

White Paper

Report ID: 112322

Application Number: HD-51907-14

Project Director: Caroline T. Schroeder (cschroeder@pacific.edu)

Institution: University of the Pacific

Reporting Period: 5/1/2014-12/31/2016

Report Due: 12/31/2016

Date Submitted: 12/30/2016

Coptic Scriptorium: A Corpus, Tools, and Methods for Corpus Linguistics and Computational Historical Research in Ancient Egypt

White Paper

National Endowment for the Humanities Office of Digital Humanities

Grant HD-51907-14

Director: Caroline T. Schroeder, University of the Pacific

Co-Director: Amir Zeldes, Georgetown University

Project website: copticscriptorium.org

December 30, 2016

Table of Contents

- [1. Project Description](#)
 - [2.1 Digitized Coptic Corpus in Multiple Formats and Visualizations](#)
 - [2.2 Digital and Computational Tools](#)
 - [2.3 ANNIS Database](#)
 - [2.4 Collaborative Platform](#)
 - [2.5 Documentation](#)
 - [2.6 Web application to provide easier user access for reading and citing visualizations of textual data](#)
 - [2.7 Symposium and workshop \(“Digital Coptic 2,” March 2015\)](#)
 - [2.8 Public tutorial and workshop](#)
 - [2.9 Articles and conference papers to distribute the results of our work](#)
 - [Peer-reviewed Articles](#)
 - [Published Conference Proceedings](#)
 - [Other Conference Presentations and Lectures](#)
 - [3. Project Activities](#)
 - [4. Audiences and Impact](#)
 - [5. Evaluation](#)
 - [6. Continuation of the Project](#)
- [Appendices](#)
- [Appendix A: List of Participants](#)
- [Appendix B: Website Screenshots](#)
- [Appendix C: Data Visualizations](#)
- [Appendix D: Digital Coptic 2 Program](#)
- [Appendix E: Guidelines for Citation, Transcription, Part of Speech Tagging, Lemmatization](#)

1. Project Description

Coptic, having evolved from the language of the hieroglyphs of the pharaonic era, represents the last phase of the Egyptian language and is pivotal for a wide range of disciplines, such as linguistics, biblical studies, the history of Christianity, Egyptology, and ancient history. Coptic SCRIPTORIUM has developed the first open-source technologies for computational and digital research across the disciplines as applied to Egyptian texts. We created a digitized corpus of Coptic texts available in multiple formats and visualizations (including TEI XML), tools to analyze and process the language (e.g., the first Coptic part-of-speech tagger), a database with search and visualization capabilities, and a collaborative platform for scholars to contribute texts and annotations and to conduct research. The technologies and corpus function as a collaborative environment for digital research by any scholars working in Coptic.

During the funded period, the project also received support from an NEH Preservation and Access Division, Humanities Collections and Reference Resources Foundations Grant (PW-51672-14). This White Paper contains a holistic account of project activities, concentrating on work pertaining to the ODH grant while also including activities that intersected with our other grant. This White Paper may contain similar, or in some places identical, phrasing as in the Final Performance Report and White Paper for PW-51672-14.¹

2. Accomplishments and Grant Products

The project achieved the primary objectives we set out to accomplish in the original proposal:

1. a digitized corpus of Coptic texts available in multiple formats and visualizations
2. digital and computational tools to analyze, process, and visualize the language (e.g., the first Coptic part-of-speech tagger, converters between digital formats and fonts, etc.)
3. a database created from the texts and tools using the ANNIS search and visualization infrastructure
4. a collaborative platform for scholars to contribute texts and annotations as well as conduct research using the corpus, tools, and database
5. documentation of project tools and technology

¹ Schroeder, Caroline T., Zeldes, Amir, and Elizabeth Platte. *NEH White Paper Report for Coptic SCRIPTORIUM: Digitizing a Corpus for Interdisciplinary Research in Ancient Egyptian*. National Endowment for the Humanities, 2016. Search for grant, White Paper, and grant products at: <https://securegrants.neh.gov/publicquery/main.aspx>. Also available on our website at <http://copticscriptorium.org/download/Coptic-SCRIPTORIUM-white-paper-2016-NEH-PW-51672-14.pdf>.

In addition to these outcomes, the project also achieved the following accomplishments, which were partially co-funded by an NEH Preservation and Access, Humanities Collection and Reference Resources Foundation grant (PW-51672-14):

6. Web application to provide easier user access for reading and citing visualizations of textual data
7. Symposium and workshop on digital Coptic Studies (“Digital Coptic 2,” March 2015)

Finally, the project also succeed in accomplishing:

8. Public tutorial and workshop on Coptic SCRIPTORIUM at the International Association of Coptic Studies
9. Articles and conference papers to distribute the results of our work

2.1 Digitized Coptic Corpus in Multiple Formats and Visualizations

Coptic SCRIPTORIUM uses the tools described in 2.2 to annotate digitized Coptic texts. Due to our interdisciplinary approach, we release our data in three formats.

- XML files using the EpiDoc subset (<https://sourceforge.net/p/epidoc/wiki/Home/>) of the Text Encoding Initiative (TEI) XML standards (<http://www.tei-c.org/index.xml>). EpiDoc is the standard in the field for producing digital editions of epigraphy, ancient manuscripts, papyrology, and other ancient text-bearing objects. Our EpiDoc TEI XML files do not contain the full set of annotations available in our corpora. The annotations in these files include information about text and manuscript structure, core philological and linguistic annotations (such as part of speech and loan words), and most metadata.
- XML files in standoff annotation using the PAULA XML format (<https://hal.inria.fr/hal-00783716>). These files contain the complete dataset of text and annotations for all corpora.
- Relational database files for use in the ANNIS search and visualization infrastructure (<http://corpus-tools.org/annis/>). ANNIS is the web-based database Coptic SCRIPTORIUM uses for search and visualization of the corpora. ANNIS can also be installed on a user’s desktop; the researcher can load our relANNIS database files into this local installation.

All the files are released under a Creative Commons Attribution (<https://creativecommons.org/licenses/by/4.0/>) license (either CC-BY 3.0 or 4.0 depending on date of corpus release) except the corpora derived from the Sahidica project, which operate under Sahidica’s original more limited license for academic use only.

All files can be downloaded from our Github site’s corpus repository at github.com/CopticScriptorium/corpora. We provide many links to this GitHub repository throughout our site. The version control system inherent in Git allows researchers to access previous versions of the digital files even as we update the corpora.

We have two web-based services to read and query the texts and annotations. On a Georgetown University server, an instance of the ANNIS database provides searchable access to our corpus and annotations (see [section 2.3](#) below). We also developed in collaboration with the concurrent PW-51672-14 grant a web application to enable easier access to and citation of texts using the CTS (Canonical Text Services) citation model, available at <http://data.copticscriptorium.org>.² (See also [section 2.6 below](#).)

Currently in addition to the search capacity in ANNIS, the annotated text data is automatically serialized into a reader friendly HTML format and styled using several CSS stylesheets to produce multiple visualizations in HTML. The visualizations are generated dynamically and cached, meaning that updates to the corpus can easily be made browsable, but access to visualizations is instantaneous for readers. Potential visualizations are expandable based on the data model. Each visualization is dependent on various combinations of annotations in our data model (which is described in the following two sections). Current visualizations include:

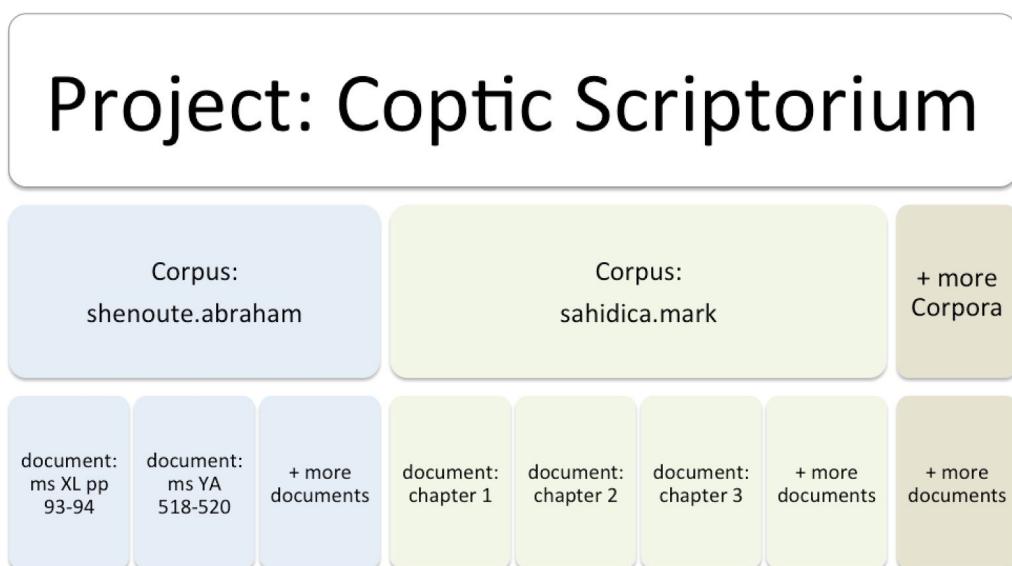
- Normalized view: the normalized Coptic text segmented into bound groups, with an English translation (when available) appearing as a pop-up when a cursor hovers over the text. Optimal for people wishing to read Coptic. Under the aegis of the NEH-DFG grant for the KELLIA project (HG-229371, see <http://copticscriptorium.org/kellia>), the normalized words now appear as hyperlinks to entries in a new online Coptic Dictionary (at <https://corpling.uis.georgetown.edu/coptic-dictionary>). Under the KELLIA grant, work is ongoing to enable proper mapping of words in our segmentation system to words segmented according to our German partners' system. There are currently occasional words where the links result in an empty search in the Dictionary.
- Diplomatic view: the diplomatic transcription of a manuscript page, which resembles the appearance of the manuscript page.
- Analytic view: an aligned visualization of the normalized Coptic text, part of speech tags, and the English translation. The manually curated Biblical corpora (currently the Gospel of Mark and 1 Corinthians) include an alignment with the Greek Bible, provided by the Society of Biblical Literature and Logos Software.
- Chapter view: Similar to the normalized view but divided into numbered chapters and verses for corpora such as the Bible which have existing canonical versification. As with the Normalized view, words link to the online Coptic dictionary.
- Histogram visualization for aggregate queries (frequency analysis for combinations of features extracted from a query).

Please see Appendix C for examples of these visualizations.

All corpora are developed and released using version control. Since this is the first project to develop natural language processing tools for the Egyptian language family, the tools and data

² Almas, Bridget, and Caroline T. Schroeder. "Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM." *Data Science Journal* 15.0 (2016): p.13. doi:10.5334/dsj-2016-013 (<http://datascience.codata.org/articles/10.5334/dsj-2016-013/>)

model have evolved over the course of the project. (For more on the multilayer data model, see the White Paper for the NEH Preservation and Access grant PW-51672-14.) Our corpus architecture consists of one or more documents within each corpus; each corpus is a logical collection of documents based on existing research standards (e.g., the shenoute.abraham corpus consists of all documents pertaining to the text by Shenoute known as *Abraham Our Father*, and the sahidica.mark corpus consists of all documents pertaining to the Sahidica online version of the Gospel of Mark). The Coptic Scriptorium project then consists of multiple corpora, with multiple documents within each corpora, as demonstrated in the example figure below:



Coptic Scriptorium Corpus Architecture

This structure is motivated in part by the functionality of the ANNIS search and visualization tool we chose to use for a database and in part by the process of digitizing and annotating the corpora in parts. We also have been working in a public, open access environment, meaning that we have released our corpora as documents have become available. We release a new version of a corpus when one or more new documents in that corpus have been digitized and annotated, or when one or more documents in that corpus have been edited due to updates in the data model, updates to the tools, or corrections to the annotations. Developing a version numbering system for a complex corpus architecture was one of the challenges of the project. After some experimentation, we now follow a system in which a new or newly edited document receives a version number which is the same as the new version number of the corpus, which also is the same as number given to the overall the project release. The overall project release consists of an archive file of all corpora files (in the three data formats described above) released on our GitHub repository at <https://github.com/CopticScriptorium/corpora/releases>.

2.2 Digital and Computational Tools

For machine processing and annotation of digital Coptic text, the project produced the following tools and technologies. The tools were developed for the Sahidic dialect of the Coptic language. Development took place primarily under the aegis of this grant (HD-51907), although the editors and annotators of corpora, who contributed data that increased the accuracy of the tools, were funded under the concurrent grant PW-51672-14. Editors used them in conjunction with manual annotation to produce the annotated text corpora (see [section 2.1](#)). All tools are available under open licenses. Documentation for these tools can be found in their respective GitHub repositories, linked below. Detailed guidelines for transcription (including tokenization), part of speech tagging, and lemmatization appear in Appendix E.

- Font and character converters: Convert text of documents transcribed in legacy ASCII fonts into the Unicode Coptic character set. Project participants produced a converter for several legacy fonts (`recode_coptic`) and one for a more complex legacy font created originally by Dirk Van Damme and Gregor Wurst (`CopticVDWtoUTFConverter`). Collaborator So Miyagawa produced another (`copticizer`) for text in the St. Shenouda Society CDs. All converters are on the [Converters repository on GitHub](#) (<https://github.com/CopticScriptorium/converters>).
- Tokenizer: Segments Coptic bound groups into words and morphs. This tool requires text input segmented according to the principles of boundedness articulated in Bentley Layton's *Coptic Grammar*.³ Available on the [Tokenizers repository on GitHub](#) (<https://github.com/CopticScriptorium/tokenizers>).
- Normalizer: Normalizes orthography and spelling of Coptic text. Available on the [Normalizer repository on GitHub](#) (<https://github.com/CopticScriptorium/normalizer>).
- Part of speech tagger: A probabilistic tagger built from the independent tool TreeTagger trained on a set of Coptic training data. The most recent version includes lemmatization data as well as part of speech annotation, and the lemma information includes Greek Loanword lemmas from the Dictionary and Database of Greek Loanwords in Coptic project (<http://research.uni-leipzig.de/ddglc/>). The original version was created with the assistance of a lexicon provided by the Corpus dei Manoscritti Copti Letterari project (<http://www.cmcl.it/>). Available on the [Part of Speech Tagger repository on GitHub](#) (<https://github.com/CopticScriptorium/tagger-part-of-speech>).
- Lemmatizer: A lexical tagger that annotates each word with its lemma, or dictionary headword. Originally developed as a stand alone tool and now incorporated into the part of speech tagger (above).
- Language of origin tagger: A lexical tagger that annotates each word or morph for its language of origin. (So if a text corpus has been segmented or tokenized to account for compounds, this tagger will annotate the loanword within the compound.) Native Egyptian vocabulary remain un-annotated.

³ Layton, Bentley. *A Coptic Grammar*. 3rd Edition, Rev. Wiesbaden: Harrassowitz, 2011. Print. Porta Linguarum Orientalium Neue Serie 20.

- Format converters:
 - SGML input/output plugin: Converts digitized text annotated with SGML or XML tags into a multilayer spreadsheet format in Microsoft Excel. Available on GitHub (<https://github.com/CopticScriptorium/converter>).
 - Annotation validation plugin: Validates multilayer annotations to ensure they conform to the data model described above. Available on GitHub (<https://github.com/CopticScriptorium/XLAddIns>).
 - EpiDoc TEI conversion tool: Converts select annotations of a multilayer text and annotations spreadsheet document into a single EpiDoc TEI-XML file. Currently operates only for Windows. Available on GitHub (<https://github.com/CopticScriptorium/XLAddIns>).

Over the course of the grant period, we updated the tools with periodic enhancements and bug fixes. In addition, we recursively updated the tools after publishing a large amount of new textual data. Since editors had manually edited the annotations, correcting any errors or omissions of the tools, using published data to update the tools resulted in greater accuracy.

Under the aegis of a bilateral NEH-DFG grant (HG-229371) of the KELLIA collaboration with German partners, the project also updated the tools with a Greek lemma list created by the [Database and Dictionary of Greek Loanwords in Coptic](#) project. In addition, funding from this grant supported development of a natural language processing pipeline, which runs any or all of the tools in a web application with an API.⁴ The pipeline can be used at <https://corpling.uis.georgetown.edu/coptic-nlp/>, and the code is available on GitHub (<https://github.com/CopticScriptorium/coptic-nlp>).

2.3 ANNIS Database

ANNIS is an open source, browser based search and visualization platform for linguistic corpora, which we have re-purposed and adapted in places in order to grant intuitive and flexible access to the annotations underlying our data model and visualizations. During the current project period, several improvements were made to the infrastructure to support our project, in particular:

- Embeddable visualization links to grant easy access to views of whole documents from external applications, including our web application for close reading purposes (rather than quantitative searches) at data.copticscriptorium.org
- Support for HTML templates to customize the appearance of specific annotation layers within generic visualizations

⁴ Amir Zeldes and Caroline T. Schroeder. “An NLP Pipeline for Coptic.” In: *Proceedings of the 10th ACL SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH2016)*. Berlin, 146–155. doi:10.18653/v1/W16-2119.

- Hyperlinking in both document and grid views. This feature is now used to link search results and archived ANNIS corpora in the repository to the live Coptic Dictionary Online at <https://corpling.uis.georgetown.edu/coptic-dictionary/>

The multilayer data model and the metadata model used in ANNIS for our corpora were partially developed under Preservation and Access Grant PW-51672-14; see the White Paper section 2.2 for more details.⁵

2.4 Collaborative Platform

Coptic SCRIPTORIUM's GitHub repository and website have developed into a collaborative platform where project editors contribute transcriptions and annotations to the project.

Annotators/editors funded by Grant PW-51672-14 forked the corpora, and then used the tools developed under this ODH Startup Grant to annotate the corpora. In addition participants and other collaborators have contributed to the development of the tools through the GitHub site.

Project co-PIs and editors have identified the need for a text transcription editor (with version controlled commits to a server, such as a university server or a GitHub repository) and a multi-layer editing tool (also with version controlled commits to a server) to enable annotations and editing in the multilayer model. Inline editing and annotation with our complex data model is not feasible. We are pursuing the development of these two tools under the NEH-DFG grant for the KELLIA project (HG-229371). (See also [Section 6](#) Continuation.)

2.5 Documentation

Documentation, including licensing information, for each tool appears in the GitHub repository with the tool, either in a separate file for Guidelines or in the ReadMe document for the repository. We also have created a wiki for project data models, instructions, and workflows at wiki.copticscriptorium.org, and a blog announcing news, corpus and tool releases, and other information at blog.copticscriptorium.org. We have a YouTube channel with video tutorials (<https://www.youtube.com/playlist?list=PLDN6-yDloRwf5HNCwZOAbbSfh1iEqOJT>). Our project website includes a tutorial with sample queries for the ANNIS database (<http://copticscriptorium.org/ANNIS-tips.html>), and a page with instructions for how to cite our work (<http://copticscriptorium.org/citation-guidelines.html>); also see Appendix E for our Citation Guidelines . In addition, all NEH project reports appear on a reports page of the website (<http://copticscriptorium.org/reports>). All of these resources can be accessed easily from links on our main Documentation page. The documentation was produced under this grant as well as the concurrent NEH-DFG grant for the KELLIA project (HG-229371), and the NEH Preservation and Access, HCRR Foundation grant (PW-51672-14).

⁵ Schroeder, Zeldes, and Platte, *op. cit.*

2.6 Web application to provide easier user access for reading and citing visualizations of textual data

We developed a web application that provides the most recent version of our data in all the formats currently available using the CTS (Canonical Text Services) URN system. Researchers can enter the the URN for a document or set of documents and retrieve the latest version(s).

When the URN resolves, the web service returns with links to all all visualizations for the relevant document(s), link(s) to the ANNIS search tool, and links to data downloads in all available formats (PAULA, TEI, relANNIS) on GitHub. The web service uses the ANNIS API to present the data in manner more accessible to researchers who wish to read or browse documents rather than search across corpora. Code is deployed at

<http://data.copticSCRIPTORIUM.org> and released open source under Apache 2.0 and Creative Commons Attribution 4.0 (CC-BY 4.0) licenses on GitHub

(<https://github.com/CopticSCRIPTORIUM/cts>). Funding from this grant as well as the NEH Preservation and Access grant PW-51672-14 supported its development. See Appendix B for screenshots.

2.7 Symposium and workshop (“Digital Coptic 2,” March 2015)

The project hosted a two-day symposium and workshop at Georgetown University in March 2015. On day one, scholars in Coptic Studies and Digital Humanities presented papers on ongoing digital projects or research using digital and computational methods. During discussion sessions, participants focused on how to collaborate across projects and technologies. On day two, Coptic Scriptorium participants led a tutorial on how to use our project resources; we also solicited feedback from the participants about the future direction of the project and suggestions generally. Many of the presentations plus part of the tutorial are available on our [YouTube channel](#). This event was co-funded by an NEH Preservation and Access, Humanities Collection and Reference Resources Foundation grant (PW-51672-14). The program and schedule appear in Appendix D.

2.8 Public tutorial and workshop

At the 2016 Congress of the International Association of Coptic Studies, held at Claremont Graduate University in July 2016, we held a tutorial and workshop about the project. Project Participants Amir Zeldes, Caroline T. Schroeder, and Rebecca Krawiec directed it. We introduced participants to our natural language processing tools, to searching and querying in the ANNIS database, and to reading and citing documents. We also received feedback from participants regarding the project.

2.9 Articles and conference papers to distribute the results of our work

The project produced the following articles and conference papers during the grant period. Some of these articles and papers document research conducted under the aegis of the NEH Preservation and Access grant PW-51672-14. For completeness and accuracy, we include all the papers and articles produced during the grant period, even if research was supported in part by the other grant.

Peer-reviewed Articles

Almas, Bridget, and Caroline T. Schroeder. "Applying the Canonical Text Services Model to the Coptic Scriptorium." *Data Science Journal* 15 (2016): p.13. doi:10.5334/dsj-2016-013 ([link](#))

Schroeder, Caroline T., and Amir Zeldes. "Raiders of the Lost Corpus." *Digital Humanities Quarterly* 10.2 (2016). ([link](#))

Zeldes, Amir, and Caroline T. Schroeder. "Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities." *Digital Scholarship in the Humanities* 30.suppl 1 (2015): i164–i176. doi:10.1093/llc/fqv043. ([link](#))

Published Conference Proceedings

Zeldes, Amir. "Duplicitous Diabolos: Parallel witness encoding in quantitative studies of Coptic manuscripts." Presented at Symposium on Cultural Heritage Markup, Washington, DC, August 10, 2015. In *Proceedings of the Symposium on Cultural Heritage Markup*. Balisage Series on Markup Technologies, vol. 16 (2015). doi:10.4242/BalisageVol16.Zeldes01. ([link](#))

Schroeder, Caroline T. "Shenoute in Code: Digitizing Coptic Cultural Heritage for Collaborative Online Research and Study." *Coptica* 14 (2015): 21-36. ([link](#))

Other Conference Presentations and Lectures

Schroeder, Caroline T. "The Future of Biblical Scholarship in a Digital Age." Plenary Presentation for the Catholic Biblical Association. Santa Clara. August 2016.

----- "Coptic Scriptorium: A Digital Platform for Research in Coptic Language and Literature." Congress of the International Association of Coptic Studies. Claremont, Ca. July 2016.

Krawiec, Rebecca S. "Charting Rhetorical Choices in Shenoute: *Abraham our Father and I See Your Eagerness* as case-studies." Congress of the International Association of Coptic Studies. Claremont, Ca. July 2016.

- Luckritz Marquis, Christine. "Reimagining the *Apophthegmata Patrum* in a Digital Culture." Congress of the International Association of Coptic Studies. Claremont, Ca. July 2016.
- Zeldes, Amir. "A Quantitative Approach to Syntactic Alternations in Sahidic." Congress of the International Association of Coptic Studies. Claremont, Ca. July 2016.
- Schroeder, Caroline T. "Preserving Coptic Cultural Heritage for the Digital Future," Seventeenth St. Shenouda-UCLA Conference of Coptic Studies. Los Angeles. July 17-18, 2015.
- Platte, Beth. "Coptic SCRIPTORIUM: Data from the Desert," Linking the Big Ancient Mediterranean. University of Iowa. June 6-8, 2016.
- Zeldes, Amir. "Tagging the Desert Fathers: Part of Speech Analysis in Sahidic Coptic Corpora." 43rd Annual North American Conference on Afroasiatic Linguistics (NACAL2015), 13-15 February 2015. Washington, DC.
- Schroeder, Caroline T. "Tag, You're It: Creating a Richly Annotated Coptic Digital Library," Society of Biblical Literature Annual Meeting. San Diego. November 2014
- , "DH Technologies for the Study of Coptic Language and Literature," Brown University Library, Digital Lab, September 30, 2014.
- Schroeder, Caroline T. and Amir Zeldes. "Digitizing the Dead and Dismembered: DH Technologies for the Study of Coptic Texts." DH2014. Lausanne, Switzerland. July 2014.
- , "Tagging Shenoute." North American Patristics Society Annual Meeting. Chicago. May 2014.
- , "Digital Coptic: Building an Online Environment for the Study of Coptic Literature." Center for Tebtunis Papyri, University of California, Berkeley. May 2014.

3. Project Activities

The primary project activities were:

- Development of tools described in [section 2.2](#)
- Application of tools to develop the annotated corpora (partially conducted by editors on the concurrent grant PW-51672-14), see [section 2.1](#) and [section 2.3](#). Editors or senior editors transcribed and/or annotated and/or translated the text. A senior editor reviewed and edited each document prior to publication to ensure editorial review.
- Software development on the document web application described in section by contractors (Archimedes and DaveBSoft).
- Digital Coptic 2 Workshop
- Writing Documentation

- Project management: coordinating editors, software development, publicity for project, budget, and communication between project participants
- Conducting research and writing articles and conference papers (see [section 2.9](#))
- Project meetings in San Francisco, CA (May 2014); Washington DC (March, 2015); Stockton, CA (May 2015); Washington DC (December 2015); Claremont CA (July 2016); Washington DC (December 2016). Some of the travel for these meetings was funded by the PW-51672-14 and HG-229371 grants.
- Advisory Board meetings conducted over Skype and email 2-3 times per year. Additionally, project co-directors consulted with Board members individually regularly throughout the grant period.

Project results were publicized at conferences and in articles (see [section 2.9](#)), on our project website and blog, and on social media.

4. Audiences and Impact

The audience consists primarily of academics (faculty, researchers, and graduate students) in Linguistics, Egyptology, and Religious Studies. These groups are the most prominent audiences at presentations and conferences and have provided the most anecdotal feedback to project directors. Additionally, undergraduates studying Coptic have used the site, according to anecdotal reports from faculty at other institutions. Social media interactions also indicate that non-academics with interest in Coptic, as well as members of the Coptic Orthodox laity in the American and European diaspora, visit the project site and follow its progress.

Our corpora have been forked from GitHub and/or used in research by several scholars. Four researchers or projects have forked our corpora, including the Classical Language Toolkit (an aggregator of natural language processing tools and corpora for ancient languages). Paul Dilley (University of Iowa) and So Miyagawa (graduate student, U. of Göttingen) used corpora and tools for their papers at our March 2015 symposium and at the International Association of Coptic Studies. The [Text Alignment Network](#) uses our edition of the Coptic New Testament in its online, aligned six-language New Testament. Rebecca S. Krawiec has used our corpora in her forthcoming article “Reading Abraham in the White Monastery: Fathers, Sources, and History,” to appear in [REDACTED]

[REDACTED]

[REDACTED]

According to Google Analytics, our main web domain copticscriptorium.org has been visited by over 12,000 unique users since it was launched in 2014, with just over 5000 repeat visitors. Most site views come from the United States; other significant audiences are in Germany, the UK, Brazil, Egypt, Japan, Canada, and Russia. The project blog, launched just over a year ago, has had over 2000 unique visitors. By comparison, approximately 200 scholars attended the 2016 Congress for the International Association of Coptic Studies.

5. Evaluation

The project was continually evaluated internally and periodically externally. In email correspondence and at project meetings project participants evaluated our progress and outputs. Advisory board meetings were held multiple times a year, over email and via Skype, since board members live in 3 different time zones. Feedback from our board meetings was incorporated into the project. We also consulted with board members individually about specific issues regarding linguistics, data curation, sharing data with other projects (such as papyri.info and the DDGLC), and translation. At two workshops (in Washington DC in 2015 and in Claremont, CA in 2016) we solicited comments, suggestion, and feedback. We also welcome feedback via social media, email to our project, and GitHub issues and pull requests.

The three primary concerns expressed by outside users were 1) a need for more digitized, annotated corpora in order to conduct more research; 2) greater accuracy of our tools; 3) difficulty of use of tools for the scholarly community without deep programming skills. Project participants agree with this assessment. Since this grant funded the creation of the tools and a small start-up corpus, however, we believe we have achieved the grant's objectives, and that addressing concerns #1 and #2 will take significantly more resources (both in time and funding). We have begun to address #3 by providing more documentation, YouTube video tutorials, and developing a web application for the natural language processing toolset (an online NLP pipeline discussed at the end of section 2.1 above). The NLP pipeline makes the toolset more accessible via a web interface and an API; it nonetheless still requires some technical skill to apply to one's research. This latter issue is a tension in many DH projects — the expectation that digital tools will make basic research processes faster, more accessible vs digital tools will enable more complex research (and thus can be complex to implement). We are currently looking at multiple models for applying the tools to develop more digitized corpora. Whereas previously, a single editor would both digitize and annotate a single text document, we are now looking at workflows that would allow some contributors to digitize text and others to use the tools to annotate. We hope this will allow us to move more quickly toward the goal of an expanded corpora. We also are pursuing funding for expanding the corpora and refining the tools.

6. Continuation of the Project

Now that we have built the core tools and a pilot corpus for digital Coptic studies, we are well-positioned to expand the corpus and refine the toolset. In particular, we intend to seek support for and continue preliminary work on the Coptic Treebank, which will enable high quality automatic parsing. This development will be of great value in itself, since it will allow us to add syntactic analysis to all of our resources, which in turn gives us access to questions like 'which verbs is a particular noun the subject or object of?' (e.g. finding all actions carried out by 'the

Devil', or all modifiers of 'blasphemy', etc.), and allows the linguistic examination of Coptic syntax across authors, genres, time and in the future, potentially also dialects.

Beyond querying syntactic information per se, we plan to use parsed data as a gateway technology to the development of Named Entity Recognition (NER) and Entity Linking for Coptic. This development will allow us to find occurrences of the same people and places mentioned in different ways, find entities across documents, and link our resources to other projects handling entities that overlap with ours.

Both the project PIs (Schroeder and Zeldes) will continue to work on the project as part of their ongoing research. Other participants, such as Rebecca Krawiec (Canisius College), and Christine Luckritz Marquis (Union Presbyterian Seminary) will as well.

Our tool development led to the aforementioned successful joint NEH-DFG grant proposal so that German digital Coptic scholars could implement our tools for annotating their texts. Our Digital Coptic workshop in our pre-NEH funded phase and our second symposium and workshop in March 2015 brought together scholars from other institutions and projects to discuss collaborations, standards, and shared challenges. The conversations initiated at these meetings led to the KELLIA project, which received funding for Göttingen, Münster, the Berlin-Brandenburg Academy, and other German partners to collaborate with us and scholars at the University of Minnesota and University of Iowa.

Appendices

Appendix A: List of Participants

Caroline T. Schroeder, the University of the Pacific

Amir Zeldes, Georgetown University

Elizabeth Platte, Reed College, Digital Humanities Specialist and Project Manager (2015-); editor and encoder/annotator (2013-)

Rebecca S. Krawiec, Canisius College, senior editor and encoder/annotator, translator (2013-)

Christine Luckritz Marquis, senior editor and encoder/annotator, translator (2014-)

So Miyagawa, University of Göttingen, editor and encoder/annotator, translator (2014-)

Elizabeth Davidson, Southern School of Energy and Sustainability, editor and encoder/annotator (2015-)

Dana Robinson, Creighton University, editor and encoder/annotator, translator (2016-)

Shuo Zhang, architecture and infrastructure (2015-)

Emma Manning, architecture and infrastructure (2016-)

Dave Briccetti, programmer and consultant (2015-)

Anthony Alcock, University of Kassel, translator (2015)

Eliese-Sophia Lincke, Humboldt University (2014-)

Bridget Almas, Perseus Digital Library, consultant for SCRIPTORIUM (2014-)

David Sriboonreuang, University of the Pacific student, intern and project manager (2015)

Lauren McDermott, the University of the Pacific student; TEI encoder and HTML programmer (2013-14)

Janet Timbie, the Catholic University of America, editor and annotator (2013)

Luke Hollis, Archimedes Digital, consultant and programmer for canonical referencing system (2014-2015)

Yanrui Liu, M.A., University of the Pacific, repository and website management (2014-2015)

Edwin Ko, Georgetown University, annotation interface development (2014)

Alex Dickerson, the University of the Pacific, student; TEI encoder and programmer (2013)

Advisory Board:

Alain Delattre, Assistant Professor, Department of Languages and Literatures, Université libre de Bruxelles; Papyl.info.

Eitan Grossman, Assistant Professor, Department of Linguistics and the School of Language Sciences, Hebrew University.

Robin Imhof, Humanities Librarian and Associate Professor, University Library, the University of the Pacific.

Appendix B: Website Screenshots

Main website and portal at www.copticscriptorium.org:

The screenshot shows the homepage of the Coptic Scriptorium. At the top, there is a dark navigation bar with links to HOME, TOOLS, CORPORA, DOCUMENTATION, NEWS, KELLIA, and ABOUT. Below the navigation bar, the logo "COPTIC SCRIPTORIUM" is displayed in red and black, along with the Coptic text "παλαιότερον σχάρι". To the right of the logo is a small illustration of a deer. The main content area has a parchment-like background. On the left, there is a text block describing the platform's purpose and features, followed by a bulleted list of tools and resources. On the right, there is a section for "Latest news" with a link to "December 2016 corpus release (v 2.2.0)". Below that is a note about server updates and a link to "link". At the bottom of the page, there is a footer with the text "Coptic Scriptorium is a collaborative, digital...".

Web application for citation and access using CTS URNs at data.copticscriptorium.org:

The screenshot shows the "Corpora & URN Resolver" section of the Coptic Scriptorium web application. At the top, there is a dark navigation bar with links to "COPTIC SCRIPTORIUM HOME" and "ABOUT". Below the navigation bar, the logo "COPTIC SCRIPTORIUM" is displayed in red and black, along with the Coptic text "παλαιότερον σχάρι". To the right of the logo is a search bar with the placeholder "Enter URN...". Below the search bar is a menu bar with dropdowns for "CORPUS", "AUTHOR", "MSNAME", "ANNOTATION", and "TRANSLATION". The main content area has a parchment-like background. It contains a brief description of the service, a note about reading Coptic texts, documentation on part-of-speech tagging, and citation guidelines. There is also a "Resources" section with links to documentation and citation guidelines.

Coptic Scriptorium Corpora & URN Resolver

Coptic Scriptorium provides Coptic texts for reading, analysis, and complex searches. The texts are citable and accessible through stable URNs, such as urn:cts:copticLit:shenoute.fox for Shenoute's work *Not Because a Fox Barks*. This application will provide the most recent version of our documents in the formats currently available for each text. If you know the URN for the material you seek, please enter it in the box above. You may also find documents by using the filters on the menu above.

If you wish to read Coptic texts, you can view individual documents online (in HTML) in various visualizations. We also provide links to our corpora in our search tool ANNIS, as well as data files in TEI XML, PAULA XML, and ANNIS formats. This web application provides the most recent version of the data. Previous versions are available on GitHub.

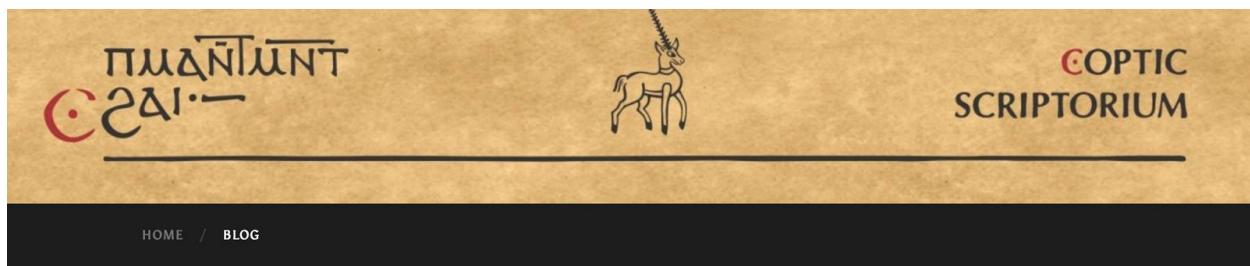
Documentation beneath the text provides information about it, including the version and date of publication online.

Resources

[Documentation on our part of speech tagging](#)

[Citation guidelines](#) for citing these documents

Project blog at blog.copticscriptorium.org:



December 2016 corpus release (v 2.2.0)

DECEMBER 16, 2016 / CTSCHROEDER / 0 COMMENTS

We are happy to release the following new and revised documents to our corpora. A copy of the official release notes is below. The data is available for download from [GitHub](#) in TEI XML, PAULA XML, and relANNIS formats. The corpora can be viewed and accessed at data.copticscriptorium.org, and they all can be queried in [ANNIS](#). We plan for another release with more documents in blog.copticscriptorium.org/2016/12/16/december-2016-corpus-release-v-2-2-0/

Search form

RECENT POSTS

[December 2016 corpus release \(v 2.2.0\)](#)

[Server updates – part of site down tonight](#)

[NEH White Paper \(Preservations and Access Grant\) published](#)

Project Wiki at wiki.copticscriptorium.org:

Coptic Scriptorium Wiki

Trace: • start

Log In

Recent Changes Media Manager Sitemap

Search

Coptic Scriptorium Wiki

<http://www.copticscriptorium.org>

Current Guidelines

- Annotation layer names
- Diplomatic transcription guidelines
- Tokenization guidelines (in Section 4 of the [Transcription Guidelines](#))
- Part of speech tagging guidelines
- Lemmatization guidelines

KELLIA Guidelines in Process

- Coptic Unicode standards and guidelines for Coptologists
- Controlled Vocabularies for Linked Open Data

Contributor/Annotator Tools and Processes

Table of Contents

- Coptic Scriptorium Wiki
 - Current Guidelines
 - KELLIA Guidelines in Process
 - Contributor/Annotator Tools and Processes
 - Transcription
 - Natural Language Processing Pipeline
 - Language Processing Tools
 - Metadata
 - Administrative Documentation and Calendar
 - March 2015 Workshop Resources

ANNIS search and visualization interface landing page:

Example Query	Description	
Q norm="NOΥΤΕ"	search for the normalized word ΝΟΥΤΕ	shenoute.a22
Q pos="NPROP"	search for proper names	shenoute.abraham.our.father
Q figure	search for figures in the papyrus	doc.papyri
Q gap_reason="illegible"	search for illegible gaps in the text	doc.papyri
Q norm="ѧՐԱ"	search for the normalized word ѧՐԱ	apophthegmata.patrum
Q norm="τοπος"	search for the normalized word τοπος	shenoute.eagerness
Q morph="peq"	search for words containing the morpheme peq	shenoute.eagerness
Q pos="FM"	search for words tagged as foreign material	shenoute.eagerness
Q norm="NOΥΤΕ"	search for the normalized word ΝΟΥΤΕ	shenoute.abraham.our.father
Q pos="NPROP"	search for proper names	shenoute.a22
Q pos="VSTAT"	search for stative verb forms	apophthegmata.patrum
Q lang="Greek"	search for words of Greek origin	shenoute.a22
Q norm="ѧՆՈՔ"	search for the unit "ѧՆՈՔ"	sahidica.nt
Q norm="սօֆիա"	search for the normalized unit "սօֆիա"	sahidica.nt
Q lang="Greek"	search for automatically detected Greek lexemes	sahidica.nt
Q norm="ՆԿՈՏՔ"	search for the normalized word ՆԿՈՏՔ	besa.letters
-	search for independent personal pronouns	

ANNIS search and visualization tool query results:

norm="NOΥΤΕ"

52 matches in 4 documents

Visible: scriptorium

Corpus List Search Options

Base text v

Displaying Results 1 - 10 of 52

Result for: norm="NOΥΤΕ"

1 [Path: shenoute.a22 > a22.YA421-428 \(tokens 316 - 329\)](#) left context: 5 right context: 5

ΜΠΧՈՅԻC · **ԵՐԵՊԻՆՈՒԹԵ** ԾՈՎՆ ԵՅՆՈՑ ԻՆԻԽՈՅԵ

- annotations (grid)
- analytic view (document)
- diplomatic text (document)
- normalized text (document)

2 [Path: shenoute.a22 > a22.YA421-428 \(tokens 374 - 384\)](#) left context: 5 right context: 5

ԻՆԵՐԵՊԻՆՈՒԴԵՆԱՇԽՁ. ԵՅՈԼ ԾԻՇՈՏՈՅ

- annotations (grid)
- analytic view (document)
- diplomatic text (document)
- normalized text (document)

3 [Path: shenoute.a22 > a22.YA421-428 \(tokens 475 - 486\)](#) left context: 5 right context: 5

ՆՈՅՆ ԱՄ ԵՐԵՊԻՆՈՒԹԵ ՆԱԴՈՏՆՀԻ ԵՅՈԼ

- annotations (grid)
- analytic view (document)
- diplomatic text (document)
- normalized text (document)

4 [Path: shenoute.a22 > a22.YA421-428 \(tokens 643 - 654\)](#) left context: 5 right context: 5

ԹԻՊՈՅ ԵՅՈԼ **ԵՐԵՊԻՆՈՒԹԵ** ԵՅՈՎԻԵ ԵՐՅՈՈՒՐ

- annotations (grid)
- analytic view (document)
- diplomatic text (document)
- normalized text (document)

Appendix C: Data Visualizations

Normalized visualization

Normalized Text

[XH204] ερωτανταφορ αφωκακ εβολ αν ετεντοκ πε πρημαλ μπιμμιωνας
σηνχενροογ εγω , ερεπιμογιτρρε ετεανοκ πε πρημαλ μπεχριστος , †coouγ
χεεκ†ογβη αν , αλλα εκ†ογβεινкоуց ετογիշ { It's not when the fox cries out, which is you,
ρωμε ερօկ նտօկ նուպեկեատ լուածօլօս էտօյ †coouγ oh servant of Mammon, in voices that shout,
մուպեչեատ էտօյիշ նշտօյ իհօց էտօյկօ նշիլլ լուայ սոլուգինօց that the lion, which is I, the servant of Christ, is afraid.
մմայ վլու ենջ նդշ նտակխօօս , լոյլու մպեմա , լոօօյ մպեմա ձիպարայել մմօյ ,
†coouγ լար ըպէտիամա և բարակ էտօյ . նետեօյնտայպեօօյ մնութայ էտօյ
շնուղմ կադպենտակխօօկ երօ և լու ձեալի ննէկնօյթ շնուցրադր այա ձեալթրեյմօյր
մպէկամա մնութայ շնուցրադր այա ձեալթրեյմօյր
ննէկմիմօյ ւշնութայ շնուցրադր այա ձեալթրեյմօյր
ննէկմիմօյ ւշնութայ շնուցրադր այա ձեալթրեյմօյր
Shenoute, *Not Because a Fox Barks*, ed. Amir Zeldes and Caroline T. Schroeder. Trans. Amir Zeldes. *Coptic Scriptorium*. urn:cts:copticLit:shenoute.fox.monbxh_204_216. v. 1.5, 13 May 2106. http://data.copticscriptorium.org/urn:cts:copticLit:shenoute.fox.monbxh_204_216.

Diplomatic Visualization

Diplomatic Edition

xh205	
15	θρωπαντβαλορ'
	ασκακεβολ'
	αν`ετεντοκ
	πε`πέμχαλ
	ηπιμαμμωνας'
20	ζησενηροογ
	εγοω',ερε
	πιμογ'ιττρε'
	ετεανοκπε'
	πέμχαληπε
25	χξ,†cooyn
	χεεκ†ογβηι
	αν',αλλαεκ†
	ογβεισετογ
	ηχηνεхреи
ctianos'.ic	Νοεγρατεηη
ον'ρωφεεροк	Μπτληстис'
ηтокмнпек	φооинннете
ειωттдлдво	ογнтаγисног
5 λοс'εтouгh'	5 м'катаанен
ηрнткетгk	такадооq'ероi
շելուշ'ероq'	еволхеаіqиn
ηтooуշaоy	ннекноуте
мнпeгeиwт'	зноусрaфт
10 εтouгhнж	10 аγωжeаіtреy
тouгeтоуgкo'	моyp'мпек
ηртнγероq'	сaфmнпек
ммнттeрoиue	тoиpеeрoиuн'

Shenoute, *Not Because a Fox Barks.* Coptic SCRIPTORIUM.

urn:cts:copticLit:shenoute.fox.monbxh_204_216.

Analytic Visualization

Analytic Visualization

ACOND ART N V ADV NEG CREL PPERI COP ART N PREP ART
ερωαντι τι βαλωρ αωμακακ εβολ αν ετει ΝΤΟΚ πε πι γμχαλ μι πι
NPROP PREP ART N CREL PPERI VSTAT PUNCT CFOC ART N V CREL PPERI COP
μαμμωνας ρη γενι γροογ ει γι οω , επει πι μογι τρρε ετει ανοκ πε
ART N PREP ART N PUNCT
πι γμχαλ μι πει χριστος ,

It's not when the fox cries out, which is you, oh servant of Mammon, in voices that shout, that the lion, which is I, the servant of Christ, is afraid.

PPERS V CONJ CFOC PPERS V PREP PPERO NEG PUNCT CONJ CFOC PPERS V PREP NPROP CREL
τι κοογν ρει ει κι τι ογβη i αν , αλλα ει κι τι ογβει ιησογ ετ
VSTAT PREP ART N PUNCT
ι ογης ρη νει χρειστιανος .

I know it's not against me you fight, but against Jesus who dwells inside the Christians.

NPROP ADV V PREP PPERO PPERI PREP PPOS N ART N CREL VSTAT PREP
ιησογ ον ρωαγε εροι κι ΝΤΟΚ μηνι πεκι ειωτ πι διαβολος ετι ογης ηγητι
PPERI CREL PPERS V PREP PPERO PUNCT
κι ετι κι γελπιζε εροι η .

Jesus still conquers you yourself and your father the devil who dwells inside you, for which you hope.

PPERI IMOD PPERO PREP PPOS N CREL VSTAT PREP PPERO N CREL PPERS V PREP N
ΝΤΟΟΥ γωι ογ μηνι πεγι ειωτ ετι ογης ηγητι ογι ιησογ ετι ογι κω νι γθη
PPERI PREP PPERO PUNCT
ι γι εροι η .

And they too on their part, with their father who dwells inside them, Jesus, on whom they depend.

Shenoute, *Not Because a Fox Barks*. Coptic Scriptorium.

urn:cts:copticLit:shenoute.fox.monbxh_204_216.

Chapter visualization

Sahidica Chapter View

1: ἀνοκ δε ςω ντεριει φαρωτη ναснηγ нтдлei շноյжисе аn нважде н нсофia еизω
ερωтн нтмнтмнтрe мпнouтe .

2: мпимесеge гар дeтcoуn нллaг n_g When I came to you, brothers, I didn't come
with excellence of speech or of wisdom,
proclaiming to you the testimony of God.

3: ἀνοκ ςω нтдlei φаrωtн շn оүмнтбов . мноуշотe . мноуշтвт eнаwвq .

4: аγω пaмaжe мnpaтaмjoeiо нtaqmoшpe аn շnoյpiөe ncofia нvажde . аlla
շnoյoуѡnq eboл mпneгma շibom .

5: дeкaдe ennetnpiстic фapie շnoյpiөe ncofia npwme аlla շnoյbom нtepiноутe .

6: enmадe дe noycofia շnntelelioc oycofia дe enтapciamon an te oýde ntanapxow
an te mpeiaim on nai etnaoywscq .

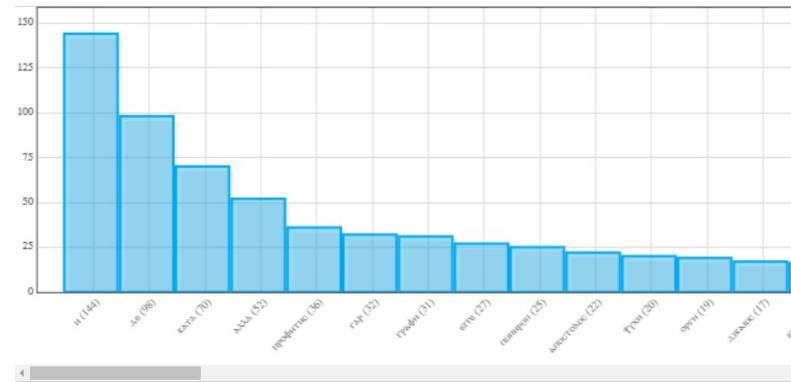
7: аlla enmадe noycofia нtepiноутe շnoյmystriion . tаi etxhi тentapnoутe
porxс eboл շatxн nnaiim epeнеooу .

1 Corinthians 2 (Sahidica version), ed. Elizabeth Platte et al., taken from J. Warren Wells' Sahidica edition. Trans. World English Bible. *Coptic Scriptorium*. urn:cts:copticLit:nt.1cor.sahidica_ed:2. v. 1.1, 22 May 2105. http://data.copticscriptorium.org/urn:cts:copticLit:nt.1cor.sahidica_ed:2.

Histogram visualization for frequency analysis from a query in the ANNIS database

Showing histogram of top 100 results, see table below for complete dataset.

linear scale
 log₁₀ scale



289 items with a total sum of 1337 (query on shenoute.eagerness, shenoute.fox, shenoute.abraham.our.father, sh

rank	#1lemma	count
1	H	144
2	Δ.ε	98

Appendix D: Digital Coptic 2 Program

Digital Coptic 2 Program

Thursday, March 12: Symposium

9 am-6 pm
Georgetown University
Poulton Hall, Room 230
1421 37th St. NW
Washington, DC 20057

Presentations are each 20 minutes long followed by 10 minutes for questions and discussion.

9:00 Coffee and Arrivals

9:15 Welcome

Amir Zeldes, Georgetown University

9:30 New and Expanding Digital Projects in Coptic Studies

Chair: Caroline T. Schroeder, the University of the Pacific

Adventures in Crowd-Sourcing Papyri - The Resurrecting Early Christian Lives DH Project, Philip Sellew, the University of Minnesota

Website Galleries of the White Monastery Candle Room Manuscript Fragments: Challenges of Digitization and Classification, Mary K. Farag, Yale University

The Digital Edition of the Coptic-Sahidic Old Testament and its planned Virtual Manuscript Room (VMR), Frank Feder, Akademie der Wissenschaften zu Göttingen

Digitizing Language Contact: Lexicography and Technological Perspectives at the Database and Dictionary of Greek Loanwords in Coptic (DDGLC), Frederic Krueger and Katrin John, Universität Leipzig

Discussion (30 minutes)

12:00-1:00 Lunch

1:15 Digital and Computational Research in Coptic Language and Literature

Chair: Elizabeth Platte, Valparaiso University

Coptic Scriptorium: Current Possibilities and Future Directions, Amir Zeldes,
Georgetown University, and Rebecca S. Krawiec, Canisius College

Coptic Scriptorium beyond the Manuscript: Tokenization and Corpus Analysis, Paul
Dilley, the University of Iowa

Synthesis, Boundness, and Clitics in Sahidic Coptic, So Miyagawa, Kyoto University

Discussion (30 minutes)

3:15 Coffee Break

3:30 Digital Humanities and Eastern Christian Traditions beyond Coptic Studies

Chair: Christine Luckritz Marquis, Union Presbyterian Seminary

Digital Preservation and Oral History of Displaced Syriac Speakers in the Middle East,
Robin Darling Young, the Catholic University of America

A New XML Exchange Format for Aligning Translations, Quotations, and Other Versions
of Texts, Joel Kalvesmaki, Dumbarton Oaks

Ex uno pro pluribus: Digitization, cataloging, and study of Eastern Christian manuscript
collections at the Hill Museum & Manuscript Library, Adam Carter McCollum, Hill
Museum and Manuscript Library

Discussion (30 minutes)

5:30 Concluding Remarks and Wrap-Up

Friday March 13

9 am-6 pm
Georgetown University
Poulton Hall, Room 230
1421 37th St. NW
Washington, DC 20057

A day-long workshop on Coptic SCRIPTORIUM for the SCRIPTORIUM team, collaborators, contributors, and those interested in becoming collaborators or contributors. We will discuss SCRIPTORIUM technologies, how to contribute and annotate text corpora for the project, future directions, and possible collaborations. A final agenda will be distributed in March. Space is limited. Please indicate on the registration form if you wish to join us, and we will confirm your attendance.

Appendix E: Guidelines

The following documents are attached:

- Citation Guidelines
- Transcription Guidelines
- Part of Speech Tagging
- Lemmatization Guidelines

Citation Guidelines

When researching our corpora for a future publication, **please note the date and version number of the documents or corpora while you are conducting your research.** We update our corpora regularly and recommend all citations include the version number and date of the resources used, as described below. (If you conducted research in the past and did not note the version and date of the corpus at that time, then please cite the date you accessed the corpus.)

Examples from the Chicago Manual of Style format appear below; please modify as appropriate for other style formats.

Cite this project

We recommend upon first citation of the project, text, query, tool, or other resource, you cite the full project as well as the individual resource.

First citation

Caroline T. Schroeder, Amir Zeldes, et al., *Coptic Scriptorium*, 2013-[current year], <http://copticscriptorium.org>.

Bibliography

Schroeder, Caroline T., Amir Zeldes, et al., *Coptic Scriptorium*. 2013-[current year]. <http://copticscriptorium.org>.

Cite a corpus or corpora

First citation

Coptic Scriptorium, [corpus name], [corpus URN], [version number], [date]. [http://data.copticscriptorium.org/\[corpus URN\]](http://data.copticscriptorium.org/[corpus URN]).

E.g.,

Coptic SCRIPTORIUM, *apophthegmata.patrūm Corpus*, urn:cts:copticLit:ap, v. 1.5, 4 October 2015. <http://data.copticscriptorium.org/urn:cts:copticLit:ap>.

Subsequent citations

Coptic SCRIPTORIUM, [corpus urn].

E.g.,

Coptic SCRIPTORIUM, urn:cts:copticLit:ap.

Bibliography

Coptic SCRIPTORIUM. [*corpus name*]. [corpus URN]. [version number], [date]. [http://data.copticscriptorium.org/\[corpus URN\]](http://data.copticscriptorium.org/[corpus URN]).

E.g.,

Coptic SCRIPTORIUM. *apophthegmata.patrūm Corpus*. urn:cts:copticLit:ap. v. 1.5, 4 October 2015. <http://data.copticscriptorium.org/urn:cts:copticLit:ap>.

Cite an individual document

First citation

Author, *Ancient title* [chapter.verse if available], ed. [annotators], trans. [translation]. Coptic SCRIPTORIUM. [document URN]. [version number], [date]. [http://data.copticscriptorium.org/\[document URN\]](http://data.copticscriptorium.org/[document URN]).

E.g.,

Sahidic Apophthegmata Patrum 6, ed. Paul Lufter et al., trans. Paul Lufter and Amir Zeldes. *Coptic SCRIPTORIUM*. urn:cts:copticLit:ap.6. v. 1.1.0, 21 May 2015.
<http://data.copticscriptorium.org/urn:cts:copticLit:ap.6>.

Besa, *Letter to Aphthonia* 1.1, ed. Amir Zeldes and Caroline T. Schroeder, trans. Amir Zeldes. *Coptic SCRIPTORIUM*. urn:cts:copticLit:besa.aphthonia. v. 1.3.0, 28 May 2015.
<http://data.copticscriptorium.org/urn:cts:copticLit:besa.aphthonia>

Subsequent citations

Author, *Ancient title (abbreviated)* [chapter.verse if available], *Coptic SCRIPTORIUM*, [document urn].

E.g.,

Sahidic AP 6, *Coptic SCRIPTORIUM*, urn:cts:copticLit:ap.6.

Besa, *Aphthonia* 1.1, *Coptic SCRIPTORIUM*, urn:cts:copticLit:besa.aphthonia.

Bibliography

Author. *Ancient title*. Ed. [annotators]. Trans. [translation]. *Coptic SCRIPTORIUM*. [document urn]. [version number], [date]. [http://data.copticscriptorium.org/\[document URN\]](http://data.copticscriptorium.org/[document URN]).

E.g.,

Sahidic Apophthegmata Patrum. Ed. Paul Lufter et al. Trans. Paul Lufter and Amir Zeldes. *Coptic SCRIPTORIUM*. urn:cts:copticLit:ap.6. v. 1.1.0, 21 May 2015.
<http://data.copticscriptorium.org/urn:cts:copticLit:ap.6>.

Besa, *Letter to Aphthonia*. Ed. Amir Zeldes and Caroline T. Schroeder. Trans. Amir Zeldes. *Coptic SCRIPTORIUM*. urn:cts:copticLit:besa.aphthonia. v. 1.3.0, 28 May 2015.
<http://data.copticscriptorium.org/urn:cts:copticLit:besa.aphthonia>

Citing and linking to a query

Links to queries in ANNIS are possible; however, these links remain stable only until the corpus is revised or updated. We recommend **citing the project** as well as providing information about **the query, query link, and date and/or version accessed**. We also recommend using the download option in ANNIS to **download query results**. Researchers may then save (or even publish on their own websites under the proper license) the raw data for future reference.

Coptic Scriptorium Diplomatic Transcription Guidelines

Version: 1.2_2016.8.26

Caroline T. Schroeder¹& Amir Zeldes²

1. University of the Pacific
2. Georgetown University

1. Preamble

This document details guidelines for transcribing a diplomatic edition of a manuscript in Sahidic Coptic according to the Coptic Scriptorium project scheme. The diplomatic transcription currently requires extensive manual annotation, due to the complexities of processing a diplomatic text in which no word breaks exist in the original and yet words and even morphemes span across line, column, and page breaks.

The transcription procedure assumes familiarity with basic paleography and traditional manuscript transcription following the Leiden conventions.

(<http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden>)

The diplomatic transcription also utilizes XML (eXtensible Markup Language) -like tagsets, including some of the TEI (Text Encoding Initiative) XML markup language, although the resulting document is **not** a valid XML document. Wherever possible, the EpiDoc subset of TEI XML is utilized for element nomenclature. EpiDoc TEI conventions were created by and for epigraphers and have come to be a standard in markup of ancient texts, epigraphic or otherwise.

(<http://sourceforge.net/p/epidoc/wiki/Home/>) In contrast to TEI, Scriptorium utilizes no milestone XML tags (e.g., <cb>). Instead, all tags are span annotations (e.g., <cb>This is a column of Coptic text.</cb>).

We recommend using an XML editor such as Oxygen to ensure the encoding is well-formed and well-structured.

The aim is twofold: 1) to achieve a transcription that documents the text and visualization of the manuscript as closely as possible to the original; 2) to provide a text file that can be processed by various digital tools and software, such as a tokenizer, a part-of-speech tagger, or the ANNIS database infrastructure (<http://www.sfb632.uni-potsdam.de/annis/>; Zeldes et al. 2009). Coptic Scriptorium has bundled some of these tools in a [Natural Language Processing web service](#).

The resulting transcription itself does not resemble a traditional text of a diplomatic edition. The markup ensures optimization for processing and search using such tools and software. For examples of the diplomatic editions visualized in HTML generated from the post-ANNIS transformations, see corpora at data.copticscriptorium.org. Valid EpiDoc TEI XML versions of the documents are also provided from this site.

2. Character Encoding

Texts are encoded using the UTF-8 (Unicode) Coptic language character set. The freely available Antinoou font and Coptic-English keyboard created by Michael Everson in cooperation with the International Association of Coptic Studies is the standard (<http://www.evertype.com/fonts/coptic/>). Unicode characters in the private use area are not recommended.

2.1 Alphanumeric Characters

Characters follow the orthography of the manuscript.

Mark oversize characters with XML tagging. Do not use uppercase version of the character.

2.2 Punctuation and Decoration

Punctuation and decoration follows the manuscript as closely as possible within the Unicode character set. Not all decoration and punctuation can be encoded using characters; deviations or documentation that can't be keyed in is instead typically indicated in a note element.

Notes on individual specific punctuation characters:

For the character ` that occasionally appears at the end of words in some manuscripts, use U+2CFF. Example:

πΕΜΜΟΝ`ΤΕ

2.3 Accentuation and Supralinear Strokes

Accentuation and supralinear strokes follow the orthography of the manuscript. Some manuscripts have binding strokes between letters (e.g. ȝñ) whereas others in the case of the same word might only provide a stroke over a single letter (e.g., ȝñ). The diplomatic transcription follows the conventions of the manuscript, even if the manuscript is internally inconsistent or contains what seem to be errors.

Notes on encoding individual specific accents, strokes, etc, using the Coptic-English keyboard for Antinoou (for MacIntosh):

- (as in ȳ) the supralinear stroke above only one letter: type the letter followed by Unicode U+0304 (; on keyboard)
- (as in ȳñ) the binding stroke between two letters: type first letter then U+FE24 (< in the Coptic-English keyboard) then second letter then U+FE25 (> in the Coptic-English keyboard), i.e. m<n> on a Mac using the Coptic-English keyboard
- (as in ȳñt) binding stroke over three letters: type the first letter then U+FE24 (< on a Mac using the Coptic-English keyboard) then second letter then U+FE26 (: [i.e. shift+;] on a Mac using the Coptic-English keyboard) then third letter then U+FE25 (> on a Mac using the Coptic-English keyboard), i.e. m<n:t>

~ (as in ôY) circumflex combining two letters: U+1DCD (keystroke shift+option+/ on a Mac using the Coptic-English keyboard) typed between the letters, so ōu (o then shift+option+/ then u)

For squiggly curved or jagged strokes over etas, use a regular circumflex rather than a dot or line or trema (): type the letter followed by U+0302 (option+3 on the keyboard)

Tremas (i, ii): type the letter followed by U+0308 (option+7 on the keyboard)

3. Text Divisions

A diplomatic transcription aims to preserve the formatting of the original text. Line breaks, column breaks, and page breaks as they appear in the manuscript are all documented.

3.1 Line Breaks

All line breaks in the transcription should follow the line breaks of the manuscript. Editors may manually encode line breaks using the tags <lb></lb>. However, if you plan to use the [Coptic NLP web service](#) to further annotate your text, you may use the “Enter” or “Return” key to produce a line break in the text file of the transcription. Selecting the option for “meaningful line breaks” in the NLP web service will insert encoding for the line breaks.

3.2 Column Breaks

All column breaks in the transcription should follow the column divisions in the manuscript. Columns are wrapped in span annotations using the <cb></cb> tagset.

Fig. 1: Opening column break tag, corresponding to beginning of manuscript column

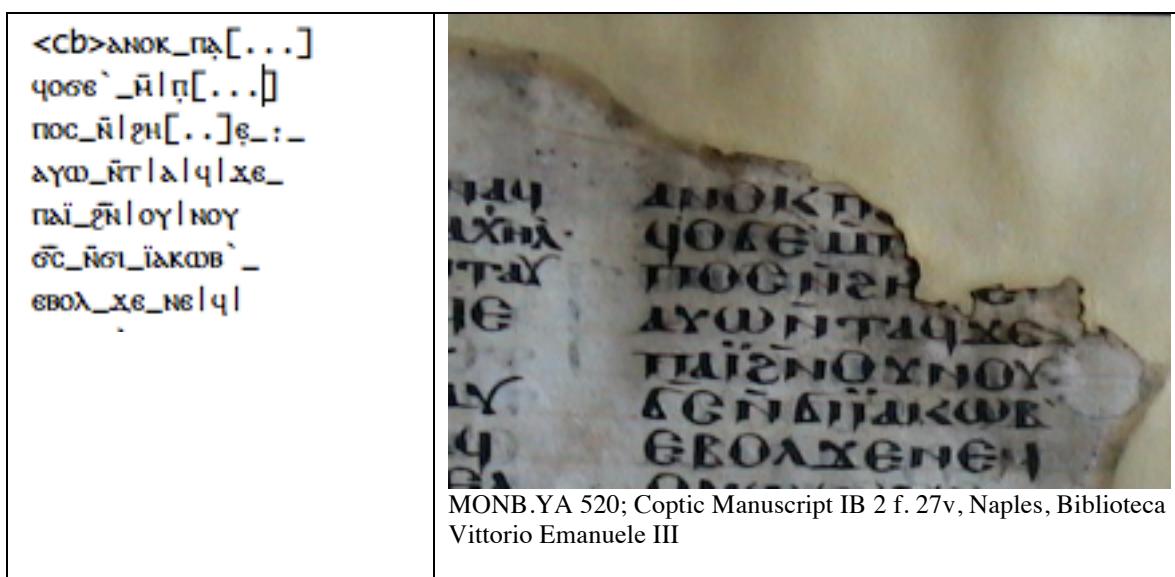


Fig. 2: Closing column break tag, corresponding to the end of a manuscript text column

<pre> φηρε` _ε φω πε_ῆνο η` _† να ηογ_ · · ἀγω_ ἰακωβ` _νε φι χω_ ἰηο c_ να c_ πε_ ε φι ογολc_ η σητ_ · · ξε_ ηη` _· ανῆ π ξοεic_ </cb> </pre>	
MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples, Biblioteca Vittorio Emanuele III."	

3.3 Page Breaks and Numbering

All page breaks in the transcription should follow the page divisions in the manuscript.

Page numbering in the transcription reflects the page numbering in the original manuscript codex. Codex sigla in the example below are two-letter codes following the White Monastery codex siglum list created by Tito Orlandi (Orlandi 2002; also <http://www.cmcl.it/>). Page breaks are wrapped in TEI compatible span annotations using the `<pb></pb>` tagset with the `xml:id` element. The entire page of text (including the relevant column tags) should be wrapped with these tags. Thus `<pb xml:id="YA518">` is the opening tag for page 518 in White Monastery codex YA (MONB.YA). The `xml:id` should not contain spaces. (Thus, `xml:id="YA518"` not `xml:id="YA 518">`.)

Fig. 3: Closing and opening page break tags indicating the end of one page and beginning of the next. (Note: the opening tag for the first page and closing tag for the second page are not visible here but are required.)

<pre> ονος` _ε να φω φι · · ἀγω_ π ετ ηηγ_ · · επολ_ ἔη_ ηεκ` _· · </cb></note></pb> <pb xml:id="YA519"><note note="page number φι barely visible in upper right"><cb>[....]φ_ · · ιαακ` · [....]φ_ π ετ να [κλη]ρονομει_ ηηο κ_ · · θτβε_ παι_ πῶ_ α c ρρο c_ ἔη_ ηογ </pre>

The location and Coptic numeration of the page number is currently documented in a note element. (See Figure 3 above).

4. Word Segmentation, Spacing, and Tokenization

Sahidic Coptic bound groups are formed by several words and/or morphemes attaching together. A word refers to one noun, preposition, article, etc. One complex word can be comprised of multiple morphemes, including affixes such as $\alpha\tau$, MNT , or peq , or compound words, such as complex numbers (e.g. -teens) and verbs formed with p . One bound group may include multiple prepositions and objects, or a verbal auxiliary + subject + infinitive, or even more words and morphemes strung together (generally speaking clitics). The copula, which some might consider a clitic, remains unbound. Coptic SCRIPTORIUM follows the practices in Bentley Layton's grammar (Layton 2011) for word, morpheme, and bound group segmentation.

Examples of individual words comprised of one morpheme:

सωτम

नोवे

გჰთ

Examples of individual words comprised of multiple morphemes:

मन्तात्सωत्म

ρεψρнове

Examples of bound groups comprised of words with multiple morphemes:

τμντατ्सωтм

μπρεψрнове

πρμнշт

ҳенταγձիտ

Note: if a project wishes to annotate on the morpheme level (i.e. internal analysis of units like MNT) and not just on the *word* level, the morphemes need to be tokenized. Coptic SCRIPTORIUM annotates on the word level and then provides additional annotation on the morpheme level for compound words and words with affixes. (See section 4.4 for more information.)

In most manuscripts, no spaces between words or bound groups are provided.

Sometimes a diacritical mark, such as ` does appear, but word segmentation following diacritics and punctuation does not always correspond with contemporary segmentation practices (such as Layton or Till (1960)). More study of this marking is required.

4.1 Word Segmentation

SCRIPTORIUM diplomatic transcription marks word segmentations according to Layton's conventions (Layton 2011). The transcriber inserts a unique character, such as an underscore ("_"), after each Coptic bound group, even when the end of the bound group falls at the end of a line.

Likewise, all punctuation is followed by an underscore.

- (1) ετεῖςμανλ_ (word ends at end of line)
- (2) πε_νψακλη (two words, in which the second bound group flows into line 3)
- (3) ρονομει_μ (the bound group continues from line 2, is followed by an underscore)
- (4) μοκ_αν_· (punctuation followed by an underscore)

These underscores are not and do not need to be visualized in HTML transformations of the diplomatic editions; they are nonetheless essential for processing the text, since they demarcate breaks between bound groups and will enable searches and visualizations of a word-segmented text.

We do not recommend using spaces to demarcate bound groups and punctuation, since spaces may occur elsewhere in the document (such as inside XML tags), and lead to confusion during automatic processing.

4.2 Spacing

Encoding of blank space is preferred to using the space key. The encoding should match spaces in the manuscript. Consequently, if the manuscript provides no spaces between words or punctuation, the diplomatic transcription contains no spaces. Where there are significant spaces in the manuscript that the transcriber wishes to draw attention to, the transcription should encode a space using TEI XML tags in order to visualize the white space in the manuscript. Encode the word, morpheme, or punctuation next to the white space, as in these examples:

- (1) <hi rend="1_space_right">·</hi> will visualize one space to the right of the ·
- (2) ἦ<hi rend="1_space_right">βονεος</hi> will visualize one space to the right of he n t
- (3) χω<hi rend="2_space_right">ϊ</hi>_<hi rend="1_space_right">·</hi>_ετβε_ will visualize two spaces to the right of i and one space to the right of the ·

It is important to make sure that attributes are surrounded by straight, not curly quotes (i.e. " on both sides).

4.3 Tokenization of Words

If one wishes to manually segment bound groups into words, one can do so using the pipe character (“|”).

- (1) γιτμ|π|νοΥτε_ (preposition|article|noun)
- (2) ετε|ιcμανλ_ (converter|noun)
- (3) πε`_n|q|na|κλη (word_auxiliary|subject pronoun|future marker|verb (verb continues to line 4))
- (4) ρονομει`_μ

The NLP web service contains a tokenizer that will take as input bound groups and provide as output word segmentation with pipes. Coptic SCRIPTORIUM’s standalone tokenizer tool will do the same.

4.4 Tokenizing and Annotating Morphemes below the Word Level

To conduct research on the morpheme level in compound words or other words that contain multiple morphemes, the words will need to be tokenized and annotated below

the word level and on the morpheme level. In Coptic SCRIPTORIUM, text is annotated on the word level for the part of speech (see [SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic](#)) and other characteristics, such as language of origin. Tokenizing and annotating on the morpheme level allows for additional search, visualization, and research capabilities.

Examples of individual words comprised of multiple morphemes, tokenized on the morpheme level:

word	MNT&TCWTM			
morpheme	MNT	&T	CWTM	

word	РeçPNOBé			
morpheme	Рeç	P	NOBé	

Examples of bound groups comprised of words with multiple morphemes:

bound group	TMNT&TCWTM			
word	T	MNT	&T	CWTM
morpheme	T	MNT	&T	CWTM

bound group	MПРЕÇPNOBé			
word	M	P	РЕÇPNOBé	
morpheme	M	P	Рeç	NOBé

bound group	ПРМНГHT			
word	P	M	N	GHT
morpheme	P	M	N	GHT

Compound words that involve an article or affixed personal pronoun to the second item of the compound typically are tokenized as bound groups comprised of multiple words, not as one word comprised of multiple morphemes.

Examples of bound groups containing compound words with articles or pronouns on the second unit of the compound:

bound group/compound	ρξηαρ		
word	ρ		ξηα
<i>no tokenization & annotation on the morpheme level below the word level</i>			

bound group	μπετηρπιμεεγε				
word	μπε	τη	ρ	η	μεεγε
<i>no tokenization & annotation on the morpheme level below the word level</i>					

(where πιμεεγε is considered to contain multiple words, not morphemes below one word level)

[Note: the part-of-speech tagger developed by Coptic Scriptorium operates on the *word* level, not the sub-word morpheme level. So, ρξοτε is tagged as one V, μντατсωтм as one N, etc.]

Transcription conventions for segmenting morphs should utilize a unique character, such as a dash or hyphen. E.g.:

τ|μντ-ατ-сωтм

μ|η|ρεη-ρ-νοвє

If you plan to use Coptic Scriptorium's NLP web service, you may transcribe the Coptic in bound groups with no pipes or morphemes. The NLP web service's tokenizer can provide as output segmentation with pipes between words and dashes between morphs. Likewise, Coptic Scriptorium's stand-alone tokenizer can output words with segmented morphs. The webservice can further automatically annotate the segmented words and morphs for part of speech, language of origin, and lemma.

5. Rendering and Leiden Transcription Conventions

Coptic SCRIPTORIUM uses Leiden and Leiden+ conventions for transcribing manuscripts. The encoding follows the EpiDoc guidelines. Not all Leiden documentation is currently XML encoded as Leiden+, however.

5.1 Characters Highlighted, Raised, Lowered, or Set Apart in Some Way

Characters that are raised, lowered, or printed in different colors or styles are encoded using the TEI XML element `<hi>` with the `rend` attribute. Letters written above the line are encoded: `<hi rend="superscript">`. Characters written below the line are encoded: `<hi rend="subscript">`. Letters in a different color ink are encoded with the color ink, e.g., `<hi rend="red">`. It is possible to combine these annotations, e.g. `<hi rend="red subscript">`. Coptic SCRIPTORIUM currently encodes large, tall (the letter stretches above the line), long (letter stretches below the line), thin, superscript, subscript, and colors. Any additional information can be provided in a note element. To encode two attributes, use a space (not a comma) between the two attributes.

Example

γιτάπιν<hi rend="superscript"><note note="o is directly above the γ">o</note></hi>γ
ΠΠΙΕΤΝΔΝΟ<hi rend="large">γ</hi><hi rend="long thin">ϙ</hi>_-

Diplomatic Visualization

γιτάπινꝝ (ANNIS) or
γιτάπιν\o/ꝝ (EpiDoc XSLT)
ππιετνδνοꝝ.

Other encodings are colors (red, brown, green, etc.) “ekthetic” should be used for characters that are part of the ongoing text but written to the left of the margin line. See below, in which the Π is encoded `<hi rend="red large ekthetic">Π</hi>`

έταηκέβολ·-ήτε
5 πεθερίον γητ·-
Παίδεαψχωνογμα

`hi@rend` cannot contain more than five words as per Epidoc guidelines and may contain only alphanumeric characters. (No punctuation. So `<hi rend="long, thin">ϙ</hi>` is invalid.)

5.2 Damaged Characters

Characters that are damaged but restored based on context are marked with an underdot. Coptic SCRIPTORIUM uses the diacritical character . (Unicode U+0323). These characters are not currently encoded in TEI XML using the EpiDoc tagset for Leiden+. Coptic SCRIPTORIUM uses the underdot character rather than annotation to designate this information.

5.3 Lacunae and Lost Characters

Lost lines and characters (lacunae) are indicated using square brackets, as in the Leiden conventions. They may be encoded using the EpiDoc tagset, but it is not required. See EpiDoc guidelines for more details (“EpiDoc Guidelines: Lost Characters, Quantity Unknown”; “EpiDoc Guidelines: Editorial Restoration: Characters Lost but Restored by Modern Editor”; “EpiDoc Guidelines: Lost Characters, Quantity Approximate”; “EpiDoc Guidelines: Lost Characters, Quantity Known”; “EpiDoc Guidelines: Erased and Lost”; “EpiDoc Guidelines: Lacunas, Other Units”).

- (1) Example encoded using the gap element:

```
<gap reason="lost">  
[ ]_  
[ ]_  
[ ]_  
</gap>
```

- (2) Unencoded gaps (no XML elements):

[.....]𠀠[...]

5.4 Other

Other rendering information is encoded either according to EpiDoc conventions or recorded as information within a note element. See the cheatsheet for Leiden+ conventions in EpiDoc at

<https://sourceforge.net/p/epidoc/code/HEAD/tree/trunk/guidelines/msword/cheatsheet.doc?format=raw> and http://papyri.info/docs/leiden_plus. See also the full list of text transcription guidelines here <http://www.stoa.org/epidoc/gl/latest/app-alltrans.html>.

Transcribing in Oxygen or a similar XML editor is recommended, to ensure tags are well-structured.

6.0 File Format and Document Preferences

Documents are transcribed in a text editor such asTextEdit. Document preferences are set to UTF-8 encoding without byte-order Mark (BOM). (E.g., inTextEdit 1.7.1 for MacIntosh, in the File-->Preferences menu, click on “Open and Save,” and select “Unicode (UTF-8)” for Opening files and Saving files.)

Bibliography

An up-to-date bibliography can be found at the project’s Zotero page:

https://www.zotero.org/groups/coptic_SCRIPTORIUM/items/collectionKey/8IHTW3NZ

Bodard, Gabriel. “EpiDoc Appendix: Glossary: Leiden, Leiden-plus.” *Appendix: Glossary*. 18 Jun. 2013. <<http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden>>.

“Corpus Dei Manoscritti Copti Letterari.” *CMCL - Studies in Coptic Civilization*. 11 Sep. 2012. <<http://cmcl.aai.uni-hamburg.de/>>.

“EpiDoc Guidelines.” *EpiDoc Guidelines*. 25 May 2013. <<http://www.stoa.org/epidoc/gl/dev/>>.

- . *EpiDoc: Epigraphic Documents in TEI XML*. 25 May 2013.
<<http://sourceforge.net/p/epidoc/wiki/Home/>>.
- “Everttype: Antinoou.” *Everttype: Antinoou - A Standard Font for Coptic* 2012. 29 May 2013. <<http://www.everttype.com/fonts/coptic/>>.
- Layout, Bentley. *A Coptic Grammar*. 3rd Edition, Revised. Wiesbaden: Harrassowitz, 2011. Print.
- Orlandi, Tito. “The Library of the Monastery of Saint Shenute at Atrięe.” *Perspectives on Panopolis: An Egyptian Town from Alexander the Great to the Arab Conquest*. Leiden: Brill, 2002. 211–231. Print.
- Till, Walter C. “La séparation des mots en Copte.” *Bulletin de l’Institut français d’archéologie orientale* 60 (1960): 151–70.
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian “ANNIS: A Search Tool for Multi-Layer Annotated Corpora.” *Proceedings of Corpus Linguistics 2009* (2009) : n. pag. 10 Sep. 2012. <<http://ucrel.lancs.ac.uk/publications/cl2009/>>.

SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic

Amir Zeldes¹ & Caroline T. Schroeder²

1. Georgetown University

2. University of the Pacific

Version: 1.1.8_2016.08.31

1. Preamble

This document details guidelines for part-of-speech tagging Sahidic Coptic according to the Scriptorium project scheme. The tagging procedure assumes the text has already been normalized to the orthography and morpheme based segmentation described in the Scriptorium tokenization guidelines, which are closely related to the conventions found in Layton's (2004) grammar. In case of doubt we refer to Layton (2004) as well as Shisha-Halevy (1988).

As in all tagging projects, the aim is to achieve a practicable compromise between linguistic accuracy/usefulness, speed and reliability of human tagging, and performance of automatic tagging software. This means that in many cases concepts that are linguistically distinct are not distinguished since they are difficult to tell apart in practice in many cases, or determining some distinctions is too costly in terms of annotation time. Additionally, the project is using the CMCL lexicon, kindly provided by Prof. Tito Orlandi, which has its own, much more detailed scheme, so that in some cases the categories used here are chosen to be derivable from the CMCL scheme (see <http://cmcl.let.uniroma1.it/>).

There are two proposed tagsets, a coarse tagset with fewer tags for projects wishing to save annotation time, and a finer tagset with more detailed subcategories for some of the coarse grained tags, which is also expected to yield lower accuracy in automatic tagging. Links to the latest training models are provided from the Scriptorium website and have been tested and developed using the freely available TreeTagger (Schmid 1994, see <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>).

2. Tagsets

The two tagsets described below are compatible with each other in that the fine-grained tagset uses the same overarching categories of the coarse one, but with further categories distinguished. The tag names are built ‘hierarchically’, so that additional letters in the name of a tag specify a special type of the superordinate category, e.g. all pronoun tags being with P, though not all tags with P are pronouns, as in PREP for prepositions.

In the coarse-grained list below, tags that have multiple fine-grained variants are followed by [*] (this is **not** part of the tag within the course-grained tagset).

Additionally, both tagsets admit certain cases where a single form contains two categories and must therefore be assigned two tags. This results in special underscore separated **portmanteau tags**, which are described in Section 2.3.

2.1 Coarse-Grained Tagset

Tag	Name	Examples
A[*]	Auxiliary tripartite base	ᾳ[q], με[q], τρε[q], ...
ADV	Adverb	εβολ, ον, πως
ART	Article	π(ε), τ(ε), ν(ε), γεν, κε
C[*]	Converter	ε, ετε, νε, ...
CONJ	Conjunction	αγω, η, μη, και, ειτε, ...
COP	Copula	πε/τε/νε
EXIST	Existential/possessive	ογν/μη
FM	Foreign material	παρα τογτο
FUT	Future	να
IMOD	Inflected modifier	τηρ[q], γωω[τ], ...
N[*]	Noun	αθητ, ρωμε, αρχη, ...
NEG	Negation	η, αη, τη[σωτη]
NUM	Numeral	ογα, σναγ, ...
PDEM	Pronoun, demonstrative	πει/παι, τει/ται, νει/ναι
PINT	Pronoun, interrogative	ογ, νημ
PPER[*]	Pronoun, personal	η, ει, τη, ανοκ, ανγ, ...
PPOS	Pronoun, possessive	πεη, τετη, πογ, πα, πωι, ...
PREP	Preposition	ετве, γн, η, μμο[q], ...
PTC	Particle	Δε, ήσι, ρε, ...
PUNCT	Punctuation	., . . .
UNKNOWN	Unknown morph, lacuna	β_ _ _, _ _ oc, _ _ _, ...
V[*]	Verb	σωτη, σωτη, σοτη, ειρε, ο, αρι, ...
VBD	Verboid	ηανογ[q], πεχα[q], πεχε, ...

2.2 Fine-Grained Tagset

For descriptions of the added fine-grained tags, marked in cursive type, see the coarse tag descriptions below.

<i>AAOR</i>	<i>AJUS</i>	<i>ANY</i>
<i>ACAUS</i>	<i>ALIM</i>	<i>AOPT</i>
<i>ACOND</i>	<i>ANEGAOR</i>	<i>APREC</i>
<i>ACONJ</i>	<i>ANEGRJUS</i>	<i>APST</i>
<i>ADV</i>	<i>ANELOPT</i>	<i>ART</i>
<i>AFUTCONJ</i>	<i>ANEGPST</i>	<i>CCIRC</i>

<i>CFOC</i>	NEG	PREP
<i>CPRET</i>	<i>NPROP</i>	PTC
<i>CONJ</i>	NUM	PUNCT
<i>COP</i>	PDEM	UNKNOWN
<i>CREL</i>	PINT	V
<i>EXIST</i>	<i>PPERI</i>	VBD
<i>FUT</i>	<i>PPERO</i>	<i>VIMP</i>
<i>IMOD</i>	<i>PPERS</i>	<i>VSTAT</i>
N	PPOS	

2.3 Portmanteau tags

In certain cases, one indivisible form corresponds to what normally constitutes two categories. This can happen either because of a phonological merger of two units, or because the formal marker of one category can be ‘zero’, i.e. have no form at all (usually in the case of 2nd person singular feminine forms). Portmanteau tags currently supported by the SCRITORIUM tools are:

tag	example	notes
AOPT_PPERS	εφε(сωтм)	Personal pron. within optative ε_ε. Note that εφε(сωтм) for 2nd pers. sg. fem. is also AOPT_PPERS, but nominal εφε(πρωμε сωтм) is only AOPT.
ACOND_PPERS	εφωαн(сωтм)	Personal pron. within conditional ε_φωαн. Note that εφωαн(сωтм) for 2nd pers. sg. fem. is also ACOND_PPERS, but nominal εφωαн(πρωμε сωтм) is only ACOND.
ACONJ_PPERS	τα(сωтм)	Truncated conjunctive 1st person (instead of ηταсωтм)
ANEGPST_PPERS	ηπε(сωтм)	Fused negative past 2nd pers. sg. fem. form.
APST_PPERS	ηρ(сωтм)	Fused positive past 2nd pers. sg. fem. form.
CCIRC_PPERS	ερε(сωтм)	Fused circumstantial 2nd pers. sg. fem. form.
CFOC_PPERS	εφε(сωтм)	Fused focalized 2nd pers. sg. fem. form.
CPRET_PPERS	ηερε(сωтм)	Fused preterit 2nd pers. sg. fem. form.
CREL_PPERS	ετερε(сωтм)	Fused relative 2nd pers. sg. fem. form.
IMOD_PPERO	ηηινηммo	The 2nd pers. sg. fem. form of ‘yourself’ (not to be confused with ηηιнηммo(q) etc.)
PREP_PPERO	ερo	Any preposition where a 2nd pers. sg. fem. is realized as zero (not to be confused with ερo(q) etc.)
V_PPERO	(q)ητ	Verb forms with a fused 1st pers. sg. object, e.g. ητ ‘bring me’ from ειнe ‘bring’, where the presuffixal form ητ- is merged with the 1st pers. object marker -τ)

Note that in all cases, coarse grained tags can be substituted for fine grained ones, e.g. CCIRC_PPERS and CFOC_PPERS both become C_PPER. Further combination tags are not ruled out and new ones will therefore be added if they are determined to be necessary.

2.4 Part of speech in conversion

In rare cases, a part of speech may appear in a syntactically unusual position. For example, an adverb or preposition may follow an article if they begin a phrase that is treated as a nominal phrase syntactically: the word *εβολ* is tagged as an ADV, although in the sequence *օγ|εβολ չմ|Ռ|ՀՈՎՄԱ* ‘one (which is) out of the body’, it appears to behave like a noun. We consider such cases of ‘conversion’ between categories to be a syntactic phenomenon, and we therefore continue to tag *εβոլ* morphologically as an adverb.

An exception to this rule is the tagging of verbal infinitives following an article. In essence, almost any Coptic infinitive may be used as a noun, for example *ՈՒԹՅՈՒՆ* ‘the call’. Cases such as these are widespread and are tagged as nouns, not as verbs, when the infinitive is used in this way.

3. Guidelines

The following guidelines describe the recommended assignment of part of speech tags to segmented morphemes. Fine-grained tags are given in the section describing the corresponding coarse-grained tag. In each example, the area corresponding to the tag under discussion is underlined. Vertical lines (‘pipes’) are used to segment morphemes for added clarity only.

3.1 Auxiliaries (A)

Auxiliaries include all conjugation bases in the tripartite patterns described in Layton (2004:251-290). These include both negative and positive variants and cover all lexical material preceding the subject noun or pronoun, e.g.:

- (1) Ճ|ՉՈՒԹԻ (3rd person masculine past tense)
- (2) ՃՐԵ|ՉՈՒԹԻ (2nd person feminine past tense, with zero subject)
- (3) ՃՌՈՒ|ՉՈՒԹԻ (negative past tense)

Note that when used with pronominal subjects, the optative and conditional conjugation encompass the subject pronoun, leading to a portmanteau tag like AOPT_PPERS (or A_PPER in the coarse grained tagset):

- (4) ԵՎԵ/AOPT_PPERS ՉՈՒԹԻ (optative and 3rd pers. masc. pronoun)
- (5) ԵՎՈԱՆ/ACOND_PPERS ՉՈՒԹԻ (conditional and 3rd pers. masc. pronoun)

Fine-Grained Tags

The different individual fine-grained tags cover all distinct conjugation bases, making auxiliaries the largest fine-grained tag group. They are divided as follows:

APST	Auxiliary, past	Ճ
ANEGPST	Auxiliary, negated past	ՃՌ(Ե)

ANY	Auxiliary, ‘not yet’	ῆπατ(ε)
AAOR	Auxiliary, aorist	ῳδ, ὠδρε
ANEGAOR	Auxiliary, negated aorist	με(ρε)
AOPT	Auxiliary, optative	ε[χ]ε, ερε
ANEGOPT	Auxiliary, negated optative	ῆνε
AJUS	Auxiliary, jussive	μαρ(ε)
ANEJJUS	Auxiliary, negated jussive	ῆπιτρε
APREC	Auxiliary, precursive (‘after’)	ῆτερ(ε)
ACOND	Auxiliary, conditional	ε[χ]ωλη, ερωλη
ALIM	Auxiliary, limitative (‘until’)	ῳαντ(ε)
ACONJ	Auxiliary, conjunctive	ῆ(τε)
AFUTCONJ	Auxiliary, future conjunctive	ταρ(ε)
ACAUS	Auxiliary, causative	τρε

Note that the irregular negation με in με-ῳδε ‘it is not appropriate’ is also tagged NEG and not as ANEGAOR.

3.2 Adverbs (ADV)

Adverbs include indeclinable native Egyptian and Greek lexemes that modify verbs and other phrases as in the following examples.

- (6) τααγζανε ῆμοق εματε/ADV ‘I shall glorify him greatly’
- (7) πετ|ῆμιαγ/ADV ‘the one (who is) there’
- (8) ῆπρμογ κακωσ/ADV ‘don’t die badly’

The first part of ‘complex prepositions’ is also tagged as an adverb, as in the following examples:

- (9) εβολ/ADV matchCondition/PREP ‘from, out of’ (lit. ‘out in’)
- (10) ερογν/ADV matchCondition/PREP ‘in towards’ (lit. ‘inside at’)

This does not apply to etymologically complex one-word prepositions derived e.g. from nouns for body parts (see the tag PREP for details), nor is the initial ε in words such as εβολ separated from the adverb (see segmentation guidelines).

3.3 Articles (ART)

Articles include definite articles, indefinite articles and article-like words such as κε/σε ‘other’. The following examples illustrate some variants:

- (11) η/ART ρωμε/N ‘the man’
- (12) τε/ART κληρονομια/N ‘the inheritance’
- (13) ογ/ART νομοс/N ‘a law’

- (14) ἡΕΝ/ART ἔργα/N '(some) deeds'
(15) κε/ART πονηρός/N 'another wicked one'

Note that possessive pronouns like περι are not tagged as articles (see PPOS) and relative articles like πλέτ are segmented to contain a relative converter (see C and CREL).

Articles followed by a noun beginning with Σ and consequently spelled ο or φ e.g. οε ‘the way’ are normalized and tokenized as τ and γε before part-of-speech tagging, so that τ etc. can be tagged as an article alone (see segmentation guidelines).

3.4 Converters (C)

The class of converters, which is syntactically heterogeneous, is described in Layton (2004: 319-366). It includes four types of converters which have several realizations depending on their syntactic environment. In the coarse tagset, all converters are tagged as C, allowing for lower error rates in automatic tagging (especially by removing the distinction between circumstantial and relative conversions, which can be ambiguous). The examples below are for the four fine grained classes:

CCIRC	Converter, circumstantial	ε, ε[ά], ερε
CFOC	Converter, focalizing (a.k.a. 2 nd tenses)	ε, ερε, ετε, ήτ[ά], εντ[ά]
CPRET	Converter, preterite	νε, νερε
CREL	Converter, relative	ετε, ετ, ήτ[ά], εντ[ά], ετερε

Note that a following conjugation base is segmented separately from the converter (cf. segmentation guidelines), e.g.:

- (16) ητ/CREL α/APST ψιμού ‘which he blessed’

The converter includes only ήτ, while α is a separate auxiliary base. The fused second person singular feminine form preceding a future marker, νερα ('you(F) would'), is tokenized into norms and tagged as follows: νερ/CPRET_PPERS α/FUT. Note that the normalized form of the future marker in this case remains α, but the lemma is να.

3.5 Conjunctions (CONJ)

Conjunctions are indeclinable words of Greek and Egyptian origin which link phrases and clauses. No distinction is made between subordinating conjunctions which introduce clauses ('because', 'lest') and coordinating conjunctions which connect phrases (e.g. 'and', 'or').

- (17) αγω/CONJ αιειβεγ ‘and I became thirsty’
(18) ειδω πιος κε/CONJ μηποτε/CONJ ταειβε ‘saying [that:] lest I become thirsty’

In the first example, the coordinating conjunction **ѧ** ‘and’ appears. Note that it is still tagged as a conjunction even if the first coordinated phrase is missing. In the second example, two consecutive conjunctions appear: **ѧ** ‘that, saying’ introduces the direct speech and the Greek origin **μηποτε** ‘lest’ is a conjunction within the direct speech clause. Also note that the word **ѧ**, originally derived from **ѧ** ‘say’ is not considered a verb in this usage.

3.6 Copulas (COP)

Copulas are markers in so-called nominal sentences which express predications of the sort A is B. The copula forms are **πε/τε/νε**. The tag COP is given also to copulas following a verbal clause for focalizing emphasis (i.e. ‘it is the case that...’), as illustrated below.

- | | |
|--|--|
| (19) ογαειν <u>πε</u> /COP | ‘he <u>is</u> a doctor’ |
| (20) ιερτωβε <u>μπλοεις</u> <u>πε</u> /COP | ‘(it <u>is</u> that) he prayed to God’ |

In the latter example, it is less obvious that **πε** is the copula, as its predicate is formally a clause and the form never changes its gender or number (i.e. as **τε/νε**; this is also referred to as ‘invariable **πε**’). Though the English translation cannot convey the presence of the copula adequately, these types of cases are still tagged as COP (see Layton 2004:223).

3.7 Existentials (EXIST)

Existentials include the unique lexemes **ογῆ** and **μῆ** in both pure existential and possessive forms, positive and negative, illustrated in the following examples.

- | | |
|---|--|
| (21) <u>ογῆ</u> /EXIST ογά εφεινε μμοκ | ‘ <u>there is</u> one who is like you’ |
| (22) <u>μῆ</u> /EXIST γῆγάλ εφδοε επεφδοεις | ‘ <u>there is</u> no servant who is above his
master’ |

The same tag is also used for the indefinite durative present and the fixed phrase **ογῆ σοι** ‘be able’ literally ‘there is power’.

- | | |
|--|---|
| (23) <u>ογητα</u> /EXIST η/PPERΟ ῆμαγ /ADV μπενειωτ αβραχαμ | |
| | ‘we have Abraham our father’, lit. ‘exists to us ... of Abraham...’ |
| (24) <u>μῆ</u> /EXIST σοι ητε τε γραφη βωλ εβολ | ‘scripture <u>cannot</u> be broken’ |

Note that the possessor pronoun is segmented apart from **ογητα** and tagged as a pronoun, and the accompanying **ῆμαγ** is an adverb.

3.8 Foreign Material (FM)

Foreign material includes text that is lexically and syntactically from a foreign language. It is distinct from loan words. Loan words are lexical entries that originate in another language (e.g., Greek, Latin) but are used in Coptic with Coptic syntax. Foreign material

consists of words, especially multiword expressions, with foreign syntax. The writer has momentarily switched languages rather than embedded a loan word into a Coptic construction

- (25) ογ παρα τογτο/FM νογ εвол αн ȝмпсѡмѧ тe ‘it is therefore not part of the body’

3.9 Future Marker (FUT)

The future marker **на**, derived from the verb ‘go’ is not considered an independent verb form when introducing a second verb and marking future tense. The following example illustrates the construction.

- (26) + **на**/FUT ȝотвек ‘I will kill you’

In rare cases, forms other than **на** can be considered for the future marker, e.g. **а** in:

- (27) нeр/CPRET_PPERS **а**/FUT ϲѡω ‘you would despise’ (2nd pers. fem.)

Contractions of multiple **и** are usually restored in the normalization, so that a diplomatic sequence like **тетнaрпmeечe** ‘you will think’ are usually normalized and only then tagged as follows:

- (28) тетн/PPERS **на**/FUT p/V

3.10 Inflected modifiers (IMOD)

Inflected modifiers are a somewhat heterogeneous class of suffixally inflecting non-verboids, including the quantifier **тhр**= ‘all of’, the focus particle **օγaa(t)=** ‘only’ and the reflexive **иmинmо=** ‘oneself’ (see Layton 2004: 118-123 and contrast the tag VBD). The suffix itself is tokenized apart and tagged as PPERO. These items are tokenized apart even within larger phrases, as in the second examples below.

- (29) **анок ȝѡѡ**/IMOD т/PPERO ‘I, as for me / me too’
(30) **ε π тhр**/IMOD **q** ‘in all of it, at all, wholly’

If the suffix is a 2nd pers. sg. fem. realized as zero, a portmanteau tag is assigned:

- (31) **иmинmо**/IMOD_PPERO ‘yourself (2nd pers. sg. fem.)’

3.11 Nouns (N)

The tag N is used for all nouns, common and proper, though the fine-grained tagset offers the specific tag NPROP for proper nouns.

- (32) **πен ειѡт**/N ‘our father’
(33) **антѡниoc**/NPROP ‘Antonius’

Note that verbal infinitives in the durative patterns and elsewhere, though technically and etymologically nominal in nature, are nevertheless tagged as verbs in order to facilitate the retrieval of verbal lexemes across constructions.

- (34) † πιστεύε/V επινούγτε ‘I trust in God’

3.12 Negations (NEG)

The tag NEG is used for independent negative items that are not part of an auxiliary base. The following lexemes are given the tag NEG: ν, αν, τ̄μ and μπρ (negative imperative marker). The first two can occur in the same sentence, in which case one NEG tag is used for each. The third negates infinitives and is tokenized separately from the verb and surrounding auxiliaries. The fourth is also a separate token and is not considered a verb form or part of the verb ειπε (this also applies to its lemmatization as an independent item, see lemmatization guidelines)

- (35) ν/NEG κηληρονομει τ̄μοκ αν/NEG ‘he will not inherit you’
 (36) εγωδαν τ̄μ/NEG σωτ̄μ ‘if they do not listen’
 (37) μπρ/NEG μογ κακως ‘don’t die badly!’

Note that the irregular negation με in με-ωψε ‘it is not appropriate’ is also tagged NEG and not as ANEGAOR.

3.13 Numerals (NUM)

The tag NUM is given to numerals and numerical constituents of complex numerals, as well as suffixed numerals as in the last example below.

- (38) τ̄ογ/NUM νοεικ ‘five (loaves) of bread’
 (39) χογτ/NUM αρτε/NUM ‘twenty-four’
 (40) ν|σεπ снаγ/NUM ‘two times, twice’

Note that the indefinite article ογ ‘a, one’ preceding a noun is tagged as ART, not NUM. Letters being used as numbers are considered NUM (including an alpha preceding a noun for the quantity ‘one’)

3.14 Demonstrative pronouns (PDEM)

The demonstrative pronouns, both attributive to the noun and substituting for a noun are tagged as PDEM.

- (41) ν ται/PDEM ξε ‘in this way’
 (42) ται/PDEM τε τ ξε ‘this is the way’

3.15 Interrogative pronouns (PINT)

This tag is used for the interrogative pronouns ογ ‘what’, νιμ ‘who’, των ‘where’, αω ‘which’, ογηρ ‘how much’. This is also true when they are used in complex phrases, as in the examples below.

- (43) ετβε/PREP ογ/PINT ‘what for, why?’
(44) ε/PREP των/PINT ‘where to?’

Note that the item νιμ is tagged PINT even when used after a noun to mean ‘some, any’.

3.16 Personal pronouns (PPER[*])

Personal pronouns generally receive the tag PPER, with three subtypes in the fine-grained subset for subject pronouns (PPERS), object pronouns (PPERL) and independent pronouns (PPERI).

- (45) α γ/PPERS σωτή ερο κ/PPERL ‘he heard you’
(46) ετβηητ ς/PPERL ‘for her’

Note that ‘object’ pronouns include objects of prepositions and all suffixed pronouns except the subject markers of verboids of the type [νανογ]q, [πεχα]q etc., which are tagged as PPERS.

- (47) πεχα γ/PPERS ‘he said’

The independent pronouns are reserved for emphatic uses and nominal sentences, including nominal sentence subject forms like ανΓ ‘I’ and the full forms of the type ανοκ ‘I’.

- (48) ανοκ/PPERI γωω τ/PPERL ανΓ/PPERI πεq γῆγαλ
‘I, as for me, I am his servant’

Also note that possessive pronouns like πεq ‘his’ are not segmented and receive a separate tag, PPOS.

3.17 Possessive pronouns (PPOS)

Much like demonstratives, all possessive pronouns, both attributive and standing in for a noun are tagged as PPOS. The personal suffix at the end of the pronoun is not separated, rather the entire forms, including πεq ‘his’, πα ‘my’ and ‘the one that belongs to’, πογ ‘your (fem.)’, πωι ‘mine’ etc. The following example illustrates these different types of possessives:

- (49) τα/PPOS πα/PPOS con τωι/PPOS τε ‘the one of my brother is mine’

This tag only applies to prefixal, article-like possessives. Suffix possessives, such as πατ^q ‘his foot’ are not tagged PPOS, but rather PPERO.

3.18 Prepositions (PREP)

This tag is used for all prepositions in both independent, prenominal states and presuffixal forms (which are tokenized apart from following suffixes). Note that prepositions that are historically derived from univerbized phrases but are now unsegmentable are tagged as one preposition, but complex prepositions involving a separable adverb are given two tags, ADV and PREP (cf. the tag ADV). Additionally, the *nota relationis* and accusative marker ο/ϊμο is regarded as a preposition. The following examples illustrate these principles.

- (50) ετβε/PREP ογ ‘for what? why?’
(51) εβολ/ADV γῆ/PREP ‘from, out of’ (lit. ‘out in’)
(52) εχῆ/PREP ‘upon, on account of’ (from ‘to head of’)

Also note that 2nd pers. sg. fem. objects often lead to portmanteau tags, e.g.:

- (53) ιμο/PREP_PPERO ‘you (2nd pers. sg. fem. accusative)’

If in doubt as to whether a lexicalized combination is considered a single preposition, please refer to the formatted CMCL lexicon supplied with the project’s tokenization module. This lexicon will be updated with future versions of the guidelines to accommodate dubious cases as they arise.

3.19 Particles (PTC)

The class of particles contains all indeclinable words that do not belong to one of the other classes, most notably and frequently the apposition marker νοι ‘that is...’ and a large number of, mostly Greek origin, sentence modifying particles that tend to appear in the second, Wackernagel position as they do in Greek as well (e.g. Δε, Γαρ).

3.20 Punctuation (PUNCT)

All punctuation marks, including periods at any height in the line, commas (including punctuation added in editions when annotating edited texts) or even question marks, colons etc. if they are used, are all given the uniform tag PUNCT. If decorations are tokenized (tildes, clusters of dots etc.), they may also be tagged as PUNCT, though refer to the tokenization guidelines for recommendations on normalizing text before tagging.

3.21 Unknown, damaged and lost items (UNKNOWN)

The tag UNKNOWN is given to fragmentary word forms damaged or missing beyond the ability to reach a reliable part-of-speech assignment. It is understood in the case of larger lacunae that the string used to encode the visible part of a word may in fact contain

several words. In cases where it is clear where word divisions occur, multiple tokens with corresponding UNKNOWN tags are given.

- (54) ε[...]/UNKNOWN ‘?’
(55) ε[...]/UNKNOWN η[...]/UNKNOWN ‘?’

Generally UNKNOWN tags are given even if the range of possibility is limited, i.e. even if we are certain a damaged morpheme is either an article or a possessive pronoun, an uncertain case is usually tagged as UNKNOWN.

3.22 Verbs (V[*])

The coarse tag V is given to all lexical verb forms that are not conjugation bases, also not including verboids, which receive a separate tag even in the coarse tagset due to their distinct syntax (see the tag VBD). In the fine-grained tagset, normal verb forms (V) are distinguished from stative verb forms (VSTAT) and imperatives (VIMP) as shown in the examples below. Note that verbal infinitives in the durative present are still tagged as verbs, although they are historically nominalized in this position, whereas nominalized infinitives following an article are understood as nouns, as in the last example. Verbs are tagged as VIMP only when they appear in the specific imperative form.

- (56) α η εωτή/V ερο κ ‘he heard you’
(57) † οβε/VSTAT ‘I am thirsty’
(58) αχι/VIMP ε ‘say it!’
(59) ξη η κοογη/N η η νογτε ‘in the knowledge of God, the knowing of God’

Also note that in rare cases, object pronouns that are realized as zero will lead to portmanteau tags, e.g.:

- (60) τετη/PPERS ητ/V_PPERO ‘you bring me’

Since ητ as the presuffixal form of εινε ends in τ, the object pronoun -τ ‘me’ is subsequently dropped. However the portmanteau tag reflects the presence of a grammatical object.

For compound verbs (see §180 in Layton), the entire compound is considered “a single unit in boundness, syntax, and meaning.” Therefore, the entire compound is tagged V. The components of the compound may be annotated further on a morph level annotation. (See Transcription guidelines for more information on bound groups, morphemes, and word segmentation.) Common examples include compound verbs formed with †-, p-, and χι-.

- (61) ετ/CREL πνοβε/V
(62) ε/CCIRC κ/PPERS †cbω/V

The basic criterion for identifying compound verbs is the absence of an article: *ρνοβε* ‘to sin’ is considered as single, compound verb (which can still be analyzed morphologically into two units *ρ+νοβε*, perhaps like English ‘sin-ify’, if there were such a word). However *ρ η μεεγε* ‘to think’ looks exactly like any verb + definite noun phrase combination, and is therefore tagged as three units despite being a common lexicalized combination: it comprises a verb, an article and a noun.

Exceptions: Some object nouns cannot appear as definite, or are made definite other than by an article. These include objects with *ηιμ* ‘some, any’, *λαλγ* ‘something’ and *ζοιηε* ‘some (ones)’, number words, as well as verbal objects with a suffixal possessive pronouns, such as *κε ρατ q* ‘set one’s foot’ (the foot is definite). Even though they may appear next to a verb without an article, these are tokenized and tagged apart from the verb (for possessed objects, the possessive is its own token, tagged PPERO, not PPOS).

Compound verbs containing a specific imperative form are also considered VIMP:

- (63) *αριχμελ/VIMP* ‘serve!’ (imperative of compound *ρεμελ*)

3.23 Verboids (VBD)

The category VBD is given to a small class of suffixally inflected predicates described in Layton (2004: 297-304), including the common *πεχε-*/*πεχα-* ‘say’, *ηανογ-* ‘be good’ etc., but not including possessive existentials of the type *ογντε-* (see the tag EXIST). The personal suffix following a VBD is tagged as its subject, i.e. PPERS (or simply PPER in the coarse tagset).

- (64) *πεχα/VBD q/PPERS* ‘he said’

- (65) *ηανογ/VBD c/PPERS* ‘she/it is good’

For the form *μεωδα-* note that two analyses exist. When it is declinable and literally means ‘X does not know’ (also prenominal *μεωθε-*), then it is VBD. When it is the lexicalized adverb form *μεωδακ* meaning ‘maybe’ (etymologically from ‘you never know’), it is a single unit, tagged ADV. Note that the latter form does not agree with the addressee if they are not masculine singular. Contrast the following examples from Layton (2004:303):

- (66) *μεωδα/VBD q/PPERS* *ηπηλαγ ετηηλακω ησωφ ηπηκомос*
‘he does not know when he will leave the world’
- (67) *μεωδακ/ADV* *τηασω γατετηγтн* ‘maybe I’ll stay with you’

In the latter example, the addressee is plural (*τηγтн*), but the form remains *μεωδακ*, indicating that it is an unanalyzed adverb.

4. References

- Layton, Bentley (2004), *A Coptic Grammar*. Second Edition, Revised and Expanded.
(*Porta linguarum orientalium* 20.) Wiesbaden: Harrassowitz.
- Schmid, Helmut (1994), Probabilistic part-of-speech tagging using decision trees.
Proceedings of the Conference on New Methods in Language Processing. Manchester,
UK, 44–49.
- Shisha-Halevy, Ariel. 1988. *Coptic Grammatical Chrestomathy. A Course for Academic
and Private Study*. (*Orientalia Lovaniensia Analecta* 30.) Leuven: Peeters.

Coptic SCRIPTORIUM – Lemmatization Guidelines

Version 1.1.0 / 2016-11-02

Amir Zeldes

Preamble

The purpose of lemmatization is to facilitate finding variant and inflected forms that are related to the same lexical entry, roughly equivalent to a dictionary entry. However in many cases, it may be unclear what the underlying, uninflected form of a word is: is the lemma of the pronoun ‘me’ defined as ‘I’ (i.e. the nominative form)? Should the lemma of ‘us’ then be ‘we’? Alternatively we could put all personal pronouns under one lemma: then ‘we’, ‘us’, ‘I’, and ‘me’ all belong to the same lemma, but which form should be taken for the common lemma?

There can be many arguments for and against certain practices. In these guidelines we attempt to give a set of instructions for Coptic which is: a. easy to apply consistently and b. useful for searching purposes.

Guidelines by Part-of-Speech Class

Articles and copulas

Articles are lemmatized according to the non-assimilated, simple short form of the corresponding masculine singular article (if distinct). This means that the lemma of **π**, **πε**, **τ**, **τε**, **η**, **ηε** and **μ** (assimilated form of **η** before a labial consonant) is for all of the above **π**. For indefinites **ογ** and **ζεη** there is no special masculine form, but the singular lemma **ογ** is taken for the plural **ζεη** and also for the variant spelling **γ**.

Copulas follow a similar rule: the lemma for all three number/gender forms (**πε/τε/ηε**) is **πε**.

Pronouns

Personal pronouns

Lemmas are mainly helpful where they deliver added value over searching for plain strings. It is therefore useful to give common lemmas for each of the personal forms: first person, second M/F, ... Given that the SCRIPTORIUM part-of-speech guidelines already distinguish subject and object pronouns, it is considerably more useful to group subjects and objects of the same person together, while not distinguishing the different forms (e.g. **†**, **ι** for first person) which can be found using a plain-text search anyway. We therefore annotate the following personal pronouns (SCRIPTORIUM tags in PPER*, i.e. PPERS,

PPER0, PPERI) with the following lemmas, based on the independent stressed pronoun forms (note that lemmatization is based on normalized forms without supralinear strokes or other diacritics; cf. transcription and normalization guidelines):

Person	Lemma	Pronoun forms
1st sg.	ΔΝΟΚ	ΔΝΟΚ, ΔΝΓ, †, Ι, ΝΤ, Τ, Δ
2nd sg. masc.	ΝΤΟΚ	ΝΤΟΚ, ΝΤΚ, Κ, Γ, ΤΚ
2nd sg. fem.	ΝΤΟ	ΝΤΟ, ΝΤΕ, ΤΕ, ΤΡ, Ρ, Ε
3rd sg. masc.	ΝΤΟΩ	ΝΤΟΩ, Ζ
3rd sg. fem.	ΝΤΟΣ	ΝΤΟΣ, Σ
1st pl.	ΔΝΟΝ	ΔΝΟΝ, ΔΝ, Ν, ΤΝ, ΤΝ
2nd pl.	ΝΤΩΤΝ	ΝΤΩΤΝ, ΝΤΕΤΝ, ΤΝ, ΤΗΥΤΝ
3rd pl.	ΝΤΟΟΥ	ΝΤΟΟΥ, Υ, ΟΥ, ΣΕ, ΣΟΥ

The pronoun lemmas alone therefore primarily give access to search by person (1st, 2nd...); to cross-reference these with the form, e.g. independent pronoun, cross-reference the POS annotation (in ANNIS: pos="PPERI"). For a specific subform (e.g. ΔΝΓ not ΔΝΟΚ) use the form search norm="ΔΝΓ".

Possessives, interrogatives and demonstratives

Interrogative pronouns are each equivalent to their own lemma, i.e. ΟΥ is lemmatized ΟΥ and ΝΙΜ as ΝΙΜ.

Possessive, and demonstrative pronouns are lemmatized to their own normalized form, but with one modification: non-masculine singular determiners are given the masculine form, i.e. the lemma of ΠΕΩ is ΠΕΩ, the lemma of ΠΑ is ΠΑ etc., but the lemma of ΤΕΩ and ΝΕΩ is also ΠΕΩ. Similarly, the lemma of ΠΕΙ and ΝΕΙ is ΠΕΙ, and the lemma of ΠΑΙ and ΤΑΙ is ΠΑΙ. This allows an easier search for all possessives (in ANNIS: pos="PPOS", finds ΠΕΩ, ΠΕΩ, ΝΟΥ ...), all third person plural possessives (lemma="ΠΕΩ", finds ΠΕΩ, ΤΕΩ and ΝΕΩ) and all third person plural possessives of feminine objects (norm="ΝΕΩ"), and similarly for demonstratives.

Adverbs, particles and conjunctions

Adverbs, particles and conjunctions are always given their own normalized form as a lemma. This includes Greek adverbs in -ως, which are lemmatized as such, e.g. ΣΩΛΩΣ has the lemma ΣΩΛΩΣ.

Nouns

Nouns are given their dictionary form as a lemma. For most nouns, singular and plural forms are identical, meaning there is no dilemma. For nouns with irregular plural forms, the singular form is taken as a lemma, e.g. ΣΩΒ 'deed' is the lemma of both singular ΣΩΒ and plural ΣΒΗΣ, and similarly, possessed forms like ΤΟΟΤ(Ω) are lemmatized to the absolute form, i.e. ΤΩΡΕ. In order to find irregular forms, one can then simply search for

nouns whose lemma is different from the noun form (in ANNIS: lemma != norm). The same rules apply to proper nouns, though these rarely occur in the plural.

For nouns which only occur in the possessed form, if both prenominal and presuffixal forms exist, the prenominal is taken as the lemma, e.g. *ενα-* and *ενε-* ‘(one’s) will’ are lemmatized as *ενε*. If only a presuffixal form exists, it is taken as the lemma as well, e.g. *ναιατ-* ‘blessed is...’ has the lemma *ναιατ*.

Nouns that have related masculine and feminine forms are considered separate lemmas. For instance, the noun *ουρε* ‘son’ is its own lemma, and the separate noun *οιεερε* ‘daughter’ also has a separate lemma (which is *οιεερε*). Similarly, Greek words in -οc are considered separate from related words in -οn, e.g. *πονηρος* ‘wicked person’ is its own lemma, and so is the separate *πονηρον* ‘wicked deed/thing’ an independent lemma.

Verbs

Verbs are lemmatized to the form of the absolute infinitive. This means that special prenominal or presuffixal forms are lemmatized to their respective dictionary entries, e.g. *сотп-* and *сетп-* are lemmatized as *сотп* ‘choose’. The same applies to stative and imperative forms, which are lemmatized to the dictionary entry, e.g. *κηт[†]* has *κωт* as a lemma and *Ճ* has *ειрε*. Likewise for prenominal forms, *сетп* and *р* are lemmatized as *сотп* and *ειрε*.

Note that auxiliaries are not lemmatized to their etymological verbs, i.e. the lemma of the past tense *Ճ-* is not *ειрε* but *Ճ*. Additionally, the negative imperative marker *мп* is lemmatized as *мп* as well, as it is considered to be a form of negation independent from the verb *ειрε*. However, the negative imperative of *ειрε* itself, *мпωр* IS lemmatized as *ειрε* (since it is a morphological imperative of *ειрε* itself, and functions as part of its paradigm with the sense ‘to do’).

For fused verb-object forms like *нт* ‘bring me’, see Portmanteau Tags.

Prepositions

Prepositions are lemmatized to their standard form **before noun phrases**. Therefore the lemma of *ε-* and *εпo-* is *ε*. For preposition forms containing a second person singular feminine pronoun (realized as zero), e.g. *εхω* ‘on you (fem.)’, *нсω* ‘behind you (fem.)’ etc. see Portmanteau Tags.

Existential and possessive predicates

The existential predicates are lemmatized as *օyn* ‘there is’ and *մn* ‘there isn’t’ (again note that lemmatization does not contain supralinear strokes). Like auxiliaries, the related possessive predicates are lemmatized using their form before a nominal subject: *օynтe*

and မျှ။ Note that forms with two မျှ, such as မျှမျှ and မျှမျှ။ etc. are considered to be orthographic variants of the forms with one မျှ, and should therefore be not only lemmatized with one မျှ, but also normalized to such forms.

Auxiliaries, negations and future marker

Auxiliaries are generally lemmatized to their form when preceding a nominal subject. Attention should be paid to auxiliaries sometimes ending in -ခဲ့: in normalized orthography, this is generally present before a nominal subject. The lemma of မာရေး- and မာပံ့- (jussive) is မာရေး, and the lemma of မာနတ်- and မာနတ်။ is မာနတ်။

However, the lemmas of auxiliaries that sometimes contain an intermediate pronoun do not contain that pronoun when they occur uninterrupted: the lemmas of ခုစွာ၏ (conditional) and ခြော (optative) remain ခုစွာ၏ and ခြော။ These receive the tags ACOND and AOPT respectively. For cases with an intervening pronoun, which receive different tags, see Portmanteau Tags.

Negative morphemes such as ဘဲ, ဘေး and ဘုံး are their own lemmas (the form မျှ before a labial is also lemmatized as မျှ). The negative imperative marker မြတ် is lemmatized as itself (မြတ်), and NOT as ခြော။

The future marker is given its own lemma မှာ. Note that the lemma remains so whenever a future marker is separately identified, even if the diplomatic realization is assimilated and reduced to မှာ, e.g. in complex forms like တော်မှာ ‘you will... (pl.)’ or နော်မှာ ‘you would have (fem. sg.)’.

Converters

Like auxiliaries, converters are lemmatized to their form before a nominal subject, viz.:

CCIRC/CFOC:	ခြော
CREL:	ခုခြော
CPRET:	နော်ခြော

For second person singular feminine ခု/ခြော (lemma="ခြော_NTO") see Portmanteau Tags.

Inflected modifiers

Modifiers of the type ဆုံး-, မီးမီးမီး-, မာယား-, တို့- are lemmatized to their form before the **third person masculine singular** pronoun သူ. Thus မီးမီးမီး- and မီးမီးမီး- are lemmatized as မီးမီးမီး. The portmanteau form မီးမီးမီး (yourself, fem. sg.) is lemmatized မီးမီးမီး_NTO (see Portmanteau Tags).

Portmanteau Tags

Some fused items receive a so-called portmanteau tag representing two categories at once. For example, the form εψωαν is considered to contain a conditional auxiliary and a subject pronoun: pos="ACOND_PPERS". In order to facilitate finding such cases regardless of the pronoun in use, in tags containing a conjugation base and a personal pronoun the form is lemmatized using both lemmas, separated by an underscore. For example, the lemmas of ειψωαν, εψωαν and εψωαν are εψωαν_ΑΝΟΚ, εψωαν_ΝΤΟC and εψωαν_ΝΤΟq respectively. The lemma of εψωαν remains εψωαν (pos="ACOND"), unless it contains a second person feminine singular subject, in which case the lemma is εψωαν_ΝΤΟ according to the rule above.

For the past tense second person singular feminine form ἀπ the lemma is similarly α_ΝΤΟ (pos="APST_PPERS"). The form μμινμμο (yourself, fem. sg.) is identical to the base of other personal forms, but is lemmatized μμινμμο_ΝΤΟ just like other forms containing a personal pronoun.

The same principle applies to prepositions: forms containing a second person singular feminine pronoun (realized as zero) are given portmanteau lemmas, e.g. εξω ‘on you (fem.)’ has εξν_ΝΤΟ, ονω ‘behind you (fem.)’ has ονα_ΝΤΟ etc.

For circumstantial or focalizing converter + second person feminine singular, the lemma ερε_ΝΤΟ is used (and similarly preterit νερε_ΝΤΟ and relative ετερε_ΝΤΟ).

Verbs containing an object pronoun, such as ιντ ‘bring me’ are lemmatized using the base form of the verb and the pronoun’s lemma: εινε_ΑΝΟΚ.

FAQ

What is the lemma of...

Form	Norm	Lemma	Notes
γωω(η)	γωω	γωω	
χεκας	χεκαας	χεκαας	just a normalization issue
καταρο(η)	καταρο	κατα	
μμιντα(η)	μντα	μντε	
ογει	ογει	ογα	'one' (fem.) is considered an inflected form
σινογιοом	σινογиом	σινογωм	complex plural still gets lemmatized as sg.
саво	саво	сво	
тсаво	тсаво	тсаво	note this is one norm unit
сүнкера	сүгкера	сүгкера	Greek words should normalize r for н

(π)τηρό	τηρό	τηρό	when used as a noun, this is one norm
ναιάτ(ς)	ναιάτ	ναιάτ	
νι	νι	νι	has a separate lemma from πει/τει/νει

What about units with complex morphology?

The morph layer is not a type of lemmatization. We lemmatize the whole word, which may contain inflected forms of stems below the word level. Thus the lemma of ρχρειά is ρχρειά, not ειρέχρειά.