Alpheios and the Gəʿəz Morphological Parser

The TraCES project (https://www.traces.uni-hamburg.de/) is developing a morphological parser for Gəʿəz, or Classical Ethiopic, the language used in the Ethiopic and Eritrean Manuscript tradition. This is a living manuscript tradition, with a production which continues to these days of remarkable parchment books and scrolls, mainly with religious contents.

The parser was designed from the beginning to be able to integrate with Alpheios in order to give access to this parser and the related resources (a reference dictionary and a corpus of morphological features) to anyone on the web and for any webpage that includes Gəʿəz text. The simple fact that Alpheios can both be embedded in websites and installed as a plug-in increases the potential to spread the use and usefulness of the resources curated by the project to a much broader group of users.

The morphological parser is based on one XQuery module written by Dr Pietro Liuzzo running in eXist-db and a series of tables provided by the project Team (https://www.traces.uni-hamburg.de/en/team.html), led by Prof. Alessandro Bausi, and indexed in the same module. The code, the source tables and the XML format used by the parser are all accessible here https://github.com/TraCES-Lexicon/lexicon . They are not standard, but we will standardize them as soon as the logic is satisfactory.

The parser performs two distinct tasks: the first one is an attempt to carry out from the string a morphological analysis of the word; the second one is a search in a corpus of hand annotated text, which the TraCES team developed using the GeTa tool developed by Dr Cristina Vertan for the project at the Hiob Ludolf Centre for Ethiopian Studies (details of how the parsing is done will be described in a chapter of a forthcoming book by Liuzzo).

The parsing attempts several possible analyses and then validates each possibility to the reference lexicon for Gəʿəz, August Dillmann's *Lexicon Linguae Aethiopicae* online (http://betamasaheft.eu/Dillmann), which has been made into a web application based on TEI data by the same project. Each possible parsing might be found in Dillmann looking at one of several roots attested there. In parallel, for the searched word and for any token of it or lemma in Dillmann which is returned with the data, the parser also searches the TraCES annotations in a TEI Features Structure format exported from GeTa and loaded into the parser application to be indexed and searched. In this way, it is possible that where the logic fails, the parser will match an occurrence with a morphological annotation in the TraCES corpus and will be able to return some context-based information. When both succeed, then the user will have several results to check one against the others. Also results which do not validate for the parser are preserved, since they can help especially in the analysis of unrecorded lexical entries. Dillmann's *Lexicon* is being updated by the team and contains more and more lemmas, but the completeness will presumably never be achieved as the whole corpus is extremely large, we continue to study new texts and the dynamic connection to the resource means that the results of the parser are checked against the latest available knowledge, not only against the state of it in 1865. We hope that more and more contributors will participate in this effort to make it a resource more and more valuable for linguistic investigation.

The XQuery module serves three response formats. One is the format defined by the Alpheios Schema, the other a local XML response format and the third is an HTML visualization of the same results presented in the XML. This interface is available at http://betamasaheft.eu/morpho.

We plan to use Alpheios also as embedded library specially to help the reading of the online *Lexicon Linguae Aethiopicae*. This reference work was written in Latin in 1865 and contains references to Arabic, Hebrew, Biblical Aramaic, Syriac, Coptic, Greek and some terms in other languages. Less and less often scholars can read all of these languages and we started by linking what could be linked to tools like the ones offered by Perseus. This would mean leading the user in a new tab or page where the string was used as a parameter for a search and had to be done separately for each language. Having the Alpheios library embedded means we can avoid all this going around and access inside the application the reading aids, dictionaries, inflection tables, etc. which are available via Alpheios, including our own and hopefully in the future also support for Hebrew, Syriac, Coptic etc.

The morphological parser is now in an experimental stage, it does not answer 100% correctly, especially on nouns. We will possibly not be able to achieve that ever, but at this stage we are not even claiming the 10%, although some internal testing has succeeded to the 75% of cases provided to return at least one correct result among a series of possibilities. Offering this parser already as an experimental feature available via Alpheios at this very early stage we hope will allow us to receive constructive feedback on it to improve it more and more. At the time of writing development is actively taking place within the project and also performance should improve, while we look forward to implementing support for inflections tables.

Bausi, A. and E. Sokolinski, eds, 2016. *150 Years after Dillmann's Lexicon: Perspectives and Challenges of Gəʿəz Studies*, Supplement to Aethiopica, 5 (Wiesbaden: Harrassowitz Verlag, 2016).

Dillmann, C. F. A. 1865. *Lexicon linguae aethiopicae, Cum indice latino. Adiectum est vocabularium tigre dialecti septentrionalis compilatum a W. Munziger* (Lipsiae: T. O. Weigel, 1865).