Data Modelling for Historical Corpus Annotation

Cristina Vertan

Research Group „Computerphilologie“, University of Hamburg

Abstract

In this paper we argue that data modelling is not just a matter of adapting humanities data to a given tool but more the reverse process: analyse the data, built a data model and if necessary built the adequate tool which fully supports the data model and the user requirements. We will illustrate this for the case of the annotation of a diachronic corpus for classical Ethiopic texts.

1 Introduction

Digitalization campaigns during the last ten years made available a considerable number of historical texts. The first digitalization phase concentrated on archiving purposes; thus the annotation was focused on layout ad editorial information. The TEI standard developed 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.

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

For modern languages there are meanwhile established standards and rich tools which ensure an easy and error –prone annotation process. In this contribution we want to illustrate the challenges and special requirements connected with the annotation of historical texts, and argue that in many cases the data-model is so complex that corpus, respectively language tailored tools have still to be developed.

The annotation of historical texts has to consider the following criteria:

- The text to be annotated may change during the annotation. Several scenarios may converge to this situation:

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2018 in
M. Burghardt & C. Müller-Birn (Hrsg.): INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin
 https://doi.org/10.18420/infdh2018-17

- original text is damaged and only the deep annotation and interpretation of neighbouring context can provide a possible reconstruction;
- The text is a transliteration[1] from another alphabet. In this case transliterations are rarely standardised (also because historical language was not standardised and phonetical changes like insertion of vowels, doubling of consonants are subject of the interpretation of the annotator and assignment of one part-of-speech or other.
- The documents are a mixture of several languages and OCR performs low.
- The annotation has to be done at several layers: text structure, linguistic, domain –specific. Annotations from different levels my overlap.
- All annotation should consider a degree of impreciseness and vague assertions have to be marked. Otherwise interpretations of doubtable events are falsified 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-automatic annotation process are more than desirable.

The annotation tool has to be user-friendly as annotators do not have often deep IT-skills

As none of the current widespread annotation tools (Bollman & Petran & Dipper 2014), (de Castilho et. al. 2016) fulfils all criteria above, many projects alter the data model, i.e. features of language or of the text respectively domain are not included in the annotation model. This has consequences on the analysis and interpretation process.

In this paper we argue that data modelling is not just a matter of adapting humanities data to a given tool but more the reverse process: analyse the data, build a data model and if necessary build the adequate tool which fully supports the data model and the user requirements. We will illustrate this for the case of the annotation of a diachronic corpus for classical Ethiopic texts (Vertan & Ellwardt & Hummel 2016).

[1] Transcription=rewritting of text according to some predefined rules independent on morpho-syntactic features (can be done automatically). Transliteration is the process of adapting the transcription dependent of morphological and syntactical functions of the words. In our case this can be e.g. the duplication of a consonant or the insertion/deletion of a ə.

2 Special Requirements for the Annotation of classical Ethiopic

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

Gəʿəz has its own alphabet developed from the couth 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 written always). This leads also to the problem that morphemes borders 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: /be·tu/ the /u/ is a pronominal Suffix (his house) and the tokenisation is bet-u

Such annotation can be done only at the transcription level. Annotations at other levels (e.g. Text Divisions, Edition –Mark-ups) 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 the morphological analysis and disambiguation. Thus the annotation tool has to be robust to changes of the text during the annotation process. This is a very important feature but also a big challenge for any annotation tool.

A diachronic language analysis 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. The annotation tool must present the user the annotation options in a readable way. (e.g. CorA builds list with all possible combinations of features and values, procedure which I this case is from the user point of view impossible)

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

The GeTa (Gəʿəz Annotation) is based on a tree oriented data structure which we will present in the next section.

3 Underlying Data Model

The data model of the GeTa Tool follows an object-oriented approach. Each object can be located by a unique Id. There are two types of objects:

- Annotated Objects namely: Graphical Units, Tokens, Gəʿəz-characters and Transcription-letters.
- Annotation Objects (spans) which are attached to one or more Annotation-Objects; these are: morphological annotations, text divisions, editorial annotations.

Links between Annotated- and Annotation-Objects are ensured through the Ids. In this way the model enables also the annotation of discontinuous elements (e.g. a Named Entity which does not contain adjacent tokens).

A Graphical Unit (GU) represents a sequence of Gəʿəz-characters ending with the Gəʿəz-separator (፡). The punctuation mark (።) is considered always a GU. Tokens are the smallest annotatable units with an own meaning, for which a lemma can be assigned. Token objects are composed of several Transcription-letter objects.

E.g. The GU- Object ወይቤሎ፡ contains:

- The 4 Gəʿəz –letter objects; ወ, ይ, ቤ, ሎ. Each of these objects contains the corresponding Transcription-letter objects, 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 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 Gəʿəz-character object ሎ.

Each Token-0bject records the Ids of Transcription-letter object which he contains.

Morphological annotation objects are attached to one Token-object. They consist of a tag (the PoS e.g. Common Noun) and a list of key-value pairs where the key is the name of the morphological feature (e.g. number). In this way the tool is robust to addition of new morphological features or PoS tags.

As the correspondences between the Gəʿəz-character and the transcriptions are unique, the system stores just the labels of the Transcription-letter objects. All other object labels (Token, Gəʿəz-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 3 we present the entire data model, including also the other possible annotation levels.

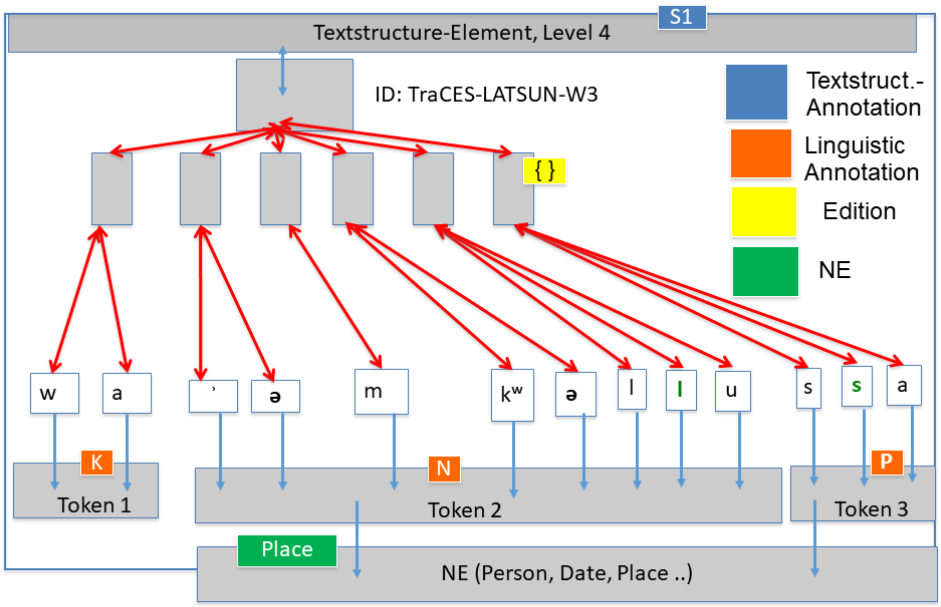

Figure 1. The GeTa Data -model

4 Interoperability and further work

GeTa is a tailored tool for annotation of Gəʿəz texts which enables a deep fine-grained linguistic annotation as well as annotation at other levels. The controlled semi-automatic annotation speeds up the mark-up process but at the same time leaves entirely full-control to the annotator. Units annotated or tokenised automatically are marked in such way that the user knows any time if a manual check is necessary. E.g.: automatic generated tokens are displayed in italic, automatic annotated tokens are marked in red.

A number of well-defined changes on the transcription enables the transliteration at any stage during the annotation process. The annotation tool is written in Java 1.8 and is platform independent. The genuine format of the output is JSON. We implemented export functionalities to plain text (TXT) and TEI/XML so that the results can be imported easily to other analytic tools like Voyant[2]. A special convertor to ANNIS -format[3] was implemented (Druskat and Vertan

2 https://voyant-tools.org/

3 http://corpus-tools.org/annis/

2018), so that the annotated corpus can be analysed with the powerful mechanism of ANNIS. The corpus will be freely accessible for further research through the ANNIS-Installation provided by the Zentrum für Sprachkorpora[4]. The TEI-export will be used for integration with the data available in the project Beta maṣāḥəft[5].

The tool is able to handle also Gəʿəz texts written with the South-Arabic alphabet with right-left writing direction (early inscriptions). Further work concerns a complete check and adaptation of all functionalities for this alphabet, as well as for non-vocalized versions of Gəʿəz texts.

The annotated texts will be used as training material; rules for transliteration, tokenisation and annotation may be extracted and used for a more advanced automatic annotation process.

The underlying data mode for GeTa was refined and adapted for two other completely different scenarios: the annotation of classical Maya database of inscriptions and texts and the computer-based analysis of original and translation in three languages of historical documents from the 18th century (Vertan & v. Hahn & Dinu 2017). These two implementations provide also mechanisms for handling uncertainty. The framework is thus enough generic to handle completely different data sets.

References

Bollmann, Marcel & Petran, Florian & Dipper, Stefanie & Krasselt, Julia. (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)*. Gothenburg, S. 86-90.

Druskat, Stephan &Vertan, Cristina. (2017). Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora. In Vogeler, Georg (ed.): *Kritik der Digitaler Vernunft Konferenzabstracts*, Köln. S. 270-273.

Eckart de Castilho & Richard & Mújdricza-Maydt, Éva & Yimam, Seid Muhie & Hartmann, Silvana & Gurevych, Iryna & Frank, Anette & Biemann, Chris. (2016): *A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures*. In: *Proceedings of the LT4DH workshop at COLING 2016*, Osaka. S. 76-84.

Vertan, Cristina & Ellwardt & Andreas & Hummel & Susanne. (2016). Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte. In: *Proceedings der DHd-Konferenz 2016*, http://www.dhd2016.de/abstracts/vortr%C3%A4ge-061.html [last accessed 25.09.2017].

Vertan, Cristina & Hahn, Walther von & Dinu, Anca (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*, Varna, S. 24-31.

4 https://corpora.uni-hamburg.de/hzsk/

5 https://www.betamasaheft.uni-hamburg.de/