märchens*. Hg. von Karl Eimermacher. München: Hanser.

Schumacher, Mareike. 2024a. 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.

———. 2024b. Methodenbeitrag: Netzwerkanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 6. Netzwerkanalyse (30. August). doi: 10.48694/fortext.3759, https://fortext.net/routinen/methoden/netzwerkanalyse.

———. 2024c. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

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.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

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.

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.

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.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

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).

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.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

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.