Interim Report

HD-51907

Coptic SCRIPTORIUM: A Corpus, Tools, and Methods for Corpus Linguistics and Computational Historical Research in Ancient Egyptian

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

Institution: University of the Pacific

November 29, 2014

Coptic SCRIPTORIUM has met all the projected goals for the grant period thus far. This report summarizes the milestones originally planned for each of the two major project phases up to now then narrates our progress on meeting those milestones. Additional activities are described at the end of the report. In addition, co-Director Dr. Amir Zeldes has taken a position at Georgetown University; the grant activities now take place primarily at the University of the Pacific in Stockton, California, and Georgetown University in Washington, D.C.

Planned Phase One (May – Fall 2014) Goals and Activities:

- Tool development of existing tools: tokenizer, part-of-speech tagger
- Development of new tool(s): lemmatizer.
- Planning for online platform for users to add text and/or annotations
- A one week working session to address linguistic issues; update documentation; review the progress of tools and technology development and their effectiveness in processing the digitized text

Planned Phase Two (Fall 2014-Winter/Spring 2015) Goals and Activities:

- Internal evaluation of technologies and methodologies by testing and applying tools to digitized corpus
- Test expanded corpus produced and imported into the ANNIS search and visualization infrastructure.
- Continued development of lemmatizer and begin development of other tools (e.g., named-entity tagger, collation navigation and visualization tool).
- Meeting of Advisory Board in Fall 2014
- Refinement of existing tools and methods based on internal evaluation and Board evaluations
- Discussion and planning regarding creating an online interface for researchers to contribute text and/or annotations to the corpus.
- Second one week joint working session for directors

The SCRIPTORIUM team is ahead of schedule on the development of tools and technologies. We updated the tokenizer (which segments Coptic texts into morphemes) to be more accurate and to add an additional feature: the tokenizer now can process transcriptions of manuscripts that contain Coptic morphemes and words that cross a line or column or page break in the manuscript, without disrupting those annotations. We have also updated the normalizer and the language of origin tagger for accuracy. The last two have been tested on additional corpora. During the remainder of Phase Two, the new feature of the tokenizer will be implemented and tested on corpora being digitized by the Coptic SCRIPTORIUM's PW-51672-14 grant. We are in conversation with the Database and Dictionary of Greek Loanwords in Coptic (DDGLC) project in Germany on the study of Greek loan words in Coptic to exchange material: they will share their lemma list with us and we will provide them with digitized texts containing Greek loan words. Project Directors began development of a lemmatizer during the summer 2014 and will continue in 2015. Project Director Schroeder

has also attended the Berkeley Prosopography Services workshops on prosopography and Digital Humanities in Fall 2014; these conversations are informing our planning for our named-entity tagger.

Live update: A student at Georgetown University is expanding the ANNIS search and visualization tool's API so that the links to the HTML visualizations of data on the website will be linked directly to data visualizations generated out of ANNIS, instead of static HTML pages manually copied from ANNIS and uploaded to the website's server each time we update the data.

Additionally, we have our own domain (www.copticscriptorium.org) and have moved all our workspace, data, and tool repositories to GitHub (http://github.com/CopticScriptorium). A student at the University of the Pacific funded by an internal Pacific grant set up the project account and repositories. Much of the work is in public repositories, but the text editing is in private repositories before publication.

The online interface for collaborative corpus annotation is underway. All project contributors (including the Editors/Annotators/Digitization Contributors funded by grant PW-51672-14 and new volunteers) work from our GitHub repositories. A second Georgetown student is working on another aspect of the planned online interface for annotation. He is expanding on the open source spreadsheet EtherCalc to develop an online spreadsheet for annotating the corpora that can save versioned forms of our multilayer data. This will make it easier for people to submit editorial content and annotations to our corpora online.

September 15-19, 2014, the project Directors held a working session at Georgetown University, where much of the work described above was evaluated, conducted, or planned. They also attended the NEH ODH Project Directors' meeting on September 15, 2014.

The Advisory Board met virtually in an email consultation about the project in November 2014. Board members indicated that our outreach to other scholars in our subject field and in Digital Humanities was effective and that we are on track with our tools and technologies. They also indicated that our corpus development and data curation efforts are on track. For future activities, one board member recommended providing additional, more detailed documentation for searching the digital corpus. Another recommended additional features to language of origin tagging (providing Greek lemmas for Greek loanwords) and reconsidering decisions about tagging Greco-latin, Greco-hebrew, etc., as Latin and Hebrew instead of Greek). This was a joint meeting to discuss activities for both the ODH grant and the Preservation and Access Grant (PW-51672-14)

Additional Activities

In addition to the originally planned goals and milestones, we have several further achievements:

- Schroeder has presented conference papers about the project at the North American Patristics Society annual meeting in Chicago (May 2014), DH 2014 in Lausanne (July 2014), and the UCLA-St. Shenouda Society annual Coptic Studies conference (July 2014).
- Zeldes used the corpora to teach a Coptic course at Humboldt University; students used the projects' technologies to prepare Coptic text for publication with Coptic SCRIPTORIUM as their final projects.
- Project Directors have submitted two co-authored articles, for special issues of the *Journal of Digital Humanities* and the *Journal of Digital Scholarship* in the Humanities.
- Zeldes has submitted a paper to the upcoming North American Conference on Afroasiatic Linguistics.
- Schroeder met with the staff at the Perseus Digital Library on site at Tufts University in September, 2014, to talk about their workflow and management processes and received some good advice and suggestions.

We also have realized some cost savings due to decisions at the University of the Pacific to underwrite more of Dr. Schroeder's salary than originally planned in the grant proposal. We are reallocating some of the salary, benefits, and indirect costs saved for two activities planned for Phases 2 & 3:

- A 2-day symposium and workshop on Digital Humanities and Coptic Studies at Georgetown University in March, 2015. One day will be an open symposium and discussion on topics in the field, including presentations by other scholars in the field and conversations about common questions, issues, technologies, and standards. The second day will be devoted to Coptic SCRIPTORIUM: contributing to the project, annotating texts, problem solving, and future directions.
- Hiring a contractor recommended by Perseus Digital Library senior programmer, Bridget Almas (who is consulting on Coptic SCRIPTORIUM's NEH Preservation and Access: Foundations grant PW-51672-14) to implement the data curation standards developed under the Preservation and Access: Foundations grant.