Interim Report

HG-229371
Koptische/Coptic Electronic Language and Literature International Alliance (KELLIA) Project

Directors:
Caroline T. Schroeder, University of the Pacific
Amir Zeldes, Georgetown University

Institution: University of the Pacific

May 31, 2016

Introduction

This is the second interim report of the KELLIA project. It demonstrates the progress the US partner has made on the five outcomes proposed in the original bilateral grant since the last interim report, which was filed November 28, 2015. Although this report will refer work done on all five outcomes outlined in the original proposal, including work at German partner institutions, this report is intended primarily to discuss the progress of the US partner and should not be considered a complete report on the activities of German partners. The US partner is primarily responsible for outcomes two and four of the proposal. Work on most outcomes is ahead of or on schedule.

Outcome 1 - Milestones for data standards

Work has continued on establishing metadata standards for Coptic editions and manuscripts. The German partners of the bilateral grant are primarily responsible for this work. U.S. PI Caroline T. Schroeder has been coordinating with Uwe Sikora (one of our partners at the University of Göttingen) in his survey of data and metadata standards. Mr. Sikora is preparing recommendations to make to the full KELLIA group about the standards.

We have also created a wiki (http://wiki.copticscriptorium.org/doku.php?id=kellia:utf8:coptic_unicode_utf-8_standards_and_guidelines_for_coptologists) to discuss and establish standards and best practices in Coptic Unicode, which is hosted and managed by the US partner. The wiki is publicly accessible with a limited editorial group of interested Copticists and unicode specialists. The wiki is still in its infancy, but we expect that it will grow over the next several months.

Outcome 2 - Server based batch conversion tools

Several critical improvements have been made, most notably the development of a new Natural Language Processing (NLP) pipeline, which was introduced in the previous interim report.

We have created a preliminary solution to allow annotators to manually check automated tokenization prior to running the other annotation tools, which were developed using the NEH ODH Start Up grant (HD-51907). The NLP interface (https://corpling.uis.georgetown.edu/coptic-nlp/) now provides an option for "just piped and dashed morphemes" for output, so annotators can correct segmentation errors and run the rest of the pipeline on this corrected input. Because the proper function of the other tools (normalizer, part of speech tagger, language of origin tagger, lemmatizer, and morphological analysis) depends on tokenization, the ability to stop the pipeline at this point should result in less manual annotation at the end of the process and better output from the automated annotation tools. This new process is currently being tested by editors annotating corpora working under the Preservation and Access Foundations grant (PW-51672-14)..

The annotation tools available through the NLP pipeline now also integrate the Greek loanword list supplied by our partner, the Database and Dictionary of Greek Loanwords in Coptic (DDGLC). The addition of the DDGLC word list affects the tokenizer, language of origin tagger, lemmatizer, and

morphological analysis. We have since released a new machine-annotated New Testament Corpus using the updated tools with greatly improved results.

Finally, we are working to integrate the Thesaurus Linguae Aegyptiae's (TLA) Coptic lexicon into our textual annotations. The addition of the TLA Coptic lexicon to our annotations will make our corpora much more useful to scholars and students of Coptic alike, and we plan to have a pilot version ready for the Congress of the International Association of Coptic Studies (IACS) in July.

Outcome 3 - integration of linguistic tools and methods to produce collaborative digital editions

We have completed initial testing of the transcription of literary texts using the Virtual Manuscript Room (VMR): the US partner tested several *Apophthegmata*, while our German collaborator tested one of Besa's letters. Based on preliminary results, we have added an option to the NLP to stretch milestones to accommodate the VMR's XML output, which uses unary tags. Discussion of the next steps in integrating the VMR, NLP, and annotation tool continues.

Outcome 4 - development of a web-based, multi-layer annotation tool for collaborative text annotations and stand-off markup

As stated in the November 28 interim report, we chose the open-source web-based spreadsheet software EtherCalc to develop a multi-layer, collaborative annotation tool and created an API to stream data into EtherCalc. This means that automatic analyses from our NLP pipeline can be automatically imported into EtherCalc. We have since tested the process of annotating a literary text in EtherCalc using output from the NLP pipeline with good results. We are continuing the work of adapting EtherCalc to fit our project and to export data.

Outcome 5 - Sharing, linked data, and textual re-use

Preliminary work has been done on automatic syntactic analysis tools and entity recognition. First entity tagging tests have been done on select *Apophthegmata* as well as Shenoute's "Not Because a Fox Barks" as the gold standard for these tools. Using entity lists taken from Coptic SCRIPTORIUM corpora, Zeldes used xrenner (https://corpling.uis.georgetown.edu/xrenner/#) to recognize classes of entities. Certain texts also have initial syntactic analysis in the form of dependency trees. Automatic entity and syntax annotation is improving and will form the basis of future work on textual re-use and data exchange.

Events

The KELLIA advisory board held a virtual meeting in January. In addition, project members have consulted with members of the advisory board as needed on an individual basis.

No scheduled events took place since the last report in November. We are currently planning for our next KELLIA workshop, which will be held in Claremont, California on July 23 and 24, immediately preceding the International Conference of Coptic Studies. Attendees will include representatives from KELLIA partners (Coptic SCRIPTORIUM, Coptic Old Testament, the Institute for New Testament Textual Research, the Thesaurus Linguae Aegyptiae), as well as other Copticists conducting digital research.

Additionally, Schroeder and DH Specialist and Project Manager Elizabeth Platte will be attending a symposium at the University of Iowa sponsored by the Big Ancient Mediterranean Project (led by Sarah Bond and Paul Dilley). We will be meeting with them and other DH projects focused on the ancient world to advance our goals for entity recognition and linked data.

Paper resulting from this project phase

Zeldes, Amir and Schroeder, Caroline T. (submitted) “An NLP Pipeline for Coptic”. Submitted to the *10th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities at the annual conference of the Association for Computational Linguistics*. Berlin.