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, 2015

Introduction

Our first interim report for the project KELLIA traces our current status with regard to the work packages defined in out grant application on the U.S. side of this bilateral grant. We give an overview of meetings, decisions, important events and software and data releases. Although we will refer to both sides of our bilateral activities, this report is not meant as an exhaustive account of progress on the German/DFG bilateral component of our project.

Grant activities took place individually at the co-PIs' institutions in Stockton, CA and Washington, DC, as well as at **a bilateral workshop in Göttingen**, hosted by the German partners, and a project meeting at the **Society of Biblical Literature in Atlanta**. In most work packages, the project is currently ahead of schedule. The report is ordered along the lines of the five outcome areas defined in the project proposal.

Individual Progress – Outcome Areas

Outcome 1 - Milestones for Data Standards

Substantial progress has been made on researching existing metadata and encoding standards in cooperation with German partners, who are primarily responsible for this outcome. This progress is in large part due to the workshop in Göttingen, which took place in September (see Events for the detailed agenda discussed at this workshop).

For the majority of relevant materials, an online spreadsheet has been introduced by the partners in Göttingen, which is being worked on from both sides to achieve uniform nomenclature. Data from main existing resources in and beyond the field of Coptic studies have been or are set to be linked with our data sources, notably the Trismegistos project's TM numbers (http://www.trismegistos.org/), which have already been collected, and corresponding keys from the CMCL's (Corpus dei Manoscritti Copti Letterari) Clavis Patrum Copticorum. Some outstanding work is still planned, led by the German partners, including a discussion of which of the collected resources will be represented in the TEI XML representation of project data, as well as a publically available whitepaper describing our metadata standards.

Outcome 2 – Server based batch conversion tools

We have advanced on several points in producing server based conversion tools:

- Within the annotation process, we have released a new Natural Language Processing (NLP) pipeline for Coptic, which supplies both a human readable Web interface to create linguistically analyzed data automatically, and a machine readable REST API for batch annotation of data. The pipeline runs the tools developed in our NEH ODH Start Up grant (HD-51907) and Preservation and Access Foundations grant (HD-51907): tokenizer, normalizer, part of speech tagger, language of origin tagger, and lemmatizer. The resources are at: http://corpling.uis.georgetown.edu/coptic-nlp/
- We have just released a new component of our natural language processing tools automating morphological analysis, further reducing annotation effort.

These new resources mean a very substantial reduction of manual involvement in adding new resources to our collection or updating existing ones. Some batch automation of the search engine indexing using our

search engine ANNIS (http://corpling.uis.georgetown.edu/annis/scriptorium/) remains to be done in this outcome area.

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

In November we were able to expose the linguistic annotation pipeline mentioned in outcome 2 as an API. As a result, German partners working on the VMR software have now begun integrating our language processing resources as remote tools by addressing the API. There is as yet no release of these new software developments, and tests are ongoing.

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

At our meeting in Göttingen (see Events below) the American co-PIs presented their choice of software infrastructure to begin the development of a multilayer, collaborative annotation tools for the Coptic data. As an underlying software engine we have selected EtherCalc (https://ethercalc.net/), a freely available, open source Web spreadsheet software similar to Google spreadsheets, but which can be run directly on a server managed by the project, as well as being open to further development.

In the first phase of the project, we have created an API which programmatically streams our linguistically analyzed data into the Web spreadsheet, allowing us to display an analyzed text automatically in a collaborative view that is editable to project partners in different locations.

At present, this development is not yet capable of exporting the edited data, and has not been adapted to the specific needs of our project annotators, so that there is still much work to be done. However this proof of concept is already a very promising first step, which we will continue to work on in the coming year.

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

This work area is still in its preliminary stages. We have identified automatic syntactic analysis tools as being the primary prerequisite to enabling work on entity recognition, for entities which can subsequently be linked to standard identifiers. Entities and a syntactic analysis in itself can be leveraged textual re-use recognition, so that these goals go together hand in hand.

As a result, we are currently developing guidelines for syntactic analysis and testing software for its automation. We expect that demonstrable results are still some months away.

Events

KELLIA Kickoff Workshop – Göttingen, September 7-11, 2015

Participants:

- **CoptOT Project:** Diliana Atanassova, Frank Feder

- **KELLIA:** Heike Behlmer, Troy Griffitts, So Miyagawa, Ulrich Schmid, Caroline T. Schroeder, Uwe Sikora, Amir Zeldes
- **GCDH:** Marco Büchler
- **TLA:** Maxim Kupreyev, Simon Schweitzer, Sebastian Richter
- **DDGLC:** Katrin John, Sebastian Richter
- **INTF:** Siegfried Richter

Agenda

- Presentation of NLP pipeline
- Discuss API for communication with VMR
- Discuss segmentation of Coptic words:
    - Subword morphemes (mnt-, r- & co)
    - Bound groups (using Till vs. Layton's convention)
    - Word boundaries (esp. handling of compounds)
- VMR and annotation tools (Carrie, Amir, Ulrich, Troy, So)
    - TEI subsets for annotation
    - VMR and Scriptorium teams tag subsets / standards
    - Using the VMR for Scriptorium data
    - Other annotation interfaces (EtherCalc prototype)
    - Github connection (committing VMR material/EtherCalc)
- Lemmatization
    - Show new lemmatizer
    - Discuss use of TLA lemma list
    - Technical integration
- Metadata scan - Uwe (+Carrie, Amir, Heike, So)
    - Persistent URNs – progress at data.copticscriptorium.org
    - Trismegistos and CTS URN
    - Authority files and vocabularies
- Linked data - parsing, entities (people, geo, …), coreference
- Versification and indexation
- Collaboration with TLA
    - Crum in XML
    - TLA in JSON, couchDB and their software which runs on Java
    - Funk's lemma list
    - The Funk corpus
    - License issues
    - Hyper-lemmatization for diachronic research

Schedule

Monday, September 7
5:45 pm Meeting with Marco Büchler
7:30 pm Dinner at APEX, Burgstraße

Tuesday, September 8
Simon, Maxim (TLA); CoptOT; Kellia
09:00

- Presentation of the new BTS
    - Show new TLA lemmatizer
    - Discuss use of TLA lemma list
    - Lemmatization

Afternoon: Metadata roundtable

Wednesday, September 9
Morning: Presentations from Maxim (TLA), Katrin John (DDGLC); Siegfried (INTF); CoptOT; Kellia

Afternoon: Discussion:
- Segmentation of Coptic texts
- Use of the VMR, Coordination and Cooperation of CoptOT and INTF

Thursday, September 10
Presentations:
Sebastian (TLA; DDGLC); Siegfried (INTF); CoptOT; Kellia

Friday, September 11
General Discussion and Summary
Decisions and Next Steps
Dissemination

SBL Meeting 2015, Atlanta, November 20

Participants:

- **KELLIA:** Troy Griffitts, So Miyagawa, Elizabeth Platte, Ulrich Schmid, Caroline T. Schroeder, Uwe Sikora, Amir Zeldes
- **CoptOT:** Frank Feder
- **Coptic SCRIPTORIUM:** Rebecca S. Krawiec
- **Guests:** Janet Timbie, Christian Askeland

Agenda:
- So Miyagawa presenting work with Uwe Sikora, Tiffany Ziegler, on metadata
    - 2 Databases in development: a database of Coptic manuscripts; a database of existing databases of Coptic manuscripts
    - Sahidic MS references (Biblia Coptica and Schmitz-Mink-Richter)
- Literary text test in VMR – work on transcribing literary texts in VMR by So Miyagawa with help from Ulrich and Troy
- Progress on NLP (Amir Zeldes)
    - New morphological analysis component
- TLA lexicon update from Frank Feder
- Transcription guidelines update from Frank Feder and Christian Askeland
- Planning for future work and workshop in Claremont at the Congress of the International Association of Coptic Studies (July 2016)