Interim Report

HD51907

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, 2016

This report summarizes project goals and their realizations from the period since our previous report, dated May 31, 2015, up to May 31, 2016. The project has met its planned objective and has also developed in new directions with some additional activities, outlined below. Feedback from both our advisory board and users of our deliverables indicates positive implementation of our objectives.

Planned goals for the phase June 2015 - May 2016:

- Completion of development for tools begun in the previous phase. The relevant tools are:
    - Language of origin detection (recognition of Greek and other loanwords in Coptic texts)
    - Automatic normalization for diplomatic manuscript material
    - Automatic lemmatization
    - Coptic morphological analyzer
- Preliminary work on the development of a collaborative annotation interface
- Release of new corpus versions (see below for details)
- Write and disseminate tool documentation
- Develop linguistic and historical research questions using the corpus

We report on completion of these goals below.

Tool development

With the contribution of the Greek lemma list from the Database and Dictionary of Greek Loanwords in Coptic (DDGLC, Leipzig/Berlin) and in collaboration with our German partners in the bilateral NEH-DFG KELLIA project grant (HG-229371), we were able to implement a more robust version of the language of origin recognizer from the previous project phase. According to the evaluation in Zeldes & Schroeder (submitted), language of origin recognition now has an accuracy of over 99%. We will continue to update lexical resources for the language of origin recognizer, but for now we consider this component complete.

The normalization tool was also extended to take advantage of the mounting amount of training data gather in the project. The same paper puts its accuracy at 98.01%, using a combination of deterministic rules and statistical information.

Automatic lemmatization is new in this project phase, with a current accuracy of 97.23%, based on perfectly normalized text. The development of lemmatization also required new guidelines (see documentation below), for which training is ongoing. We also retro-fitted our existing corpora with the new lemmatization possibilities.

The existing Coptic morphological analyzer was extended with a mechanism to use not only deterministic rules, but also training data from our corpus and some heuristics, especially for recognizing complex verbs. A quantitative evaluation remains outstanding, but the tool's performance is intuitively fairly satisfactory.

Preliminary work on an annotation interface

As part of this objective, we explored the use of collaborative online spreadsheets, annotating some test data with the open source EtherCalc editor (https://ethercalc.net/), and installed our own version on our server. Our experiences have shown that this tool has potential, although some adjustments may be required for its successful application. We have gotten in touch with the developers of EtherCalc regarding particular issues, and plan to keep working with our German partners on this and other relevant tools as part of the agenda for the KELLIA grant.

Release of new corpus versions

Under our NEH Preservation and Access Grant (PW-51672-14), we edited and annotated corpora with the refined tools. We released several new datasets, from the works of Shenoute of Atripe (specifically the work *I See Your Eagerness* and a new fragment of *Acephalous Work 22*) and the *Apophthegmata Patrum*. We also expanded our documentary papyrus corpus by a further sample, and re-released our automatically processed Sahidic New Testament, using the latest and by now more accurate versions of our tools.

Linguistic research using corpora

In addition to papers dealing with the development of our tools, co-PI Amir Zeldes will be presenting a paper using the corpora at the Congress of the International Association for Coptic Studies (IACS) c, titled “A Quantitative Approach to Syntactic Alternations in Sahidic”. This is the first linguistic paper using large amounts of partly automatically analyzed text to answer questions about grammatical phenomena in Coptic which had previously been described as not well understood or ‘optional’ (e.g. in Bentley Layton’s authoritative Coptic grammar). In addition, Paul Dilley (University of Iowa) will be presenting a paper on “distant reading” of Coptic Literature using tools and corpora we have developed.

Dissemination of documentation

Besides continuously updating our documentation, our most recent addition in this project phase to the set of guidelines laid out by the project is the production of detailed lemmatization guidelines for searchable Coptic digital corpora. These guidelines are consonant with our part of speech tagging guidelines, and together allow users to find specific content words in specific uses, regardless of inflections or different ways in which they are spelled. We also added a wiki (http://wiki.copticscriptorium.org) and reorganized the documentation section of our website to be more user-friendly.

Other activities

Third working session

The PIs met in Washington, DC in December 2015 to discuss the further development of the tools mentioned above, the design of a repository for the data (currently implemented at http://data.copticscriptorium.org) and joint publications (see Papers below).

The development of the web application has been a significant additional activity beyond the originally projected goals of the project. As described in our May 2015 interim report, the web application implements a citation and referencing system developed by consultant Bridget Almas under the PW5167214 grant. This web application resolves stable URNs referencing documents and text groups, so that researchers can cite our data with stable, well-curated references. The application also pulls visualizations of the data from our online ANNIS database 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. We deployed the system as beta last summer. A new contractor (Dave Bricetti) has been fixing remaining bugs and improving the performance of the application.

Board meeting

Our Advisory Board meeting was held in January 2016, although we also consulted with Board members throughout the report period on relevant issues individually. The Board expressed satisfaction with the current tools and datasets, and expressed an interest in expanding both general public and especially student contributions to our datasets. The development of annotation interfaces and dissemination of documentation have both kept this goal in mind, and we are exploring possibilities for integrating data collection into classroom settings.

The Board also stressed the importance of avoiding duplication of digitization efforts. Our collaboration with German partners under the KELLIA bilateral grant (NEH-DFG HG-229371) is one important step in this direction. We are also organizing a Coptic SCRIPTORIUM panel and workshop on digital methods for Coptic at the IACS Congress in July at Claremont.

Papers resulting from this project phase

Schroeder, Caroline T. "The Digital Humanities as Cultural Capital: Implications for Biblical and Religious Studies," *Journal of Religion, Media, and Digital Culture* 5:1 (2016).

Zeldes, Amir (to appear) "A Quantitative Approach to Syntactic Alternations in Sahidic". *International Association for Coptic Studies*. Claremont, CA.

Zeldes, Amir and Schroeder, Caroline T. (2015) "Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities". *Digital Scholarship in the Humanities* 30(1), 164-176.

Future plans and outlook

As this grant draws to a close, some of the international standardization activities and development of collaborative annotation tools naturally flow into the goals planned for the bilateral KELLIA grant. We recognize that the tools developed during the final phase of this project will require a commitment to maintenance and further development as required in order to ensure the continuing availability of searchable Coptic data online.

We are also excited by the potential for interdisciplinary quantitative work on Coptic that this project has opened up. We are now in a position to share and collaborate using Coptic primary sources through established and well documented standards and best practices. Ensuring their dissemination, training new-comers to the field and regulating the interpretation and expansion of our guidelines remain important challenges for the future.

While our priorities for this funding phase and ongoing work in the KELLIA project remain consolidation of our resources and the expansion of our current literary corpus of Sahidic, we are considering plans for further deepening our data by working on semantic analysis of Coptic data (including named entities, work on geographical information and linking to other projects), as well as looking to new kinds of Coptic texts that our work could make more accessible, including data in new genres, and potentially further dialects of Coptic, beyond the classical Sahidic texts we have been working on so far.