An End-To-End Integration of Automatic Annotations into CATMA

Thomas Bögel[*] **Marco Petris**[†] **Jannik Strötgen**[*] **Michael Gertz**[*]

[*] Institute of Computer Science, Heidelberg University
{boegel, stroetgen, gertz}@uni-hd.de
[†] Institute for German Studies, University of Hamburg
marco.petris@uni-hamburg.de

1 Introduction

Natural Language Processing offers solutions for predicting linguistic annotations at different levels of complexity. Thus, it seems obvious and – in general – a good idea to apply these methods to the Humanities in order to automate laborious manual annotations and to facilitate a deeper text analysis understanding. Apart from the purely technical aspect of developing suitable models, however, additional challenges for NLP in the Humanities arise: in order to be used as part of an analysis tool, humanists often desire justifications and explanations of automatic annotations. Just implementing a black-box approach, evaluating it intrinsically and returning the presumably best results to the user is not sufficient. In this paper, we suggest a transparent way of presenting the results of a NLP pipeline in a collaborative setting. This gives the user the possibility to judge the results directly within an already existing annotation interface and potentially use them for individual analysis tasks.

We will first present individual components that are combined with each other, namely the collaborative annotation tool CATMA and UIMA as a processing pipeline for Natural Language Processing. We will then show our end-to-end integration of UIMA into CATMA and its advantages.

2 CATMA integration

CATMA[1] is a flexible, collaborative annotation tool for literary scholars. So far, it integrates three functional and interactive modules, namely the tagger, the analyzer, and the visualizer. While the tagger module is a graphical interface to allow the easy creation of manual annotations in texts using flexible tag sets (including feature structures, overlapping annotations, etc.), the analyzer component offers a wide range of possibilities to query a document collection or single documents, e.g., for frequently occurring patterns. Finally, the visualizer module can be used to explore a document collection, e.g., by generating distribution charts of the analysis results. In this paper, we present an extension to CATMA, which was developed in the context of the heureCLÉA project[2] - the integration of a UIMA-based text processing pipeline for the automatic creation of tag annotations created by natural language processing tools.

UIMA (Unstructured Information Management Architecture)[3] is a wide-spread framwork for developing and using natural language processing pipelines. One of its key characteristics is that it allows the easy combination of tools that have initially not been built to be used together. All UIMA components rely on the same data structure - the Common Analysis Structure (CAS) - there are three types of components: collection readers, analysis engines, and CAS consumers. The collection readers task is to access the source of the documents that are to be processed and to initialize a CAS object for each document. Then, the analysis engines perform linguistic processing of the data and stand-off add annotations to the CAS object. The subsequently called analysis engines can access the annotation results of the earlier components, i.e., they can perform more complex tasks. Finally, a CAS consumer performs the final processing of the CAS object.

[1] Website: http://www.catma.de/clea

[2] http://heureclea.de/

[3] Website: http://uima.apache.org/

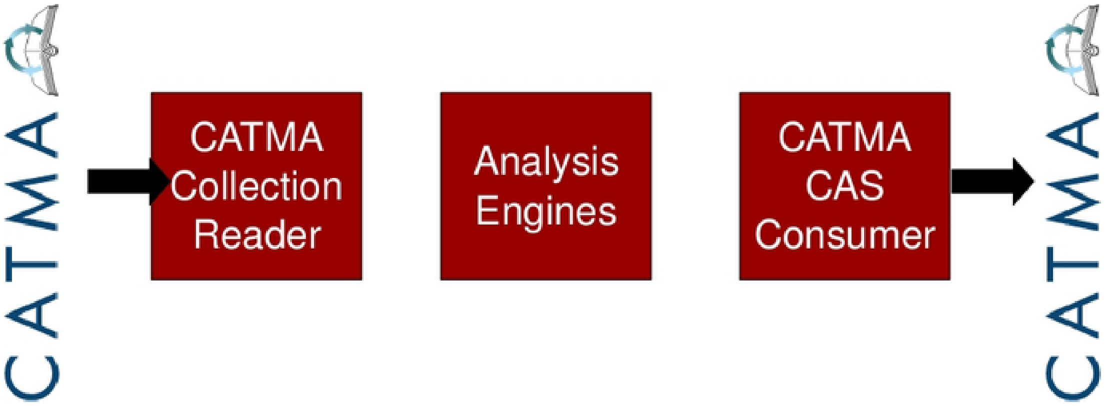


Figure 1: End-To-End architecture of combining the collaborative annotation platform CATMA with the automatic text processing pipeline UIMA.

In our case, the pipeline architecture is set up as depicted in Figure 1. The Collection Reader accesses the documents directly from CATMA and returns annotation information back to CATMA. However, the actual key feature of our development is that the user can directly access the automatic processing feature within the CATMA interface. That is, the user can select the types of annotations that shall be added to her document or document collection automatically. This significantly decreases the boundary for users not familiar with applying NLP tools for automatic processing of textual data, i.e., for typical CATMA users who are often literary scholars or students of the Humanities.

Nevertheless, our implementation is not a black box solution that only adds annotations that the user has to accept. In contrast, we are currently working on integrating a user feedback interface that will allow the initialization of user parameters based on the users feedback in the form of accepted or rejected annotations.

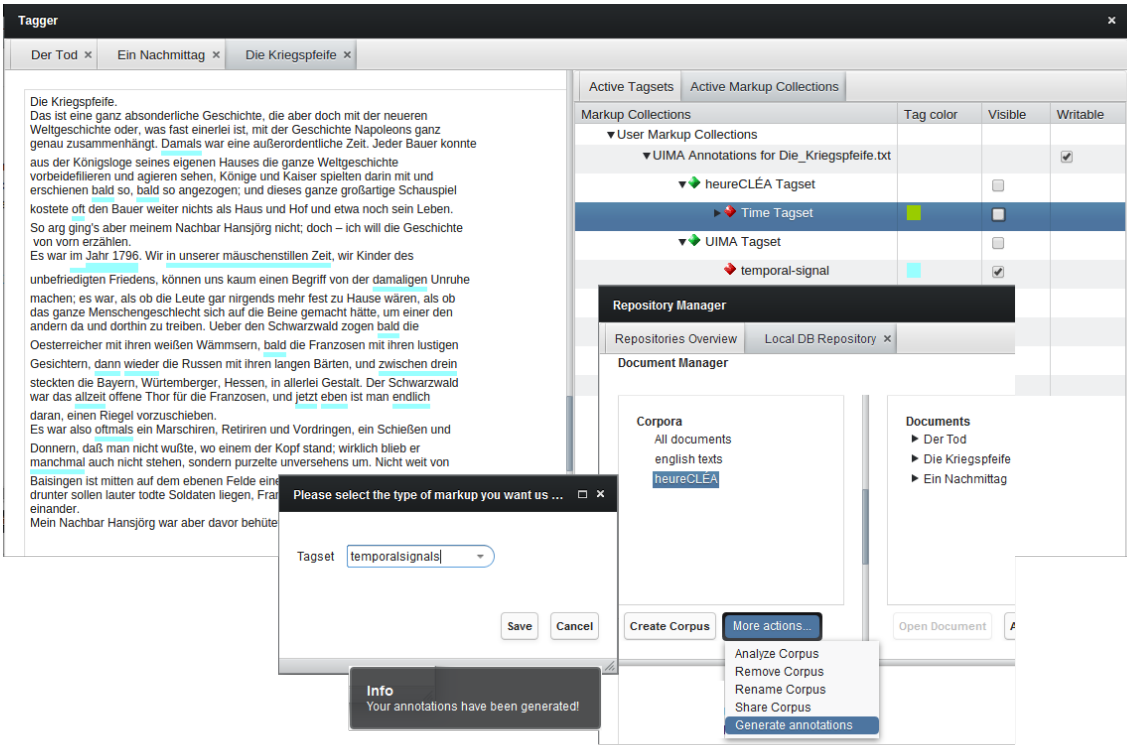
3 Research Workflow within CATMA

The advantage of a direct integration of UIMA into CATMA is best illustrated with an example: in order to analyse the temporal structure of documents (such as order phenomena), many linguistic aspects need to be taken into account. Temporal signals, e.g., calendrical, deictic or relational temporal expressions (Lahn and Meister, 2008), offer a hint for temporal phenomena of order. As manual annotation for these basic linguistic phenomena is laborious, we are currently developing a machine learning system for predicting temporal signals. Figure 2 shows the possibility to create and directly inspect automatic annotations directly within the CATMA interface. With one click, the prediction of our NLP pipeline for temporal signals – or other annotations such as date and time expressions (Strötgen and Gertz, 2013) – can be shown. Note that the system output can easily be compared to any manual annotation as the type systems are completely independent. This flexibility allows scholars to focus on complex phenomena of the text with the possibility of automating simpler annotations. All automatic annotations are, however, non-obtrusive and completely changeable and reversible to give the choice of the level of automation to the user.

References

Lahn, S. and J. C. Meister (2008). *Einführung in die Erzähltextanalyse*. Stuttgart: JB Metzler.

Strötgen, J. and M. Gertz (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation 47*(2), 269–298.



Die Kriegspfeife.
Das ist eine ganz absonderliche Geschichte, die aber doch mit der neueren
Weltgeschichte oder, was fast einerlei ist, mit der Geschichte Napoleons ganz
genau zusammenhängt. Damals war eine außerordentliche Zeit. Jeder Bauer konnte
aus der Königsloge seines eigenen Hauses die ganze Weltgeschichte
vorbeidefilieren und agieren sehen, Könige und Kaiser spielten darin mit und
erschienen bald so, bald so angezogen; und dieses ganze großartige Schauspiel
kostete oft den Bauer weiter nichts als Haus und Hof und etwa noch sein Leben.
So arg ging's aber meinem Nachbar Hansjörg nicht; doch – ich will die Geschichte
von vorn erzählen.
Es war im Jahr 1796. Wir in unserer mäuschenstillen Zeit, wir Kinder des
unbefriedigten Friedens, können uns kaum einen Begriff von der damaligen Unruhe
machen; es war, als ob die Leute gar nirgends mehr fest zu Hause wären, als ob
das ganze Menschengeschlecht sich auf die Beine gemacht hätte, um einer den
andern da und dorthin zu treiben. Ueber den Schwarzwald zogen bald die
Oesterreicher mit ihren weißen Wämmsern, bald die Franzosen mit ihren lustigen
Gesichtern, dann wieder die Russen mit ihren langen Bärten, und zwischen drein
steckten die Bayern, Würtemberger, Hessen, in allerlei Gestalt. Der Schwarzwald
war das allzeit offene Thor für die Franzosen, und jetzt eben ist man endlich
daran, einen Riegel vorzuschieben.
Es war also oftmals ein Marschiren, Retiriren und Vordringen, ein Schießen und
Donnern, daß man nicht wußte, wo einem der Kopf stand; wirklich blieb er
manchmal auch nicht stehen, sondern purzelte unversehens um. Nicht weit von
Baisingen ist mitten auf dem ebenen Felde eine
drunter sollen lauter todte Soldaten liegen, Fran
einander.
Mein Nachbar Hansjörg war aber davor behüte

Tagset: temporalsignals

Info
Your annotations have been generated!

Figure 2: Screenshot showing automatic annotations within CATMA.