Multi-level Digital Annotation of Ethiopic Texts*

Susanne Hummel, Vitagrazia Pisani, and Cristina Vertan, Universität Hamburg

The GeTa tool has been developed at Hamburg to address the challenge of tokenization and multi-level annotation of Ethiopic texts, with the aim of further computer-assisted analysis of the morphology and lexicography of the Gəʿəz language. The paper illustrates the workflow of linguistic annotation with the help of the tool.

1. Introduction

Although of major importance for the study of Christian Orient, the Gəʿəz language (also known as Classical Ethiopic) has been so far neglected by the new research trends in Digital Humanities. While some Gəʿəz texts exist in digital form, there are no tools to assist their linguistic analysis. The project *TraCES*: *From Translation to Creation*: *Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages* aims at addressing this desideratum by the development of a complex annotation tool which allows the production of coherent, reliable, and extensive linguistic data. The tool (called GeTa for Gəʿəz Text Annotation) is used to annotate a pre-selected corpus of texts: several texts belonging to different periods and genres of Ethiopic literature, original and translated have been singled out.[1] Each text is (in full or in part) annotated at different levels. The main level is formed by the detailed linguistic (part-of-speech) annotation ('deep annotation' in the project's terminology), where each word is linked to the corresponding dictionary entry. We also annotate named entities such as persons, places, dates, titles of work, or offices. Furthermore, we mark up the text structure (e.g. parts, chapters, sentences, verses). Special features related to the edition, like editorial intervention such as conjectures, are marked upon occurrence.

The GeTa tool and the data will be made freely available to enable a systematic, diachronic analysis of the Gəʿəz language, including its lexicography, morphology, and style.

In this paper we focus on the workflow of linguistic annotation, and discuss the requirements and challenges posed by the annotation process for the tool development. We also briefly present the tool's components and the underlying data structure.

* The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme, grant agreement no. 338756 (ERC Advanced Grant TraCES).

1 See the project note by Eugenia Sokolinski in this issue.

2. Challenges of the Gəʿəz language for digital tools

As of today, from the computational linguistics point of view, the Gəʿəz language belongs to the group of 'very low-resource languages',[2] i.e. languages with a significant lack of resources (corpora, lexicons, terminological data bases, semantic networks) and tools. Often, low-resource languages can be helped by adapting tools and materials existing for other languages within the same family. In the case of Gəʿəz, this was not possible. Better-resource Semitic languages (such as Hebrew or Arabic) use a different writing system (right-to-left consonantal writing against the left-to-right syllabic writing for Ethiopic). Amharic uses the same writing system, but its morphological structure differs in many aspects from that of Gəʿəz.[3] Besides, all these languages are still low resource, and the available tools and data are very limited.

A number of tools claim to be language independent. They incorporate data from very large language corpora, so that linguistic features can be elicited, and learnt, from the data. This statistical paradigm cannot be followed for the moment for Gəʿəz as there exists no significant corpus for Classical Ethiopic. Additionally, machine learning methods perform best when the number of features to be learnt is limited. This is not the case of Gəʿəz, for which we have identified 33 part-of-speech tags that can be accompanied by various features, the number of possible combinations going in several hundred (see § 3 below).

An additional challenge is the absence of an electronic dictionary (lexicon) for Gəʿəz. Usually a dictionary is the first digital resource to be developed for a language. Lexicons give important information about the lemma, the root, and morphological features. The *TraCES* project has to build up lexicon and annotated corpus in parallel. This means that before a word in the corpus can be linked to the lexicon, unless it is already present in the initial word-list, the corresponding lemma (with the morphological information, translation, examples) has to be created.

A fully automatic annotation is therefore impossible for Gəʿəz at this stage. We adopt a two-stage workflow: (1) at a first stage, texts are manually supplied with detailed linguistic annotation ('deep annotation'). The process is facilitated by a controlled semi-automatic component (batch annotation, see § 3 below); (2) at a second stage, the annotated corpus will be used as training material for a machine learning algorithm. The complete architecture, including the links to the lexicon component, is illustrated in fig. 1.

2 See Maegaard et. al 2006 for the definition of a minimum set of resources and tools which are necessary to insert a language on the digital map.

3 For further details and a morphosyntactic tagset for the Amharic see Krzyżanowska 2017.

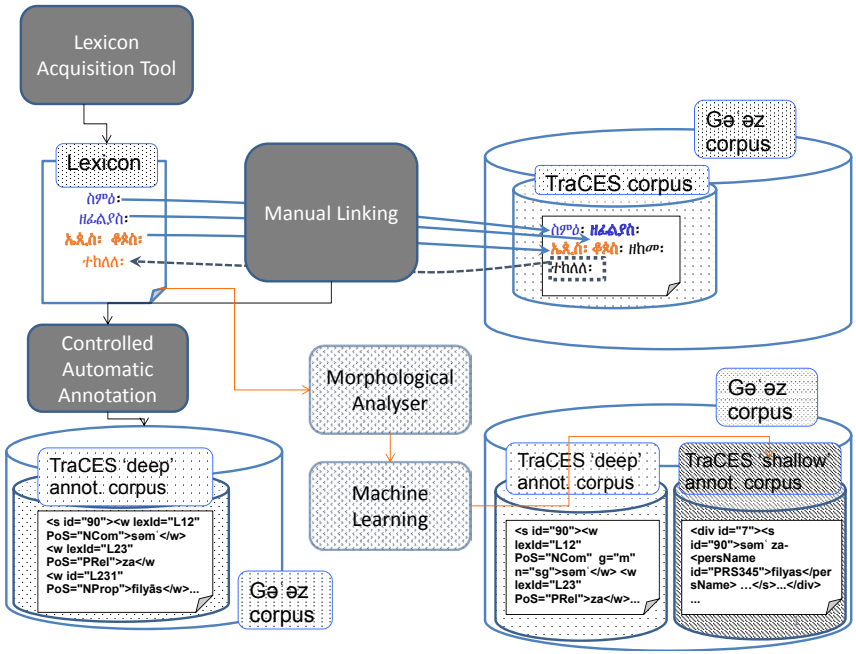

Fig. 1. *TraCES* modules for linguistic annotation.

3. Gəʿəz specific requirements for deep annotation and workflow

Over the past years, several language-independent and/or language customizable annotation tools have become available, including, just to name two examples, WebAnno[4] and CorA.[5] It would have been nice to be able to use them, at least as a starting point. However, certain features of Gəʿəz—in combination with the high scientific demands of the project—made the use of these or other existing tools impossible. In the following, we explain the reasons and the choices behind the decisions taken for the annotation workflow and the design of the GeTa tool.

As mentioned above, the project part-of-speech tag set is particularly fine grained and consists of 33 different tags (PoS); for many of them additional morphological features must be provided.[6] The PoS are divided into six main categories, of which some have further subdivisions: (1) nominals: nouns (2 subdivisions), pronouns (10 subdivisions), numerals (2 subdivisions); (2) verbs; (3) existentials (affirmative and negative); (4) particles: adverbs (2

4 De Castilho et. al. 2016; <https://webanno.github.io/webanno/>.
5 Bollman et. al. 2014; <https://www.linguistics.rub.de/comphist/resources/cora/>.
6 For an overview of the tag set and an introduction to the applied annotation principles (in particular to the complex noun annotation), see Hummel and Dickhut 2016.

subdivisions), prepositions, conjunctions, interjections, further particles (9 subdivisions); (5) foreign material; (6) punctuation.

The linguistic annotation is conducted mainly on morphological criteria, but not solely, as morphologically identical forms need to be disambiguated in the context of syntax and semantics. As the examples below show, disambiguation is required at all stages of the annotation: during the process of transliteration, of tokenization, and of assigning the correct PoS tag.

Due to the lack of training material on the one side and the large number of linguistic features on the other, unsupervised machine learning approaches performing automatic tagging were not suitable for our corpus. We opted in a first stage for a semi-automatic workflow as shown in fig. 2. The annotated corpus from this stage will serve as training material for machine learning.

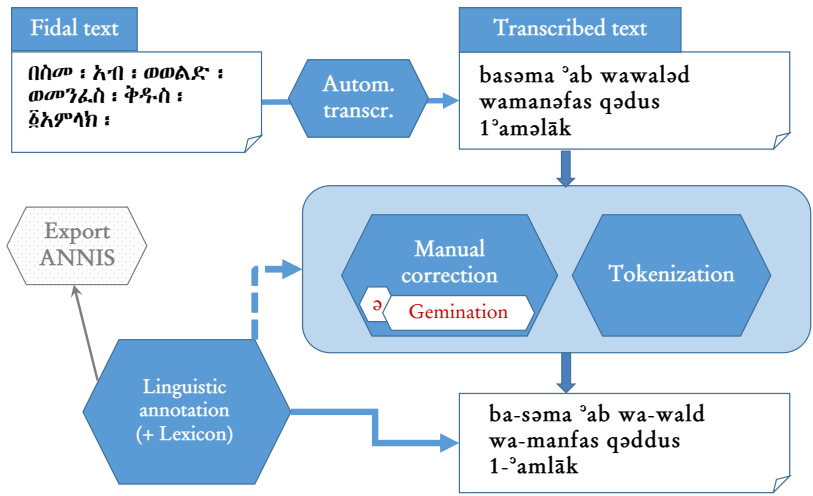

Fig. 2. Annotation workflow.

Because of the syllabic script and the detailed linguistic features to be annotated, any text processing requires a transcription of the Ethiopic text.[7] Therefore, the annotation tool must handle in parallel the text in its original script (*fidal*) and its transcription (respectively, corrected transliteration). Both windows are kept synchronized during all tasks. The transcription is conducted automatically and must be then manually corrected. Corrections concern primarily the presence or absence of the sixth-order vowel (*ə*) be-

7 The graphic unit (GU) **ቤታ ፡** *betā* consists of two syllables *be* and *tā*, but it has to be tokenized as *bet-ā* ('her house'), which would be impossible on the Ethiopic script.

tween consonants and the gemination of consonants. Both phonetic features are not reflected in the Gəʿəz script, and therefore cannot be implemented automatically. The corrections are performed in a semi-automatic way: having corrected one graphic unit (GU),[8] the annotator can decide whether or not the same correction applies to all other identical occurrences in the text. Sometimes the decision on the correct transliteration can be taken only after a morphological analysis; thus the tool must be able to handle later corrections without losing annotations.

A typical example of manual correction required in connection with a geminated consonant is the form **ይትቃተሉ ፡** of the verb **ቀተለ ፡**, *qatala*, 'to make war'. This form is ambiguous, as it can stand for two different verbal moods. It corresponds either to an imperfect indicative (third person, masculine, plural)—in which case the correct transliterated form would be *yətqāttalu*, with the reduplication of the second radical (*t*), which is the phonetic feature of the imperfect in Gəʿəz—or to a jussive (third person, masculine, plural), which requires the transliteration without germination: *yətqātalu*. This example shows clearly that the correct transliteration can be achieved only after a close analysis taking syntax and semantics into account.

Another classic example illustrates the disambiguation of the epenthetic vowel (*ə*), again demonstrated with a verb, here with the meaning 'to save'. The form **ያድኅን ፡** can stand either for an imperfect (third person, masculine, singluar) or for a jussive (third person, masculine, singular). In the transliteration, however, the presence or absence of the sixth-order vowel (*ə*) after the first radical (*d*) differentiates the two verbal moods: with the vowel, *yādəḫən*, the imperfect, without the vowel, *yādḫən*, the jussive.

Linguistic annotation also involves an identification of independent tokens ('tokenization'). We split each complex GU into its smallest analysable units ('tokens'), to which one can assign a PoS. During this process, the annotator, too, needs to resolve ambiguous forms. Identical GUs may carry different meanings and consequently may be split into a varying number of tokens and assigned different PoS.

For example, **ገብሩ ፡** *gabru* may be translated as 'they did'; in this case it would be considered a single token, with the PoS 'verb' (perfect, third person, masculine, plural). In a different context, the same GU **ገብሩ ፡** *gabru* may carry the meaning 'his servant'. In this case, it consists of two tokens, each to be assigned a different PoS: *gabr* 'common noun' (nominative, pronominal state, masculine by pattern, singular by pattern), and *-u* pronominal suffix (third person, masculine, singular).

8 We define a graphic unit (hereafter GU) as a sequence of characters separated by a word divider (፡), or by a punctuation sign (።); the latter is a GU in its own right.

Based on the tokenized transliteration, we eventually conduct the proper linguistic annotation: assigning the PoS together with its features and values. The annotator is able to check and, if required, to correct or adjust the transliteration and tokenization work done so far. Disambiguation is, however, necessary also during this process. It concerns in particular tokens that can have a different meaning depending on the context.

The token **ከመ ፡** *kama* or the token **ኀበ ፡** *ḫaba* may be annotated as 'conjunction' if they precede a verb, or as 'preposition' if they precede a noun or a pronoun. The prefix **ለ** *la-* occurs in the most cases in the function of a preposition, but it can also function in a final clause as a conjunction if it is attached to a jussive verb form.

Finally, each token is linked to a lemma of the newly established digital lexicon.

4. 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:

1. Annotated Objects: GUs, tokens, Gəʿəz characters and transcription letters.
2. Annotation Objects (spans) which are attached to one or more Annotated Objects: 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).

For example the GU-object **ወይቤሎ ፡** contains the 4 Gəʿəz-character objects **ወ**, **ይ**, **ቤ**, **ሎ** (for synchronization reasons, we consider the word separator **፡** as property attached to the Gəʿəz-character object **ሎ**). Each of these objects contains the corresponding Transcription-letter objects:

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*

During the transliteration and tokenization phase, three Token objects are built: *wa*, *yəbel*, and *o*. Each Token object records the IDs of Transcription-letter objects it contains. Finally, the labels '**ወይቤሎ**' and '*wa-yəbel-o*' are attached to the initial GU object.

Morphological annotation objects are attached to one Token object. They consist of a tag (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 when adding new morphological features or PoS tags.

As the correspondences between the Gəʿəz character and the transcription are unique, the system only stores 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, so that the system uses less memory and remains error proof during the transliteration process. In fig. 3 we present the entire data model, hinting also at the other possible annotation levels.

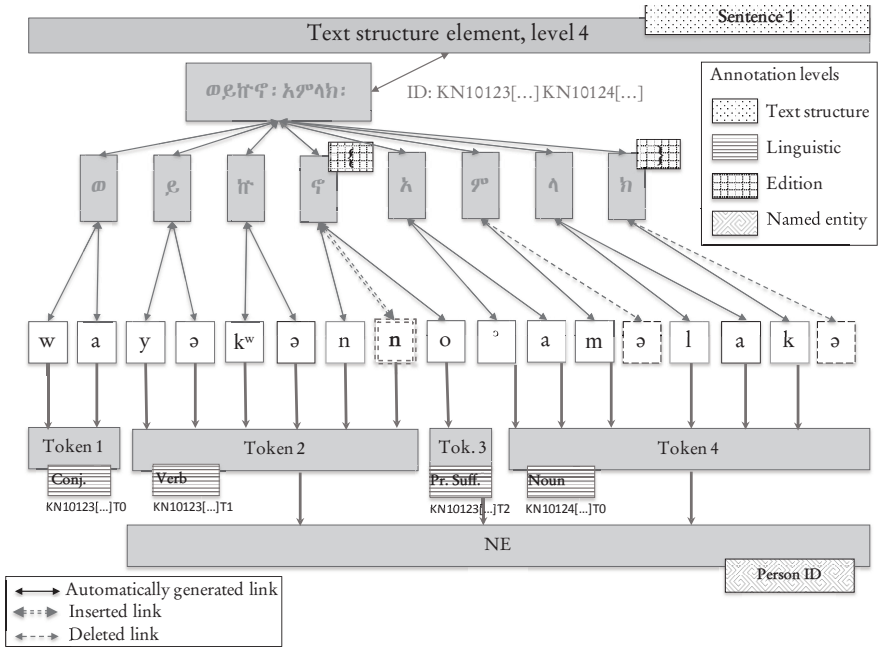

Fig. 3. GeTa data model.

5. 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 facilitates the mark-up process but at the same time leaves the full control entirely to the annotator. Units annotated or tokenized automatically are highlighted, so that the user knows anytime if a manual check is necessary. For example automatically generated tokens are displayed in italic, automatically annotated tokens are marked in red.

Corrections to the transcription, as they were described above, can be performed at any moment 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 and visualization applications like Voyant Tools.[9] A special convertor to ANNIS[10] format has been implemented, so that the annotated corpus can be analysed with the powerful mechanism of the ANNIS visualization tool. The corpus will be freely accessible for further research through the ANNIS installation provided by the Hamburger Zentrum für Sprachkorpora.[11] The TEI export will be used for integration with the data available in the project *Beta maṣāḥəft*.[12]

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

Rules for transliteration, tokenization, and annotation may be extracted from the annotated texts and used for a more advanced automatization of the annotation process.

References

Bollmann, M., F. Petran, S. Dipper, and J. Krasselt 2014. 'CorA: A Web-based Annotation Tool for Historical and Other Nonstandard Language Data', in K. Zervanou, C. Vertan, A. van den Bosch, and C. Sporleder, eds, *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (Gothenburg 2014), 86–90, <doi:10.3115/v1/W14-0612>.

De Castilho, R.E., E. Mújdricza-Maydt, Seid Muhie Yiman, Silvana Hartmann, I. Gurevych, A. Frank, and C.Biemann 2016. 'A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures', in Erhard W. Hinrichs, Marie Hinrichs, and Thorsten Trippel, eds, *Proceedings of the LT4DH workshop at COLING 2016* (Osaka, 2016), 76–84, <http://aclweb.org/anthology/W16-4011.pdf>.

Druskat, S. and C. Vertan 2018, 'Nachnutzbarmachung von Forschungsdaten und Tools am Beispiel altäthiopischer Korpora', in G. Vogeler, ed., *Kritik der digitalen Vernunft. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 26.02.–02.03.2018 an der Universität zu Köln, veranstaltet vom Cologne Center for eHumanities (CCeH) (*Köln: Universität zu Köln, 2018), 270–273.

9 <https://voyant-tools.org/>.

10 <http://corpus-tools.org/annis/>; see Druskat and Vertan 2018.

11 <https://corpora.uni-hamburg.de/hzsk/>.

12 <https://www.betamasaheft.uni-hamburg.de/>.

Hummel, S. and W. Dickhut 2016. 'A Part of Speech Tag Set for Ancient Ethiopic', in A. Bausi and E. Sokolinski, eds, *150 Years after Dillmann's Lexicon: Perspectives and Challenges of Gəʿəz Studies*, Supplement to Aethiopica, 5 (Wiesbaden: Harrassowitz Verlag, 2016), 17–29.

Krzyżanowska, M. 2017. 'A Part-of-Speech Tagset for Morphosyntactic Tagging of Amharic', *Aethiopica. International Journal of Ethiopian and Eritrean Studies*, 20 (2017), 210–235.

Maegaard, B., S. Krauwer, K. Choukri, and L. Damsgaard Jørgensen 2006, 'The BLARK Concept and BLARK for Arabic', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)* (Genoa: LREC, 2006), 773–778 <http://lrec-conf.org/proceedings/lrec2006/pdf/521_pdf.pdf>.