Toolbeitrag: CATMA

Mareike Schumacher [1]

1. Universität Regensburg

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3761
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2019-04-15 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.

Der Workflow von CATMA: Import einer Textdatei zum Beispiel im TXT-, TEI oder PDF-Format, Annotation mit eigens angelegten Tagsets, Analyse und Visualisierung z. B. als Distribution Graph oder Double Tree der Text- und Annotationsdaten. Der Weg zurück aus den Visualisierungen zum Text ist ebenfalls möglich, sodass der Workflow zirkulär sein kann.

- **Systemanforderungen:** Webbasiertes Tool, über den Browser (z. B. Chrome, Firefox, Safari) nutzbar
- **Stand der Entwicklung:** Derzeit Version 6.2; seit 2008 stetig weiterentwickelt
- **Herausgeber:** CATMA-Team der Universität Hamburg
- **Lizenz:** kostenfrei zugänglich
- **Weblink:** https://catma.de
- **Im- und Export:** Importformate: DOC, PDF, XPDF, HTML (vgl. Markup Language), HTM, RTF, TXT (vgl. Reintext-Version), TEI, XML2 (vgl. XML), DOCX, ZIP; Downloadformate: XML2, UTF-8 (vgl. Unicode/UTF-8) Plaintext (für Primärtextdokumente), JSON (für CATMA-Annotationen (vgl. Annotation))
- **Sprachen:** Sprachunabhängig: Hebräisch, Arabisch, Deutsch, Englisch, Französisch etc. (Spracheinstellung beim Hochladen des Dokuments. Alle Schriftsprachen stehen zur Auswahl)

1. Für welche Fragestellungen kann CATMA eingesetzt werden?

CATMA (kurz für Computer Assisted Text Markup and Analysis) ist ein im Browser laufendes Tool, das die manuelle Annotation (Jacke 2024a) und Analyse von Texten digital unterstützt und dabei den traditionellen philologischen Workflow zum Vorbild hat. Taxonomiebasierte Textarbeit, die „top-down“ und theorie- wie kategoriengeleitet verfährt, ist damit ebenso möglich wie die „bottom up“ verfahrende und zirkuläre hermeneutische Forschung, die erst im Zuge der Exploration konkreter Texte ihre spezifischen Beschreibungsterme und -Kategorien entwirft und präzisiert. CATMA kann darum für eine große Vielfalt an Forschungsansätzen genutzt werden. Eine mögliche Fragestellung wäre: Wie wird das Motiv des Doppelgängers in E.T.A. Hoffmanns *Die Elixiere des Teufels* dargestellt und inwiefern lässt es sich der Thematik der Persönlichkeitsspaltung zuschreiben?

2. Welche Funktionalitäten bietet CATMA und wie zuverlässig ist das Tool?

Funktionen:

- „undogmatische“, d. h. dynamisch erweiterbare, nicht notwendig nach einem fix vorgegebenen Schema verfahrende, Annotation von Textdokumenten; Kernmerkmale sind dabei
  - freie Annotation nach individuell definierten Kategorien
  - Mehrfachannotation einzelner Wörter und Passagen

  - überlappende Annotation
  - widersprüchliche Annotation
- Entwickeln eigener Annotationskategorien (Tags) und deren Systematisierung in Taxonomien (Tagsets) (vgl. Tagset)
- kollaboratives Annotieren (Jacke 2024b) in Echtzeit
- Analyse von Text- und Annotationsdaten für Einzeltexte und Textsammlungen
- Natürlichsprachliche Entwicklung von Analyseabfragen (Queries) (vgl. Query) mit dem Query Builder
- Visualisierung von Text- und Annotationsdaten für Einzeltexte und Textsammlungen (vgl. Korpus)
- Halbautomatische Annotation von Wort- oder Phrasengruppen
- Automatisches POS-Tagging (vgl. POS) deutschsprachiger Textsammlungen (in CATMA 5.0)
- Automatische Annotation von Zeitformen und Zeitausdrücken in deutschsprachigen Textsammlungen (in CATMA 5.0)

Zuverlässigkeit: CATMA wird seit 2008 kontinuierlich weiterentwickelt. Derzeit sind parallel die Versionen 5.0 und 6.0.6 nutzbar. Das webbasierte Tool braucht nicht auf dem eigenen Rechner installiert zu werden, ist sehr performant und zuverlässig. Die Funktionen der automatischen Annotation in Version 5.0 können allerdings je nach Umfang der Korpora relativ viel Zeit in Anspruch nehmen.

3. Ist CATMA für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	✓
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	✓
Leichter Einstieg	teilweise
Handbuch vorhanden	✓
Handbuch aktuell	✓
Tutorials vorhanden	✓
Erklärung von Fachbegriffen	✓
Gibt es eine gute Nutzerbetreuung?	✓

CATMA wurde entwickelt, um geisteswissenschaftlich-hermeneutische Arbeitsweisen zu unterstützen. Die im Zentrum dieser Arbeitsweisen stehende manuelle Annotation ist daher auch in CATMA zentral und die entsprechenden Funktionen sehr intuitiv bedienbar. Der große Funktionsumfang des Tools macht es in Version 5.0 teilweise unübersichtlich. Für CATMA 6 wurde darum die Benutzeroberfläche (vgl. GUI) komplett überarbeitet, sodass das Tool intuitiv bedienbar ist und der Einstieg leicht fällt. Auch das Handbuch (Manual) wurde für Version 6 vollständig überarbeitet. Tutorials für alle Funktionen werden derzeit erarbeitet und sukzessive veröffentlicht.

4. Wie etabliert ist CATMA in den (Literatur-)Wissenschaften?

CATMA ist ein in den digitalen Geisteswissenschaften gut etabliertes Tool. Bisher wurden über 6200 Korpora in CATMA hochgeladen, die von mehr als 9500 registrierten Nutzer*innen einzeln oder kollaborativ untersucht werden. 9270538 Annotationen wurden manuell oder automatisch mit Hilfe von CATMA erstellt. Neben Forscher*innen mit Schwerpunkt in den digitalen Geisteswissenschaften arbeiten auch traditioneller forschende Textwissenschaftler*innen aufgrund der Nähe zur nicht-digitalen Arbeitsweise mit CATMA.

5. Unterstützt CATMA kollaboratives Arbeiten?

Ja, in jeder Version von CATMA können Texte simultan von mehreren Personen annotiert werden. Auch Annotationstaxonomien (vgl. Tagset) und (annotierte) Textdokument (vgl. Annotation; Korpus) können direkt aus dem Tool heraus geteilt und kollaborativ genutzt werden.

6. Sind meine Daten bei CATMA sicher?

Ja. CATMA ist ein webbasiertes Tool, das auf Servern (vgl. Server) des Rechenzentrum der Universität Hamburg läuft. Textdaten sind nur in einem geschützten Login-Bereich einsehbar. Für den Login ist eine Registrierung mit einer gültigen Email-Adresse oder über ein Google-Konto notwendig. Die CATMA-Login-Daten werden ebenfalls auf Servern im Hamburger Rechenzentrum gespeichert und nicht an Dritte weitergegeben. Bei Verwendung des Google-Logins werden die CATMA-Daten nicht automatisch für Google verfügbar und CATMA kann ausschließlich

auf die Login-Daten Ihres Google-Kontos zugreifen. Die Nutzung von CATMA ist darum aus urheberrechtlicher Perspektive unbedenklich.

Externe und weiterführende Links

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

Bibliographie

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Marie Flüh und Jan Horstmann. 2020. CATMA. 11. November. https://zenodo.org/records/4353618 (zugegriffen: 29. April 2022).

Jacke, Janina. 2024b. 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.

———. 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.

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.

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.

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.

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.

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.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

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.

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.

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.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.