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 28, 2017

Report Authors:

Elizabeth Platte, Reed College

Caroline T. Schroeder, University of the Pacific

Amir Zeldes, Georgetown University

Introduction

This report details the activities of the US members of the KELLIA grant since the last report was filed on May 30, 2017. This report will also reference work done by German partners as part of the bilateral grant, but it is intended as a report on the progress made by US partners.

During this period, US partners Caroline T. Schroeder (University of the Pacific); Amir Zeldes (Georgetown University); Elizabeth Platte (KELLIA Digital Humanities Specialist and Project Manager); and Rebecca Krawiec (Canisius College) met with German partners and representatives or related projects at the 2017 KELLIA meeting at the University of Göttingen June 19 through 23, 2017. Some results from this meeting will be mentioned in the following report.

Outcome 1 - Milestones for data standards

Outcome 1 is primarily the responsibility of German partners. So Miyagawa, Uwe Sikora, and Tiffany Ziegler of the University of Göttingen circulated a drafts of a white papers concerning data curation, encoding, linked data, and metadata standards. Miyagawa presented on the recommendations at the KELLIA meeting in June.

Outcome 2 - Server-based batch conversion tools

As mentioned in the May 30 interim report, the US partner Amir Zeldes, along with graduate students Shuo Zhang and Emma Manning, have created GitDox, a version-controlled annotation interface for transcribing and annotating texts (see Outcomes 3 and 4 for more information about recent progress on GitDox). The GitDox admin interface now includes options for bulk upload of SGML files into either the XML or spreadsheet editor modes, which has allowed the project to enter all previously published texts into GitDox.

Additionally, GitDox now supports bulk download of documents. Documents can be filtered by corpus and status, allowing a user to select only documents ready for publication for download. Documents can be downloaded in either XML or spreadsheet mode. Documents in spreadsheet mode can be downloaded in custom configurations. The bulk download feature streamlines the conversion process for publication of texts in our database (ANNIS) and in our archival formats (TEI XML, relANNIS, PAULA XML) on GitHub.

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

As of this summer, a new, experimental tokenizer created by US Partner Amir Zeldes is available using the web interface for Coptic SCRIPTORIUM's NLP tools (https://corpling.uis.georgetown.edu/coptic-nlp/). In initial tests, this tokenizer reduced the error rate by approximately 50%. Once the experimental tokenizer proves stable, it will be integrated into the GitDox workflow.

Since the last report, GitDox, the annotation interface developed by Zeldes, Zhang, and Manning (See outcome 2, above), has been updated and now supports the entire transcription and annotation process. Annotators are now able to access both tokenization (already available as of the last report) and other analysis (normalization, lemmatization, and part of speech and language of origin tagging) from Coptic SCRIPTORIUM's NLP service directly from GitDox. The option to send a transcribed text through the NLP now links directly to the spreadsheet mode of GitDox (see outcome 4 below).

In addition, Zeldes and Manning have created more robust validation for XML schemas, spreadsheet annotations, and metadata in GitDox. Individual documents can be validated within the annotation interface, with validation errors for both metadata and either the XML schema or spreadsheet annotations, depending on the mode of the document. Validation for all documents is also available from the document list. Icons on the document list appear in green for passed validations and red for failed validations; a list of validation errors appears when a user hovers over the icon. Validation rules for both spreadsheet annotations and metadata are configurable via the administration interface.

Due to these updates to GitDox, as well as the batch import and export functions outlined in outcome 2, above, Coptic SCRIPTORIUM has been able to migrate all previously published texts to GitDox, and we are now using GitDox exclusively for both preparing new material and updating published material. This month (November 2017), Coptic SCRIPTORIUM completed the first publication of texts transcribed and edited in GitDox (see outcome 5, below, for more information about this publication).

Finally, as of this summer, GitDox has also been used by non-Coptic projects, including at the 2017 Linguistics Summer Institute. Extensive documentation for GitDox is available at a standalone website for the interface (https://corpling.uis.georgetown.edu/gitdox/), while documentation for using GitDox to transcribe and annotate Coptic SCRIPTORIUM texts is available on our wiki (http://wiki.copticscriptorium.org/doku.php?id=gitdox_workflow).

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

The May 30 interim report noted that the web-based spreadsheet program EtherCalc had been integrated into GitDox; the link between the XML editing mode and spreadsheet editing mode via the NLP pipeline is now functional. GitDox validation for the spreadsheet includes cell highlighting and is more robust than validation previously available to Coptic SCRIPTORIUM for spreadsheet editing. The possibility of jointly editing spreadsheets without conflicts as well as more thorough validation streamlines the annotation and editing processes. (See also Outcome 2)

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

In June, a word frequency feature using data from Coptic SCRIPTORIUM corpora was added to the online Coptic lexicon. The lexicon was prepared at the Berlin-Brandenburg Academy of Sciences by

German partner Frank Feder, with significant support from Maxim Kupreyev and other collaborators, and the online interface was created by US partner Amir Zeldes and graduate student Emma Manning as part of the KELLIA project. Users of the lexicon can now hover over the frequency icon at the top of the entry to see the frequency of the word and lemma per 100,000 as well as the rank of the work in Coptic SCRIPTORIUM corpora.

On August 26, 2017, Coptic SCRIPTORIUM released version 2.1 of the Coptic Treebank (https://corpling.uis.georgetown.edu/coptic-treebank/), which now includes over 10,000 tokens. Three letters of Besa have been added to the treebank, including On Lack of Food, which was transcribed by So Miyagawa. Treebank annotation was done by US partner Amir Zeldes and contractor Elizabeth Davidson.

The May 30 interim report outlined plans to link queries to geographic entities in Coptic SCRIPTORIUM to Pleiades (https://pleiades.stoa.org/) through Pelagios Commons (http://commons.pelagios.org/). US partner Elizabeth Platte has prepared publicly available RDF and VoID files following Pelagios standards for all corpora prior to the Old Testament release in June (https://github.com/CopticScriptorium/pelagios-dataset-summary). Query links were formulated to be easy to update after future Coptic SCRIPTORIUM data releases; Platte will document the process for updating the files on the Coptic SCRIPTORIUM wiki. The current data are already viewable in Pelagios' geographic visualization, Peripleo (http://peripleo.no5.at/ui#selected=http%3A%2F%2Fcorpling.uis.georgetown.edu%2Fannis%2Fscriptoriummy-dataset). The same data will be visible in Pleiades via the Pelagios API shortly, pending an update to Pelagios 3 planned for December 2017.

Since the last interim report, Coptic SCRIPTORIUM has released new corpus data on two occasions. In June 2017, we released the automatically annotated Coptic Old Testament, which was mentioned as forthcoming in the May 30 report. This corpus is based on the version made available by the the version of the available texts kindly provided by the CrossWire Bible Society SWORD Project thanks to work by Christian Askeland, Matthias Schulz and Troy Griffitts.

On November 21, 2017, Coptic SCRIPTORIUM announced release version 2.4.0 of our corpus data. This release included material provided by Alin Suciu (Pseudo-Theophilus, On the Cross and the Thief); David Brakke (Shenoute, Some Kinds of People Sift Dirt); and Diliana Atanassova (the Canons of Apa Johannes, http://coptot.manuscriptroom.com/web/apa-johannes). The release also includes an excerpt of the Martyrdom of Saint Victor the General, an out-of-copyright text originally published by E. A. Wallis Budge and provided in digitized form by the Marcion project (http://marcion.sourceforge.net/). Finally, we added several apophthegms to Apophthegmata Patrum (Sayings of the Desert Fathers) corpus. Some of these documents are available thanks to the work of new annotators, including those who attended the NAPS workshop in May 2017, which was described in the previous interim report. Together, these texts add over 6,000 tokens to our data.

We have begun adding versification to released texts in order to enable better citation of electronic texts. In most cases, no prior versification scheme exists. Our guidelines on versification of texts is as follows. Texts with prior versification (i.e., Pseudo Theophilus with chapters designated by

Suciu) should follow existing versification. Verses generally should follow sentence structure. Long compound or complex sentences may be separated into multiple verses at linguistically appropriate junctures (conjugations, new main verb clause or subject/verb pairing, etc.). Verses may be clustered into chapters; for texts digitized directly from manuscript transcriptions, chapter divisions typically follow the rendering of ekthesis in manuscripts where applicable. (Ekthetic letters protrude into the left margin and often are also written in a larger size.) These guidelines do not cover all circumstances and will need to be revised as we release more material. In addition, we have not yet addressed the versification of texts or manuscripts with large lacunae.

Events

On June 19 to 23, 2017, US partners Amir Zeldes, Caroline T. Schroeder, Elizabeth Platte, and Rebecca Krawiec attended the 2017 KELLIA meeting at the University of Göttingen. We met with our German partners to provide updates on our respective progress on the bilateral grant, and we also had the opportunity to hear from members of other relevant projects, including, eTrap led by Dr. Marco Büchler (http://www.etrap.eu/) and PATHs led by Dr. Paola Buzi (http://paths.uniroma1.it/). The schedule for KELLIA 2017 is attached as Appendix 1. US partners also held an internal meeting during that time, with Christine Luckritz Marquis (Union Presbyterian Seminary) joining remotely. The agenda for that meeting is attached as Appendix 2.

On November 6-10, Caroline T. Schroeder met with Global Philology Project and Open Greek and Latin Project members at the University of Leipzig (Humboldt Chair of Digital Humanities): Dr. Gregory Crane, Dr. Matthew Munson, Dr. Monica Berti, and Dr. Giuseppe Celano.

At the annual Society of Biblical Literature/American Academy of Religion meeting in Boston (November 2017), Caroline T. Schroeder met with project contributors Dr. Christine Luckritz Marquis and Dr. Rebecca Krawiec and new advisory board member Dr. Janet Timbie to discuss project progress and future plans. Schroeder also met with KELLIA collaborator Dr. Frank Feder for updates on KELLIA work in Germany and the U.S.

Presentations resulting from this project phase

Amir Zeldes, "Web-based Natural Language Processing Pipeline and Online Spreadsheet for Coptic Digital Humanities," Georg-August University, Göttingen, 19 June 2017

Amir Zeldes, "Coptic Treebanking," Georg-August University, Göttingen, 19 June 2017

Caroline T. Schroeder, "Web-based Natural Language Processing and Annotation Technologies for Ancient Languages," Leipzig, 6 November 2017

Appendix 1: KELLIA 2017 Meeting Agenda, June 19 to 23, 2017

In attendance: Diliana Atanassova; Heike Behlmer; Julian Bogdani; Marco Büchler; Paola Buzi; Julien Delhez; Frank Feder; Troy Griffitts; Rebecca Krawiec (Coptic SCRIPTORIUM); Theresa Kohl; Maxim Kupreyev; So Miyagawa; Matthew Munson; Elizabeth Platte (Coptic SCRIPTORIUM); Sebastian Richter; Malte Rosenau; Ulrich Schmid; Caroline T. Schroeder (Coptic SCRIPTORIUM); Laura Slaughter; Agostino Soldati; Alin Suciu; Alberto Winterberg; Amir Zeldes (Coptic SCRIPTORIUM)

Monday, June 19

10:00-12:00

Lagardehaus: guided visit forKELLIA participants by CoptOT staff; Internal Coptic SCRIPTORIUM meeting; (simultaneously: CoptOT Steering Committee Meeting in Heyne-Haus)

12:00-13:30 lunch

13:30-15:00 Agenda; Discussion of progress from last meeting

15:00-15:30 Coffee Break

15:30 Discussion with eTrap members (Text re-use) OCR of Coptic Texts (So, Kirill, and Marco)

Tuesday, June 20

9:00–9:30 NLP pipeline and online spreadsheet (Amir)

9:30-10:00 Coptic Treebank (Amir)

10:00-12:00 Exchange formats and text segmentation in digital editions: discussion

12:00-13:30 lunch

13:30-14:00 Metadata standards and exchange (So, Beth, Ulrich, Troy)

14:00-15:00 News from CoptOT: Infrastructure, Base-Texts, and Metadata (Frank, Malte, Ulrich, Troy) & Discussion

15:00-15:30 Coffee Break

15:30 Topic modelling in the Greek NT (Paul, via Videolink)

Wednesday, June 21

9:00–10:00 Discussion about the Coptic Dictionary Online

10:00-10:30 Presentation of the actual status of the Coptic Lemma List and BTS (Maxim)

10:30 DDGLC (Katrin via Skype)

12:00-13:30 lunch

13:30-14:00 PAThs presentation (Paola)

14:00-14:30 Digital Infrastructure of PAThs (Julian) & Discussion

15:00-15:30 Coffee Break

15:30 Colophons of Biblical and other Mss (Agostino) & Discussion of future cooperation

Thursday, June 22

9:00-10:30 C(anonical)T(ext)S(ervice) and CapiTainS Suite (Matt) & Discussion

10:30-12:00 Discussion about Unicode

12:00-13:30 lunch

13:30-15:00 How to construct a Coptic wordnet? (Laura) & Discussion
15:00-15:30 Coffee Break
15:30 Internal Coptic SCRIPTORIUM meeting

Friday, June 23
10:00-12:00 Discussion of results; further steps; final report for KELLIA
12:00-13:30 lunch

Appendix 2: Coptic SCRIPTORIUM internal meeting, June 22, 2017

Participants

Rebecca Krawiec (in Göttingen)
Beth Platte (in Göttingen)
Carrie Schroeder (in Göttingen)
Amir Zeldes (in Göttingen)
Christie Luckritz Marquis (remote)

Future "big picture" priorities

- Corpora
- Development effort
- Collaborations (PATHs, Global Philology)
- Outreach
- Publications
- Versification

Priorities regarding specific features/issues in current projects

- GitDox (controlled vocabularies, spreadsheet features)
- Wiki updates
- Website update
- XRenner updates
- Metadata: Add links to manuscript images in public data (museum/lib websites) when available
- Online Pepper converter

General check-in

Discussion of additional future goals

Action items based on meeting discussions