Semi-Annual Performance Progress Report

Report ID: 2891720
Application Number: HAA-261271-18
Project Director: Amir Zeldes (amir.zeldes@georgetown.edu)
Institution: Georgetown University
Reporting Period: 9/1/2018 -2/28/2019
Report Due: 3/31/2019
Submission Date: 4/6/2019 9:14:29 PM

Accomplishments

What were the major goals of the project?

List the major goals of the project as stated in the approved application or as approved by the agency, including the technological objectives of this effort. Describe the proposed technical approach to obtain those goals. If the application listed milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion. Generally, the goals will not change from one reporting period to the next and are unlikely to change during the final reporting period. However, if NEH approved changes to the goals during the reporting period, list the revised goals and objectives. Also explain any significant changes in approach or methods from the agency-approved application or plan.

* Corpus data acquisition and prioritization * Standard normalization and morphological analysis * Linked metadata and geographical data linking with the PATHS project * Ongoing extension of the Coptic Treebank * Pilot Named Entity Recognition annotation

What was accomplished under these goals?

For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results or key outcomes, including major findings, developments, or conclusions (both positive and negative); and/or 4) other achievements. Include a discussion of stated goals not met. As the project progresses to completion, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

Corpus data acquisition and prioritization ========== A major goal of the current project is to scale up the high quality, manually curated seed corpus developed in our previous projects, into a mature, broad coverage resource an order of magnitude larger than the previously available data. This goal is strategic and qualitatively transformative, as high coverage is a prerequisite for a number of applications, including the study of intertextualilty, textual reuse and the transmission history of ideas, as well as large scale quantitative linguistic study of language change in Coptic internally and vis-a-vis Ancient Egyptian, or language contact studies examining the role of Greek in the formation of Coptic as an independent language stage. A major challenge in realizing a large scale corpus of Coptic is the high level of diversity in text encoding standards, not only technically (in the sense of XML formats), but also in word segmentation practices, orthographic normalization, and encoding of document structure (e.g. treatment of punctuation). As a first step towards growing our corpus in the first months of the current project, we have adapted the existing Sahidic Coptic Old Testament materials, provided by courtesy of the Digital Edition of the Coptic Old Testament project (CoptOT, http://coptot.manuscriptroom.com/), and converted them into our format and normalization scheme. In addition to the CoptOT data, we increased the size of our holdings for works by Shenoute of Atripe (Abraham Our Father and Acephalous Work 22), released updated versions of our manually curated Gospel of Mark and 1 Corinthians, and the Martyrdom of Victor the General (see our blog at http://blog.copticscriptorium.org/ for individual release notes). As a result, our publicly available data has seen an increase from some 546,000 tokens at the beginning of the project, to over 810,000

tokens at the end of the report period, our largest six month increase to date, owing mainly to the newly released Old Testament corpus. We are hoping to clear the 1 million mark at the end of our first year. We also made arrangements with the PATHS project in Rome (http://paths.uniroma1.it/) to publish and annotate digital texts produced by that project. Standard normalization and morphological analysis ========== The Old Testament pilot required the development of some tools for standardization and analysis which are some of the first products of this grant period: Conversion scripts and translation merging scripts, which add the freely available Septuagint English translation by L. C. Brenton. A new detokenizer module, which adjusts word separation from the CoptOT standard to the linguistic standard used by Coptic Scriptorium following Layton's (2011) grammar. The module is also informally referred to as a 'Laytonizer', and ensures that users of our search interface will find words separated homogeneously using the same uniform standard as our other datasets. The original data, which was morphologically unanalyzed, was segmented and tagged using our NLP tools, leading to the release of a new version of our morphological analyzer. This module forms the basis for the machine learning tokenizer planned as a centerpiece of our infrastructure in this project phase. The accuracy of the new morphological analyzer is currently assessed at 94.5% (proportion of complex word forms that receive a perfect segmentation), a substantial improvement over a score of 90.21% that we reported using our older rule-based system in 2016. While we plan to improve further, we are already encouraged by these results, which prompted us to test the system on other benchmarks. As we initially envisioned in the grant proposal, our tools are being written with the possibility of being applied to other languages in mind. In a paper at the annual workshop of the Association for Computational Linguistics' Special Interest Group on Morphology and Phonlogy (SIG MORPHON), we achieved a new state of the art on segmenting Hebrew using the same segmenter we developped for Coptic (Zeldes 2018, see publications). Performing much better on Hebrew thanks to the larger dataset available and the comparative simplicity of Hebrew segmentation, we achieved 98.19% accuracy, nearly 4% better than the previous state of the art of 94.25%. Linked Metadata and Geographical Data ========== We began conversations with the PATHS project in Rome, which has recently published an online Archaeological Atlas for Coptic Literature (https://atlas.paths-erc.eu/). The Atlas contains entries and stable identifiers for geographic places, textbearing objects (including the codices containing texts already published in Coptic Scritporium), authors, and literary works. We are discussing how to link back and forth between our projects so that people studying Coptic literature have access to the geographical context of the literature through PATHS and the digitally annotated text through Coptic Scriptorium. We will also explore the possibility of annotations of PATHS ids and links within our texts for named geographic entities (see more on entity recognition generally above). Coptic Treebank ========== Work expanding the syntactically annotated Coptic Treebank is ongoing, and vital to our goal of supplying high quality parses and entity recognition for our data. Syntax trees allow us to find out 'who did what to whom' in texts automatically (e.g. find the subjects and objects of certain verbs, verbs associated with particular agents, and the span of words belonging to each mention of a nominal expression). From some 18,400 tokens at the beginning of this project phase, we have increased the size of the gold annotated dataset to 27,000 tokens, and are currently planning the release of the new data in the Universal Dependencies project (UD, version 2.4), which releases comparable treebanks in 76 languages (https://universaldependencies.org/). Our next milestone and goal for the subsequent UD 2.5 is to exceed 30,000 tokens. This benchmark will be particularly meaningful, as it is the threshold for participating in international competitions on syntactic parsing, such as last year's CoNLL shared task (http://universaldependencies.org/conll18/). Being included in the shared task would benefit Coptic infrastructure immensely, since it would mean that any parsing system developed for other languages, including English, would also be scored based on how well it

does for Coptic, effectively putting pressure on parser developers to also consider Coptic as a target. Named Entity Recognition ========== We are currently in the process of developing and testing guidelines for entity annotation in running Coptic text, constituting the first such project to our knowledge. Our pilot corpus is still very small, containing around 5,000 tokens at present, and as a result we do not yet have any evaluation results on automatic NER accuracy. As the dataset grows, we plan to set aside test and development partitions and run experiments using automatic systems. Linking to the Coptic Dictionary Online and Multiword Expressions ========== Though not a stated goal for the first two quarters of the project, we are continuing work on the Coptic Dictionary Online (https://corpling.uis.georgetown.edu/coptic-dictionary/), whose interface was designed in Georgetown as part of the KELLIA project, in collaboration with our German partners at the University of Göttingen and the Academies of Science of Göttingen and Berlin-Brandenburg (BBAW). Compatibility and interoperability with the dictionary, which is fast becoming one of the most popular online services of the project, is an important goal and a part of our commitment to Linked Open Data standards and data exchange with outside projects. In the current phase we have developed a pilot for a multiword expression recognition module, which allows us to move beyond linking of individual words in our dataset to the dictionary. Multiword expressions have entries in the Coptic Dictionary Online, which can now be reached from our corpora next to links for individual component words. Conversely, adding the multiword expression tags will allow us to collect frequency information for such entries, which has so far been unavailable. Adding these frequencies is a work item for the coming months.

How were the results disseminated to communities of interest?

If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities for the purpose of enhancing public understanding and increasing interest in learning and careers in the humanities.

We disseminated our results at our blog (http://blog.copticscriptorium.org/) and website (http://copticscriptorium.org), through the Digital-Coptic mailing list, social media, and a number of conference talks and publications: Miyagawa, So, Zeldes, Amir, Büchler, Marco, Behlmer, Heike and Griffitts, Troy (2018) "Building Linguistically and Intertextually-Tagged Coptic Corpora with Open Source Tools". Proceedings of JADH2018. Tokyo, Japan. Schroeder, Caroline T. (2019) "Understanding Space and Place through Digital Text Analysis". Third PAThs International Conference: Coptic Literature in Context. The Contexts of Coptic Literature: Late Antique Egypt in a dialogue between literature, archaeology and digital humanities. Sapienza University, Rome. Zeldes, Amir (2018) "A Characterwise Windowed Approach to Hebrew Morphological Segmentation". In: Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Brussels, Belgium, 101-110. Zeldes, Amir and Abrams, Mitchell (2018) "The Coptic Universal Dependency Treebank". In: Proceedings of the Universal Dependencies Workshop 2018. Brussels, Belgium, 192-201.

Participants and Other Collaborating Organizations

What individuals have worked on the project?

Provide the following information for: (1) project director(s) (PDs); and (2) key personnel. Provide the name and identify the role the person played in the project. Indicate the number of person-months (Calendar, Academic, Summer), rounding off to a whole month, that the individual worked on the project (a person-month equals approximately 160 hours of effort). Show the most senior role in which the person has worked on the project for any significant length of time For example, if an undergraduate student graduated, entered graduate school, and continued to work on the project, show that person as a graduate student, preferably explaining the change in involvement. Describe how this person contributed to the project and with what funding support. If information is unchanged from a previous progress report, provide the name only and indicate “no change.” Identify the person’s state, U.S. territory, and/or country of residence. If unknown, so indicate. State whether this person has collaborated internationally. If the participant was U.S.-based, state whether this person collaborated internationally with an individual located in a foreign country, and specify whether the person traveled to the foreign country as part of that collaboration, and, if so, what the duration of stay was. The foreign country(ies) should be identified. If the participant was not U.S.-based, state whether this person traveled to the United States or another country as part of a collaboration, and, if so, what the duration of stay was. The destination country should be identified.

Most work was accomplished by the co-PIs and only paid staff as of this phase of the project, Amir Zeldes (Georgetown University) and Caroline T. Schroeder (University of the Pacific). A number of students also worked on the project at Georgetown University, specifically Luke Gessler and Mitchell Abrams, both PhD students in Linguistics at GU. A pilot on entity annotation was begun with a PhD student at Catholic University of America, Lance Martin. Collaborators from our previous bilateral KELLIA project are also actively contributing, including Rebecca Krawiec, Elizabeth Platte, Elizabeth Davidson and Christine Luckritz Marquis.

Enter the number of project staff (project directors and key personnel) from each state or territory in the lists below.

Special Instructions:To assist us with understanding the geographic reach of our funded projects, please list all project staff in a table using the following column headers: Staff Member Last Name, Staff Member First Name, Title or Role, Institution, State or Territory, Country.

California -Caroline T. Schroeder

District of Columbia -Amir Zeldes

Impact

What was the impact on the development of the principal discipline(s) of the project?

If there is nothing significant to report during this reporting period, state "Nothing to

Report."

Describe how findings, results, techniques that were developed or extended, or other products from the project made an impact or are likely to make an impact on the base of knowledge, theory, and research in the principal disciplinary field(s) of the project. Summarize using language that a lay audience can understand.

How the field or discipline is defined is not as important as covering the impact the work has had on knowledge and technique. Make the best distinction possible, for example, by using a field or discipline, if appropriate, that corresponds with a single academic department.

The project is still in a preliminary phase, but the availability of new tools to make diverse types of Coptic language input more uniformly accessible is already making an impact. This includes making standardized data available to researchers, the release of new and more accurate natural language processing tools for researchers, and improvements to the Coptic Dictionary Online, all of which impact researchers in areas touching on early Christianity, history and linguistics of Egypt in the first millennium.

What was the impact on other disciplines?

If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the findings, results, techniques that were developed or improved, or other products from the project made an impact or are likely to make an impact on other disciplines.

Our tools turned out to be highly effective at analyzing other languages, raising the state of the art in morphological segmentation of Hebrew, despite being developed to benefit Coptic studies. Linking our data to other projects also widens the scope of impact of the project for other disciplines using geographical data.

What was the impact on teaching and educational experiences?

If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the project made an impact or is likely to make an impact on teaching and educational experiences. For example, did the project develop and disseminate new educational materials; lead to ideas for new approaches to course design or pedagogical methods; or develop online resources that will be useful for teachers and students and other school staff?

Special Instructions:Please also describe students who may have worked on your project, what role they played, mentoring activities, and how this work advanced their education.

Students at Georgetown actively participated in the development of annotated materials and software, such as annotations interfaces and corpus creation, allowing them to gain practical experience with digital humanities topics, as well as the domain of Coptic studies.

What was the impact on society beyond specialists in the humanities?

If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of the academic world and specialists in the humanities on areas such as: improving public knowledge, skills, and abilities; changing practices; or improving social, economic, or civic conditions.

Special Instructions:Please also describe students who may have worked on your project, what role they played, mentoring activities, and how this work advanced their education.

Nothing specific to report (yet!), but ultimately this project aims to make Coptic accessible online, which is the cultural heritage language of over 15 million Copts worldwide, including many in the USA. Just as one might expect to be able to read the Iliad in Greek in the 21st century with linked translations and linguistics analysis, we aim to bring Coptic literature into the public domain in the most accessible way possible.

Enter the total number of attendees.

Special Instructions:To better understand the reach of our grantees' activities, we ask that you please list any workshops, conferences, or other events held during this reporting period and list the number of external attendees (individuals not responsible for grant activities) as well as team members who attended the event.

None in this phase.

Changes/Problems

Changes in approach and reasons for change

Describe any changes in approach during the reporting period and reasons for these changes. Remember that significant changes in objectives and scope require prior approval of the agency.

Actual or anticipated problems or delays and actions or plans to resolve them

Describe problems or delays encountered during the reporting period and actions or plans to resolve them.

Changes that had a significant impact on expenditures

Describe changes during the reporting period that may have a significant impact on expenditures, for example, delays in hiring staff or favorable developments that enable meeting objectives at less cost than anticipated.

Change of primary performance site location from that originally proposed

Identify any change to the primary performance site location identified in the proposal, as originally submitted.

The project plan originally scheduled an in-person meeting of domestic project stakeholders together with the project directors -this meeting has now been moved to the week of May 27th 2019, due to scheduling difficulties, but we do not foresee any delays or problems for project goals as a result, and frequent video conferences in the interim have so far been sufficient for project coordination.

Special reporting requirements

Respond to any special reporting requirements specified in the award terms and conditions, as well as any award-specific reporting requirements

Project Directors are reminded that the award terms and conditions require an acknowledgment of federal funding agency support for any product (including World Wide Web pages) based on or developed under this award. Indicate whether the product included an acknowledgement of support, and describe how NEH support was or is acknowledged.

None

Project Outcomes

Describe any project outcomes in the space provided.

In this project phase we presented the following papers: Miyagawa, So, Zeldes, Amir, Büchler, Marco, Behlmer, Heike and Griffitts, Troy (2018) "Building Linguistically and Intertextually-Tagged Coptic Corpora with Open Source Tools". Proceedings of JADH2018. Tokyo, Japan. Schroeder, Caroline T. (2019) "Understanding Space and Place through Digital Text Analysis". Third PAThs International Conference: Coptic Literature in Context. The Contexts of Coptic Literature: Late Antique Egypt in a dialogue between literature, archaeology and digital humanities. Sapienza University, Rome. Zeldes, Amir (2018) "A Characterwise Windowed Approach to Hebrew Morphological Segmentation". In: Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Brussels, Belgium, 101-110. Zeldes, Amir and Abrams, Mitchell (2018) "The Coptic Universal Dependency Treebank". In: Proceedings of the Universal Dependencies Workshop 2018. Brussels, Belgium, 192-201.

Grant Products

Conference Paper/Presentation

Conference Paper/Presentation

Conference Paper/Presentation

Conference Paper/Presentation