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

May 31, 2015

Coptic SCRIPTORIUM has met most of the projected goals for the grant period thus far. This report summarizes the milestones originally planned for the major project phases since our last report (dated November 29, 2014) and then narrates our progress on meeting those milestones. Additional activities as well as challenges and future plans are described at the end of the report. As stated in the November 2014 report, co-Director Dr. Amir Zeldes has taken a position at Georgetown University; the grant activities take place primarily at the University of the Pacific in Stockton, California, and Georgetown University in Washington, D.C.

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

(Some of the work for Phase Two was completed prior to the November 2014 interim report and is documented there. We report on activities since November 2014 below.)

- 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).
- 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 with most of the development of tools and technologies planned for Winter/Spring 2015:

> *Evaluation, testing, and refinement of tools improved in Fall 2014*: We successfully tested the tokenizer (which segments Coptic texts into morphemes), normalizer, and language of origin taggers updated in Phase One on corpora digitized by the Coptic SCRIPTORIUM's PW-51672-14 grant. Regarding our language of origin tagger, our Advisory Board recommended reconsidering our decision to tag loan words with the oldest possible language of origin. (For example: most Hebrew or Aramaic words probably came into Coptic via Greek, because these words typically appeared in the Greek Bible and then migrated into Coptic; we tag them as Hebrew or Aramaic, not Greek, Greco-Hebrew, or Greco-Aramaic.) The board did not advocate changing the practice but asked us to think more about it and consider a change. After discussion with the entire project team and other Coptic scholars at a meeting in March 2015, we decided not to change our practices; especially for personal names and place names, we decided that tagging these loan words as Greco-Hebrew, or Greco-Aramaic (or simply Greek) was adding another layer of interpretation to the annotation process. We decided instead

to add more documentation about our editorial decisions and how to easily search for all loan words regardless of language of origin in our search tool.

Tokenizer update: We have made two major updates to the tokenizer to increase its accuracy and flexibility; we first introduced changes to tolerate orthographic variation and diacritics, as well as word-internal XML markup to allow tokenization of marked-up diplomatic transcriptions. We then used the the existing data in our corpora (data that had been manually corrected by encoders/digitizers under the PW-51672-14 grant) and fed that data back into the latest 3.0 version of the tokenizer to introduce a training data component that learns from our annotators' most common tokenization and correction practices.

Lemmatizer, entity taggers, and collation navigation tool: The development of the lemmatizer, entity taggers, and collation navigation tool was placed on hold while we pursued the development of three other technologies, described below.

HTML data visualizations: A doctoral student at Georgetown University (Shuo Zhang) completed the expansion of the ANNIS search and visualization tool's API under Zeldes' supervision, so that the HTML visualizations of data on our website will be drawn directly and automatically from data visualizations generated by ANNIS; previously we had been manually copying the HTML visualizations from ANNIS and publishing them as static HTML pages each time we updated the data. This data is now being ingested via an API used by the stable referencing software described below.

Stable referencing web application: The creation of a web application to implement the citation and referencing system developed by consultant Bridget Almas was completed under the PW-51672-14 grant. This web application implements a data curation model that resolves stable URNs referencing documents and text groups, so that researchers can cite our data with stable, well-curated references. The application also pulls the aforementioned visualizations of data from ANNIS so that users can easily read the texts and visualizations, even without going into ANNIS's search interface. This process ensures that our search software, ANNIS, and representations of manuscripts for close reading on our website remain in sync. This system is currently being tested and will be deployed at http://data.copticscriptorium.org later this summer. The developer is Luke Hollis, of Archimedes Digital; he was referred to us by Bridget Almas of the Perseus Digital Library.

TEI XML converter: We currently provide our data in three primary formats to facilitate interdisciplinary research: PAULA XML (used by computational linguists), TEI XML (used by Digital Humanists), and relANNIS (the database files, efficiently indexed for our search and visualization tool, ANNIS). We have had a tool to convert most of the data (but not the metadata) from our multi-layer data model into TEI XML; the TEI XML metadata had been encoded manually. This spring, we expanded the converter to cover the metadata as well as the inline data. Now the exact same data model feeds our PAULA, TEI, and relANNIS files. Examples of TEI-XML files generated from this converter can be found on our public corpora repository on GitHub (e.g., file versions dated May 22, 2015 at https://github.com/CopticScriptorium/corpora/tree/master/AP/apophthegmata.patrum_TEI).

We had a one-week project directors' meeting/hackathon in Stockton, CA, in May 2015. During this week we: tested the latest implementation of the new web application; finalized the TEI converter; edited our existing text data files to be compatible with the data model for the web application and the TEI converter; published new and revised documents in ANNIS, PAULA, and TEI formats; discussed possible avenues for developing an online interface for contributors to add digitized texts or annotations (discussing models such as the Virtual Manuscript Room developed in Germany and SOSOL used by projects such as papyri.info); presented our project's progress via Skype to a team of scholars creating print critical editions of an important Coptic corpus (the writings of monastic author Shenoute). For more information on the data models and the newly published documents, see the interim report for Coptic SCRIPTORIUM's PW-51672-14 grant from the Division of Preservation and Access.

We have also started Phase 3 of the project, beginning work on the following milestones outlined in our proposal for Phase 3 (Winter/Spring-Summer/Early Fall 2015):
• Edit documentation (drafts composed in process in earlier phases) for the tools.
• Publicly release expanded multi-platform corpus, tools, and first version of SCRIPTORIUM platform at end of period.

We have expanded the documentation section of our website (http://copticscriptorium.org/documentation.html) and added a public wiki at http://corpling.uis.georgetown.edu/wiki/doku.php. As discussed above, we have already released some corpora in multiple formats from the same base data model and have released improved versions of the tools.

Additional Activities

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

- A one-week project directors' meeting in Washington, DC, in March 2015, in conjunction with *Digital Coptic 2: A Symposium and Workshop* at Georgetown University (http://copticscriptorium.org/workshop2015/index.html). The meeting, symposium, and workshop were funded by Coptic SCRIPTORIUM's PW-51672-14 grant from the Division of Preservation and Access. (See the May 2015 interim report for more details on the directors' meeting and the symposium and workshop)
- Three of the papers at *Digital Coptic 2* utilized Coptic SCRIPTORIUM technologies and data. Paul Dilley, Asst. Prof. of Religious Studies and Classics at the University of Iowa, presented a research paper that applied our tokenizer to Coptic texts and then performed a stylistic analysis of those texts. So Miyagawa, a graduate student in linguistics at Kyoto University, presented a paper on synthesis, boundness, and clitics;

his dataset came in part from our online corpus. SCRIPTORIUM team members Rebecca Krawiec (Canisius College) and Amir Zeldes gave a paper on the use of SCRIPTORIUM for interdisciplinary research (history, Religious Studies, philology, and linguistics).

- Feedback and questions from other Coptic scholars during the workshop at *Digital Coptic 2* is being incorporated into our project planning.
- Revised our guidelines for tagging Coptic (http://copticscriptorium.org/download/tools/scriptorium_tagset_documentation.pdf)
- Project Directors' co-authored article for a special issue of the *Journal of Digital Scholarship* in the Humanities has been accepted for publication.
- Zeldes presented the paper "Tagging the Desert Fathers: part of speech analysis in Sahidic Coptic corpora" to the North American Conference on Afroasiatic Linguistics.

Challenges and Future Plans

Implementation of the web application, training other users in our tools and technologies, and providing detailed documentation have been the most challenging aspects of this period. Patience and perseverance have proven to be the best strategies. The additional time spent has been rewarding; we previewed an early version of the web application to the scholars involved in Shenoute print critical edition project and members of the team working on the PW-51672-14 grant, and they found it very useful. We anticipate we may request an extension on the timeline of this grant in order to complete our planning for the additional tools and to provide fuller documentation of the project's technologies, standards to cite and reference our data, and use cases for research.