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

November 30, 2016

Introduction

This is the third interim report of the KELLIA project, which covers progress made since the last interim report was filed on May 31, 2016. This report details work on the the five outcomes proposed in the original bilateral grant in order, including the outcomes for which German partner institutions are primarily responsible. However, it will focus on the work done by the US partners and should not be considered a complete report on the activities of German partners. The US partners are primarily responsible for outcomes two and four of the proposal. Work on most outcomes is ahead of or on schedule.

KELLIA held the second of the international workshops proposed in the original grant on July 23 and 24, 2016, before the International Congress of Coptic Studies in Claremont, California. Representatives from German and US partner institutions, as well as affiliated projects in the US, were present. A detailed report of the workshop is available in the activities section below, but this report also refers to the Claremont workshop at several points in the description of progress.

Outcome 1 - Milestones for data standards

During the KELLIA workshop in Claremont (see Activities, below), Uwe Sikora of the University of Göttingen presented an overview of metadata standards and suggested the TEI and EpiDoc standards for digital Coptic projects. German partners are currently exploring options to integrate metadata transfer schemes into the Virtual Manuscript Room (VMR).

In the process of creating an XML converter to ensure XML obtained from the VMR is compatible with the Natural Language Processing pipeline for Coptic (see Outcome 3, below), Sikora and Elizabeth Platte, the DH Specialist and Project Manager on the US team, have begun to review metadata presented by the VMR and in Coptic SCRIPTORIUM's TEI-XML files. They have also discussed authorities for metadata and tagged entities within annotated corpora. Platte and US PI Caroline T. Schroeder will continuing to coordinate with Sikora about encoding and metadata standards.

Outcome 2 - Server based batch conversion tools

See the discussion in Outcome 3 of an XML converter to convert data output from the Virtual Manuscript Room into other formats (such as EpiDoc TEI XML).

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

US partners continued testing using the VMR to index and transcribe literary (non-Biblical) texts, and expanded testing to two more transcribers to explore the project management capabilities of the VMR. However, testers found that the VMR was not easily adaptable to the different genre of texts and to the Coptic SCRIPTORIUM data model. For this reason, US partners, in collaboration with German partner Uwe Sikora, are working to create an XML converter to facilitate moving data from the VMR through the Natural Language Processing tools created by US partners and detailed in the last interim report. Preliminary XML mapping has been completed and initial testing has begun. Testing and refinements will continue throughout Winter 2017.

Since it is not practical to use the VMR to transcribe text for Coptic SCRIPTORIUM, US partner Amir Zeldes and graduate student Shuo Zhang at Georgetown University have created an online XML editor that can be configured to use the TEI XML schema used by Coptic SCRIPTORIUM, as well as provide version control system support for data archival and sustainability. The editor, called GitDOX (Git Datastorage Online XML editor), uses Github directly to store and share files online, thereby removing the need to install version control software for individual users, and allowing an administrator to update the XML schema centrally whenever needed.

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

Work has continued to integrate the web-based spreadsheet EtherCalc into our annotation process. The initial testing mentioned in the previous interim report of May 2016 revealed a critical issue with EtherCalc, which has since been resolved. We plan to test the newest version of EtherCalc next, and if all issues are now resolved, to move forward in using it as a basis for an online multi-layer annotation tool.

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

During the current project phrase, KELLIA achieved a significant goal: the release of an online Coptic lexicon linked to Coptic SCRIPTORIUM corpora in ANNIS (https://corpling.uis.georgetown.edu/annis/scriptorium), the search and visualization interface. The lexicon is encoded in TEI-XML, and entries represent Coptic words of all dialects with translations available in English, German, and French. It was developed at the Berlin-Brandenburg Academy of Sciences by German partner Frank Feder, with significant collaboration from Maxim Kupreyev. Sonja Dahlgren, Julien Delhez, Frank Feder, Lena Krastel, Maxim Kupreyev, Tonio Sebastian Richter, and Anne Sörgel contributed to compiling lexical information.

US partner Amir Zeldes and graduate student Emma Manning, both of Georgetown University, created a standalone website (https://corpling.uis.georgetown.edu/coptic-dictionary/), which is available as a pilot on the Georgetown server and was introduced at the KELLIA workshop in Claremont in July. The search capability allows users to search for Coptic words of all dialects, as well as reverse lookup of English, German, or French translations. Each entry also includes a link to a scanned version of W.E. Crum's seminal Coptic Dictionary, available on the Digital Edition of the Coptic Old Testament page from partners at the University of Göttingen.

The online Coptic dictionary has been integrated into Coptic SCRIPTORIUM's ANNIS interface. Entries in the online dictionary are linked to Coptic SCRIPTORIUM corpora though the ANNIS icon. Normalized editions of Coptic SCRIPTORIUM texts also link individual words back to entries in the online dictionary.

US partner Amir Zeldes and contractor Elizabeth Davidson are continuing to expand the Coptic treebank based on Coptic SCRIPTORIUM corpora. The syntactical analysis in the treebank is foundational to further work on automatic entity tagging in the corpora, since syntactic annotation allows us to recognize the word borders for covered by entity mentions and to resolve their discourse relations to each other (e.g. apposition, pronominalization and subsequent mention). The latest version of the treebank was released as part of the Universal Dependencies

project in version 1.4 in November (see http://universaldependencies.org/). The Treebank now contains 5220 tokens completely annotated for syntactic structure and grammatical function (see https://corpling.uis.georgetown.edu/coptic-treebank/ for the latest status).

US partner Elizabeth Platte, in collaboration with German partner Uwe Sikora, has been identifying opportunities for linked data across digital Coptic projects and other digital projects focused on the ancient world. Platte has been working to collect information from tagged locations from Coptic SCRIPTORIUM data to link to the Pleiades Project (https://pleiades.stoa.org/), an open online gazetteer of the ancient world which provides URIs for places and locations. Pleiades is an attractive platform for linked open data for many reasons, including the ability to add locations. Platte is therefore developing a spreadsheet to collect the information necessary to add locations mentioned in Coptic corpora and metadata but not yet available in Pleiades. This spreadsheet will be posted on a wiki available to KELLIA members.

Finally, German partner So Miyagawa, in collaboration with US partner Elizabeth Platte, has created a website for the KELLIA project which will be publicly available shortly. The website provides links The website is hosted on a server at the University of Göttingen, and the code is available on a GitHub repository, which can be accessed by members of all KELLIA-associated projects. The website can therefore be easily updated as necessary and will is sustainable in the long term.

Events

The Advisory Board for the U.S. KELLIA partners met virtually over email in the fall. In addition, individual board members were consulted on specific matters (such as linguistics questions or citation practices) as needed.

Elizabeth Platte and US P.I. Caroline Schroeder attended Linking the Big Ancient Mediterranean (https://www.lib.uiowa.edu/bam/linking-the-big-ancient-mediterranean-conference-june-6-8-2016/), a symposium co-organized by the US affiliated partner Paul Dilley at the University of Iowa. Platte and Schroeder presented a paper during the symposium on Coptic SCRIPTORIUM, focusing on entity recognition and opportunities for linked data. They made valuable connections with other digital projects working in the ancient world.

The second KELLIA Workshop, as proposed in the original grant application, was held in Claremont, California on July 23 and 24, immediately preceding the International Congress of Coptic Studies. Attendees included members of Coptic SCRIPTORIUM, the Digital Edition of Coptic Old Testament, the Institute for New Testament Textual Research, the Thesaurus Linguae Aegyptiae, and the Database and Dictionary of Greek Loanwords in Coptic, as well as affiliated projects. The program from the workshop is included below, as Appendix 1, and further information about products of the workshop are available in the discussion of outcomes above. We also continued conversations about standards for transcribing and segmenting Coptic for digital and computational work.

Finally, US KELLIA members met both in-person and via Skype during the Society of Biblical Literature annual meeting on November 20. They discussed the challenges of adapting the VMR to the Coptic SCRIPTORIUM data model and work on the converter and presented an initial preview of the transcription editor described above (Outcome 3).

Papers resulting from this project phase

US KELLIA members presented papers in two panels at the International Congress of Coptic Studies (ICCS) in Claremont, California (July 25-30). The program for these panels follows.

Coptic Digital Studies, David Brakke, chair

- Prof. Dr. Caroline Schroeder, Coptic SCRIPTORIUM: A Digital Platform for Research in Coptic Language and Literature
- Dr. Christine Luckritz Marquis, Reimagining the Apophthegmata Patrum in a Digital Culture
- Prof. Amir Zeldes, A Quantitative Approach to Syntactic Alternations in Sahidic
- Dr. Rebecca Krawiec, Charting Rhetorical Choices in Shenoute: Abraham our Father and I See Your Eagerness as case-studies

Coptic Digital Humanities, Caroline T. Schroeder, chair

- Dr. Paul Dilley, Coptic Scriptorium beyond the Manuscript: Towards a Distant Reading of Coptic Texts
- Mr. So Miyagawa and Dr. Marco Büchler, Computational Analysis of Text Reuse in Shenoute and Besa
- Mr. Uwe Sikora, Text Encoding – Opportunities and Challenges
- Ms. Eliese-Sophia Lincke, Optical Character Recognition (OCR) for Coptic. Testing Automated Digitization of Texts with OCRopy

Additional conference papers and presentations:

Zeldes, Amir and Schroeder, Caroline T. (2016) "An NLP Pipeline for Coptic". In: Proceedings of LaTeCH 2016 - The 10th SIGHUM Workshop at the Annual Meeting of the ACL. Berlin, 146-155.

Platte, Elizabeth and Schroeder, Caroline T. “Coptic Scriptorium: Data from the Desert.” Linking the Big Ancient Mediterranean Conference, University of Iowa, June 6-8, 2016.

KELLIA members also hosted two workshops at the IACS, a Workshop on Coptic Fonts & Coptic Bible led by Christian Askeland and Frank Feder and Digital Tools for Beginners (Workshop on Coptic SCRIPTORIUM), led by Caroline T. Schroeder, Amir Zeldes, and Rebecca S. Krawiec.

Appendix 1: KELLIA Workshop program

2016 KELLIA meeting
Saturday, July 23 and Sunday, July 24
Burkle Building, Room 12
Claremont Graduate University

Each portion of the program will consist of a short, informal presentation by the listed speaker(s) followed by discussion.

Participants:
Heike Behlmer, Georg-August-Universität Göttingen
Paul Dilley, University of Iowa
Frank Feder, Akademie der Wissenschaften zu Göttingen
Troy Griffitts, Akademie der Wissenschaften zu Göttingen
Christine Luckritz Marquis, Union Presbyterian Seminary
Rebecca Krawiec, Canisius College
Maxim Kupreyev, Berlin-Brandenburg Academy of Sciences and Humanities
So Miyagawa, Georg-August-Universität Göttingen
Beth Platte, Coptic SCRIPTORIUM
Sebastian Richter, Freie Universität Berlin (unable to attend due to flight cancellation)
Caroline Schroeder, University of the Pacific
Melissa Harl Sellew, University of Minnesota
Uwe Sikora, SUB Göttingen
Alin Suciu, Akademie der Wissenschaften zu Göttingen
Amir Zeldes, Georgetown University

Saturday, July 23

9:00 Welcome

9:15 Uwe Sikora: Data model
So Miyagawa: Coptic OCR with a neural network modelled OCR engine and high-performance computing

10:15 Break

10:30 Maxim Kupreyev: TLA Lexicon update

11:30 Amir Zeldes: Web Application for the TLA lexicon

12:30 lunch

1:45 Amir Zeldes: NLP Pipeline and spreadsheet editor

2:45 Troy Griffitts: VMR and satellite sites

3:45 Break

4:00 Frank Feder: Transcription guidelines (developed with C. Askeland)

Sunday, July 24

9:00 Welcome

9:15 Frank Feder: Report on OT Base text

10:15 Break

10:30 Beth Platte/Amir Zeldes: Entities

11:30 Sebastian Richter: Database and Dictionary of Greek Loanwords in Coptic (note: due to a flight cancellation, Sebastian Richter was unable to attend, but Dylan Burns of the DDGLC generously presented a note on the project in his absence.)

12:30 lunch

1:45 Paul Dilley: Big Ancient Mediterranean/Iowa Canon of Coptic Authors and Works project report

2:45 Break

3:00 Melissa Harl Sellew: Ancient Lives project report

4:00 Wrap up/plan for next year