**TEXTANNOTATOR: A Browser-based Framework for Annotating Textual Data in Digital Humanities**

Giuseppe Abrami  
Goethe University Frankfurt  
abrami@em.uni-frankfurt.de

Philipp Helfrich  
Goethe University Frankfurt  
helfrich@cs.uni-frankfurt.de

Elias Rieb  
Goethe University Frankfurt  
s0294036@stud.uni-frankfurt.de

Alexander Mehler  
Goethe University Frankfurt  
mehler@em.uni-frankfurt.de

**Abstract**

In the *Digital Humanities* (DH), scholars are supported in their research by the use of digital methods to process increasingly large amounts of data. Here, the task of computer science is to contribute methods for automatically (pre-)processing this data. In *Natural Language Processing* (NLP), preprocessing depends on the methods used and the quality of the underlying models. For this purpose, especially for historical texts, suitable models are rarely available for pre-processing. For this reason, the training of models in DH is becoming increasingly important, which however requires a significant amount of training data.

In order to meet this requirement, we present a tool for annotating natural language texts to enable and simplify the creation of training data – the so-called TEXTANNOTATOR. TEXTANNOTATOR is a browser-based annotation tool, implemented with the JavaScript framework ExtJS\(^1\) and the visualization library d3.js\(^2\), that can process and annotate NLP documents in UIMA \(^3\) (*Unstructured Information Management applications*) format. In order to create UIMA conform documents, TEXTANNOTATOR is supported by the so-called TEXTIMAGER \(^5\), which allows for processing texts in different languages using a wide range of NLP methods (part of speech tagging, named entity recognition etc.). In addition, TEXTIMAGER enables the visualization of text analyses in an various, interactive ways to make the annotation process as versatile as possible. The result of TEXTIMAGER’s preprocessing are XMI documents that are utilized by UIMA which contain all information as stand-off annotations. Based on this tool set, TEXTANNOTATOR contains a set of annotation modules for

- annotating RST structures \(^4\) (classic and cascaded),
- annotating propositional content using the *PropositionAnnotator* (see Fig. 1) and for
- utilizing different knowledge bases (Wikidata, Wikipedia, Wiktionary, GeoNames, GermaNet and the German National Library (GND)) by beans of the so-called *KnowledgeBaseLinker* (see Fig. 2).

In addition, TEXTANNOTATOR is connected to the RESOURCEMANAGER \(^2\), which allows for storing results in XMI format, in a MongoDB (using UIMA DATABASE INTERFACE \(^1\)), and for publishing the results via the CLARIN Virtual Language Observatory\(^3\).

TEXTANNOTATOR is a framework that offers several annotation tools for generating UIMA-compliant training data that can be used to train classifiers in support of the needs of humanities scholars. However, a tool depends on constructive feedback from potential users, which we request by means of discussions during the conference.

---

2. https://d3js.org/
3. https://www.clarin.eu/content/virtual-language-observatory-vlo
Figure 1: Exemplary visualization of the *PropositionAnnotator*. From the predicate, relations to other tokens are created based on semantic roles.

Figure 2: Exemplary visualization of the *KnowledgeBaseLinker*. A visualization inspired by Babelfy\(^5\) was chosen, but with the extension of element annotation. At the same time, implicit relations, such as from Wikidata, are automatically displayed (edges at the bottom of the image).

References


