

# GeTa - A Tool for the Controlled Semi-Automatic Multilevel Annotation of Classical Ethiopic

**Cristina Vertan**

University of Hamburg  
VogtKölln Strasse 20, 22527 Hamburg  
cristina.vertan@uni-hamburg.de

## Abstract

Preservation of the cultural heritage by means of digital methods became extremely popular during last years. After intensive digitization campaigns the focus moves slowly from the genuine preservation (i.e digital archiving together with standard search mechanisms) to research-oriented usage of materials available electronically. This usage is intended to go far beyond simple reading of digitized materials; researchers should be able to gain new insights in materials, discover new facts by means of tools relying on innovative algorithms. In this article we will describe the workflow necessary for the annotation of a diachronic corpus of classical Ethiopic, language of essential importance for the study of Early Christianity

## Rezumat (Romanian)

Conservarea patrimoniului cultural prin intermediul metodelor digitale a devenit extrem de popular în ultimii ani. După campaniile intensive de digitizare, atenția cercetătorilor se deplasează încet de la conservarea autentică (adică arhivare digitală împreună cu mecanisme de căutare standard) la utilizarea orientată spre cercetare a materialelor disponibile pe cale electronică. Această utilizare dorește să patrundă mult dincolo de simpla citire a materialelor digitizate; Cercetătorii ar trebui să poată descoperi noi elemente, prin intermediul instrumentelor care se bazează pe algoritmi inovativi. În acest articol vom descrie fluxul de lucru necesar pentru adnotarea unui corpus diacronic în Ge'ez, limba etiopiană clasică, o limbă de importanță esențială pentru studiul creștinismului timpuriu.

**Keywords:** annotation, classical Ethiopic, south-semitic language3

## 1 Introduction

Although of major importance for the understanding of Christian Orient, the Gə'əz language was up to now somehow neglected by the new research directions in Digital Humanities. Substantial material in digital form exist, but there are no tools which allow a deep analysis of the language and the content.

Improving our knowledge of the Gə'əz language is crucial in order to refine our philological and text-critical methods as well as for advancing our understanding of thought and literature expressed in Gə'əz.

This implies a substantial enlargement of the data by:

- seizing Classical Ethiopic texts in digital form
- adding significant linguistic information
- collecting metadata
- providing tools to interpret all this information.

The project TraCES<sup>1</sup> (From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages) aims to fill this gap by providing a collection of reliable and extensive linguistic data based on annotated diachronic corpus of Gə'əz. The annotation and the developed tools will enable analysis at the level of

lexicography, morphology and style. The annotated texts belong to different periods and genres of Ethiopic literature (text-critical editions). The project employs a multidisciplinary approach, involving methods from linguistics, philology and digital humanities. Major results expected to bring Gə'əz in the digital era are:

- a (deep) annotated corpus linked with
- a lexicon (first digital lexicon for Gə'əz)
- tools for the annotation, analysis, and visualization of the corpus, and browsing the lexicon.

In this paper we will focus on the description of the annotation tool. We will explain the requirements and the challenges these requirements imply for the tool development, and we will present its components, the underlying data structure as well as the linguistic -set.

## 2 Challenges of Gə'əz language for digital tools

The digital annotation and analysis of any corpus, implies several steps:

- The identification of punctuation marks
- The identification of independent tokens (Tokenisation). By token we denote the smallest

<sup>1</sup> Funded through the ERC Research Grant 2014-2019 (<http://www.traces.uni-hamburg.de/about.html>)

unit to which one can assign a part-of-speech (PoS).

- The division of the text in sentences.
- The construction of a linguistic tag-set (PoS + possibly attached features and their values)
- The annotation of these features as well as attaching to each word a lemma, and a link to a language lexicon

The Gə'əz language belongs for the moment to the group of "very low resourced languages", i.e. languages which face a significant lack of resources (corpora, lexicons, terminological data bases, Thesemantic networks) and tools. (Maegard and Krauwer 2006) defines the minimum set of such resourced and tools which are necessary to insert one language on the digital map. Usually the problematic of (very) low resourced languages is solved through adaptation of existent material for other languages within the same family. In the case of Gə'əz this is not possible due to several issues:

- Within the semitic language family the situation is better for Arabic and Hebrew. However classical variants of these languages are as well under-resourced. The particularities of Gə'əz writing system (alphabet, left-right writing) make impossible any adaptation
- From the point of view of the writing system Amharic seem to be the best next candidate for an adaptation. Amharic lacks itself language resources and tools. Additionally the morphological structure differs in many points from that one of Gə'əz

There are a number of tools which claim to be language independent. These are tool developed with a statistical paradigm: very large language corpora are used and linguistic feature are learned from those. This paradigm cannot be followed for the moment for Gə'əz as there exist no statistically relevant Corpus for classical Ethiopic. Additionally machine learning methods are quite performant when the number of features to be learned is rather small. This is not the case of Gə'əz, for which we identified over 30 OPoS (Hummet and Druskat 2017) together with various features to be annotated.

An additional challenge is the absence of an electronic dictionary (lexicon) for Gə'əz. Usually this is the first electronic resource to be developed for a language. Lexicons give important information about the lemma, the root as well as morphological features. The TraCES project builds the lexicon and the annotated corpus in parallel. This means that there is a bidirectional link between these 2 resources: already existing lemmas are marked in the lexicon but also new found words from the corpus are inserted (together with lemma and morphological information) into the lexicon.

A fully automatic annotation process is therefore for Gə'əz impossible at this stage. We adopt a 2-stage workflow:

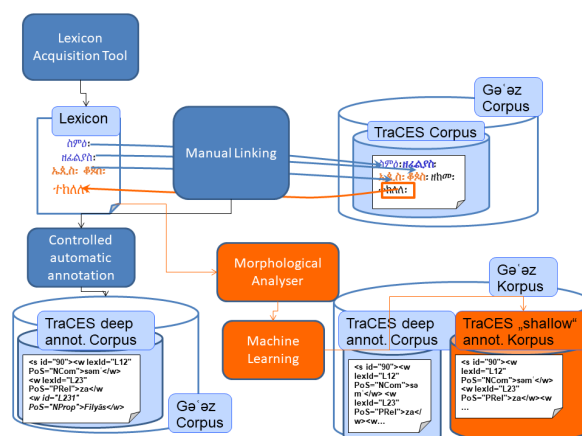


Figure 1 TraCES Modules for linguistic annotation

1. In a first stage a manual deep-annotated corpus is built. The manual Annotation is speeded-up by a controlled semi-automatic component, which will be explained in section 3
2. In a second stage the deep annotated corpus will be used as training material for a machine learning algorithm.

The complete architecture, including also the links to the lexicon component is presented in figure 1.

During last years several language-independent, respectively language customizable annotation tools were made available for researchers in humanities. Among those the most used are WebAnno (de Casthilo et. al 2014)) and CorrA (Bollmann et al. 20174) However a certain specificities of Gə'əz made not possible the usage of these tools. In this section we will list these specificities and explain how they influenced the decisions taken for Annotation.

#### i) PoS Tagset

As mentioned the final goal of the TraCES project is to provide a framework which makes possible a diachronic analysis of this language. As usually variations in language occur at the micro and not the macro level, we need to perform a deep annotation which implies: a fine-grained PoS tag-set together with very precise and detailed features for each PoS. We defined a set of 30 PoS, grouped as follows:

- Nominals
  - Nouns: Common Noun, Proper Name
  - Pronouns: Independent Personal Pronoun, Pronominal Suffix, Subject Pronoun Base, Object Pronoun Base, Possessive Pronoun Base, Demonstrative Pronoun, Relative Pronoun, Interrogative Pronoun, Pronoun of Totality Base, Pronoun of Solitude Base
  - Numerals: Cardinal Numeral, Ordinal Numeral
  - Verb

- Existentials: Existential Affirmative Base, Existential Negative Base
- Particle
  - Adverbs: Interrogative Adverb, Other Adverb
  - Preposition
  - Conjunction
  - Interjection
  - FurtherParticles: Accusative Particle, Affirmative Particle, Deictic Imperative Particle, Interrogative Particle, Negative Particle, Presentational Particle Base, Quotative Particle, Vocative Particle, Other Particle
- Foreign material
- Punctuation

The inclusion of different types of particles like Prepositions and Conjunctions or relative pronouns makes imperative a splitting of Gə'əz word units in tokens e. g.

The word unit *ḥzāfilyās* (zafilyās) will be split in *ḥ:(za)* as relative pronoun and *zāfilyās* as proper noun.

A more challenging issue is the annotation of pronominal suffixes which can be in fact marked just in the transliteration like in the following example:

The word unit *ba'āsuru* transliterated as *ba'āsuru* has the following tokens: *ba* (Preposition), *āsur* (common noun) and *u* (pronominal suffix). However the pronominal suffix *u* is part of transliteration of the Gə'əz letter (ʿ). Thus an annotation of such part of part of speech can be done only on transliterations.

The linguistic annotation is just part of a more complex annotation as several layers (text structure, editorial marks, named entities like persons, places, date) some of them being more appealing if they are inserted in the original script.

The annotation tool must handle in parallel the text in its original form (fidāl) and transliteration

ii) Transliteration process

Given the motivation under i) we need for all texts their transliterated version. Time constraints make impossible a manual transliteration. On the other hand a fully automatic transliteration cannot handle (without apriori knowledge) phenomena like disambiguation of 6<sup>th</sup> grade (ə) or gemmination. There are no clear linguistic rules which could cover all cases. Moreover, even some rules may imply linguistic information, which at the moment of the transliteration is not available to the system. Unsupervised machine learning approaches (without training material) will not perform satisfactory as we do not have any big corpus in both fidāl and transliteration.

Thus the annotation tool may support a kind of controlled semi-automatic transliteration 2 stages: first a rough transliteration, based on the general accepted

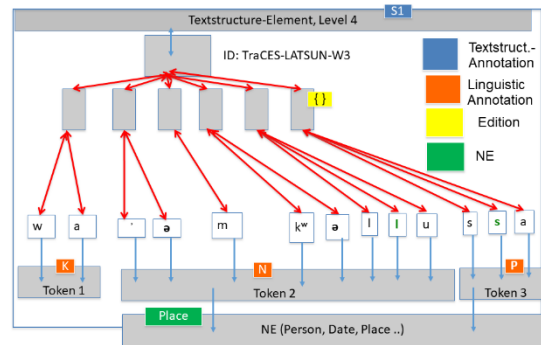
transliteration rules is performed automatically. In a second stage corrections are done in a semi-automatic manner. We will explain this in section 4.

The gemmination or disambiguation of 6<sup>th</sup> grade are linguistic motivated processes. From the technical point of view the linguistic annotation is preceded by a tokenisation process (splitting of word units in tokens). As consequence a gemmination (e.g.) may occur only after the PoS and its features are decided.

### 3 Underlying DataModel

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 Figure 2 GeTa Data -model objects:

Annotated Objects namely: Graphical Units, Tokens, Gə'əz-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 Gə'əz-characters ending with the Gə'əz-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 *ḥzāfilyās* contains

the 4 Gə'əz-letter objects ; *ḥ, z, ā, y*. Each of these objects contains the corresponding Transcription-letter objects, namely:

- *ḥ* contains the Transcription-letter objects: *w* and *a*

- ያ contains the Transcription-letter objects: *y* and *a*
- 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*, *yābel*, and *o*

Finally, the initial GU-Object will have attached two labels: ወይሎ and *wa-yābel- o*. For synchronisation reasons we consider the word separator (፣) as property attached to the Gəʼz-character object ሎ.

Each Token-Object 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 Gəʼz-character and the transcriptions are unique, the system stores just the labels of the Transcription-letter objects. All other object labels (Token, Gəʼz-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.

## References

Bollmann, Marcel and Petran, Florian and Dipper,Stefanie and Krasselt, Julia 2014: *CorA: A web-based annotation tool for historical and other nonstandard language data*, in:Kalliopi Zervanou and Cristina Vertan and Antal van den Bosch and Caroline Sporleder (Eds.), Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) Gothenburg, Sweden April 2014, 86-90.

Druskat, Stephan and Vertan, Cristina 2017, ' *Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora* ', in Gerog Vogeler (ed.) Kritik der Digitaler Vernunft Konferenzabstracts, Köln 2018, 270-273

Eckart de Castilho, Richard and Mújdricza-Maydt, Éva and Yiman, Seid Muhie and Hartmann,Silvana and Iryna and Frank, Anette and Biemann, Chris 2016, 'A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures', in Erhard W. Hinrichs and Marie 39

Hinrichs, and Thorsten Trippel (eds.), *Proceedings of the LT4DH workshop at COLING 2016*, Osaka, Japan: 76-84.

Hummel, Susanne and Wolfgang Dickhut 2016. 'A part of speech tag set for Ancient Ethiopic', in Alessandro Bausi and Eugenia Sokolinski, eds, *150 Years after Dillmann's Lexicon: Perspectives and Challenges of Gəʼaz Studies*, Supplement to *Aethiopica*, 5 (Wiesbaden: Harrassowitz Verlag, 2016), 17–29.

Krzyżanowska, Magdalena 2017. 'A Part-of-Speech Tagset for Morphosyntactic Tagging of Amharic', *Aethiopica*, 20 (2017), 210–235.

Maegaard, Bente and Krauwer, Steven and Choukri, Khalid and Jørgensen, Lars, 2006, 'The BLARK concept and BLARK for Arabic', in *Proceedings of the LREC Conference 2006*, <http://lrec-conf.org/proceedings/lrec20>