Workshop CATMA featuring GitMA and Vis-A-Vis

Schumacher, Mareike Katharina

schumacher@linglit.tu-darmstadt.de
Technical University of Darmstadt, Germany

Gerstorfer, Dominik

dominik.gerstorfer@tu-darmstadt.de
Technical University of Darmstadt, Germany

Gius, Evelyn

evelyn.gius@tu-darmstadt.de
Technical University of Darmstadt, Germany

Meister, Malte

malte.meister@tu-darmstadt.de
Technical University of Darmstadt, Germany

Münz-Manor, Ophir

ophirmm@openu.ac.il
The Open University of Israel

This CATMA (Computer Assisted Text Markup and Analysis) workshop is intended for users with some preliminary experience with manual digital annotation who (want to) operate with larger amounts of annotation data in the context of their own work or research projects. It will demonstrate the basic features of CATMA and how annotation data can be processed using either the Python package GitMA (Vauth et al. 2021) or the web-based visualization tool vis-À-vis (Münz-Manor et al. 2020). The workshop will provide answers to questions like: How do I access my CATMA annotation data via GitMA? How do I calculate the agreement between multiple annotators? How do I generate advanced visualizations of the texts and annotations in vis-À-vis? How do I visualize collaboratively created annotation data stored in multiple collections?

Annotation in CATMA

Annotation is a core cultural and research practice that has been practiced non-digitally for a very long time (cf. Moulin 2010) before being transferred to the digital in the context of the Digital Humanities. Text markup and enrichment, free-text annotation, and taxonomy-based annotation are forms of annotation that overlap to some extent (cf. Jacke 2018, § 9). All of these forms are digitally supported by CATMA. CATMA (Gius et al. 2022) is a web-based collaborative text annotation and analysis platform that has been in development since 2008. With the release of CATMA 6 in 2019, a Git-based backend was introduced for the platform. GitMA builds on top of that to facilitate access to and further processing of annotation data. For many projects already using CATMA at an advanced level, as well as for people with experience using Git and a basic knowledge of Python, this opens up a number of new possibilities for working on and with annotations. The most important of these will be introduced during this workshop.

In CATMA, with the help of self-created tagsets or tagsets provided on the forTEXT.net platform (e.g. Flüh 2020), taxonomy-based annotation can be performed individually or in teams in a project-centered manner. To begin your work with CATMA, you create a project with any number of documents to be analyzed and any number of team members who want to work on them. You can make single and multiple, overlapping or even contradictory annotations in CATMA. You can integrate open questions, interpretation approaches that have not been thought through to the end, or even exchanges with other team members into your annotation process by using annotations or the comment function. Both annotations and comments can be searched, tabulated, or visualized via CATMA's analysis functions.

We will introduce the functionalities of the CATMA GUI in the first part of the workshop.

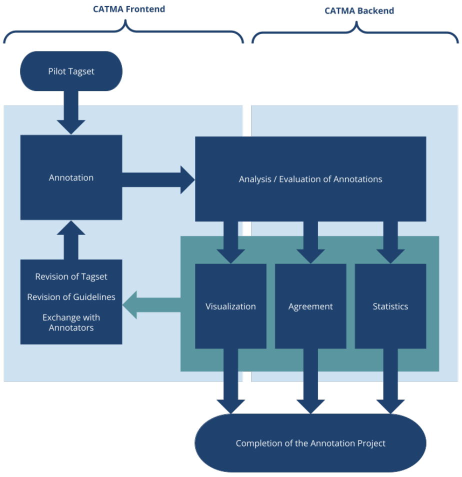

Figure 1: Workflow for annotation evaluation and revision with the CATMA backend, taught during the workshop.

Processing annotations with GitMA

Although the scope of what can be achieved with the CATMA GUI is quite large, the introduction of the Git-based backend makes the tool even more interesting for Digital Humanities practices of text analysis. The undogmatic access that was previously only provided to annotations and annotation taxonomies now extends to annotation data and its further processing (see Fig. 1).

A low barrier to entry and proximity to the established (analog) methods in the humanities remain central principles implemented in CATMA (cf. Schumacher and Gius 2022). In addition to the possibility for text analysis using diverse methods or theories of interpretation (cf. Piez 2010), adherence to best practices and standards developed within the Digital Humanities community is important. "GitMA", the combination of CATMA GUI functionalities with direct data access via Git, enables both. For example, the data created in the annotation process can be evaluated for inter-annotator-agreement. It is possible to specify one of the annotators as the 'silver standard' and measure annotations of the others against it. The detected disagreement can form the basis of a disagreement tagset, which can also be fed back into the frontend of the CATMA GUI via the backend (see Fig. 1). The same applies to the passages showing disagreement, which in turn can themselves be represented by annotations in the GUI. In addition, this workflow enables the creation of gold annotations (cf. Wissler et al. 2014). Both workflows, inter-annotator-agreement evaluation and gold standard creation, will be shown during the second part of the workshop.

Visualizing annotations with vis-À-vis

While CATMA offers basic visualizations of the annotations created within the system, vis-À-vis is a comprehensive visualization tool designed to assist scholars in recognizing patterns in annotated texts. vis-À-vis automatically imports the tagged corpora from CATMA, (re)presents them in various graphs and charts, provides comparative graphical tools for grouping the texts, and offers aggregative tools for inter- and intra-corpora relations. vis-À-vis offers five visualization possibilities: Small multiple distribution charts of annotation categories within a text, an aggregated Stacked Area chart, a Sunburst chart showing the proportions and quantities associated with each tag in the text, a force-directed network of annotation categories and a bubble visualization. In addition, a dynamic visualization of the logical structure of a tagset as well as its manifestations in specific annotated texts is offered.

The vis-À-vis backend provides easy connectivity with CATMA, connecting it to its GitLab backend through user-provided API keys. The user is prompted for a key and receives information on all of the user’s projects from CATMA. Projects can then be imported, or updated if they were imported previously. vis-À-vis uses a similar data model to CATMA, although compartmentalized: Users have Projects, which are composed of Texts and Annotations. Annotations associate parts of a text with a Tag, and tags belong to a Tagset. The last part of the workshop will demonstrate the possibilities of visualizing annotations with vis-À-vis.

Format and workshop schedule

The workshop will be offered as a half-day hands-on tutorial (four hours). With our group of experienced instructors, we will be able to provide close assistance to the participants.

Procedure:

- CATMA (90 minutes)
- Short introduction to the CATMA frontend
- Create tagsets
- Annotate documents
- Analyze annotations

Break (20 minutes)

- Access to annotation data via GitMA (55 minutes)
- Access to annotation collections, documents and tagsets
- Exploratory annotation evaluations
- Visualizing annotation data
- Inter-annotator-agreement, Silver & Gold Standard

Break (20 minutes)

- Visualizations of annotation data with vis-À-vis (55 minutes)

Target audience

Users who manage annotations with CATMA in research projects or in teaching situations, as well as anyone who needs a fast workflow between manual annotation or annotation editing and annotation evaluation. Users without prior CATMA experience are welcome, though we would kindly ask them to prepare by working through the manual annotation tutorial at https://catma.de/how-to/tutorials/manual-annotation/ in advance.

Number of possible participants

50

Technical requirements

To prepare for the workshop, participants should have already created a CATMA account (at https://app.catma.de/catma/). The required pre-installations of Git, Anaconda and GitMA (as well as its dependencies) can be avoided by using a provided Docker image. Participants should have Docker Desktop installed on their own laptop (touch devices are not supported) and bring it to the workshop.

Required previous knowledge

Participants should have basic knowledge of CATMA, Python, and Jupyter Notebooks.

Contributors

Dominik Gerstorfer

Technical University of Darmstadt, Institute of Linguistics and Literature, Residenzschloss 1, 64283 Darmstadt, Germany

Dominik Gerstorfer is doing his PhD on "Philosophical Issues in the Digital Humanities" at the University of Stuttgart. He is currently working in the KatKit project, and previously worked in the DFG project forTEXT in Darmstadt and in the Digital Humanities project CRETA in Stuttgart. Dominik studied philosophy, political science, and sociology (M.A.) at the University of Tübingen. His research interests lie in the areas of philosophy of science, formal methods, and argumentation analysis. In the context of KatKit and forTEXT, Dominik is working on intertextuality, ontologies, and the development of category systems, among other topics.

Evelyn Gius

Technical University of Darmstadt, Institute of Linguistics and Literature, Residenzschloss 1, 64283 Darmstadt, Germany

Evelyn Gius is a professor for digital philology and modern German literature at Technical University of Darmstadt. She received her PhD from the University of Hamburg with a thesis on the narrative structure of conflict narration. Her research is focused on manual annotation, operationalization, narrative theory, segmentation and conflicts. She is the PI of several Digital Humanities projects (EvENT, KatKit, CATMA, forTEXT) and serves as chair of the Digital Humanities association in the German-speaking area (DHd), co-editor of the Journal of Computational Literary Studies (JCLS) and co-editor of the book series “Digitale Literaturwissenschaft” (Digital Literary Studies).

Malte Meister

Technical University of Darmstadt, Institute of Linguistics and Literary Studies, Residenzschloss 1, 64283 Darmstadt, Germany

Malte Meister earned his computer science degree (B.Sc.) in Cape Town in 2009. As part of the final project for his diploma, he was commissioned to create the text annotation and analysis tool CATMA, for the University of Hamburg. He contributed to the team working on CATMA until early 2010, before focusing on his career in the private sector. After more than ten years of professional experience as a software developer and team leader, he decided to rejoin the CATMA development team. He has been a technical staff member at the TU Darmstadt since 2021, where he is mainly involved in the operation and further development of CATMA and related systems as part of forTEXT.

Ophir Münz-Manor

The Open University of Israel, Department of History, Philosophy and Judaic Studies, 1 University Road, Ra’anana 4353701, Israel

Ophir Münz-Manor is associate professor of Rabbinic Culture and a specialist in Jewish liturgy and liturgical poetry from Late Antiquity and the early Middle Ages. His studies focus on the intersections with contemporary Christian texts as well as questions of ritual, performance and gender in late antique Near Eastern cultures. In recent years, Prof. Münz-Manor has embarked on several projects that combine traditional literary analysis with quantitative and computerized methods from the realm of Computational Literary Studies.

He is the PI of two recent Digital Humanities projects (Critique of Pure Digital Knowledge and Algorithmic Detections of Metaphors in Pre-Modern Hebrew Poetry) and the creator of vis-À-vis. Recently, he edited a special issue of Jewish Studies Quarterly published by Mohr Siebeck on Digital Humanities and Jewish Studies and published (together with Itay Marienberg-Milikowsky) the first Hebrew annotated reader of fundamental essays in the realm of DH, translated from English.

Mareike Schumacher

University of Regensburg, Institute of Language, Literature and Culture, Universitätsstraße 31, 93053 Regensburg, Germany

Mareike Schumacher is Juniorprofessor of Digital Humanities. Before she coordinated the DFG project forTEXT (https://fortext.net), in which, in addition to the dissemination of digital routines, resources and tools to the more traditional disciplines, the further development of CATMA plays an essential role. She handed in her PhD-thesis about “Place and Space in the Novel. A contribution to Computational Literary Studies” in 2021. She is particularly interested in distant reading methods (including named entity recognition or stylometry), digital humanities theory, and the connection between digital methods and theory-based literary and cultural studies research.

Bibliography

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

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. 2021. “CATMA 6 (Version 6.5)”. Zenodo. DOI: 10.5281/zenodo.1470118. URL: https://catma.de/

Münz-Manor, O. et al. (2020) ‘ViS-Á-ViS#: Detecting Similar Patterns in Annotated Literary Text’. Available at: https://doi.org/10.48550/ARXIV.2009.02063.

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.

Schumacher, Mareike and Evelyn Gius. 2022. ‘forTEXT.net – Literatur digital erforschen’, Mitteilungen des Deutschen Germanistenverbandes, 69(2), pp. 121–126. Available at: https://doi.org/10.14220/mdge.2022.69.2.121.

Michael Vauth, Malte Meister, Hans Ole Hatzel, Dominik Gerstorfer, and Evelyn Gius. (2022). GitMA (1.4.9). Zenodo. https://doi.org/10.5281/zenodo.6330464

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