White Paper Report

Report ID: 112322

Application Number: PW-51672-14

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

Institution: University of the Pacific

Reporting Period: 5/1/2016-8/31/2016

Report Due: 8/31/2016

Date Submitted: 8/29/2016

Coptic SCRIPTORIUM: Digitizing a Corpus for Interdisciplinary Research in Ancient Egyptian

White Paper

National Endowment for the Humanities Division of Preservation and Access

Humanities Collections and Reference Resources

Grant PW-51672-14

Director: Caroline T. Schroeder, University of the Pacific

Co-Director: Amir Zeldes, Georgetown University

Project website: copticscriptorium.org

August 31, 2016

Table of Contents

1. Project Description

This grant enabled the Coptic SCRIPTORIUM project to plan for and pilot a digitized corpus of Coptic texts of importance to scholarship in biblical studies, early Christian history, and linguistics. We developed a pilot text corpus and established technical standards for the digitization of Coptic literature. This corpus represents the first online, open-access Coptic corpus digitized according to existing and emerging standards in the Humanities and in digital preservation.

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. The Coptic language has proven essential for the decipherment and continued study of Ancient Egyptian and is of major interest for Afro-Asiatic linguistics and Coptic linguistics in its own right. Coptic manuscripts are sources for biblical and extra-biblical texts and document ancient and Christian history. The pilot corpus and standards advance knowledge in these fields by increasing access to now largely inaccessible texts of broad significance, and by providing models for future work in these fields.

2. Grant Achievements and Products

We sought to address the following five objectives over the course of the grant period:

1. Digitization and transcription of texts for a digital Coptic corpus
2. Development of a corpus architecture that takes into account the fragmentary nature of the source manuscripts and the original text and codex structures
3. Development of data curation standards (including universal references) and exemplars, which take into account the complexity of the corpus architecture and metadata, and which conform with current standards in the field (such as the EpiDoc TEI-XML standards)
4. Establish standards and exemplars for linked data that are conformant with standards and ensure interoperability with other digital projects on the ancient world (such as Pleiades)
5. Planning standards, methods, and workflows to ensure interoperability of the digitized corpus in the multiple digital formats that are required for interdisciplinary work.

To meet these objectives, we achieved the following seven products or outcomes, which we describe in more detail below, referencing the above objectives:

1. A pilot digitized corpus of Coptic texts available in multiple formats and visualizations
2. Data modeling and data curation for Coptic digital text corpora
3. Digital and computational tools to analyze, process, and annotate the language

4. A searchable database created from the texts and tools using the ANNIS database infrastructure
5. Planning for a collaborative platform for scholars to contribute texts and annotations as well as conduct research using the corpus, tools, and database
6. Documentation on standards, methodologies, and workflows as well as user guides.
7. Articles and conference papers to distribute the results of our work

2.1 Pilot Text Corpus

Corpus Transcription and Digitization

Coptic SCRIPTORIUM published a pilot digital corpus of literary, biblical, and documentary Coptic texts. Rather than producing one all-inclusive corpus, we created a digital architecture of multiple corpora based on existing standards in the field for text groupings by author, work, or genre.

The pilot corpora include samples from three genres: Bible, monastic literature, and documentary sources.

The following corpora were transcribed by project participants in consultation with manuscripts or manuscript facsimiles (photographs), machine-annotated using Coptic SCRIPTORIUM's tools (described further below), manually reviewed and further annotated for metadata by project participants (see below for metadata standards), and reviewed by project senior editors:

> *Abraham Our Father* (letter by monastic author Shenoute of Atripe): This corpus is nearly complete, missing only a few manuscript pages that currently reside in Cairo, which are difficult to access. It was our first pilot corpus, created when were beginning to develop digital standards for processing and encoding Coptic. As our standards and processes evolved over the grant period, we updated the corpus. More inconsistencies exist in this corpus than others, however, due to it being our test case. Desiderata for future work includes fully revising this corpus.
>
> *Acephalous Work 22* (letter/sermon by Shenoute): This corpus includes some but not all of the known fragments of this work, including previously unpublished (even in print sources) manuscript pages.
>
> *I See Your Eagerness* (sermon/discourse by Shenoute): Most of the known fragments of this work have been digitized and published on our site.
>
> *Not Because a Fox Barks* (letter by Shenoute): Known fragments of one manuscript witness to this work have been digitized and published on our site.
>
> *Letters of Besa* (letters by monastic author Besa): Three letters have been digitized and published in partnership with two other grant projects: the bilateral NEH-DFG grant (HG-229371) of the KELLIA collaboration with German partners (2015-2018), and

KOMeT project (2013-14) funded by the German Federal Ministry of Education and Research (BMBF).

Apophthegmata Patrum (or, *Sayings of the Desert Fathers*, a collection of sayings attributed to early monastics): Thirty-six sayings or apophthegms have been digitized and published; more have been digitized with plans to publish in Fall 2016.

The following corpora were scraped from existing digitized corpora on other project websites, machine-annotated using Coptic SCRIPTORIUM's tools, manually reviewed and further annotated for metadata by project participants, and reviewed by project senior editors:

Gospel of Mark: The entire Sahidic Gospel of Mark has been published. The digitized text originates from J. Warren Wells' Sahidica project website, which is now defunct. Please see our documentation on Sahidica for more details.

1 Corinthians: The entire Sahidic 1 Corinthians has been digitized and annotated. Most chapters have been published with remaining chapters scheduled for release in Fall 2016. The digitized text originates from J. Warren Wells' Sahidica project website, which is now defunct. Please see our documentation on Sahidica for more details.

Papyri.info Coptic documentary texts: Three Coptic papyri and ostraca from open access project Papyri.info have been annotated and published.

The following corpus was scraped from an existing digitized corpus and machine-annotated using Coptic SCRIPTORIUM's tools. Project senior directors added metadata but did not review or edit text annotations for accuracy.

Sahidica New Testament: The digitized text originates from J. Warren Wells' Sahidica project website, which is now defunct. Please see our documentation on Sahidica for more details.

We have published over 59,000 Coptic words in our manually-annotated corpora and over 231,000 Coptic words in our machine-annotated New Testament.

Corpus Architecture

Coptic SCRIPTORIUM implements digital methods used in different academic fields, especially digital editions, philology, and corpus linguistics. We needed to develop a corpus architecture that suited the needs of these different methodologies.

In order to facilitate work by historians and philologists who work with individual texts, authors, and other historically contingent text groupings, we chose to create multiple corpora based on existing field-specific understandings of authors and works (as described above).

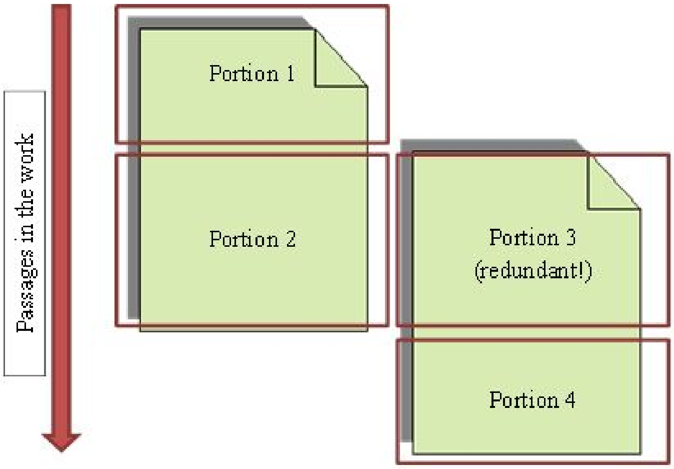
Within each corpus, we created digital documents corresponding to contiguous extant manuscript pages in a modern repository. This architecture was motivated by several concerns.

First, the manuscript witnesses (which all originate from Egypt) are fragmented and dispersed across multiple modern repositories, predominantly libraries, museums, and universities in Europe and North America.[1] This architecture allows us to publish fragment by fragment even if we are unable to publish the contents of an entire codex or all witnesses to a work. Our corpus linguistics database infrastructure also functions more efficiently with large numbers of small documents than with smaller numbers of large documents.

Our corpus architecture also accounts for parallel manuscript witnesses. Extant Coptic literary manuscripts differ from Greek and Latin biblical manuscripts or even commonly known Greek and Latin authors; Greek and Latin (especially in Biblical manuscripts) typically have many witnesses and variants with complex stemmatics. In contrast, Coptic literary manuscripts often do not have parallel witnesses, and when they do they are few or the parallel witnesses contain lacuna, meaning there are fewer witnesses for any given section of a work. (NB: Our biblical corpora right now are not based on individual manuscript witnesses but on a pre-existing digital edition. Documentary sources rarely have multiple witnesses.) Where parallel witnesses exist, we annotate one fragment in the metadata as primary (or non-redundant). Our criteria for determining the primary witness is usually that the text on the manuscript page(s) is clearer and/or contains the fewest lacunae. For other fragments with parallel witnesses, we create a digital document for the section of that fragment that contains the parallel, noting it in the metadata. We then create a separate digital document (or documents) for the section(s) of the parallel witness that are *not* redundant (see the diagram below). This decision is motivated by search functionality. This corpus architecture and annotation allows researchers to search for *all* witnesses to a text string (redundant and primary/non-redundant) or only one (search only non-redundant). We anticipate that philologists, historians, and librarians may wish to see hits for all witnesses. However, multiple hits for the same text (due to parallel manuscript witnesses) may not be desirable for computational linguists, for example when calculating vocabulary frequencies.[2]

[1] Caroline T. Schroeder and Amir Zeldes. “Raiders of the Lost Corpus.” *Digital Humanities Quarterly* 10.2 (2016). Web. 18 Aug. 2016.

[2] Amir Zeldes, “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.



Data Formats and Visualizations

Due to Coptic SCRIPTORIUM’s interdisciplinary approach, we release our data in three formats.

- XML files using the EpiDoc subset of Text Encoding Initiative (TEI) XML standards. 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. 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. 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 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.

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.
- 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.

See Appendix B for examples.

2.2 Data Modeling and Data Curation

Data Curation: Text Annotation

Our digitized texts (either scraped from existing digital corpora or manually transcribed by project contributors) have many levels of annotation. We annotate using the tools and technologies described in section 2.2 Digital and Computational Tools to create a digitized file annotated in a multi-layer format, such as a spreadsheet.

The layers of text annotation are as follows:

tok	Tokens, which are smallest possible units of text annotated. These units may be smaller than individual words, since words may be broken across a line break in a manuscript or individual letters may require annotation for color, style, or lacunae. The orthography of text reflects the original text source (e.g., a diplomatic transcription of a manuscript or the exact digitized text taken from Sahidica or Papyri.info).
orig	This layer consists of text segmented into the unit of language that corresponds to what we would consider a “word” in Coptic. The orthography reflects the original text source (e.g., a diplomatic transcription of a manuscript or the exact digitized text taken from Sahidica or Papyri.info) and includes supralinear strokes and other markings from the original source. (Text in the orig layer contains the same data as the tok layer, but segmented differently.)
orig_group	This layer consists of the same text data (using the original orthography) as in the tok and orig layers but segmented in what are known as bound

	groups in Coptic. Coptic is an agglutinative language in which linguistic units such as articles and nouns or subject pronouns and verbs are bound together. We follow the linguistic theory of Bentley Layton regarding binding of Coptic words and morphs.[3]
norm	Like the orig layer, this layer consists of text segmented into what we would consider "words" in Coptic. Spelling and orthography have been normalized in several ways. Supralinear strokes, overdots, and tremas have been removed. Normalization corrects spelling and expands abbreviations, including nomina sacra.
norm_group	This layer consists of the same bound groups as the text in the orig_group layer. The text itself, however, has been normalized with respect to spelling and orthography. The textual data matches the norm layer in content but the orig_group layer in segmentation.
pos	Part of speech tags (e.g. N for noun, ART for article, CFOC for focalizing converter, etc.).[4] This annotation annotates the normalized word (or norm layer).
morph	Words that contain units of linguistic importance smaller than the word unit are segmented into morphs. In Coptic morphs include prefixes such as mnt, at, and ref. Compound words (e.g., r-xreia, "to need") are also annotated on the morph level.[5] Note that morph units DO NOT receive parts of speech. Words that do not need annotation on on the morph level remain unannotated in this layer.
lang	Language of origin tags for loan words (Greek, Latin, Aramaic, Arabic, Hebrew,etc.). Words that are not loan words remain unannotated. This layer annotates the normalized word (or norm layer) except in the case of compounds and other words annotated on the morph level. For words annotated on the morph level, the morph is annotated. E.g., for r-xreia, only xreia receives the annotation "Greek." It can be difficult to determine how a loan word entered into the Coptic language. For example, some words of Hebrew or Aramaic origin may have entered into the language through the Greek Bible. However, such a determination is not certain. So loan word annotation reflects the *oldest* possible originary language.

[3] Bentley Layton, *A Coptic Grammar*, 3rd ed., Porta Linguarum Orientalium Neue Serie Vol. 20 (Wiesbaden: Harrassowitz, 2011).

[4] Amir Zeldes 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, *http://dsh.oxfordjournals.org/content/30/suppl_1/i164*. See our current Tagset Documentation for updated information.

[5] See our Transcription Guidelines (link and Appendix G), sections 4.3 and 4.4, for more information.

lemma	The lemma (or dictionary head word); it annotates the normalized words (“norm” layer).
note	Notes inserted into the transcription by editors.
hi@rend	Text renderings, such as color of ink, size of letters, ekthesis, white space in the manuscript, etc. Corresponds to the EpiDoc TEI-XML element and attribute hi@rend.
gap	Annotates for lacunae. Corresponds to the EpiDoc TEI-XML element gap. Uses attributes such as @reason, @unit, @quantity, and @extent. With attributes, each element+attribute annotation generates a new layer in the multi-layer data model.
supplied	Annotates for supplied text where text is missing from the original for a variety of reasons. Corresponds to the EpiDoc TEI-XML element supplied. Uses attributes such as @evidence and @reason. With attributes, each element+attribute annotation generates a new layer in the multi-layer data model.
lb@n	Annotations for line breaks, which are numbered according to the original manuscript or source. Corresponds to the EpiDoc TEI-XML element and attribute lb@n.
cb@n	Annotations for column breaks, which are numbered according to the original manuscript or source. Corresponds to the EpiDoc TEI-XML element and attribute cb@n.
pb@xml_id	Page numbers of original manuscript (not the current repository catalogue numbering). Corresponds and converts to TEI-XML element and attribute pb @xml:id.
translation	English translation. Not aligned word for word. Aligned roughly by sentence or clause.
p	Paragraph breaks for translation and normalization.
verse	Verse of the text written as number. Currently only annotated in works and documents with standard or canonical versification (such as the Bible).
chapter	The chapter of text as number. Currently only annotated in works and documents with standard or canonical chapters (such as the Bible).
sbl_greek	The Greek New Testament text. Annotation for the New Testament corpora only. Aligned by verse with the Coptic. Source is the XML Greek New Testament created by the Society of Biblical Literature and Logos Software.

sbl_apparatus	The apparatus for the Greek New Testament text. Annotation for the New Testament corpora only. Aligned by verse with the Coptic and Greek New Testament text. Source is the XML Greek New Testament created by the Society of Biblical Literature and Logos Software.

We convert each document file containing the multilayer text annotations and metadata (described in the next section) into the following formats, described in more detail in the previous section: one TEI-XML file that validates to the EpiDoc subset, standoff PAULA XML files, and relANNIS database files.

Data Curation: Metadata

Each document receives annotation for metadata according to the following fields:

corpus	Corpus name
Coptic_edition	If the text has been published before, include publication information here
Greek_source	Information about the Greek version of the text if it exists (e.g., Greek Alphabetical or Systematic Apophthegmata Patrum).
title	Title of this document (unique).
msItem_title	The name or title of the conceptual work, e.g. Abraham Our Father, To Thieving Nuns. Corresponds to the TEI XML element and attribute msItem@title
author	Author of the conceptual work (if known).
language	Language in which the text is written
annotation	Names of annotators (transcribers, editors, annotators) in comma delimited sequence
project	Name of project supporting the transcription/annotation/publication (e.g., Coptic SCRIPTORIUM, KoMET, KELLIA, etc.)
translation	Name(s) of translator(s) inserted here in comma delimited sequence. ("None" indicated if no translation published by Coptic SCRIPTORIUM.)
msName	The sigla for the original manuscript of the text in the document (if a manuscript is the original source). Use Corpus dei Manoscritti Copti Letterari codes (e.g., MONB.YA). Corresponds to TEI XML element msName.
pages_from	Beginning of page sequence of document using the pagination of the original codex (the page number of scribe but written in arabic numerals rather than Coptic). Corresponds to TEI XML element and attribute pages@from.

pages_to	End of page sequence of document using the pagination of the original codex (the page number of scribe but written in arabic numerals rather than Coptic). Corresponds to TEI XML element and attribute pages@to.
msContents_title@type	Optional annotation. Describes the type of contents (e.g., *Discourses* or *Canons* for Shenoute's corpora). Corresponds to the TEI XML element and attribute title@type within the msContents section of the TEI header.
msContents_title@n	The volume number of the type of entity annotated in msContents_title@type (e.g., 3 for Volume 3 of Shenoute's *Canons*). Corresponds to the TEI XML element and attribute title@n within the msContents section of the TEI header.
repository	The current museum/library/etc where the manuscript currently resides.
collection	The collection or department within the current repository where the manuscript currently resides.
idno	The catalogue number of the manuscript in the current repository.
version@n	The version number of this Coptic SCRIPTORIUM data
version@date	The version date of this Coptic SCRIPTORIUM data in YYYY-MM-DD format.
source_info	Information about the source of the data if it was not a transcription commissioned by Coptic SCRIPTORIUM.
license	The license for the data.
document_cts_urn	A URN that applies to the document following a data model based on the Canonical Text Services model of citation and data retrieval. (See more below.)
Trismegistos	The identification number assigned to the manuscript (or text bearing object) by the Trismegistos database, if Trismegistos includes the text-bearing object in its database.
objectType	Type of text-bearing object (e.g., codex, papyrus, ostracon, etc.). Corresponds to the TEI XML element objectType.
country	Country of origin of the text-bearing object.
placeName	City or village or place name of the original location of the text-bearing object; should be a recognizable name in a gazetteer for future linking to online GIS databases such as Pleiades. Corresponds to the TEI XML element placeName.
origPlace	The name of the place of origin of the text-bearing object, not necessarily city/town/village name (e.g., White Monastery). Corresponds to the TEI XML element origPlace.
origDate	Prose description of the date of the text-bearing object (e.g. Between 900 and 1200 C.E.). Corresponds to the TEI XML element origDate.

origDate_precision	Likelihood that the dating is accurate – usually “low”, “medium”, or “high”. Corresponds to the TEI XML element and attribute origDate@precision.
origDate_notBefore	Date of the terminum post quem (in four digits with leading zeros, e.g., 0900). Corresponds to the TEI XML element and attribute origDate@notBefore.
origDate_notAfter	Date of the terminum ante quem (in four digits with leading zeros, e.g., 1200). Corresponds to the TEI XML element and attribute origDate@notAfter.
source	If the digitized text comes from another source, the names of the editors of that source are listed here (used for Sahidica and other donated texts).
note	Additional prose note about the text and/or document.
witness	Prose description of parallel witnesses (if any) to the text in the document.
redundant	Required field with only “yes” or “no” entries. Yes annotates the file as a redundant witness (meaning another parallel is the primary witness). No annotates the file as the primary witness for the text it contains (whether or not it has a parallel). Any file with NO parallel witness is annotated as redundant=no.
previous	Contains the document CTS URN for the previous document in the corpus.
next	Contains the document CTS URN for the next document in the corpus.
endnote	Contains a note about the document that will appear in the HTML visualizations at the bottom of the visualization.

For ease of use by multiple editors, we instituted a system of semantic file names for our digital files, rather than random numeric file names. The naming system for filed in each corpus is internally consistent and includes terms or sigla derived from the standard canonical referents for the texts contained within those files.

The metadata model outlined above includes an entry for URNs minted according to the Canonical Text Services data model. We implemented this system of URNs in order to facilitate both citation of data and retrieval of text data, particularly by researchers in history, philology, or Religious Studies who would be citing passages, chapters, and documents (as opposed to ANNIS database query results). As described on the Homer Multitext project site for the CTS URN notation system, CTS URNs organize data into text groups and passages, with the following syntax:

> “Colons separate the top-level elements of a CTS URN.... The top-level elements are:
>
> 1. urn name space (required: always cts)
>
> 2. cts namespace (required: a value that can be resolved to a unique URI)

3. work identifier (required: a value registered in the designated registry)

4. passage reference (optional)

5. subreference (optional)

The general structure of a CTS URN is therefore

urn:cts:CTSNAMESPACE:WORK:PASSAGE:SUBREFERENCE?"

Components separated by a dot (period) within the work namespace indicate the textgroup, work, edition, translation, and/or exemplar.

Coptic SCRIPTORIUM has minted two CTS namespaces: copticLit (for literary works) and copticDoc (for documentary papyri and ostraca). Within the work namespace, we identify a textgroup (often but not always the author) and an edition (either the manuscript/papyrus/text-bearing object or the born digital edition). For example:

- urn:cts:copticLit:shenoute indicates all literary works in the text group of works by the author Shenoute
- urn:cts:copticLit:shenoute.abraham indicates all documents containing the literary work by Shenoute known as *Abraham Our Father*
- urn:cts:copticLit:ap indicates all literary works in the text group known as the *Apophthegmata Patrum*
- urn:cts:copticLit:ap.6.monbeg indicates the apophthegm traditionally numbered #6 in the text group known as the *Apophthegmata Patrum* as witnessed in the manuscript known as codex MONB.EG (White Monastery codex EG).
- urn:cts:copticLit:nt.1cor.sahidica:1 indicates chapter one of the book of the New Testament known as 1 Corinthians as witnessed in the digital edition Sahidica.

Text data can be retrieved using the namespaces of these CTS URNs on a web application described in the Tools section below, or through search on ANNIS. Because these URNs are location-independent, they can be used to help locate and identify our corpus data even when the data files are archived and the project website goes offline. Our data model for CTS URNs is part of a working group on persistent identifiers in the Research Data Alliance.

2.3 Searchable Database

The text corpora are released in a searchable database, the ANNIS search and visualization platform. ANNIS was developed to enable computational research in multiple languages.[6] We chose to adapt this open source infrastructure rather than recreate our own database. The primary customizations have been to install a pop-up Coptic keyboard for entering queries (if the

[6] Thomas Krause and Amir Zeldes (2016). "ANNIS3: A New Architecture for Generic Corpus Query and Visualization." *Literary and Linguistic Computing* 31(1), 118-139: fqu057. *llc.oxfordjournals.org*. Web.

researcher does not have a Coptic keyboard installed on her/his device) and embedding a Coptic webfont to users can read the Coptic characters even if they do not have a Coptic font installed on their devices.

ANNIS is installed on a Georgetown University server, and the SCRIPTORIUM instance is accessible openly on the web at https://corpling.uis.georgetown.edu/annis/scriptorium.

Researchers can search combinations of annotations and/or metadata described in the data model above. The visualizations (also described above) are generated by and within ANNIS using the annotations and CSS. For search, researchers can use the ANNIS query language or regular expressions. Results can be downloaded. Links to search queries can also be saved.

2.4 Digital and Computational Tools for Processing and Annotation of Digitized Text

For machine processing and annotation of digital Coptic text, the project produced the following tools and technologies. Development took place primarily under the aegis of the NEH grant HD-51907, which ran concurrently with PW-51672-14. Descriptions of these tools are available in the documentation accompanying the code in the corresponding GitHub repositories. The White Paper for HD-51907 will document the tools more fully.

The following tools were developed for the Sahidic dialect of the Coptic language. Editors used them in conjunction with manual annotation to produce the annotated text corpora according to the data models outlined above.

- Font and character converters: Convert text of documents transcribed in legacy ASCII fonts into the Unicode Coptic character set. Two converters were produced for several legacy fonts.
- 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.*[7]
- Normalizer: Normalizes orthography and spelling of Coptic text.
- Part of speech tagger: A probabilistic tagger built from the independent tool TreeTagger trained on a set of Coptic training data.
- 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.
- Language of origin tagger: A lexical tagger that annotates each word or morph for for its language of origin. Native Egyptian vocabulary remain un-annotated.

Over the course of the grant period, we updated the tools with period enhancements and bug fixes. In addition, we recursively updated the tools after publishing a large amount of new textual

[7] Layton, *A Coptic Grammar*.

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.[8] The pipeline can be used at https://corpling.uis.georgetown.edu/coptic-nlp/, and the code is available on GitHub.

Under the aegis of grant HD-51907, we developed tools and plugins for validating data and converting it into different formats.

- SGML input/output plugin: Converts digitized text annotated with SGML or XML tags into a multilayer spreadsheet format in Microsoft Excel. Available on GitHub.
- Annotation validation plugin: Validates multilayer annotations to ensure they conform to the data model described above. Available on GitHub.
- 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.

We developed a web application that provides the most recent version of our data in all the formats currently available using the CTS 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. Funding from this grant as well as the NEH ODH grant HD-51907 supported its development.

2.5 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. In addition participants and other collaborators have contributed to the development of the tools through the GitHub site. A goal for this grant period was to map out and plan for a more robust means for the general community of people interested in Coptic to contribute text or annotations to the project through a web service (such as Papyri.info's editor). Project co-PIs and editors have identified that such a platform would require a text transcription editor (with version

[8] 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.

controlled commits to a server, such as a university server or a GitHub repository) and a multilayer editing tool (also with version controlled commits to a server) to enable annotations and editing in the multilayer model. Inline editing and annotation with such a complex datamodel is not feasible. (See also section 5. Continuation.)

2.6 Documentation

We have developed a robust system of documentation for our data models and formats, tools, workflows, and the use of the site. Documentation for the project was conducted under the aegis of both this grant and HD-51907. The "Quick Start" Documentation page on our main website directs researchers to the most commonly used resources. The project wiki is used especially for documenting project workflows and processes for editors and contributors. ReadMe files document and provide instructions for how to use the tools and data in our GitHub repositories. We also publish grant reports and white papers on the project website.

We provide here links to some of the most important documentation for researchers:

- Sample queries for the ANNIS search and visualization tool to aid researchers new to ANNIS
- A video of a tutorial for using ANNIS during our March 2015 workshop.
- Citation guidelines with instructions for how to cite the project and corpora in publications. (Also in Appendix E of this document.)
- Part of speech tagset and tagging guidelines which explain the part of speech sannotations
- Lemmatization guidelines which explain the lemma annotations.
- Transcription guidelines for our standards and encoding practices regarding transcription of Coptic texts. (Also in Appendix G of this document.)
- Basic annotation workflow documentation for project editors and annotators. (Also in Appendix C of this document.)
- Checklist for publishing and releasing our corpora online. (Also in Appendix D of this document.)
- Frequently Asked Questions page.

2.7 Articles and Conference Papers

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 ODH startup grant HD-51907. 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: Published

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)

Peer-reviewed Article under Review

Almas, Bridget, and Caroline T. Schroeder. “Applying the Canonical Text Services Model to the Coptic SCRIPTORIUM.”

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 *Apopthegmata 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. Activities

3.1 Meetings and Workshops

In-person meetings and workshops were fundamental activities over the course of the grant period. Although project participants communicated electronically using email, Skype, the project Wiki, Google Documents, and the GitHub system of issues and commits, we found that sustained, in-person meetings for engaged conversation, training, and planning were crucial to the project’s success. We regularly updated guidelines, workflows, and protocols based on the recommendations and experiences of the project’s editors. We also utilized in person meetings to train editors in our evolving technologies and to install requisite software on their computers.

Workshop and Training in San Francisco, May 2014

In May, 2014, project co-directors Amir Zeldes and Caroline T. Schroeder met in San Francisco with project editors Rebecca S. Krawiec and Christine Luckritz Marquis. The editors were trained in the tools and technologies for annotating digital Coptic text, and existing standards for transcribing digital Coptic were refined with the participants’ input.

Digital Coptic 2: Symposium and Workshop

Coptic SCRIPTORIUM hosted a second workshop and symposium on Digital Humanities and Coptic Studies March 12-13, 2015 at Georgetown University in Washington, DC. This event followed on the workshop in May 2013 at Humboldt University, Berlin. No registration fees were required to attend. Day 1 consisted of a public symposium on Digital Humanities and Coptic Studies with 10 presentations by scholars from North America, Germany, and Japan. Day 2 was a workshop on Coptic SCRIPTORIUM. Please see Appendix F or the full program. Project editors also met to discuss developing standards for transcribing and annotating digital Coptic texts and to train project participants in the evolving tools and technologies.

Summer 2015 Editors' Hackathon and Workshop

Project director Schroeder and two editors met in person for a hackathon and workshop on text annotation in California in July 2015, with other editors and co-director Zeldes joining virtually. We discussed changes in annotation practices, improvements in annotation tools, and refinements to our transcription, normalization, and tagging guidelines.

December 2015 Directors Meeting and Editorial Meeting

At Georgetown University in December 2015, project directors met with editors to review text data for publication and update tools and technologies based on the input from the editors. One editor attended in person while others attended virtually.

July 2016 Project Meetings and Workshop at the Congress for the International Association of Coptic Studies (Claremont, Ca.)

In addition to a two-day workshop with German and American partner projects under the financing of bilateral grant HG-229371, the directors met with senior editors on the project to review the year's work and plan for how to move forward after the end of the this grant period.

3.2 Digitizing, Editing, and Annotating Coptic Texts

One of the primary activities for the grant period was digitization of pilot corpora, which are enumerated in the grant products. The project directors recruited other experts in the field of Coptic Studies to digitize texts, as well as undergraduate and masters students without knowledge of Coptic to catalogue metadata and manuscript information and set up our GitHub repositories for the texts. All project participants are listed on our website and in Appendix A. We wish to acknowledge here in particular the work of senior editors Elizabeth Platte, Rebecca S. Krawiec, and Christine Luckritz Marquis; the work of editors Elizabeth Davidson and Dana Robinson; and students Lauren McDermott, Yanrui Liu, and David Sriboonreuang. Their work was supported by the grant and/or cost-sharing from the University of the Pacific.

Our complex data model required us to develop protocols that differed from other projects, such as Papyri.info, which produce digital editions in TEI-XML but not further linguistic annotation.

Over the course of the grant period, our digitization and annotation process evolved from applying individual tools to text transcriptions to using the NLP pipeline web service described above. Editorial processes were refined, concluding with framework of steps for the annotation process. NB: at all stages, annotators save their work and commit their documents and/or changes to the appropriate project GitHub repository. The workflow accounts for several possible paths toward digitization of a document. Editors may begin with an existing digital source from the web or provided to the project by a colleague. Additionally, it describes the steps for using either the Natural Language Processing service online or individual stand-alone tools. A regularly updated and more detailed version appears on our wiki page for annotation workflow. The basic workflow can be found in Appendix C.

When publishing a digitized, annotated text in our corpora, we follow the a publication checklist that ensures peer review of each document, proper versioning of the data, and publication on all platforms and in all data formats. We provide the checklist in Appendix D.

3.3 Data curation consulting & implementation

Models for data curation, especially the CTS URN system of identifiers, were developed in consultation with Bridget Almas, lead software developer for the Perseus Digital Library. These models had to take into account fragmentary manuscript witnesses that are now dispersed in across multiple modern repositories, and set of texts that do not all currently have canonical referencing systems. Almas and Schroeder have an article currently under review documenting the rationale for using the system for Coptic SCRIPTORIUM and the details of its implementation. The web application implementing the CTS URN service was developed by contractors Luke Hollis of Archimedes Digital and Dave Briccetti of Dave Briccetti Software in consultation with Almas. Almas's consulting was conducted under PW-51672-14, while Hollis' and Briccetti's work was conducted under both PW-51672-14 and HD-51907, as it involved data curation as well as technological development of the tool to enable data curation.

Project director Schroeder also visited the Perseus Digital Library on site during Fall 2014 to learn from staff members Alison Babeu and Lisa Cerrato about their procedures, especially related to data management. We appreciate their generosity in donating their time and expertise.

3.4 Preservation and Access

We adopted a multifaceted digital preservation strategy. First, releasing our corpora open access contributes to ensuring an afterlife for digitized texts in the form of reuse and redistribution by others. We chose GitHub as our means of initial archiving and distributing the files, since so many projects and individual researchers work there, and the system of forking and sharing repositories is an essential part of the platform.

We also piloted depositing a selection of our archived data files in the LAUDATIO repository for historical corpora at Humboldt University in Berlin (see for example: http://hdl.handle.net/11022/0000-0000-4680-0). Future work on the project may include adding the rest of our corpora.

As part of our grant application, Perseus Digital Library agreed to host our data files in order to assure their preservation and continued access. Perseus is able to publish TEI XML digital text annotated in the Capitains format. We have mapped out possibilities for converting our files into this requisite forma. Such conversion and integration will require additional labor and time (see 5.0).

4. Evaluation and Advisory Board

The Coptic SCRIPTORIUM Advisory Board consists of scholars with diverse areas of expertise. Alain Delattre (University of Brussels) is a papyrologist and editor at Papyri.info. Eitan Grossman (Hebrew University) is a linguist. Robin Imhof (University of the Pacific) is a humanities librarian. Project directors held a virtual Advisory Board meeting via Skype or email every four to six months during the grant period. Their feedback on linguistic matters, digital humanities project management, data curation, and digital editing of documentary papyri were particularly valuable for the success of the project. We also consulted with individual members as needed.

Consultations with other scholars in the field proved equally important. We wish to acknowledge Stephen Emmel (Münster), David Brakke (Ohio State University), and Janet Timbie (Catholic University of America).

After the March 2015 Digital Coptic 2 Symposium and Workshop, attendees participated in an open discussion about the direction of the project. Feedback from this group informed our work over the next year.

We also welcome researchers with GitHub accounts to submit issues (whether enhancements, modifications, or corrections) to GitHub. We have established a contact email on our website, as well, for email correspondence.

5. Audience 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 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 for their papers at our March 2014 symposium and a paper 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 ”

Our main web domain copticscriptorium.org has been visited by over 11,000 users since it was launched in 2014. Most site views come from the United States; other 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 visitors. By comparison, approximately 200 scholars attended the 2016 Congress for the International Association of Coptic Studies.

6. Continuation of the Project

Now that we have established data models, standards, and procedures for developing digital Coptic corpora, we are well poised to expand our text base. The first priority will be to digitize and annotate more literature in the Sahidic dialect, since we have the tools. Next would be to expand to other dialects, such as Bohairic and Nag Hammadi. Bohairic is important historically as a liturgical language and is still used today in contemporary Coptic Orthodox Church services. It is also the more conservative of the Coptic dialects, and markedly different from Sahidic in several points, making it an important source of information for comparative and historical linguistics. The Nag Hammadi library would form a more confined corpus, since it consists of 4th century codices discovered in 1945 near the Egyptian town of Nag Hammadi. The texts, in an unusual dialect, raise questions about scribal practices and language development, and their contents (non-canonical Christian texts and philosophical texts translated into Coptic) are important for historical research and Biblical Studies.

Expansion of the corpora will also be aided adding an online, web-based transcription editor and multilayer annotation editor, which will allow the project to incorporate digitized text and transcriptions from a wider group of scholars beyond the core project staff. Development of these tools is being conducted under the KELLIA project (HG-229371). Training sessions to introduce additional researchers to the transcription tools would also be desirable in order to expand the corpora.

Further developments in annotation models for more linked open data and data sharing are also desirable. As part of the KELLIA collaboration (HG-229371), the project this summer participated in the development and release of a prototype of an online dictionary linked to our corpora by lemma. The dictionary requires additional development but is a first step. Linking other data in our text corpora (such as geographical entities or persons) to existing online resources such as Pleiades and Trismegistos remains an important desideratum. Additional funding will be required to build tools for automatic entity annotation (especially recognizing mentions of people, places and concepts), as well as for the time and labor to annotate and publish the new annotations.

Digital preservation of corpora will also continue. Converting our data for integration into Perseus will require time and labor to build the tools and to manage the integration. We have mapped out possible paths forward in a Perseus collaboration (see 3.4); development and implementation remain a future consideration.

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:

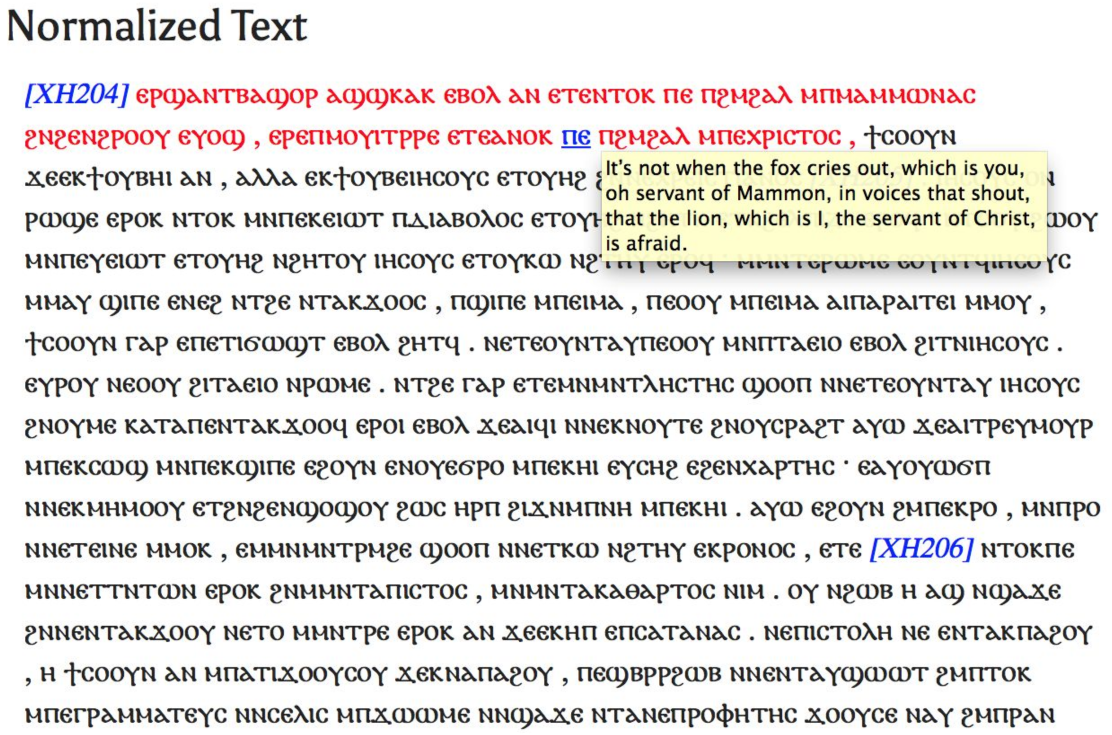
Alain Delattre, Assistant Professor, Department of Languages and Literatures, Université libre de Bruxelles; Papryi.info.

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: Visualizations

Normalized visualization



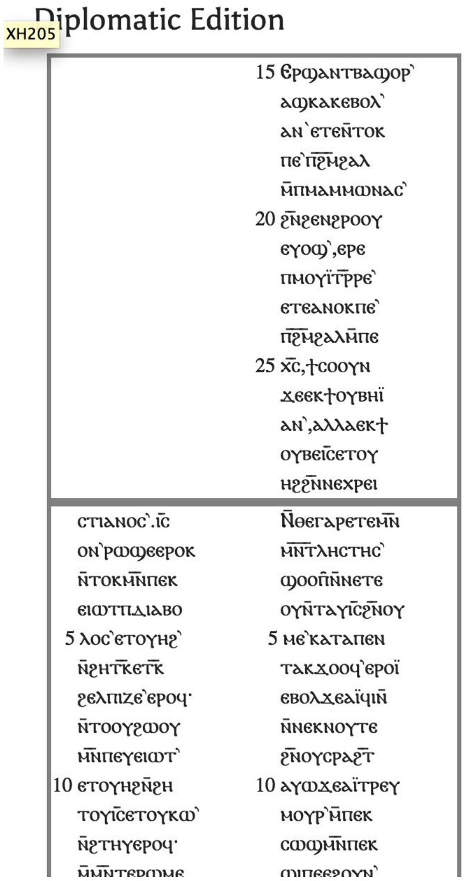
Normalized Text

[XH204] ⲉⲣϣⲁⲛⲧⲃⲁϣⲟⲣ ⲁϣϣⲕⲁⲕ ⲉⲃⲟⲗ ⲁⲛ ⲉⲧⲉⲛⲧⲟⲕ ⲡⲉ ⲡϩⲙϩⲁⲗ ⲙⲡⲙⲁⲙⲙⲱⲛⲁⲥ
ϩⲛϩⲉⲛϩⲣⲟⲟⲩ ⲉⲩⲟϣ , ⲉⲣⲉⲡⲙⲟⲩⲓⲧⲣⲣⲉ ⲉⲧⲉⲁⲛⲟⲕ ⲡⲉ ⲡϩⲙϩⲁⲗ ⲙⲡⲉⲭⲣⲓⲥⲧⲟⲥ , ϯⲥⲟⲟⲩⲛ
ϫⲉⲉⲕϯⲟⲩⲃⲏⲓ ⲁⲛ , ⲁⲗⲗⲁ ⲉⲕϯⲟⲩⲃⲉⲓⲏⲥⲟⲩⲥ ⲉⲧⲟⲩⲏϩ
ⲣⲱϣⲉ ⲉⲣⲟⲕ ⲛⲧⲟⲕ ⲙⲛⲡⲉⲕⲉⲓⲱⲧ ⲡⲇⲓⲁⲃⲟⲗⲟⲥ ⲉⲧⲟⲩ
ⲙⲛⲡⲉⲩⲉⲓⲱⲧ ⲉⲧⲟⲩⲏϩ ⲛϩⲏⲧⲟⲩ ⲓⲏⲥⲟⲩⲥ ⲉⲧⲟⲩⲕⲱ ⲛϩⲧⲏⲩ ⲉⲣⲟϥ · ⲙⲛⲛⲧⲉⲣⲱⲙⲉ ⲉⲟⲩⲛⲧϥⲓⲏⲥⲟⲩⲥ
ⲙⲙⲁⲩ ϣⲓⲡⲉ ⲉⲛⲉϩ ⲛⲧϩⲉ ⲛⲧⲁⲕϫⲟⲟⲥ , ⲡϣⲓⲡⲉ ⲙⲡⲉⲓⲙⲁ , ⲡⲉⲟⲟⲩ ⲙⲡⲉⲓⲙⲁ ⲁⲓⲡⲁⲣⲁⲓⲧⲉⲓ ⲙⲙⲟⲩ ,
ϯⲥⲟⲟⲩⲛ ⲅⲁⲣ ⲉⲡⲉⲧⲓϭⲱϣⲧ ⲉⲃⲟⲗ ϩⲏⲧϥ . ⲛⲉⲧⲉⲟⲩⲛⲧⲁⲩⲡⲉⲟⲟⲩ ⲙⲛⲡⲧⲁⲉⲓⲟ ⲉⲃⲟⲗ ϩⲓⲧⲛⲓⲏⲥⲟⲩⲥ .
ⲉⲩⲣⲟⲩ ⲛⲉⲟⲟⲩ ϩⲓⲧⲁⲉⲓⲟ ⲛⲣⲱⲙⲉ . ⲛⲧϩⲉ ⲅⲁⲣ ⲉⲧⲉⲙⲛⲙⲛⲧⲗⲏⲥⲧⲏⲥ ϣⲟⲟⲡ ⲛⲛⲉⲧⲉⲟⲩⲛⲧⲁⲩ ⲓⲏⲥⲟⲩⲥ
ϩⲛⲟⲩⲙⲉ ⲕⲁⲧⲁⲡⲉⲛⲧⲁⲕϫⲟⲟϥ ⲉⲣⲟⲓ ⲉⲃⲟⲗ ϫⲉⲁⲓϥⲓ ⲛⲛⲉⲕⲛⲟⲩⲧⲉ ϩⲛⲟⲩⲥⲣⲁϩⲧ ⲁⲩⲱ ϫⲉⲁⲓⲧⲣⲉⲩⲙⲟⲩⲣ
ⲙⲡⲉⲕⲥⲱϣ ⲙⲛⲡⲉⲕϣⲓⲡⲉ ⲉϩⲟⲩⲛ ⲉⲛⲟⲩⲉϭⲣⲟ ⲙⲡⲉⲕⲏⲓ ⲉⲩⲥⲏϩ ⲉϩⲉⲛⲭⲁⲣⲧⲏⲥ · ⲉⲁⲩⲟⲩⲱϭⲡ
ⲛⲛⲉⲕⲙⲏⲙⲟⲟⲩ ⲉⲧϩⲛϩⲉⲛϣⲟϣⲟⲩ ϩⲱⲥ ⲏⲣⲡ ϩⲓϫⲛⲙⲡⲛⲏ ⲙⲡⲉⲕⲏⲓ . ⲁⲩⲱ ⲉϩⲟⲩⲛ ϩⲙⲡⲉⲕⲣⲟ , ⲙⲛⲡⲣⲟ
ⲛⲛⲉⲧⲉⲓⲛⲉ ⲙⲙⲟⲕ , ⲉⲙⲙⲛⲙⲛⲧⲣⲙϩⲉ ϣⲟⲟⲡ ⲛⲛⲉⲧⲕⲱ ⲛϩⲧⲏⲩ ⲉⲕⲣⲟⲛⲟⲥ , ⲉⲧⲉ [XH206] ⲛⲧⲟⲕⲡⲉ
ⲙⲛⲛⲉⲧⲧⲛⲧⲱⲛ ⲉⲣⲟⲕ ϩⲛⲙⲙⲛⲧⲁⲡⲓⲥⲧⲟⲥ , ⲙⲛⲙⲛⲧⲁⲕⲁⲑⲁⲣⲧⲟⲥ ⲛⲓⲙ . ⲟⲩ ⲛϩⲱⲃ ⲏ ⲁϣ ⲛϣⲁϫⲉ
ϩⲛⲛⲉⲛⲧⲁⲕϫⲟⲟⲩ ⲛⲉⲧⲟ ⲙⲙⲛⲧⲣⲉ ⲉⲣⲟⲕ ⲁⲛ ϫⲉⲉⲕⲏⲡ ⲉⲡⲥⲁⲧⲁⲛⲁⲥ . ⲛⲉⲡⲓⲥⲧⲟⲗⲏ ⲛⲉ ⲉⲛⲧⲁⲕⲡⲁϩⲟⲩ
, ⲏ ϯⲥⲟⲟⲩⲛ ⲁⲛ ⲙⲡⲁⲧⲓϫⲟⲟⲩⲥⲟⲩ ϫⲉⲕⲛⲁⲡⲁϩⲟⲩ , ⲡⲉϣⲃⲣⲣϩⲱⲃ ⲛⲛⲉⲛⲧⲁⲩϣⲱⲱⲧ ϩⲙⲡⲧⲟⲕ
ⲙⲡⲉⲅⲣⲁⲙⲙⲁⲧⲉⲩⲥ ⲛⲛⲥⲉⲗⲓⲥ ⲙⲡϫⲱⲱⲙⲉ ⲛⲛϣⲁϫⲉ ⲛⲧⲁⲛⲉⲡⲣⲟⲫⲏⲧⲏⲥ ϫⲟⲟⲩⲥⲉ ⲛⲁⲩ ϩⲙⲡⲣⲁⲛ

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.

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



XH205

Diplomatic Edition

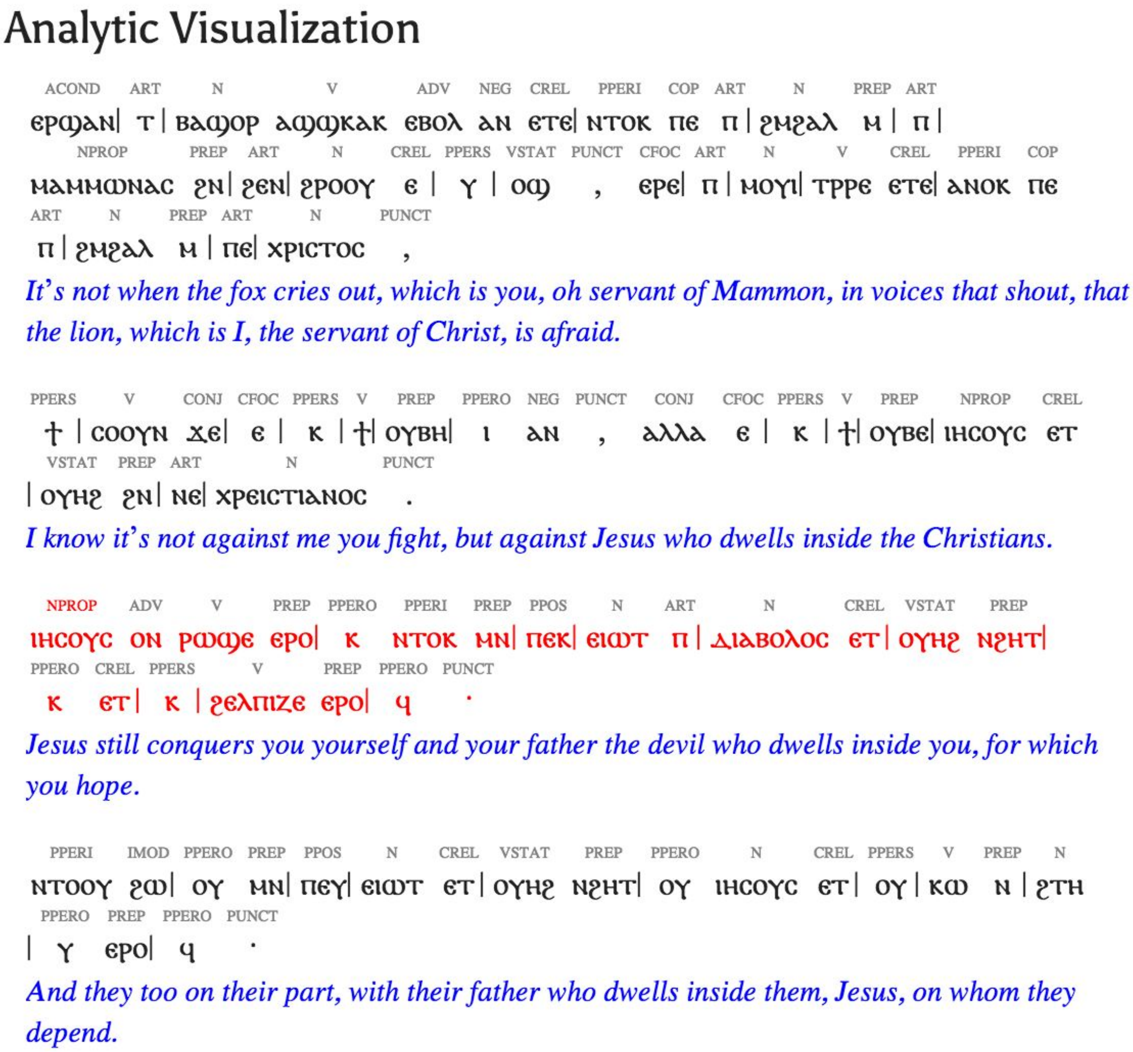
Ⲉⲣϣⲁⲛⲧⲃⲁϣⲟⲣ̀
ⲁϥⲕⲁⲕⲉⲃⲟⲗ̀
ⲁⲛ̀ⲉⲧⲉⲛ̄ⲧⲟⲕ
ⲡⲉ̀ⲡ̄ϩ̄ⲙ̄ϩⲁⲗ
ⲙ̄ⲡⲙⲁⲙⲙⲱⲛⲁⲥ̀
ϩ̄ⲛϩⲉⲛϩⲣⲟⲟⲩ
ⲉⲩⲟϣ̀,ⲉⲣⲉ
ⲡⲙⲟⲩⲓ̈ⲧ̄ⲣ̄ⲣⲉ̀
ⲉⲧⲉⲁⲛⲟⲕⲡⲉ̀
ⲡ̄ϩ̄ⲙ̄ϩⲁⲗⲙ̄ⲡⲉ
ⲭ̄ⲥ̄,ϯⲥⲟⲟⲩⲛ
ϫⲉⲉⲕϯⲟⲩⲃⲏⲓ̈
ⲁⲛ̀,ⲁⲗⲗⲁⲉⲕϯ
ⲟⲩⲃⲉⲓ̄ⲥ̄ⲉⲧⲟⲩ
ⲏϩϩ̄ⲛ̄ⲛⲉⲭⲣⲉⲓ

ⲥⲧⲓⲁⲛⲟⲥ̀.ⲓ̄ⲥ̄
ⲟⲛ̀ⲣⲱϣⲉⲉⲣⲟⲕ
ⲛ̄ⲧⲟⲕⲙ̄ⲛ̄ⲡⲉⲕ
ⲉⲓⲱⲧⲡⲇⲓⲁⲃⲟ
ⲗⲟⲥ̀ⲉⲧⲟⲩⲏϩ̀
ⲛ̄ϩⲏⲧ̄ⲕⲉⲧ̄ⲕ
ϩⲉⲗⲡⲓⲍⲉ̀ⲉⲣⲟϥ·
ⲛ̄ⲧⲟⲟⲩϩⲱⲟⲩ
ⲙ̄ⲛ̄ⲡⲉⲩⲉⲓⲱⲧ̀
ⲉⲧⲟⲩⲏϩⲛ̄ϩⲏ
ⲧⲟⲩⲓ̄ⲥ̄ⲉⲧⲟⲩⲕⲱ̀
ⲛ̄ϩⲧⲏⲩⲉⲣⲟϥ·
ⲙ̄ⲙ̄ⲛ̄ⲧⲉⲣⲱⲙⲉ

Ⲛ̄ⲑⲉⲉⲅⲁⲣⲉⲧⲉⲙ̄ⲛ̄
ⲙ̄ⲛ̄ⲧ̄ⲗⲏⲥⲧⲏⲥ̀
ϣⲟⲟⲡ̂ⲛ̄ⲛⲉⲧⲉ
ⲟⲩⲛ̄ⲧⲁⲩⲓ̄ⲥ̄ϩ̄ⲛ̄ⲟⲩ
ⲙⲉ̀ⲕⲁⲧⲁⲡⲉⲛ
ⲧⲁⲕϫⲟⲟϥ̀ⲉⲣⲟⲓ̈
ⲉⲃⲟⲗϫⲉⲁⲓ̈ϥⲓⲛ̄
ⲛ̄ⲛⲉⲕⲛⲟⲩⲧⲉ
ϩ̄ⲛⲟⲩⲥⲣⲁϩ̄ⲧ̄
ⲁⲩⲱϫⲉⲁⲓ̈ⲧⲣⲉⲩ
ⲙⲟⲩⲣ̀ⲙ̄ⲡⲉⲕ
ⲥⲱϣⲙ̄ⲛ̄ⲡⲉⲕ
ϣⲓⲡⲉⲉϩⲟⲩⲛ̀

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

ⲉⲣϣⲁⲛ| ⲧ | ⲃⲁϣⲟⲣ ⲁϣϣⲕⲁⲕ ⲉⲃⲟⲗ ⲁⲛ ⲉⲧⲉ| ⲛⲧⲟⲕ ⲡⲉ ⲡ | ϩⲙϩⲁⲗ ⲙ | ⲡ |

NPROP PREP ART N CREL PPERS VSTAT PUNCT CFOC ART N V CREL PPERI COP

ⲙⲁⲙⲙⲱⲛⲁⲥ ϩⲛ| ϩⲉⲛ| ϩⲣⲟⲟⲩ ⲉ | ⲩ | ⲟϣ , ⲉⲣⲉ| ⲡ | ⲙⲟⲩⲓ| ⲧⲣⲣⲉ ⲉⲧⲉ| ⲁⲛⲟⲕ ⲡⲉ

ART N PREP ART N PUNCT

ⲡ | ϩⲙϩⲁⲗ ⲙ | ⲡⲉ| ⲭⲣⲓⲥⲧⲟⲥ ,

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

ϯ | ⲥⲟⲟⲩⲛ ϫⲉ| ⲉ | ⲕ | ϯ| ⲟⲩⲃⲏ| ⲓ ⲁⲛ , ⲁⲗⲗⲁ ⲉ | ⲕ | ϯ| ⲟⲩⲃⲉ| ⲓⲏⲥⲟⲩⲥ ⲉⲧ

VSTAT PREP ART N PUNCT

| ⲟⲩⲏϩ ϩⲛ| ⲛⲉ| ⲭⲣⲉⲓⲥⲧⲓⲁⲛⲟⲥ .

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

ⲓⲏⲥⲟⲩⲥ ⲟⲛ ⲣⲱϣⲉ ⲉⲣⲟ| ⲕ ⲛⲧⲟⲕ ⲙⲛ| ⲡⲉⲕ| ⲉⲓⲱⲧ ⲡ | ⲇⲓⲁⲃⲟⲗⲟⲥ ⲉⲧ | ⲟⲩⲏϩ ⲛϩⲏⲧ|

PPERO CREL PPERS V PREP PPERO PUNCT

ⲕ ⲉⲧ| ⲕ | ϩⲉⲗⲡⲓⲍⲉ ⲉⲣⲟ| ϥ ·

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

ⲛⲧⲟⲟⲩ ϩⲱ| ⲟⲩ ⲙⲛ| ⲡⲉⲩ| ⲉⲓⲱⲧ ⲉⲧ | ⲟⲩⲏϩ ⲛϩⲏⲧ| ⲟⲩ ⲓⲏⲥⲟⲩⲥ ⲉⲧ| ⲟⲩ | ⲕⲱ ⲛ | ϩⲧⲏ

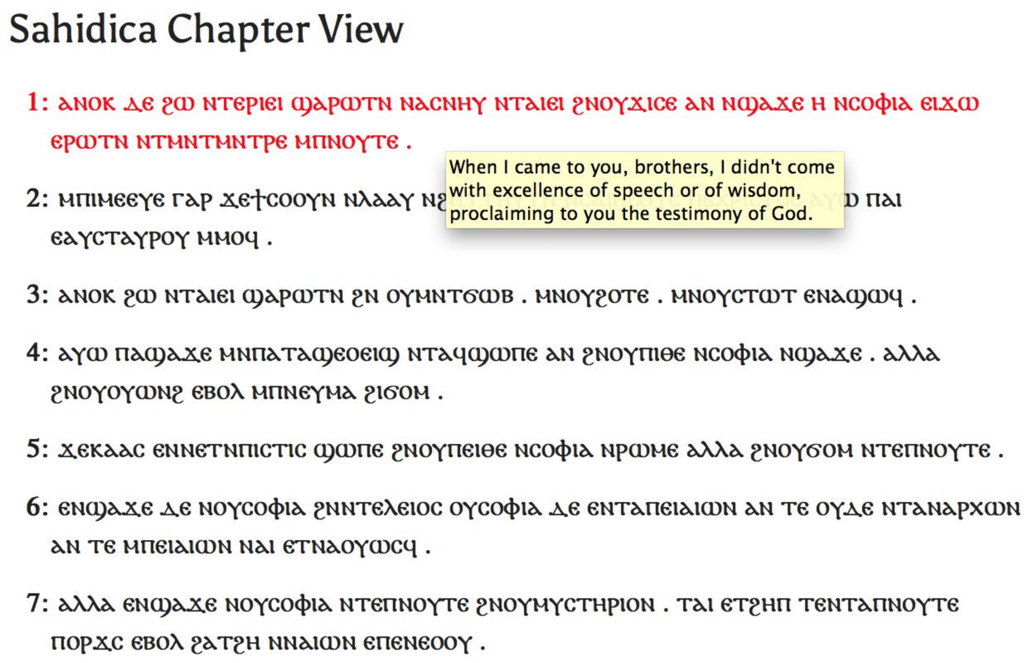
PPERO PREP PPERO PUNCT

| ⲩ ⲉⲣⲟ| ϥ ·

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: ⲁⲛⲟⲕ ⲇⲉ ϩⲱ ⲛⲧⲉⲣⲓⲉⲓ ϣⲁⲣⲱⲧⲛ ⲛⲁⲥⲛⲏⲩ ⲛⲧⲁⲉⲓ ϩⲛⲟⲩϫⲓⲥⲉ ⲁⲛ ⲛϣⲁϫⲉ ⲏ ⲛⲥⲟⲫⲓⲁ ⲉⲓϫⲱ
ⲉⲣⲱⲧⲛ ⲛⲧⲙⲛⲧⲙⲛⲧⲣⲉ ⲙⲡⲛⲟⲩⲧⲉ .

When I came to you, brothers, I didn't come
with excellence of speech or of wisdom,
proclaiming to you the testimony of God.

2: ⲙⲡⲓⲙⲉⲉⲩⲉ ⲅⲁⲣ ϫⲉϯⲥⲟⲟⲩⲛ ⲛⲗⲁⲁⲩ ⲛϩ ⲱ ⲡⲁⲓ
ⲉⲁⲩⲥⲧⲁⲩⲣⲟⲩ ⲙⲙⲟϥ .

3: ⲁⲛⲟⲕ ϩⲱ ⲛⲧⲁⲉⲓ ϣⲁⲣⲱⲧⲛ ϩⲛ ⲟⲩⲙⲛⲧϭⲱⲃ . ⲙⲛⲟⲩϩⲟⲧⲉ . ⲙⲛⲟⲩⲥⲧⲱⲧ ⲉⲛⲁϣⲱϥ .

4: ⲁⲩⲱ ⲡⲁϣⲁϫⲉ ⲙⲛⲡⲁⲧⲁϣⲉⲟⲉⲓϣ ⲛⲧⲁϥϣⲱⲡⲉ ⲁⲛ ϩⲛⲟⲩⲡⲓⲑⲉ ⲛⲥⲟⲫⲓⲁ ⲛϣⲁϫⲉ . ⲁⲗⲗⲁ
ϩⲛⲟⲩⲟⲩⲱⲛϩ ⲉⲃⲟⲗ ⲙⲡⲛⲉⲩⲙⲁ ϩⲓϭⲟⲙ .

5: ϫⲉⲕⲁⲁⲥ ⲉⲛⲛⲉⲧⲛⲡⲓⲥⲧⲓⲥ ϣⲱⲡⲉ ϩⲛⲟⲩⲡⲉⲓⲑⲉ ⲛⲥⲟⲫⲓⲁ ⲛⲣⲱⲙⲉ ⲁⲗⲗⲁ ϩⲛⲟⲩϭⲟⲙ ⲛⲧⲉⲡⲛⲟⲩⲧⲉ .

6: ⲉⲛϣⲁϫⲉ ⲇⲉ ⲛⲟⲩⲥⲟⲫⲓⲁ ϩⲛⲛⲧⲉⲗⲉⲓⲟⲥ ⲟⲩⲥⲟⲫⲓⲁ ⲇⲉ ⲉⲛⲧⲁⲡⲉⲓⲁⲓⲱⲛ ⲁⲛ ⲧⲉ ⲟⲩⲇⲉ ⲛⲧⲁⲛⲁⲣⲭⲱⲛ
ⲁⲛ ⲧⲉ ⲙⲡⲉⲓⲁⲓⲱⲛ ⲛⲁⲓ ⲉⲧⲛⲁⲟⲩⲱⲥϥ .

7: ⲁⲗⲗⲁ ⲉⲛϣⲁϫⲉ ⲛⲟⲩⲥⲟⲫⲓⲁ ⲛⲧⲉⲡⲛⲟⲩⲧⲉ ϩⲛⲟⲩⲙⲩⲥⲧⲏⲣⲓⲟⲛ . ⲧⲁⲓ ⲉⲧϩⲏⲡ ⲧⲉⲛⲧⲁⲡⲛⲟⲩⲧⲉ
ⲡⲟⲣϫⲥ ⲉⲃⲟⲗ ϩⲁⲧϩⲏ ⲛⲛⲁⲓⲱⲛ ⲉⲡⲉⲛⲉⲟⲟⲩ .

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.

Appendix C: Digitization and Annotation Editorial Workflow

The following steps describe the basic editorial workflow for digitizing and annotating texts for Coptic SCRIPTORIUM

- Begin with one of two sources for digitized text:
    1. Download or scrape digitized Coptic text from an existing source (source must be in the public domain or accompanied by the appropriate use licenses, or prior permission must have been secured)
        - → If necessary, convert the text into the Unicode Coptic character set using project converters
        - → Ensure digital text is segmented into bound groups according to project Transcription Guidelines; edit as necessary
    2. Transcribe text from another source (manuscript, manuscript facsimile, other non-digital text-bearing object) according to project Transcription Guidelines; source must be in the public domain or prior permission must have been secured.
- At this point, you may follow one of two paths:
    1. Use the Natural Language Processing (NLP) Service online
        - → Copy the digitized text into NLP Service. Select “Just piped and dashed morphemes” and run the Service. (Pipes indicate segmentation into words; dashes indicate smaller morphs.)
        - → Cut and paste the SGML output into a text file and proofread the automatic tokenization, editing as necessary.
        - → Copy the proofread SGML back into the NLP Service input window. Under “Tokenize” select “From pipes in input.” Select all annotations desired (usually all except “parse”), and run the Service.
        - → Copy and convert the SGML output into a multilayer spreadsheet format using the project’s converter.
        - → Manually proofread and edit data in existing layers.
        - → Add any missing layers manually or using other existing tools.
        - → Check layer names to ensure they conform to project standards for the data model (published on the wiki)..
    2. Process using our NLP tools individually on your local machine
        - → Run the stand alone tokenizer tool on the text file.
        - → Copy and convert the tokenized text (SGML output) into a multilayer spreadsheet format using the project’s converter.
        - → Proofread tokenization of the bound groups, editing as necessary.
        - → Run the normalizer tool on the layer of data that corresponds to Coptic words to create a normalization annotation layer.
        - → Create and or proofread the morph layer (for compounds and Coptic units below the word level)

    - → Run the part of speech tagger (including the lemmatization parameter) on the text in the normalized data layer.
    - → Run the language of origin tagger.
    - → Manually proofread and edit data in existing layers.
    - → Add any missing layers manually or using other existing tools.
    - → Check layer names to ensure they conform to project standards for the data model (published on the wiki).
- Add translation and metadata to the file. Confirm that metadata conforms to our data model (published on the wiki).
- Use Google Refine to clean and proofread text data one more time (especially for large files, or if editing multiple files at a stretch).
- Validate the file using the project's data validation spreadsheet plugin

At the conclusion of these steps, submit the file to a senior editor for peer review, typically by creating a new issue or pull request in GitHub.

Appendix D: Checklist for the Publication and Release of Corpora

□ New and revised docs are reviewed by a Senior editor. (Files returned to original editor for modifications if necessary. If no modifications (or few modifications) are necessary, Senior editor continues with the remaining checklist.

□ Senior editor adds/corrects metadata on new documents, paying particular attention to the names of the editors and the version number and date of each document. Senior editor's name is added to the list of annotators in the document metadata.

□ Senior editor checks the existing issues lists on GitHub for the corpus/corpora containing the new or newly revised document(s) to be released. Any required edits or corrections documented in GitHub issues for the relevant corpora should be made before the release.

□ Corpus metadata is added or revised by the Senior editor, paying particular attention to version number, version date, and names of annotators.

□ File(s) validated a final time using the project's data validation plugin.

□ Convert the data files into relANNIS files and install as a pre-publication corpus on the ANNIS server in order to check visualizations and/or find possible bugs in the corpus. Edit working files if necessary. Edit version number and date in metadata if necessary.

□ Convert the data files into relANNIS, PAULA XML, and TEI XML.

□ Publish the three formats to the GitHub corpora repository, and create a new release in GitHub with appropriate documentation

□ Install the relANNIS files on the public ANNIS server.

□ Refresh the CTS URN web service deployed at data.copticscriptorium.org with the new or revised corpus data in ANNIS.

□ Announce release.

Appendix E: Citation Guidelines

Citation practices for complex corpora with multiple disciplinary uses will necessarily be diverse. The following guidelines provide principles and models for citing our corpora. We anticipate modifications and additions as the project grows.

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].

E.g.,

Coptic SCRIPTORIUM, apophthegmata.patrum 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].

E.g.,

Coptic SCRIPTORIUM. apophthegmata.patrum 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].

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].

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.

Appendix F: 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 G: Transcription Guidelines

The following document includes the current version of the Transcription Guidelines for for creating a digital corpus of Coptic texts. This document is occasionally updated; a link to the most recent version can be found on GitHub, and there is a link to it on our website under documentation.

Coptic SCRIPTORIUM Diplomatic Transcription Guidelines

Version: *1.2_2016.8.26*

Caroline T. Schroeder[1]& Amir Zeldes[2]

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:
>
> ⲡⲉⲙⲙⲟⲛ⳿ⲧⲉ

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. ϩ︤ⲛ︥) whereas others in the case of the same word might only provide a stroke over a single letter (e.g., ϩⲛ̄). 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 ⲙ̄) the supralinear stroke above only one letter: type the letter followed by Unicode U+0304 (; on keyboard)
>
> ‾ (as in ⲙ︤ⲛ︥) 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 ⲙ︤ⲛ︦ⲧ︥) 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 Mmac using the Coptic-English keyboard), i.e. m<n:t>)

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

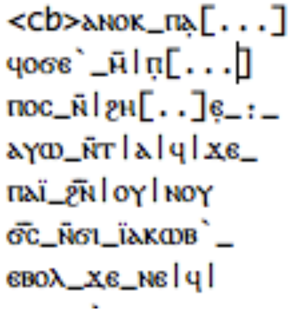
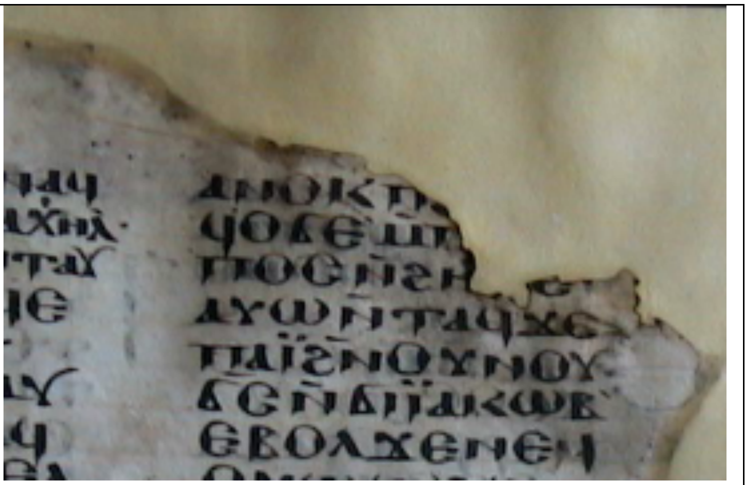
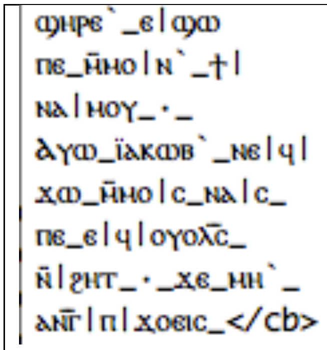
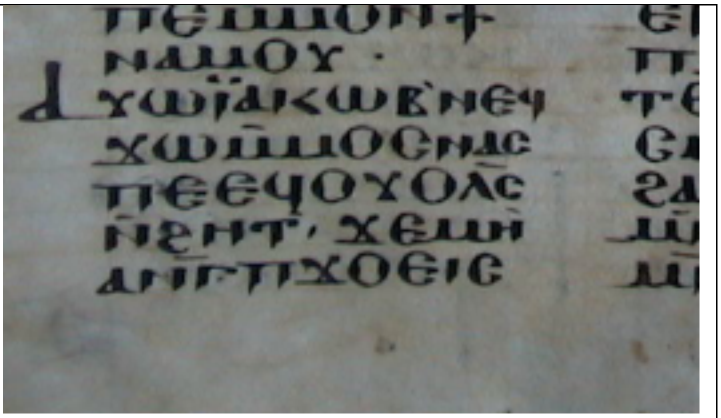
 <cb>ⲁⲛⲟⲕ_ⲡⲁ̣[...] ϥⲟϭⲉ`_ⲙ̄|ⲡ̣[...] ⲡⲟⲥ_ⲛ̄|ϩⲏ[..]ⲉ̣_:_ ⲁⲩⲱ_ⲛ̄ⲧ|ⲁ|ϥ|ϫⲉ_ ⲡⲁï_ϩ̄ⲛ|ⲟⲩ|ⲛⲟⲩ ϭ̄ⲥ_ⲛ̄ϭⲓ_ïⲁⲕⲱⲃ`_ ⲉⲃⲟⲗ_ϫⲉ_ⲛⲉ|ϥ|	 MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples, Biblioteca Vittorio Emanuele III

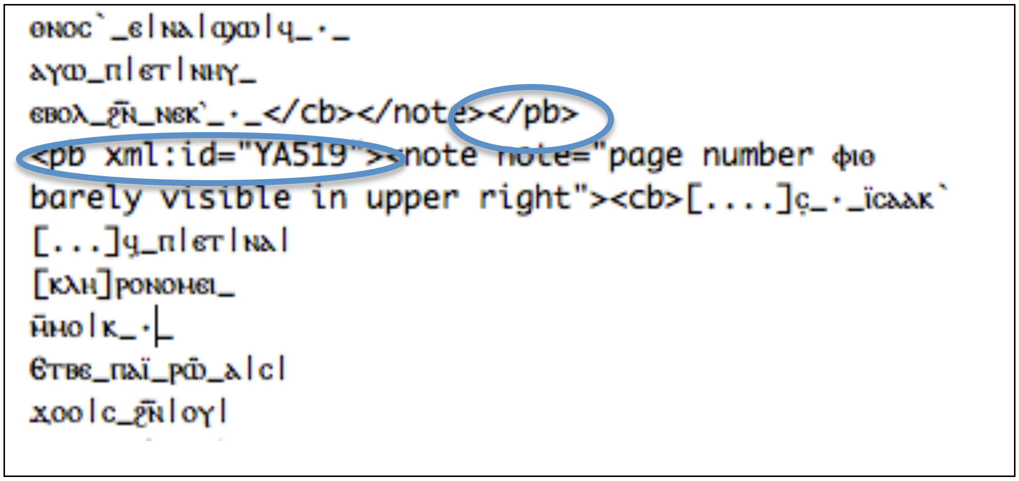
Fig. 2: Closing column break tag, corresponding to the end of a manuscript text column

 ϣⲏⲣⲉ`_ⲉ|ϣⲱ ⲡⲉ_ⲙ̄ⲙⲟ|ⲛ`_ϯ| ⲛⲁ|ⲙⲟⲩ_·_ ⲁⲩⲱ_ⲓ̈ⲁⲕⲱⲃ`_ⲛⲉ|ϥ| ϫⲱ_ⲙ̄ⲙⲟ|ⲥ_ⲛⲁ|ⲥ_ ⲡⲉ_ⲉ|ϥ|ⲟⲩⲟⲗ̅ⲥ̅_ ⲛ̄|ϩⲏⲧ_·_ϫⲉ_ⲙⲏ`_ ⲁⲛ̄ⲅ̄|ⲡ|ϫⲟⲉⲓⲥ_</cb>	 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.)



```
ⲑⲛⲟⲥ`_ⲉ|ⲛⲁ|ϣⲱ|ϥ_·_
ⲁⲩⲱ_ⲡ|ⲉⲧ|ⲛⲏⲩ_
ⲉⲃⲟⲗ_ϩⲛ̄_ⲛⲉⲕ`_·_</cb></note></pb>
<pb xml:id="YA519"><note note="page number ⲫⲓⲑ
barely visible in upper right"><cb>[....]ⲥ̣_·_ⲓ̈ⲥⲁⲁⲕ`
[...]ϥ_ⲡ|ⲉⲧ|ⲛⲁ|
[ⲕⲗⲏ]ⲣⲟⲛⲟⲙⲉⲓ_
ⲙ̄ⲙⲟ|ⲕ_·⌊
ⲉⲧⲃⲉ_ⲡⲁⲓ̈_ⲣⲱ̄_ⲁ|ⲥ|
ϫⲟⲟ|ⲥ_ϩⲛ̄|ⲟⲩ|
```

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 ⲁⲧ, ⲙⲛⲧ, or ⲣⲉϥ, or compound words, such as complex numbers (e.g. -teens) and verbs formed with ⲣ. 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:*
> ⲥⲱⲧⲙ
> ⲛⲟⲃⲉ
> ϩⲏⲧ
>
> *Examples of individual words comprised of multiple morphemes:*
> ⲙⲛⲧⲁⲧⲥⲱⲧⲙ
> ⲣⲉϥⲣⲛⲟⲃⲉ
>
> *Examples of bound groups comprised of words with multiple morphemes:*
> ⲧⲙⲛⲧⲁⲧⲥⲱⲧⲙ
> ⲙⲡⲣⲉϥⲣⲛⲟⲃⲉ
> ⲡⲣⲙⲛϩⲏⲧ
> ϫⲉⲛⲧⲁⲩϫⲓⲧⲥ
>
> Note: if a project wishes to annotate on the morpheme level (i.e. internal analysis of units like ⲙⲛⲧ) 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) ⲉⲧⲉⲓ̈ⲥⲙⲁⲏⲗ_ (word ends at end of line)
(2) ⲡⲉ_ⲛϥⲛⲁⲕⲗⲏ (two words, in which the second bound group flows into line 3)
(3) ⲣⲟⲛⲟⲙⲉⲓ_ⲙ (the bound group continues from line 2, is followed by an underscore)
(4) ⲙⲟⲕ_ⲁⲛ_·_ (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 demark 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) ⲛ̄<hi rend="1_space_right">ⲃⲟⲏⲑⲟⲥ</hi> will visualize one space to the right of he n t
(3) ϫⲱ<hi rend="2_space_right">ⲓ̈</hi>_<hi rend="1_space_right">·_</hi>_ⲉⲧⲃⲉ_ will visualize two spaces to the right of ï 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) ϩⲓⲧⲙ|ⲡ|ⲛⲟⲩⲧⲉ_	(preposition|article|noun)
(2) ⲉⲧⲉ|ⲓ̈ⲥⲙⲁⲏⲗ_	(converter|noun)
(3) ⲡⲉ̀_ⲛ|ϥ|ⲛⲁ|ⲕⲗⲏ	(word_auxiliary|subject pronoun|future marker|verb (verb continues to line 4)
(4) ⲣⲟⲛⲟⲙⲉⲓ̀_ⲙ̄	

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	ⲙⲛⲧⲁⲧⲥⲱⲧⲙ		
morpheme	ⲙⲛⲧ	ⲁⲧ	ⲥⲱⲧⲙ

word	ⲣⲉϥⲣⲛⲟⲃⲉ		
morpheme	ⲣⲉϥ	ⲣ	ⲛⲟⲃⲉ

Examples of bound groups comprised of words with multiple morphemes:

bound group	ⲧⲙⲛⲧⲁⲧⲥⲱⲧⲙ			
word	ⲧ	ⲙⲛⲧⲁⲧⲥⲱⲧⲙ		
morpheme	ⲧ	ⲙⲛⲧ	ⲁⲧ	ⲥⲱⲧⲙ

bound group	ⲙⲡⲣⲉϥⲣⲛⲟⲃⲉ				
word	ⲙ	ⲡ	ⲣⲉϥⲣⲛⲟⲃⲉ		
morpheme	ⲙ	ⲡ	ⲣⲉϥ	ⲣ	ⲛⲟⲃⲉ

bound group	ⲡⲣⲙⲛϩⲏⲧ			
word	ⲡ	ⲣⲙⲛϩⲏⲧ		
morpheme	ⲡ	ⲣⲙ	ⲛ	ϩⲏⲧ

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	ⲣϩⲛⲁϥ		
word	ⲣ	ϩⲛⲁ	ϥ
no tokenization & annotation on the morpheme level below the word level			

bound group	ⲙⲡⲉⲧⲛⲣⲡⲙⲉⲉⲩⲉ				
word	ⲙⲡⲉ	ⲧⲛ	ⲣ	ⲡ	ⲙⲉⲉⲩⲉ
no tokenization & annotation on the morpheme level below the word level					

> *(where* ⲣⲡⲙⲉⲉⲩⲉ *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, ⲣϩⲟⲧⲉ is tagged as one V, ⲙⲛⲧⲁⲧⲥⲱⲧⲙ as one N, etc.]

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

> ⲧ|ⲙⲛⲧ-ⲁⲧ-ⲥⲱⲧⲙ
>
> ⲙ|ⲡ|ⲣⲉϥ-ⲣ-ⲛⲟⲃⲉ

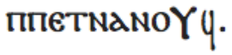
If you plan to use Coptic SCRIPTORIUM's NLP web service, you may transcribe the Coptic in bound groups with no pipes or morphemes. The NLP web service's tokenizer can provide as output segmentation with pipes between words and dashes between morphs. Likewise, Coptic SCRIPTORIUM's stand-alone tokenizer 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

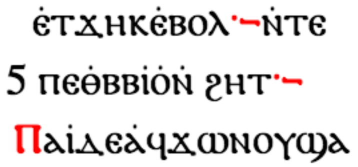
Coptic SCRIPTORIUM uses Leiden and Leiden+ conventions for transcribing manuscripts. The encoding follows the EpiDoc guidelines. Not all Leiden documentation is currently XML encoded as Leiden+, however.

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	*Diplomatic Visualization*
ϩⲓⲧⲙ̄ⲡⲛ<hi rend="superscript"><note note="o is directly above the ⲩ">o</note></hi>ⲩ	ϩⲓⲧⲙ̄ⲡⲛⲩ̊ (ANNIS) or ϩⲓⲧⲙ̄ⲡⲛ\o/ⲩ (EpiDoc XSLT)
ⲠⲠⲈⲦⲚⲀⲚⲞ<hi rend="large">Ⲩ</hi><hi rend= "long thin">ϥ</hi>_._	 ⲡⲡⲉⲧⲛⲁⲛⲟⲨϥ.

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 Ⲡ is encoded <hi rend=“red large ekthetic”>Ⲡ</hi>



ⲉ̇ⲧϫⲏⲕⲉ̇ⲃⲟⲗ·–ⲛ̇ⲧⲉ
5 ⲡⲉⲑ̇ⲃⲃⲓⲟ̇ⲛ̇ ϩⲏⲧ·–
Ⲡⲁⲓ̈ⲇⲉⲁ̇ϥϫⲱⲛⲟⲩϣⲁ

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">ϥ</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):

```
[.....]ⲡ̣ⲣ̣[..]
```

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 as TextEdit. Document preferences are set to UTF-8 encoding without byte-order Mark (BOM). (E.g., in TextEdit 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/>.

"Evertype: Antinoou." *Evertype: Antinoou - A Standard Font for Coptic* 2012. 29 May 2013. <http://www.evertype.com/fonts/coptic/>.

Layton, Bentley. *A Coptic Grammar*. 3rd Edition, Revised. Wiesbaden: Harrassowitz, 2011. Print.

Orlandi, Tito. "The Library of the Monastery of Saint Shenute at Atripe." *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/>.