Texte digital annotieren und analysieren mit CATMA 6.0

Horstmann, Jan
jan.horstmann@uni-hamburg.de
Universität Hamburg, Deutschland

Meister, Jan Christoph
jan-c-meister@uni-hamburg.de
Universität Hamburg, Deutschland

Petris, Marco
marco.petris@uni-hamburg.de
Universität Hamburg, Deutschland

Schumacher, Mareike
mareike.schumacher@uni-hamburg.de
Universität Hamburg, Deutschland

Einleitung

In diesem hands-on Workshop werden wir die Möglichkeiten der für Geisteswissenschaftler*innen entwickelten Annotations- und Analyseplattform CATMA 6.0 praktisch erkunden. Es werden keinerlei technische Vorkenntnisse vorausgesetzt. Inhaltlich werden wir uns vor allem den theoretischen und praktischen Aspekten der digitalen Annotation von (literarischen) Texten, als auch der Analyse und Visualisierung dieser Texte und der erstellten Annotationen widmen.

CATMA (Computer Assisted Text Markup and Analysis; www.catma.de) ist ein open-source-Tool, das seit 10 Jahren an der Universität Hamburg entwickelt und derzeit von über 60 Forschungsprojekten weltweit genutzt wird. Die neue Version 6.0 wird im Zuge des DFG-Projektes forTEXT (www.fortext.net) entwickelt und implementiert. Neben erweiterten technischen Möglichkeiten (wie beispielsweise die Möglichkeit der Datenversionierung und die Organisation kollaborativer Arbeit in einer Projektstruktur), bietet die neue Version ein völlig überarbeitetes, intuitiver nutzbares User-Interface, das auf Material Design als dem avanciertesten und am Markt etabliertesten Schema basiert, welches die meisten Nutzer*innen bereits internalisiert haben. Das Interface ermöglicht einen leichten Einstieg in die digitale Textannotation und -analyse, ohne dass umfangreiche technische Kenntnisse vonnöten wären, und ohne dass die Nutzer*innen mit zu vielen (Experten-)Funktionen gleichzeitig konfrontiert würden. Das gesamte Repertoire an Funktionen (wie beispielsweise kollaborative Annotation oder automatische Annotation von Textkorpora) kann dann von erfahreneren Nutzer*innen bei Bedarf genutzt werden.

CATMA unterstützt

- individuelle wie kollaborative Annotation und Analyse – Texte können privat, aber auch im Team erforscht werden;
- explorative, non-deterministische Praktiken der Textannotation – CATMA liegt ein diskursiver, diskussionsorientierter Ansatz zur Textannotation zugrunde, der auf die Forschungspraktik hermeneutischer Disziplinen zugeschnitten ist;
- die nahtlose Verknüpfung von Textannotation und -analyse in einer webbasierten Arbeitsumgebung – Analyse und Interpretation gehen nach dem Prinzip des ‘hermeneutischen Zirkels’ in CATMA damit Hand in Hand.

Von linguistischen Textanalysetools unterscheidet sich CATMA insbesondere durch seinen „undogmatischen” Ansatz: Das System schreibt mit seiner hermeneutischen Annotation (vgl. Piez 2010) weder definierte Annotationsschemata oder -regeln vor, noch erzwingt es die Verwendung von starren Ja-/Nein- oder Richtig-/Falsch-Taxonomien. Wenn eine Textstelle mehrere Interpretationen zulässt (wie es in literarischen Texten häufig der Fall ist), ist es in CATMA daher möglich, mehrere und sogar widersprechende Annotationen zu vergeben und so der Bedeutungsvielfalt der Texte Rechnung zu tragen. Mit dem sog. Query-Builder lassen sich außerdem Schritt für Schritt Textanalysen durchführen. Die Ergebnisse der Analyse können schließlich in verschiedenen Varianten visualisiert und für die literaturwissenschaftliche Interpretation und Argumentation genutzt werden.

Zudem bietet CATMA auch die Möglichkeit, bereits annotierte Texte zu verarbeiten (z.B. durch den Upload von XML-Dateien) und die in anderen Tools erstellten Annotationen anzuzeigen, mit zu analysieren und damit wissenschaftlich nachzunutzen. Außerdem lassen sich in CATMA auch automatische (z.B. POS für deutschsprachige Texte) und halb-automatische Annotationen generieren.

Manuelles und kollaboratives Annotieren

Die Annotation von Texten gehört seit Jahrhunderten zu den textwissenschaftlichen Kernpraktiken (vgl. Moulin 2010). Genauer lassen sich hier Freitextkommentare, taxonomiebasierte Annotation und Textauszeichnung unterscheiden, wobei die Übergänge häufig fließend sind (vgl. Jacke 2018, § 9). Während CATMA 6.0 auch eine Funktion für Freitextkommentare bietet, ist die taxonomiebasierte Annotation das eigentliche Kerngeschäft des Tools – wobei die Taxonomie prinzipiell undogmatisch erstellt werden

kann und die Form von sog. Tagsets annimmt, denen für kollaborative Annotationsprojekte wahlweise eine Annotations-Guideline beigegeben werden kann (vgl. auch Bögel et al).

Im Workshop werden wir den Unterschied von *Document* (der eigentliche Text), *Tagset* (die aus *Tags* – d.h. aus einzelnen Beschreibungsbegriffen – gebildete Taxonomie, mit der Texte annotiert werden) und *Annotation Collection* (die nutzerspezifische Sammlung individueller Annotationen zu einem *Document* oder einem Korpus) kennenlernen. Diese Dreigliederung ist spezifisch für CATMA und bietet eine Reihe von Vorteilen:

- Taxonomien können projektübergreifend und unabhängig von Texten und Annotationen wiederverwendet werden;
- Annotationen können als *Collections* nach unterschiedlichen inhaltlichen (z. B. nach Forschungsaspekten) oder auch organisatorischen Gesichtspunkten (z. B. nach Projektmitgliedern) gruppiert und wiederverwendet bzw. erweitert werden;
- benutzerspezifische Annotationen werden als sog. *Stand-off Markup* gespeichert und können damit wahlweise angezeigt oder ausgeblendet werden. Der eigentliche Text wird hierbei nicht verändert. Arbeitet eine Gruppe von Annotator*innen mit der gleichen Taxonomie an einem Text, lassen sich Übereinstimmungen und Widersprüche direkt und einfach erkennen (vgl. Gius und Jacke 2017), um auf interessante oder problematische Textstellen aufmerksam zu werden und die ‘Arbeit am Text’ zugleich kritisch zu reflektieren.

Analyse und Visualisierung

Neben der Annotation sind die Analyse und Visualisierung der Text- und Annotationsdaten das andere wichtige Standbein von CATMA. Hier wird *distant reading* mit *close reading* zusammengebracht, denn die zuvor manuell erstellten qualitativen Annotationen werden nun in ihrer Quantität und Verteilung hinterfragt. Dies geschieht in Zusammenhang mit „klassischen” DH-Textanalysemethoden wie dem Erstellen einer Wortfrequenzliste, der Analyse von Keywords in Context (KWIC und *DoubleTree*) oder der Distribution ausgewählter Wörter (oder eben Annotationen) im Text oder in der Textsammlung.

Neben diesen grundlegenden Funktionen, die alle per Klick ausgeführt werden können, bietet CATMA den sog. *Query Builder*, in dem komplexere Abfragen einfach per Mausklick erzeugt werden können, ohne dass tiefergehende Kenntnisse einer Abfragesprache (sog. *Query Language*) verlangt werden. Im Workshop werden wir uns dabei nicht nur den Analysefunktionen widmen, sondern auch die unterschiedlichen Visualisierungsmöglichkeiten zu den einzelnen Abfragen anschauen und hinterfragen.

Im Analysebereich können außerdem halbautomatische Annotationen erstellt werden, d.h. man annotiert wiederkehrende Wörter oder Wortgruppen auf einmal mit einem bestimmten Tag, statt dies manuell und wiederholt im Annotationsmodul zu tun.

Der Wechsel zwischen der Arbeit im Annotations- und Analyse- und Visualisierungs-Modul ist ein iterativer Prozess, der die klassisch-zirkuläre hermeneutische Interpretationsarbeit in der Literaturwissenschaft widerspiegelt (vgl. Gius, in Vorbereitung).

Ablauf

Im Workshop werden wir uns in einer Mischung aus Präsentations- und Hands-on-Phasen der textanalytischen Arbeit in CATMA 6.0 nähern. Nach einer generellen Einführung in das Tool werden die Teilnehmer*innen anhand eines vorgegebenen Beispieltextes den gesamten Workflow von der individuellen taxonomiebasierten Textannotation, über die Analyse hin zur Visualisierung und Interpretation der Text- und Annotationsdaten kennenlernen und praktisch erproben können.

Lernziele

Die Teilnehmer*innen sollen ausgehend vom digitalen Text in die Lage versetzt werden, Annotationen manuell und automatisch unterstützt zu erstellen und in Annotation Collections zu speichern, Tagsets/Taxonomien zu entwickeln und den Text alleine und in Kombination mit den Annotationen zu analysieren und zu visualisieren. Für Diskussionen und individuelle Rückfragen (theoretischer, praktischer und technischer Art) auf jedem Niveau und in Bezug auf die Projekte der Teilnehmer*innen wird ausreichend Möglichkeit bestehen.

Zeitplan

Im Workshop werden wir den Arbeitsablauf der digitalen Texterforschung praktisch kennenlernen:

- analytische Textexploration (ca. 30 Minuten)
- manuelle und automatische Annotation und Spezifikation von Annotationskategorien (ca. 40 Minuten)
- kombinierte Abfragen von Annotations- und Textdaten (ca. 30 Minuten)
- visuelle Darstellungsmöglichkeiten von Abfrageergebnissen (ca. 20 Minuten)

Beitragende (Kontaktdaten und Forschungsinteressen)

Dr. Jan Horstmann
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Jan Horstmann ist Postdoc und koordiniert das DFG-Projekt forTEXT, in dem neben der Dissemination von digitalen Routinen, Ressourcen und Tools in die klassischeren Fachwissenschaften auch die Weiterentwicklung von CATMA eine wesentliche Rolle spielt. Als Literaturwissenschaftler interessiert er sich vor allem für die neuen Perspektiven und Erkenntnispotentiale, die DH-Methoden auf literarische Artefakte bereithalten können, und forscht in diesem Sinne unter anderem zu Entsagung und Ironie bei Goethe.

Prof. Dr. Jan Christoph Meister
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Jan Christoph Meister ist Professor für Digital Humanities mit dem Schwerpunkt Literaturwissenschaft. Als ursprünglicher Erfinder von CATMA hat er etliche Forschungsprojekte zur Annotation und Visualisierung textueller Daten und der Entwicklung und Verbesserung von DH-Tools geleitet.

Marco Petris, Dipl. Inform.
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Marco Petris ist Informatiker mit starker Affinität zu geisteswissenschaftlichen Fragestellungen. Er ist von Anfang an an der Entwicklung von CATMA beteiligt und beschäftigt sich mit allen Aspekten der DH-Toolentwicklung, des Tool-Designs und der Implementierung.

Mareike Schumacher, M.A.
Universität Hamburg, Institut für Germanistik, Überseering 35, Postfach #15, 22297 Hamburg

Mareike Schumacher promoviert als digitale Literaturwissenschaftlerin über Orte und narratologische Ortskategorien in literarischen Texten, beschäftigt sich besonders mit den Methoden des distant reading (u.a. Named Entity Recognition oder Stilometrie) und ist im forTEXT-Projekt u.a. für die Dissemination in den (sozialen) Medien zuständig.

Zahl der möglichen Teilnehmer*innen

Bis zu 30 Personen.

Benötigte technische Ausstattung

Teilnehmer*innen bringen ihren eigenen Laptop mit, der mit dem Internet verbunden ist (Achtung: Touch-Devices werden derzeit noch nicht unterstützt). Am Workshop können bis zu 30 Personen teilnehmen. Neben einer stabilen Internetverbindung werden ein Beamer und eine Leinwand benötigt.

Bibliographie

Bögel, Thomas / Gertz, Michael / Gius, Evelyn / Jacke, Janina / Meister, Jan Christoph / Petris, Marco / Strötgen, Jannik (2015): *„Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative”*, in: DHCommons Journal 1.

Gius, Evelyn (in Vorbereitung): *„Digitale Hermeneutik: Computergestütztes close reading als literaturwissenschaftliches Forschungsparadigma?”* in: **Fotis Jannidis (Hrsg.)**: *Digitale Literaturwissenschaft*. DFG-Symposium 9.–13.10.2018.

Gius, Evelyn / Jacke, Janina (2017): *„The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis”*, in: International Journal of Humanities and Arts Computing 11 (2), 233–254.

Jacke, Janina (2018): *„Manuelle Annotation"*, in: forTEXT. Literatur digital erforschen. http://fortext.net/routinen/methoden/manuelle-annotation (Zugriff: 24. September 2018).

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

Piez, Wendell (2010): *„Towards Hermeneutic Markup. An Architectural Outline"*, in: Digital Humanities Conference 2010, London http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-743.html (Zugriff: 24. September 2018).