GitMA oder CATMA für Fortgeschrittene

Schumacher, Mareike
schumacher@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Meister, Malte
meister@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Gerstorfer, Dominik
gerstorfer@linglit.tu-darmstadt.de
Technische Universität Darmstadt, Deutschland

Dieser GitMA-Workshop richtet sich an fortgeschrittene CATMA User*innen mit Vorkenntnissen in digitaler Annotation, die im Rahmen der eigenen Arbeit oder von Forschungsprojekten mit größeren Mengen von Annotationsdaten operieren (wollen). Bei GitMA handelt es sich um ein Python-Package, mit dem Annotationsdaten, die in CATMA erstellt wurden, weiter verarbeitet werden können (Vauth et al. 2021). Wie greife ich über Git auf meine CATMA-Annotationsdaten zu? Wie visualisiere ich kollaborativ erstellte Annotationsdaten, die in mehreren Collections abgelegt sind? Wie berechne ich die Übereinstimmung zwischen mehreren Annotator*innen? Diese und ähnliche Fragen werden während des Workshops beantwortet.

CATMA (Gius et al. 2021) ist eine webbasierte, kollaborative Textannotations- und Analyse-Plattform, die seit 2008 an der Universität Hamburg und im Rahmen des DFG-geförderten Projektes forTEXT seit 2020 an der Technischen Universität Darmstadt entwickelt wird. Hauptzielgruppe sind traditionell-analog arbeitende Geisteswissenschaftler*innen, die über eine intuitiv bedienbare GUI Texte annotieren und analysieren können. Mit dem Release von CATMA 6 im Jahr 2019 wurde für die Plattform ein auf Git basierendes Backend eingeführt. Für zahlreiche Projekte, die bereits auf sehr fortgeschrittenem Niveau CATMA nutzen, und Interessierte aus der Digital-Humanities-Community mit Erfahrung in der Nutzung von Git und Grundkenntnissen in Python, eröffnet sich dadurch eine Reihe neuer Funktionen, die es in bisherigen CATMA-Versionen nicht gab. Einige dieser Funktionen werden im Laufe dieses Workshops vorgestellt und vermittelt.

Der Workshop bietet:

- kurze Einführung in die Nutzung von CATMA über das graphische Userinterface
- Kennenlernen der Datenstrukturen des Backends
- Zugriff auf das Backend mit Git
- Weiterverarbeitung der Daten mit Hilfe eines zur Verfügung gestellten Python-Packages

Annotation in CATMA 6

Annotation ist eine zentrale Kultur- und Forschungspraxis, die bereits seit sehr langer Zeit analog betrieben wurde (vgl. Moulin 2010), bevor sie im Rahmen der Digital Humanities ins Digitale übertragen wurde. Textauszeichnung und -anreicherung, Freitextkommentare und das taxonomiebasierte Annotieren sind Formen der Annotation, die sich zum Teil überschneiden (vgl. Jacke 2018, # 9). Alle diese Formen werden von CATMA 6 digital unterstützt. Mithilfe selbst erstellter oder auf der Plattform forTEXT.net bereitgestellter Tagsets (z.B. Flüh 2020) kann einzeln oder im Team taxonomiebasiert annotiert werden.

Eine der wichtigsten Neuerungen von CATMA 6 gegenüber früheren Versionen ist die Umstellung auf eine projektzentrierte Nutzungsarchitektur. Am Beginn der Arbeit mit CATMA steht das Anlegen eines Projektes mit beliebig vielen Dokumenten, die analysiert werden sollen, und beliebig vielen Team-Mitgliedern, die daran arbeiten wollen. Einzelne und Merfachannotationen, einander überlagernde oder überlappende Annotationen oder auch widersprüchliche Annotationen sind in CATMA durch die Speicherung der Daten als Standoff-Markup möglich. Eine weitere Neuerung im Funktionsumfang ist die Möglichkeit, Textstellen und Annotationen zu kommentieren. Offene Fragen, nicht zu Ende gedachte Interpretationsansätze oder auch der Austausch mit den anderen Team-Mitgliedern können über die Kommentarfunktion in den Annotationsprozess integriert werden. Sowohl Annotationen als auch Kommentare können über die Analyse-Funktionen von CATMA durchsucht, in tabellarische Form gebracht oder visualisiert werden. Der Umfang dessen, was über die CATMA-GUI umgesetzt werden kann, ist also recht groß. Die Einführung des auf Git basierenden Backends macht das Tool für die Digital-Humanities-Community aber noch interessanter. Der undogmatische Zugang, der bisher nur zu Annotationen und Annotationstaxonomien ermöglicht wurde, erstreckt sich nun bis zu den Annotationsdaten und der Weiterverarbeitung derselben (siehe Abbildung 1). Dieser neue Teil des CATMA-Workflows wird in diesem Workshop vermittelt werden.

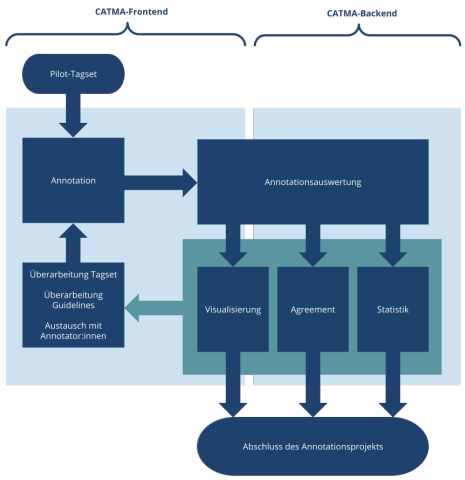

Abbildung 1: Im Workshop vermittelter Workflow zur Annotationsauswertung und -überarbeitung mit dem CATMA-Backend

Annotationen auswerten mit GitMA

Technische Niedrigschwelligkeit und Nähe zu traditionell-analogen Methoden der Geisteswissenschaften (vgl. Schumacher und Gius 2022) sind nach wie vor wichtige Grundsätze, die in CATMA implementiert sind. Doch mit zunehmender Verbreitung des Tools in den digitalen Geisteswissenschaften sind neben der Möglichkeit zu hermeneutisch-vielfältiger Textanalyse (vgl. Piez 2010) auch die Einhaltung von Best Practices und Standards, die innerhalb der Digital-Humanities-Community entwickelt wurden, von Bedeutung. Eine Verschmelzung von CATMA und dem direkten Datenzugriff über Git zu "GitMA" ermöglicht beides. Die im Annotationsprozess erstellten Daten können zum Beispiel nach der Übereinstimmung der Annotierenden untereinander (Artstein und Poesio 2008) ausgewertet werden. Es ist möglich eine der Annotationen als 'Silver Annotation' festzulegen und die anderen daran zu messen. Das festgestellte Disagreement kann zur Grundlage eines Disagreement-Tagsets werden, das über das Backend auch wieder ins Frontend der CATMA-GUI zurückgespielt werden kann (siehe Abb. 1). Dasselbe gilt für die nicht übereinstimmend annotierten Passagen, welche wiederum selbst durch Annotationen dargestellt bzw. hervorgehoben werden können. So ergibt sich ein Workflow vom Frontend zum Backend und zurück, der auch die Erstellung von Goldannotationen (vgl. Wissler et al. 2014) unterstützt.

Format und Ablauf des Workshops

Der Workshop wird als halbtägiges hands-on Tutorial angeboten.
Ablauf:
CATMA 6 (45 Minuten)

- kurze Einführung in das CATMA-Frontend
- Struktur: Tagsets, Documents, Annotation Collections

Zugriff auf Annotationsdaten über Git (30 Minuten)

- wie clone ich ein CATMA Project?
- wie update ich ein CATMA Project, um neue Annotationen zu laden?
- Installation des Packages
- Laden eines Projects
- Zugriff auf Annotation Collections, Dokumente und Tagsets

15 Minuten Pause
Explorative Annotationsauswertungen (60 Minuten)

- Annotationsdaten visualisieren
- Netzwerkanalysen von Annotationsdaten

15 Minuten Pause
Statistische Annotationsauswertungen (45 Minuten)

- Einführung in die Begrifflichkeiten Inter-Annotator-Agreement, Silver & Gold Standard
- Festlegung der Silver Annotations
- Umgang mit Annotationsspannen
- Automatische Erstellung eines Disagreement Tagsets
- Darstellung von Disagreement als Annotationen
- Manuelle Überarbeitung von automatischen Goldannotationen

Diskussion und Feedback (30 Minuten)

Zielgruppe

Nutzer*innen, die Annotationen mit CATMA in Forschungsprojekten oder Lehrsituationen managen, sowie alle, die einen schnellen Workflow zwischen Annotation bzw. Annotationsbearbeitung und Annotationsauswertung benötigen.

Zahl der möglichen Teilnehmer*innen

30

Technische Voraussetzungen

Die benötigten Vorinstallationen von Git, Anaconda und GitMA (sowie dessen Abhängigkeiten) können durch

die Bereitstellung eines Docker-Image vermieden werden. Die Teilnehmer*innen sollten Docker Desktop auf einem eigenen Laptop installiert haben (Touch Devices werden nicht unterstützt) und diesen zum Workshop mitbringen. Für die Durchführung des Workshops benötigen wir außerdem einen Beamer.

Zur Vorbereitung sollten Teilnehmer*innen außerdem schon einen CATMA-Account erstellt (unter https://app.catma.de/catma/) und sich mit der CATMA-Nutzung bekannt gemacht haben (z.B. mithilfe von der forTEXT-Lerneinheit zu CATMA 6: Manuelle Annotation mit CATMA). Wenn eigene CATMA-Annotationsdaten vorhanden sind, können diese während des Workshops analysiert werden. Für Teilnehmende, die nicht an eigenen Daten arbeiten möchten, stellen wir ein Demo-Projekt zur Verfügung, mit dem man während des Workshops arbeiten kann.

Benötigte Vorkenntnisse

Die Teilnehmer*innen sollten über grundlegende Kenntnisse der Kommandozeile, Git und Python sowie Jupyter verfügen.

Beitragende

Evelyn Gius, Prof. Dr.
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Evelyn Gius ist Professorin für Digitale Philologie und Neuere Deutsche Literatur an der Technischen Universität Darmstadt. Sie promovierte an der Universität Hamburg mit einer Arbeit über die narrative Struktur von Konflikterzählungen. Ihre Forschungsschwerpunkte sind manuelle Annotation, Operationalisierung, Erzähltheorie, Segmentierung und Konflikte. Sie ist PI mehrerer Digital-Humanities-Projekte (EvENT, KatKit, CATMA, forTEXT) und ist Vorsitzende des Vereins Digital Humanities im deutschsprachigen Raum (DHd), Mitherausgeberin des Journal of Computational Literary Studies (JCLS) und Mitherausgeberin der Buchreihe "Digitale Literaturwissenschaft".

Dominik Gerstorfer, M.A.
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Dominik Gerstorfer promoviert über "Philosophische Fragen der Digital Humanities" an der Universität Stuttgart. Derzeit ist er im Projekt KatKit tätig, zuvor war er im DFG-Projekt forTEXT in Darmstadt und im Digital-Humanities-Projekt CRETA in Stuttgart beschäftigt. Dominik hat an der Universität Tübingen Philosophie, Politikwissenschaften und Soziologie (M.A.) studiert. Seine Forschungsschwerpunkte liegen in den Bereichen Wissenschaftstheorie, formale Methoden und Argumentationsanalyse. Im Rahmen von KatKit und forTEXT beschäftigt sich Dominik u.a. mit Intertextualität, Ontologien und der Entwicklung von Kategoriensystemen.

Malte Meister, B.Sc.
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Malte Meister hat 2009 sein Informatik-Diplom (B.Sc.) in Kapstadt erworben. Im Rahmen des Abschlussprojekts für sein Diplom wurde er beauftragt, das Text-Annotations und -Analysetool CATMA, für die Universität Hamburg zu erstellen. Bis Anfang 2010 wirkte er im Team an CATMA mit, bevor er sich auf seine Karriere in der freien Wirtschaft konzentrierte. Nach mehr als zehn Jahren Berufserfahrung als Softwareentwickler und Teamleiter entschied er sich, wieder in die CATMA-Entwicklung einzusteigen. Er ist seit 2021 technischer Mitarbeiter an der TU Darmstadt und beschäftigt sich dort im Rahmen von forTEXT hauptsächlich mit dem Betrieb und der Weiterentwicklung von CATMA und den damit verbundenen Systemen.

Mareike Schumacher, M.A.
Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Residenzschloss 1, 64283 Darmstadt

Mareike Schumacher koordiniert das DFG-Projekt forTEXT (https://fortext.net), in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die traditionelleren Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Sie promovierte als digitale Literaturwissenschaftlerin über Orte und Räume im Roman, beschäftigt sich besonders mit den Methoden des *distant reading* (u. a. *Named Entity Recognition* oder Stilometrie), der Digital-Humanities-Theorie und der Verbindung von digitalen Methoden und theoriebasierter Literatur- und kulturwissenschaftlicher Forschung.

Bibliographie

Artstein, Ron, und Massimo Poesio. 2008. „Inter-Coder Agreement for Computational Linguistics“. *Computational Linguistics* 34 (4): 555–96. https://doi.org/10.1162/coli.07-034-R2 .

Flüh, Marie. 2020. „Emotionsanalyse“. In *forTEXT* . https://fortext.net/ressourcen/tagsets/emotionsanalyse .

Frey-Endres, Marcel, und Tobias Simon. 2021. *Digitale Werkzeuge zur textbasierten Annotation, Korpusanalyse und Netzwerkanalyse in den Geisteswissenschaften* . Bd. 2. Digital Philology | Working Papers in Digital Philology. Darmstadt: TUprints. https://doi.org/10.26083/tuprints-00017850 .

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh und Jan Horstmann. 2021. “CATMA 6 (Version 6.3)”. *Zenodo* . DOI: 10.5281/zenodo.1470118. URL: https://catma.de/ [letzter Zugriff 24. November 2021]

Jacke, Janina. 2018. „Manuelle Annotation“. In *forTEXT* . https://fortext.net/routinen/methoden/manuelle-annotation .

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.

Piez, Wendell. 2010. „Towards Hermeneutic Markup. An Architectural Outline“. In *Digital Humanities 2010. Conference Abstracts* , 202–5. London. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html .

Rebholz-Schuhmann, Dietrich, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, u. a. 2010. „The CALBC Silver Standard Corpus for Biomedical Named Entities – A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers“. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* . Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/888_Paper.pdf .

Vauth, Michael, Hans Ole Hatzel und Evelyn Gius. 2021. "forTEXT/catma_gitlab: 1.0.0." *Zenodo* . DOI: 10.5281/ZENODO.5669221 .

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