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.

[1] https://www.sencha.com/products/extjs/

[2] https://d3js.org/

[3] https://www.clarin.eu/content/virtual-language-observatory-vlo

[5] http://babelfy.org

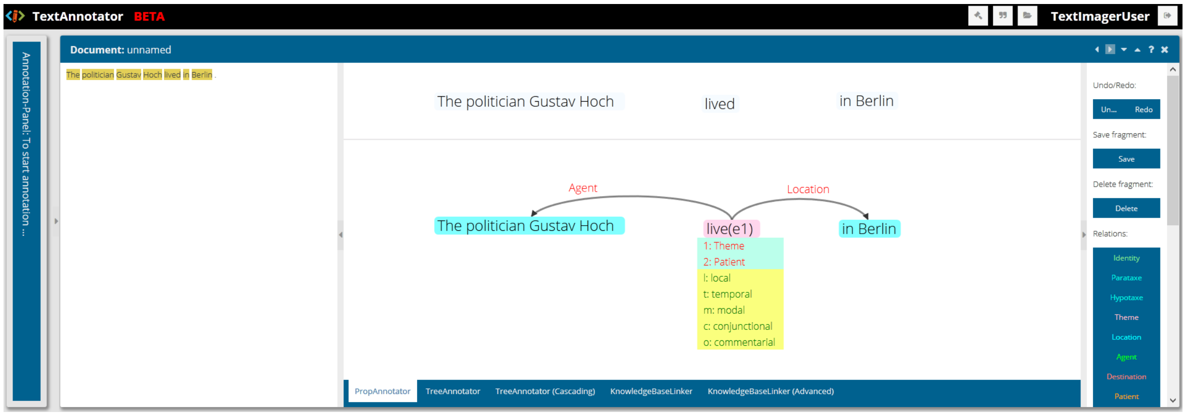
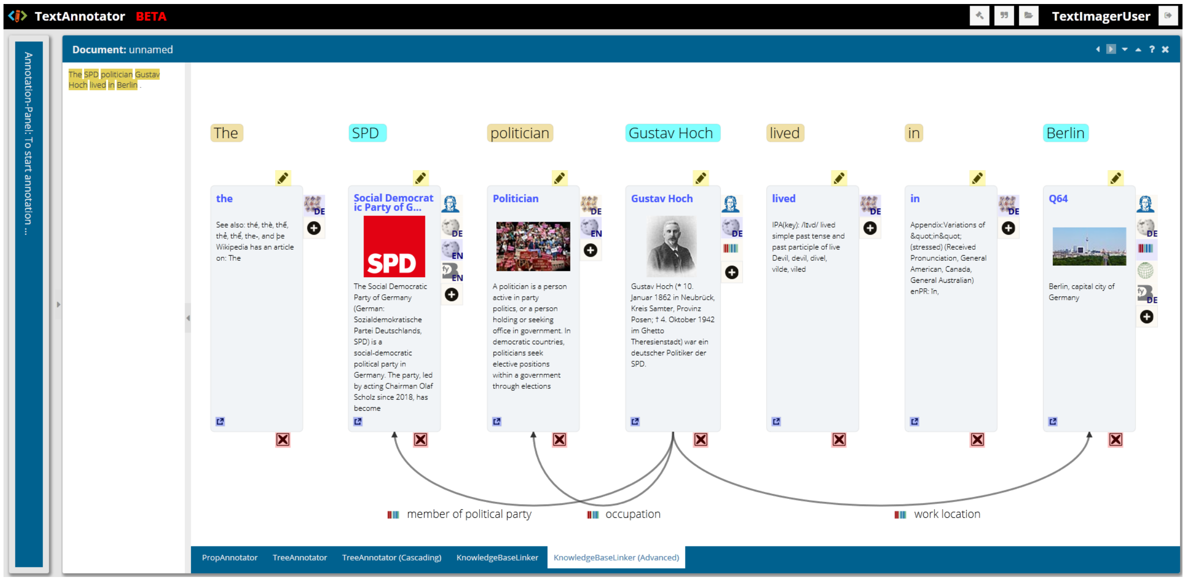


Figure 1: Exemplary visualization of the *PropositionAnnotator*. From the predicate, relations to other tokens are created based on semantic roles.


TextAnnotator BETA
TextImagerUser
Annotation-Panel: To start annotation ...
Document: unnamed
The SPD politician Gustav Hoch lived in Berlin
The
SPD
politician
Gustav Hoch
lived
in
Berlin
the
Social Democratic Party of G...
Politician
Gustav Hoch
lived
in
Q64
DE
EN
member of political party
occupation
work location
PropAnnotator
TreeAnnotator
TreeAnnotator (Cascading)
KnowledgeBaseLinker
KnowledgeBaseLinker (Advanced)

See also: thé, thè, thế, thẻ, thể, the-, and þe Wikipedia has an article on: The

The Social Democratic Party of Germany (German: Sozialdemokratische Partei Deutschlands, SPD) is a social-democratic political party in Germany. The party, led by acting Chairman Olaf Scholz since 2018, has become

A politician is a person active in party politics, or a person holding or seeking office in government. In democratic countries, politicians seek elective positions within a government through elections

Gustav Hoch (* 10. Januar 1862 in Neubrück, Kreis Samter, Provinz Posen; † 4. Oktober 1942 im Ghetto Theresienstadt) war ein deutscher Politiker der SPD.

IPA(key): /lɪvd/ lived simple past tense and past participle of live Devil, devil, divel, vilde, viled

Appendix:Variations of &quot;in&quot; (stressed) (Received Pronunciation, General American, Canada, General Australian) enPR: ĭn,

Berlin, capital city of Germany

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

[1] Giuseppe Abrami and Alexander Mehler. A uima database interface for managing nlp-related text annotations. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, May 7 - 12*, LREC 2018, Miyazaki, Japan, 2018.

[2] Rüdiger Gleim, Alexander Mehler, and Alexandra Ernst. Soa implementation of the ehumanities desktop. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany*, 2012.

[3] T. Götz and O. Suhre. Design and implementation of the UIMA common analysis system. *IBM Systems Journal*, 43(3):476 –489, 2004.

[4] Philipp Helfrich, Elias Rieb, Giuseppe Abrami, Andy Lücking, and Alexander Mehler. Treeannotator: Versatile visual annotation of hierarchical text relations. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, May 7 - 12*, LREC 2018, Miyazaki, Japan, 2018.

[5] Wahed Hemati, Tolga Uslu, and Alexander Mehler. Textimager: a distributed uima-based system for nlp. In *Proceedings of the COLING 2016 System Demonstrations*. Federated Conference on Computer Science and Information Systems, 2016.