Supporting hermeneutic interpretation of historical documents by computational methods

Cristina Vertan
University of Hamburg
Germany
cristina.vertan@uni-hamburg.de

Abstract

In this paper we will introduce a novel framework for data modelling which allows the implementation of tailored annotation tools for a specific digital humanities project. We will illustrate the generic framework model by means of two examples from completely different domains, each treating a different language: the construction of a diachronic corpus for classical Ethiopic texts and the computer-based analysis of originals and translations in three languages of historical documents from the 18th century.

1 Introduction

Digitaization campaigns during the last ten years have made available a considerable number of historical texts. The first digitaization phase concentrated on archiving purposes; thus the annotation was focused on layout and editorial information. The TEI standard includes dedicated modules for this purpose. However, the next phase of digital humanities implies active involvement of computational methods for interpretation and fact discovery within digital historical collections, i.e., active computational support for the hermeneutic interpretation.

We argue that interpretation of historical documents cannot be realised by simple black-box algorithms which rely just on the graphical representation of words but rather by:

1. Considering semantics, which implies a deep annotation of text at several layers;
2. Explicitly annotating vague information;
3. Making use of non-crisp reasoning (e.g., fuzzy logic, rough sets).

For any high-level content analysis, the deep annotation (manual, semi-automatic, or even automatic) is an unavoidable process.

For modern languages there are now established standards and rich tools which ensure an easy annotation process. In this contribution we want to illustrate the challenges and special requirements related to the annotation of historical texts; we argue that in many cases the data model is so complex that tools tailored to the corpus and/or the language still have to be developed.

The annotation of historical texts has to consider following criteria:

- The text to be annotated may change during the annotation. Several scenarios may converge to this situation:
  - Original text is damaged and only the deep annotation and interpretation of neighbouring context can provide a possible reconstruction;
  - The text is a transliteration from another alphabet. In this case transliterations are rarely standardised (also because historical language was not standardised and spelling changes like the insertion of vowels or the doubling of consonants are subject to the interpretation of the annotator and assignment of one or another part-of-speech;
  - The documents are a mixture of several languages and OCR performs poorly.
- The annotation has to be done at several layers: text structure, linguistic, domain-specific. Annotations from different levels may overlap.
- All annotations should consider a degree of imprecision, and vague assertions have to be explicitly marked. Otherwise interpretations of uncertain events may be distorted by crisp yes/no decisions. Vagueness and uncertainty may lead to different branches of the same annotation base.

- Original text and transliteration have to be both kept and synchronised.
- Historical texts lack digital resources. Historical language requires more features for annotation than modern ones. Thus a fully automatic (linguistic) annotation is in many cases impossible. Manual annotation is time consuming, so that functions allowing a controlled semi-automation of the annotation process is more than desirable.
- The annotation tool has to be user-friendly as annotators often lack extensive IT skills.

As none of the currently available annotation tools (e.g., Bollmann et al., 2014; de Castilho et al., 2016) fulfills all of the criteria listed above, many projects decide to alter the data model insted, i.e., features of language, the text, or the domain are not included in the annotation model. This has consequences on the analysis and interpretation process.

In this paper we will introduce a novel framework for data modelling which allows for the implementation of tailored annotation tools for specific DH projects. We will illustrate the generic framework model by means of two examples from completely different domains, each treating another language: first, the construction of a diachronic corpus for classical Ethiopic texts (Vertan et al., 2016) and, second, the computer-based analysis of originals and translations in three languages of historical documents from the 18th century (Vertan et al., 2017). We will present the generic model and show the derived data model for each of the two examples and we will discuss the challenges implied by the development of a new software. We will illustrate also how interchangeability with other digital resourced is assured.

2 Generic Data Model

One of the main requirements of the annotation process for historical data is the possibility of changing the base text without losing the annotation already performed. This requirement leads to the idea that the characters composing the text to be annotated and the annotation itself should be independent one from another and considered as features of an abstract object.

Our generic model is organised around the following notions

1. Annotation Information (AI)
2. Graphical Unit (GU)
3. Annotation Unit (AU)
4. Annotation Span (AS)
5. Annotation Level (AL)

An *Annotation* has two components: An *Annotation Tag* (e.g., part of speech) and an optional number of features, recorded as *[Attribute, Value]* pair (e.g., [Gender, Masculine], [Number, Plural]).

A *Graphical Unit* is the smallest unit one can select with one single operation (mouse click or key combination).

An *Annotation Unit* is any subcomponent of a GU which can hold an annotation. An Annotation Unit can include one or more other annotation units. There are cases in which the AU is identical (from the point of view of borders) with the GU.

An *Annotation Span* is an object holding an annotation and containing at least two AUs. Belonging to two GUs.

Each of these objects can have a label denominating them in the text. For example, a GU is a word in a text, an AU is each letter of the word and a sentence is modelled as an AS. In this way operations on the labels of one object (insertion deletions, replacements of characters) do not affect the already inserted annotation for the respective object.

Links between AU and/or AS objects are ensured through unique IDs. In this way, the model enables also the annotation of discontinuous elements (e.g., a named entity which does not contain adjacent tokens).

An *Annotation Level* is a list of annotation units and annotation spans. The allowed annotation tags belong to a closed list unique for each annotation level.

An annotated text contains one or more annotation levels.

One can differentiate two models which have to be defined:

1. The model for the units which have to be annotated: the graphical units and the annotation units
2. The annotation model, namely the annotation levels, the annotation information allowed for each level as well as the annotation spans

In the next sections we will illustrate through two examples how this model works in practice. In the first example in Section 3 it was implemented for deep annotation for the classical Ethiopic language. In the second example in Section 4 we will show how this framework is currently used as well

for the annotation of linguistic and factual vagueness in texts.

3 Annotation of Classical Ethiopic Texts

3.1 Particularities of Classical Ethiopic

Classical Ethiopic (Gəʿz), belongs to the south Semitic language family. Until the end of the 19th century was one of the most important written language of Christian Ethiopia. Chronologically at the beginning, the rich Christian Ethiopic literature was strongly influenced by translations from Greek and later from Arabic. Later texts develop a local indigenous style. The language plays an important role for the European cultural heritage: early Christian texts, lost or preserved badly or in fragments in other languages are transmitted entirely in classical Ethiopic (e.g., The book of Henoch) (Vertan et al., 2016).

Gəʿz has its own alphabet developed from the south Semitic script. It is a syllable script used also nowadays by several languages from Ethiopia and Eritrea (e.g. Amharic, Tigrinya). A particular feature for the Semitic language family is the left-to-right language direction. Also in contrast with most other Semitic languages it is completely vocalized (i.e., the vowels are always written). This leads also to the problem that morphemes boundaries cannot be visualised. Sometimes only the vowel within a syllable represents a part of speech and has to be tokenised and annotated (e.g., in the word ቤቱ፡ 'his house' /be·tu/ the /u/ is a pronominal suffix and the tokenisation is thus *bet-u*.

3.2 Annotation Challenges

Such annotation can be done only on the transcription level. Annotations at other levels (e.g., text divisions, editiorial markup) have to be done on the original script. This implies that original and transcription have to be fully synchronised in the annotation tool.

The transcription of the original script can follow a rule-based approach. In contrast the transliteration (e.g., doubling a consonant) can be done on the basis of the transcription, just manually. In many cases the correct transliteration can be decided only after morphological analysis and disambiguation. Thus the annotation tool has to be robust in the face of changes of the text during the annotation process. This is a very important feature but also an enormous challenge for any annotation tool.

A diachronic language analysis (as it is required in order to see the development over centuries of classical Ethiopic) can be done only if the linguistic analysis is deep. Usually changes in the language can be observed first in detail and then at a macro level. For classical Ethiopic the linguistic POS tagset has 33 elements, each with a number of features.

Given the fact that no training data exist, a manual annotation is unavoidable. However, the tool we developed provides a mechanism of controlled automatic annotation, which at one hand speeds up the process and on the other hand leaves the final decision on disambiguation to the user.

3.3 The Annotation Model

A Graphical Unit (GU) represents a sequence of Gəʿz characters ending with the Gəʿz separator ፡. The punctuation mark ፡፡ is always considered a GU. Tokens are the smallest annotatable units with a meaning of their own to which a lemma can be assigned. Token objects are composed of several transcription letter objects

For example, the GU object ወይቤሎ፡ represents also an Annotation Unit and contains the 4 Gəʿz letter objects modelled as AUs; ወ, ይ, ቤ, ሎ. Each of these objects contains the corresponding transcription letter objects modelled also as AUs, namely:

- ወ contains the transcription letter objects: *w* and *a*
- ይ contains the transcription letter objects: *y* and *ə*
- ቤ contains the transcription letter objects: *b* and *e*
- ሎ contains the transcription letter objects: *l* and *o*

Throughout the transliteration-tokenisation phase, three token objects (in our model also AUs) are built: *wa*, *yəbel*, and *o*.

Finally, the initial GU object will have attached two labels: ወይቤሎ and *wa-yəbel-o*. For synchronisation reasons we consider the word separator ፡ as property attached to the Gəʿz character object ሎ. Each Token-Object records the IDs of the transcription letter object that it contains.

Morphological annotation objects are attached to one token object. They consist of a tag (e.g., the POS "Common Noun") and a list of attribute-value pairs where the key is the name of the morphological feature (e.g., number). In this way, the tool is

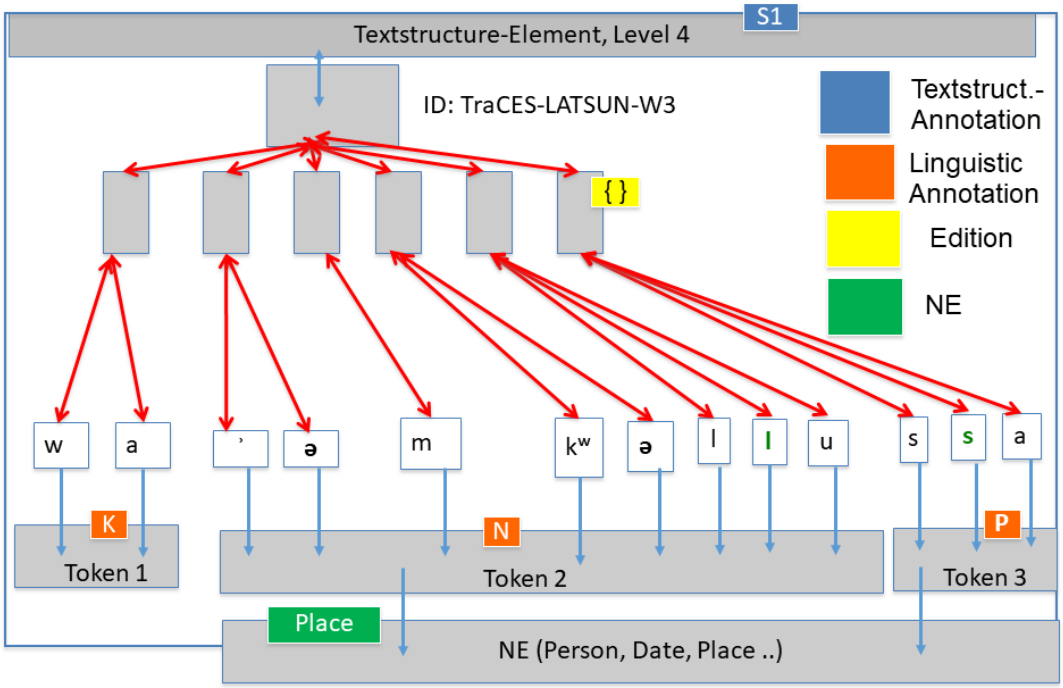

Figure 1: Annotation Model for classical Ethiopic

robust with respect to the addition of new morphological features or POS tags.

As the correspondences between the Gəʿz-character and the transcriptions are unique, the system stores just the labels of the Transcription-letter objects. All other object labels (Token, Gəʿz-character and GU) are dynamically generated throughout a given correspondence table and the Ids. In this way the system uses less memory and it remains error prone during the transliteration process. In Figure 1 we present the entire data model, including also the other possible annotation levels. The GeTa-tool implementing this model is a client-application, written in Java and distributed as open-source software.

4 Annotation of vagueness and uncertainty in historical texts

The second example discusses the annotation of historical texts from the 18th century for which we want to mark:

1. Uncertain characters or words (not entirely deciphered from the manuscript);
2. Uncertain dates, places persons and if possible their mapping on a corresponding knowledge base;
3. Vague linguistic expressions;
4. Indicators for source quotations;
5. Text structure;
6. Linguistic annotation.

We define six Annotation Levels. The Graphical Unit is a word in the text, i.e., a string delimited by spaces. Punctuation is separated in a preprocessing step as independent words.

Annotation Units are words, a single letter or a group of letters inside one word. Annotation spans will be in this case necessary for representing named entities (places, persons, etc.), text structure, or vague linguistic expressions. Especially for vague expressions it is extremely important that the model supports discontinuous elements to be part of the same annotation.

To each Annotation Span or Annotation Unit we attach Annotation Information containing Attribute–Value pairs related to the degree of uncertainty (fuzzy value), type of linguistic vagueness and source of quotation, respectively, and the

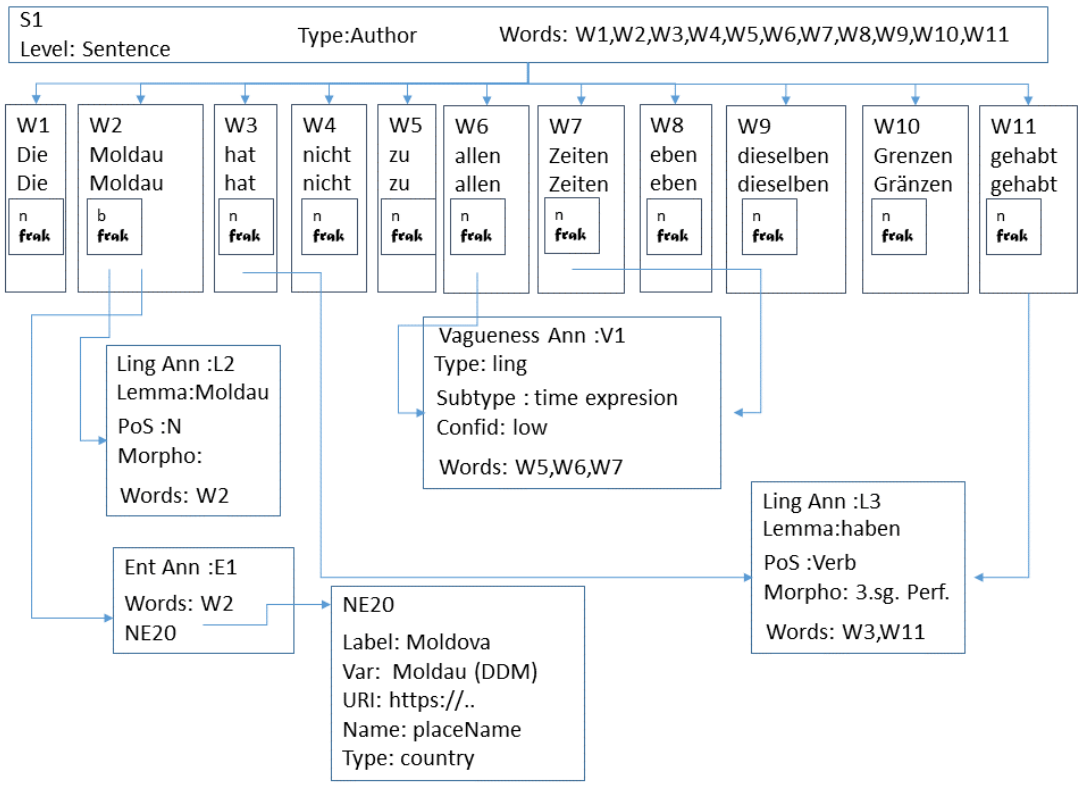

Figure 2: Annotation Model in HerCoRe

trust value of this source. An example of such Annotation is presented in Figure 2.

The aim of such annotations is not to develop an expert system in the classic way, as known from artificial intelligence. Such expert systems assume that the computer is reasoning and presents its interpretation to the user. We consider that for interpretation of historical facts such system is not reliable enough. The background knowledge necessary for producing reliable result is huge and relies often either on materials which are not available in digital form. Thus our goal is rather to make the user aware that:

1. There is a number of possible answers to one query, and
2. these possible answers may have different degrees of reliability (i.e., they are not necessarily true).

The interpretation and the final decision is left entirely to the user.

5 Conclusions

The annotation model introduced in Section 2 and exemplified in Sections 3 and 4 is flexible and supports changes of the text to be annotated during the annotation process. Of course the results of these changes must remain consistent with the annotation. This is the responsibility of the annotator (i.e., if the user changes completely the label of an Annotation Unit he must ask himself if the new label still corresponds to the annotation). In the particular examples presented in Sections 3 and 4 we encode the model as JSON objects. This allows us to keep the required storage space small and ensures fast access to the data. However, we provide export to other, in particular XML-based, formats, which ensures interoperability with other analysis tools such as ANNIS or Voyant. Further work included the implementation of the generic model for the annotation of inscriptions of Classical Maya.

Acknowledgements

This article presents work performed within two projects: the work in Section 3 was performed with the TraCES project (From Translation to Creation: Changes in the Ethiopic Lexicon and Style from Late Antiquity to the Middle Ages) supported by the European Research Council. Work performed in this project was performed together with Alessandro Bausi, Wolfgang Dickhut, Andreas

Ellwardt, Susanne Hummel, Vitagrazia Pissani, and Eugenia Sokolinski. The work in Section 4 is currently performed within the project HerCoRe (Hermeneutic and Computer-based Analysis of Reliability, Consistency and Vagueness in historical Texts) funded by the Volkswagen Foundation within the framework "Mixed Methods in Humanities"). Work reported in this section was done in collaboration with Walther v. Hahn and Alptug Güney.

References

Bollmann, Marcel, Florian Petran, Stefanie Dipper, and Julia Krasselt (2014). CorA: A web-based annotation tool for historical and other non-standard language data. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014)*, pages 86–90. URL https://aclweb.org/anthology/W14-0612.

de Castilho, Richard Eckart, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the LT4DH workshop at COLING 2016*, pages 76–84. URL https://aclweb.org/anthology/W16-4011.

Vertan, Cristina, Andreas Ellwardt, and Susanne Hummel (2016). Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte. In *Proceedings der DHd-Konferenz 2016*. URL http://www.dhd2016.de/abstracts/vortr%C3%A4ge-061.html.

Vertan, Cristina, Walther von Hahn, and Anca Dinu (2017). On the annotation of vague expressions: a case study on Romanian historical texts. In *Proceedings of the first Workshop on Language Technology for Digital Humanities in Central and (South-)Eastern Europe, in association with RANLP 2017*, pages 24–31. doi:10.26615/978-954-452-049-6_028.