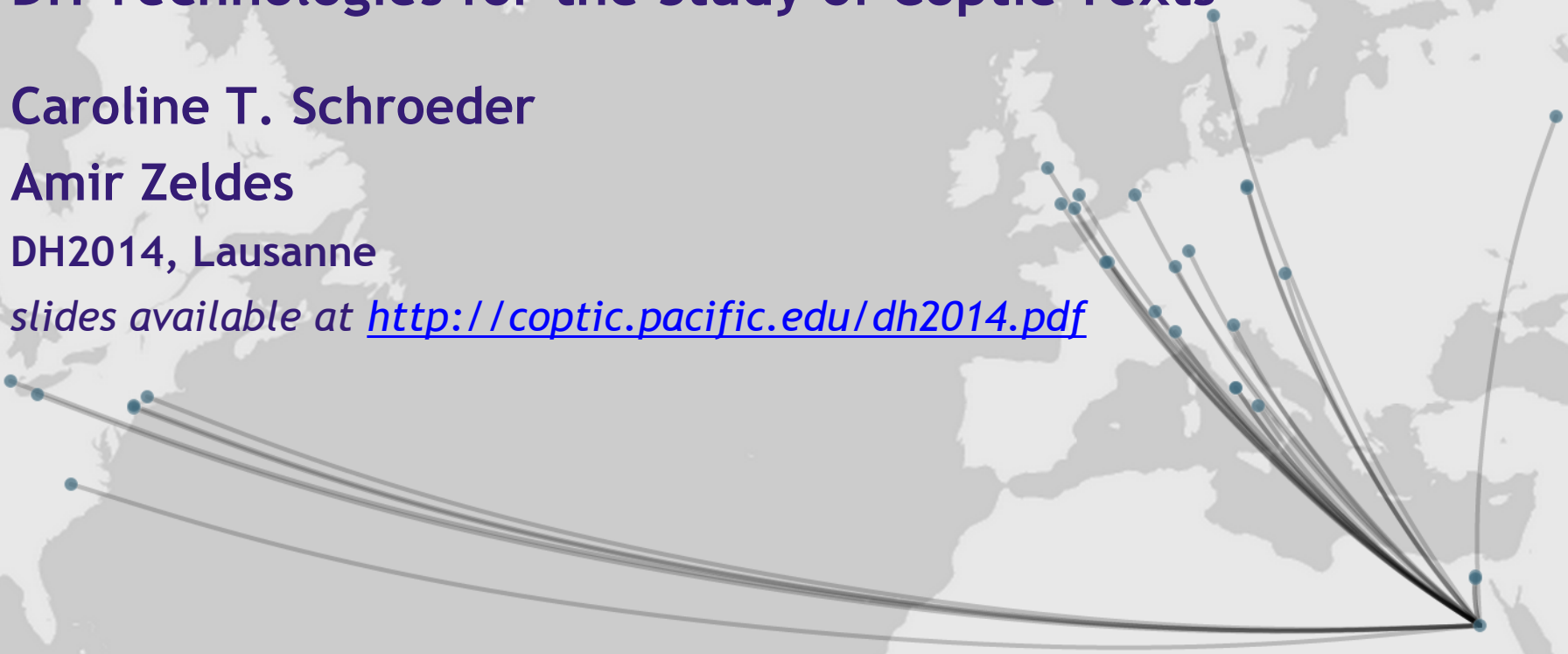
Digitizing the Dead and Dismembered: DH Technologies for the Study of Coptic Texts

Caroline T. Schroeder

Amir Zeldes

DH2014, Lausanne

slides available at http://coptic.pacific.edu/dh2014.pdf



A collaboration of Coptic SCRIPTORIUM and Project KOMET

http://coptic.pacific.edu

http://korpling.german.hu-berlin.de/komet/

NATIONAL ENDOWMENT FOR THE HUMANITIES

Federal Ministry of Education and Research

Whoever enters the monastery uninstructed shall be taught first what he must observe; and when, so taught, he has consented to it all, they shall give him 20 Psalms or two of the Apostle's epistles, or some other part of the Scripture.

And if he is illiterate, he shall go at the first, third and sixth hours to someone who can teach him and who has been appointed for this. He shall stand before him and learn very studiously with all gratitude. Then the fundamentals of a syllable, the verbs, and nouns shall be written for him, and he shall be forced to read, even if he refuses.

Rule of Pachomius, 139-40
4th century
trans. A. Veilleux

1. What is Coptic?

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

The last phase of the ancient Egyptian language family

Hieroglyphs

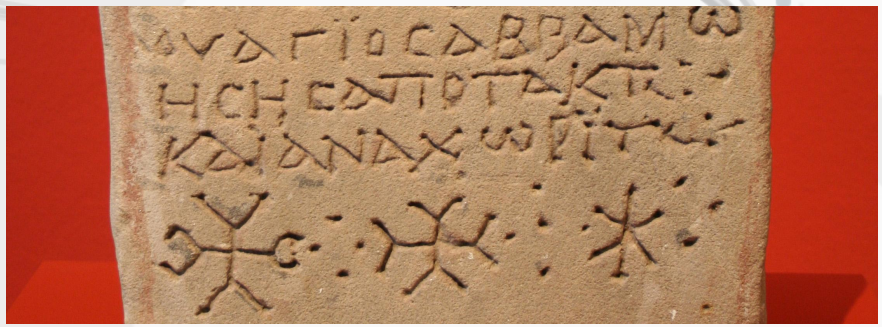
~3400-3200 BCE

Demotic

~650-400 BCE

Coptic usage

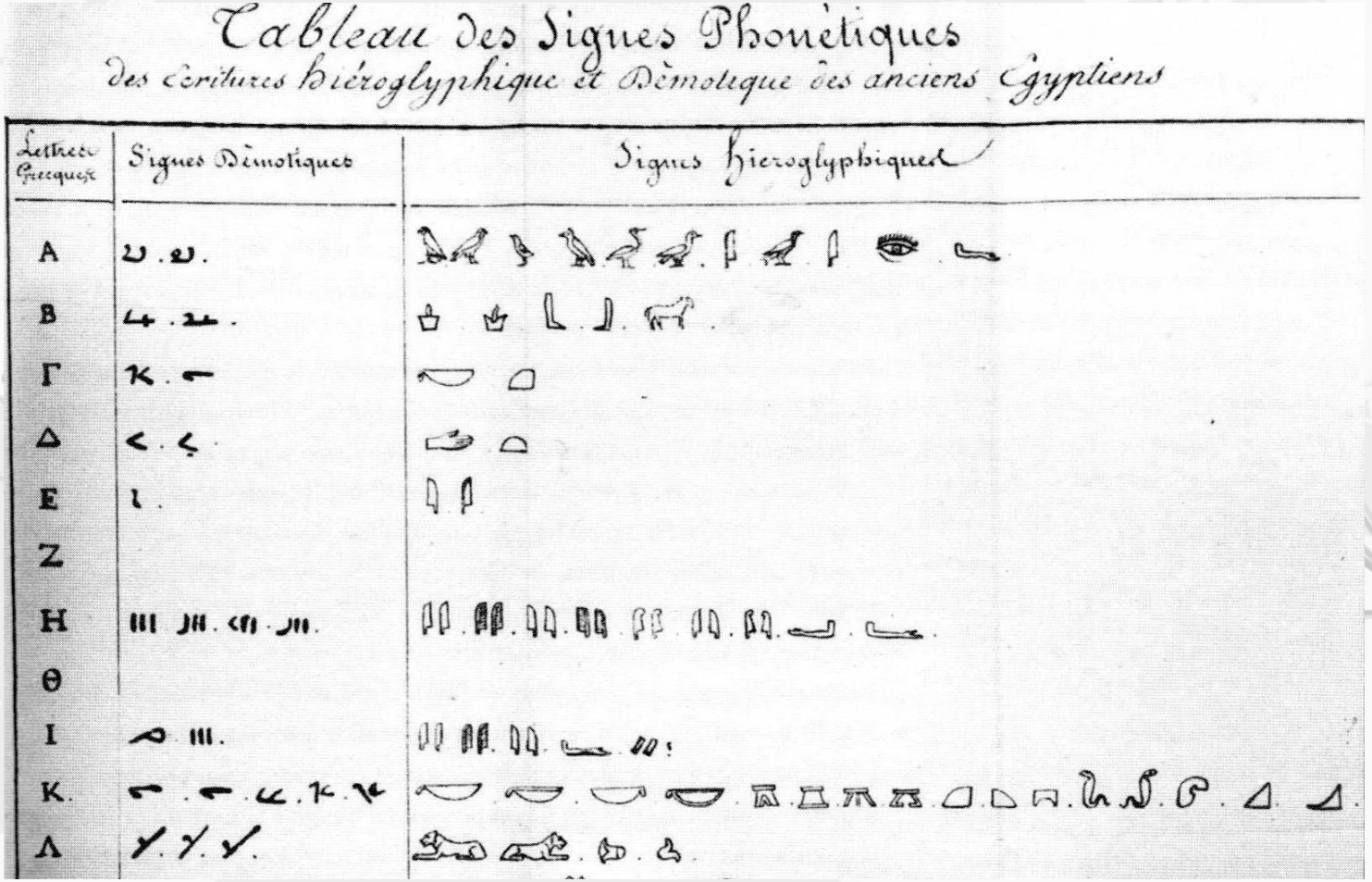
~200 CE - 1000 CE



NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

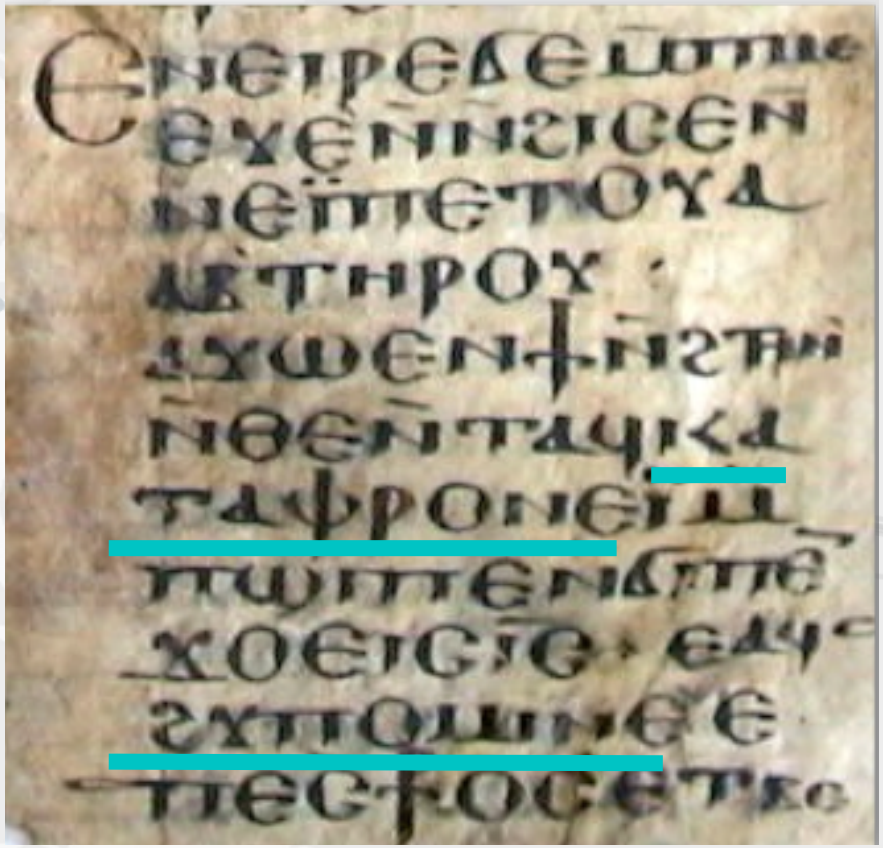
19th c. scholars used their knowledge of Coptic in the race to decipher Egyptian hieroglyphics

Chompollion's chart of Coptic-demotic-hieroglyphic parallels



The Coptic language is primarily Egyptian grammar and syntax, written in the Greek alphabet

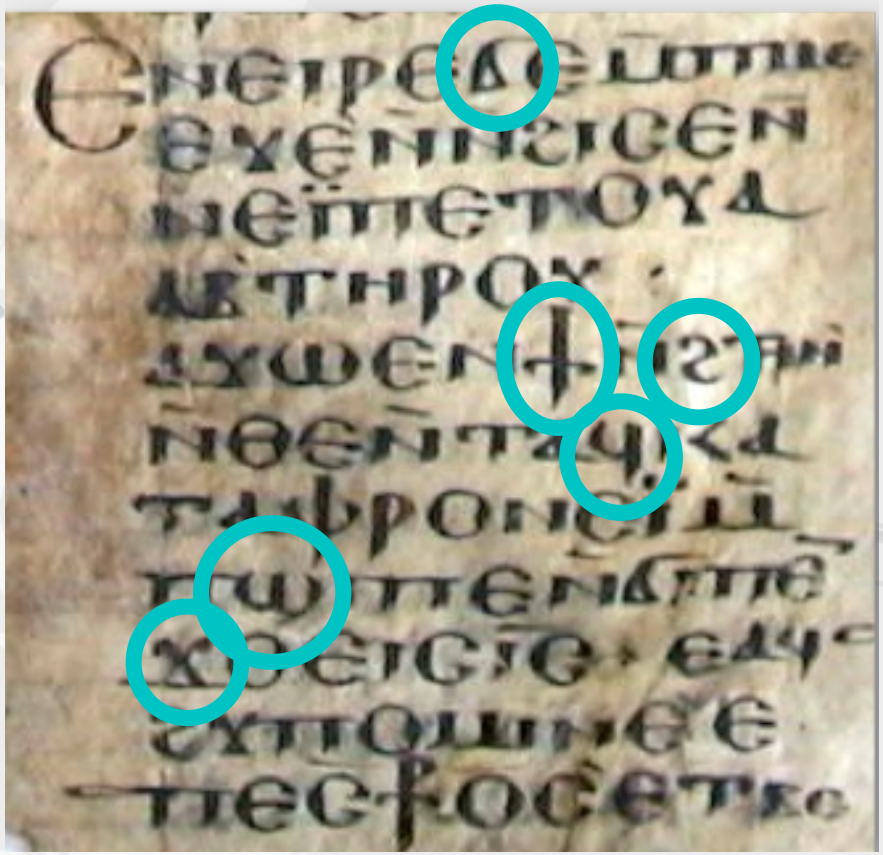
It absorbed some Greek vocabulary.



kataphronei (Greek)

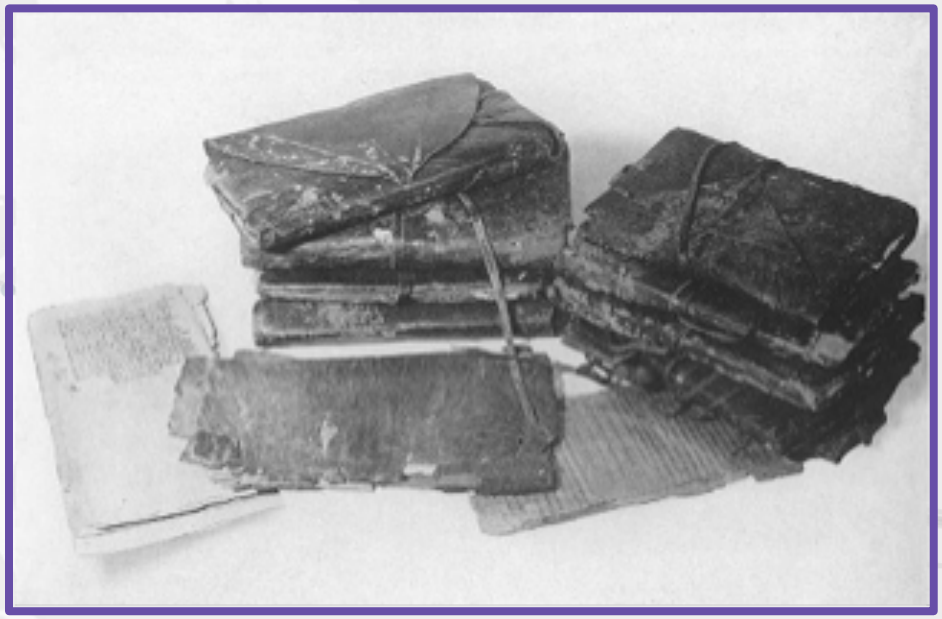
hypomine (Greek)

The alphabet also includes Egyptian characters (from Demotic)



ϭ ϯ ϩ ϥ ϣ ϫ

Coptic is used in multiple fields of research: biblical studies, history, theology and religious studies, linguistics, Egyptology….



Nag Hammadi

The Coptic "Gnostic" library discovered in the Egyptian town of Nag Hammadi in 1945, containing the apocryphal Gospels of Mary, Thomas, and Philip, among other texts

2. The Dismembered Coptic Corpus

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

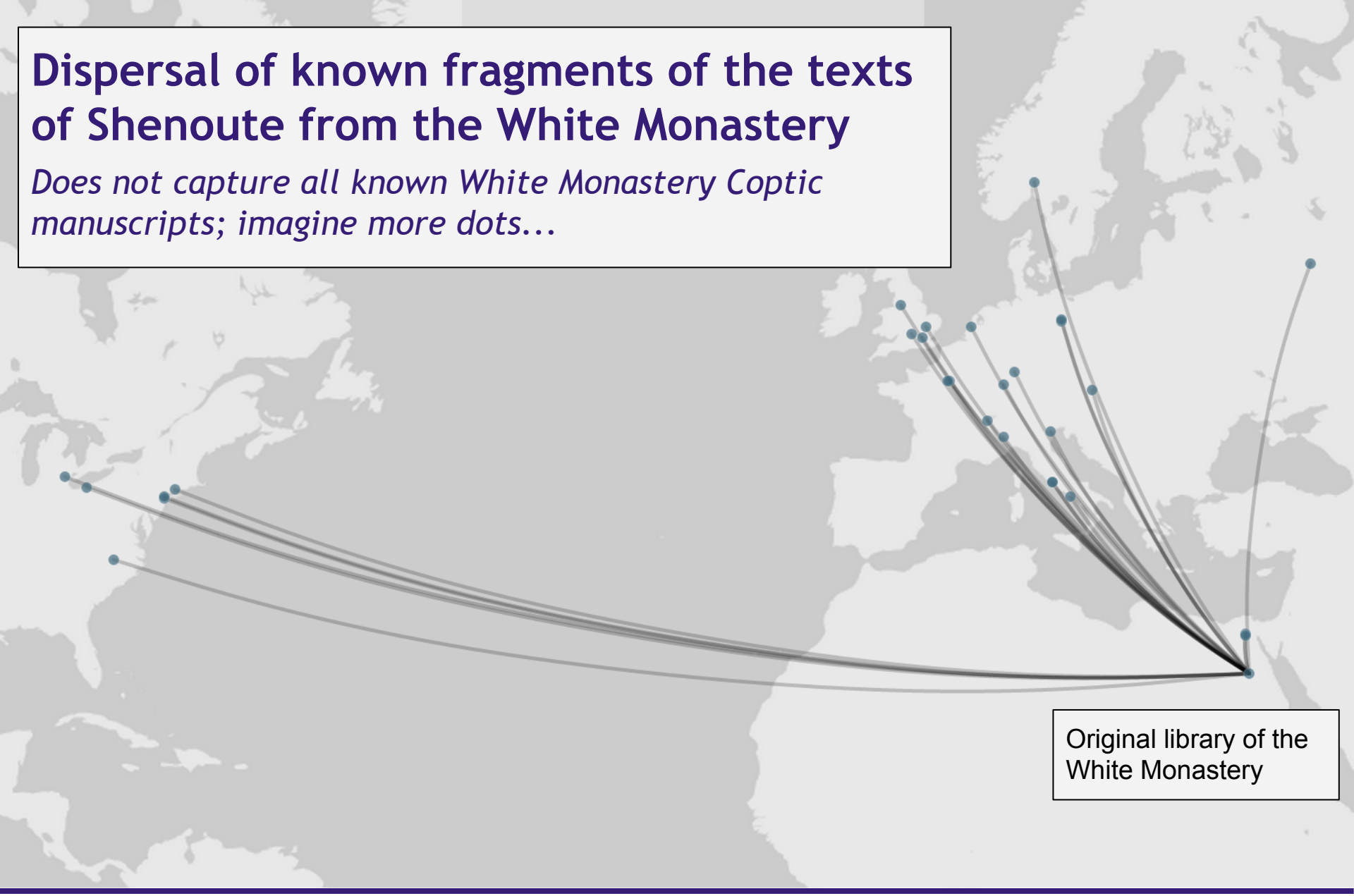
The White Monastery (or Monastery of Shenoute)

One of the most important ancient and medieval repositories for Coptic manuscripts (Sahidic dialect).

Named after its third leader, Shenoute, abbot from ca. 385 to 465 CE.

Dispersal of known fragments of the texts of Shenoute from the White Monastery

Does not capture all known White Monastery Coptic manuscripts; imagine more dots...

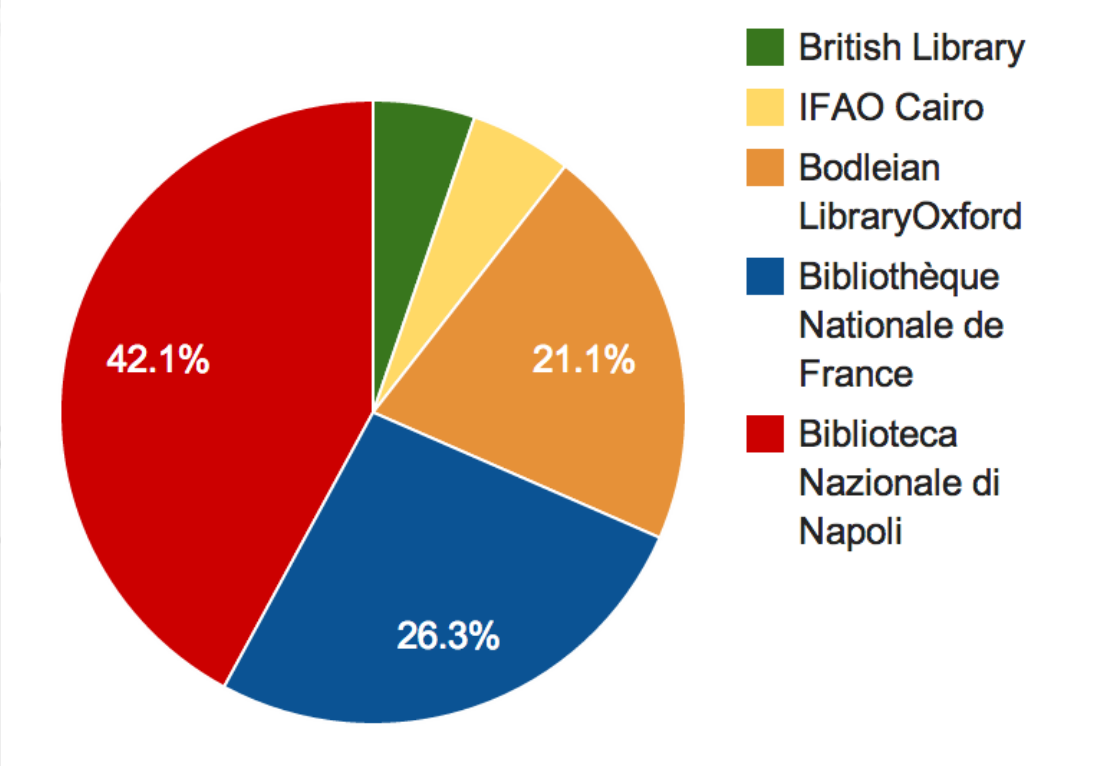


Original library of the White Monastery

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Known, surviving manuscripts of Shenoute's letter *Abraham Our Father*.

Pages from 4 codices in 6 repositories.



We have not yet digitized the Cairo pages.

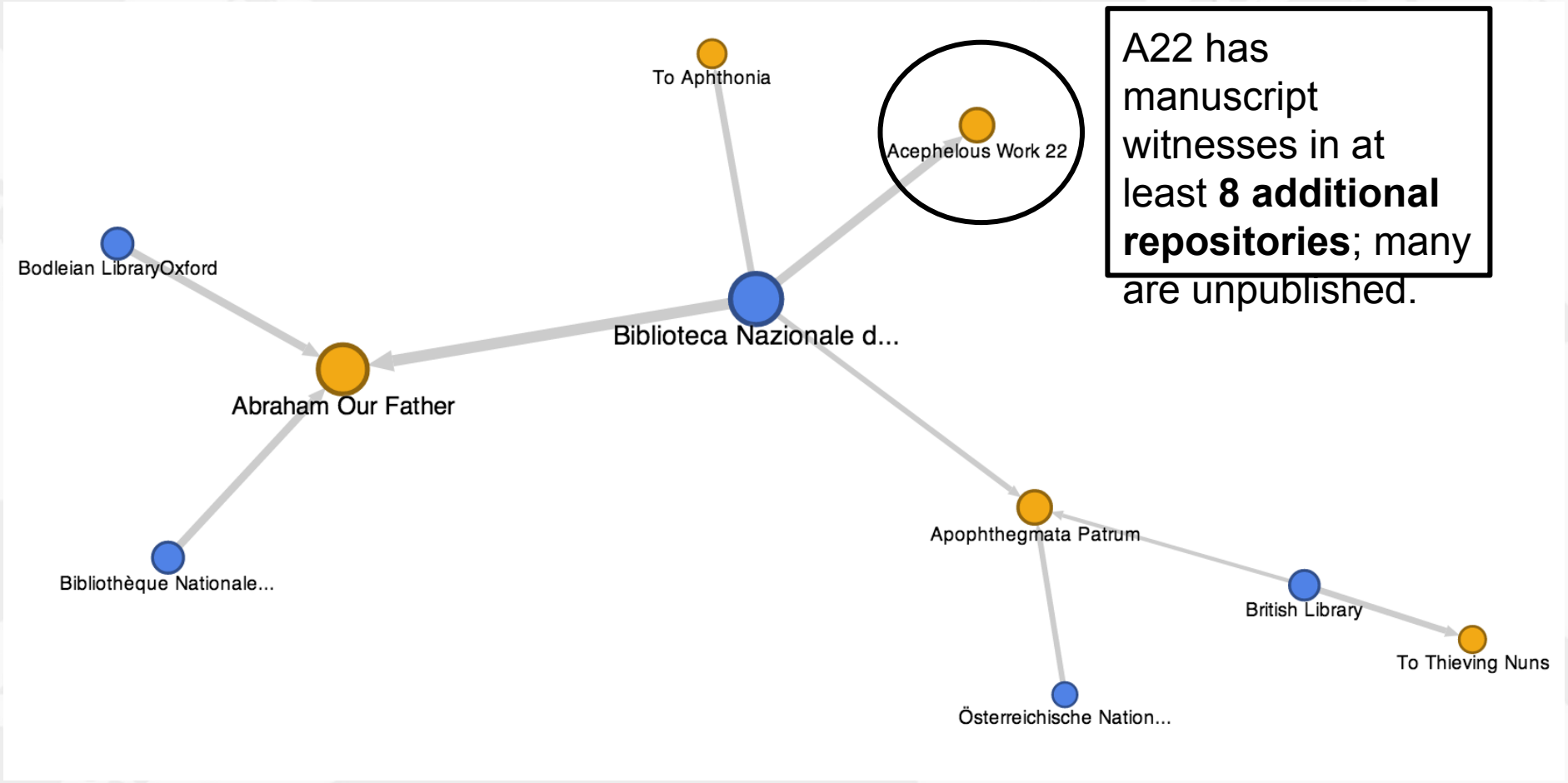
NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Pages of the non-biblical documents in Coptic SCRIPTORIUM as of June 2014

- Shenoute, Abraham Our Father
- Shenoute, Acephalous Work 22
- Besa, To Thieving Nuns
- Besa, to Aphthonia
- Select *Sayings of the Desert Fathers (Apophthegmata Patrum)*

Often, each text has pages in multiple repositories.

Network visualization of pages in Coptic SCRIPTORIUM as of June 2014

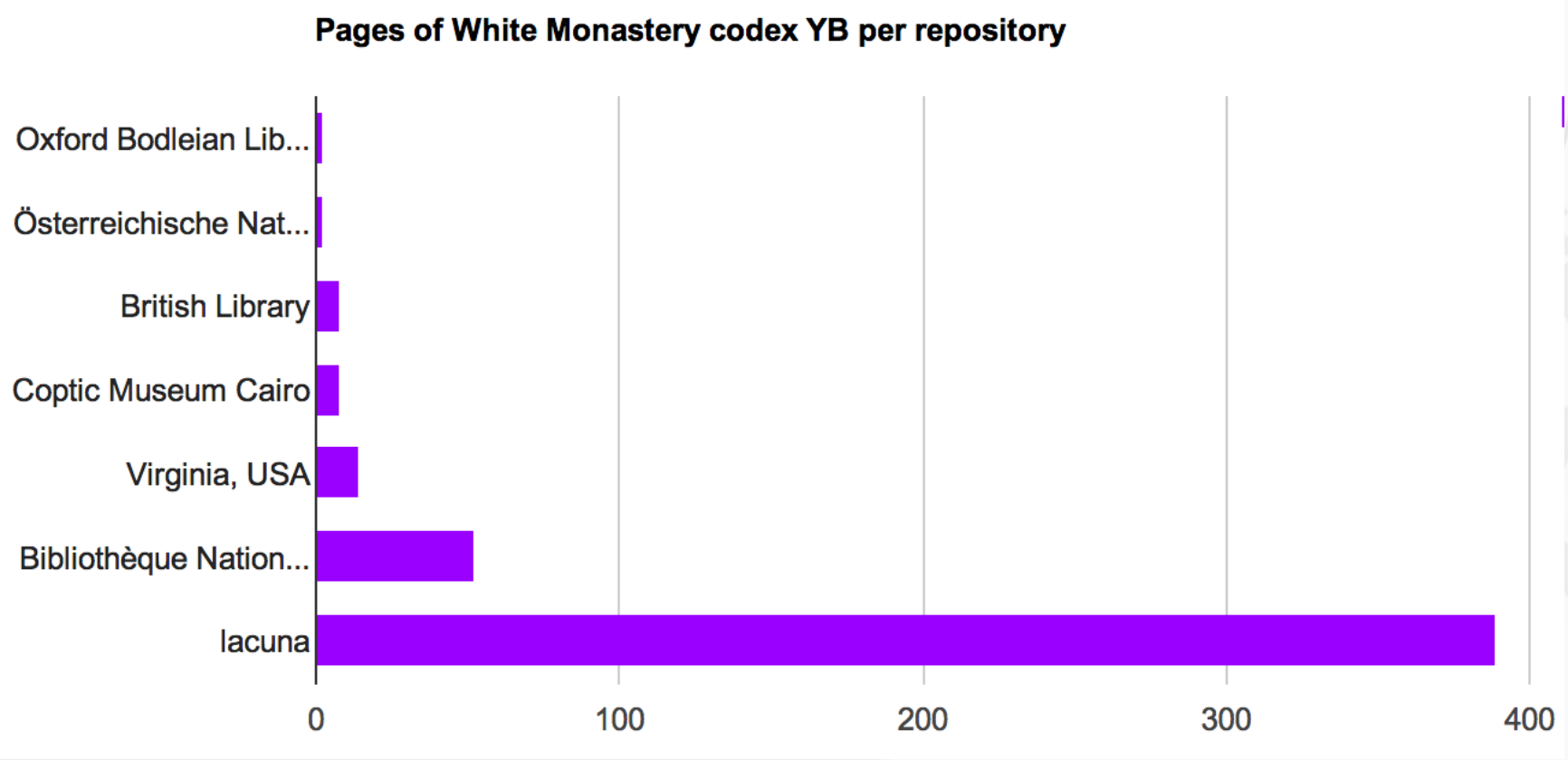


A22 has manuscript witnesses in at least **8 additional repositories**; many are unpublished.

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Typically, each codex has pages in multiple repositories

White Monastery codex YB



3. Data Models

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

What constitutes a “work”?

The letter *Abraham Our Father?*

The book: volume 3 of Shenoute’s *Canons* for monks?

Each codex that constitutes a copy of volume 3 (and *Abraham*)?

Each contiguous fragment of a codex in a modern repository?

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Do you encode each work from the perspective of the original ancient text objects (the codices) or from the perspective of the modern repository in which the manuscript now resides?

Canons Volume 3, codex YA pp.

518-20 Naples IB2 ff. 26-27

521-24 lacuna

525-30 Naples IB2 ff. 28-30

531-34 lacuna

535-40 Naples IB2 ff. 31-33

541-46 lacuna

547-50 Paris BN 130/5 ff.21-22

551-54 Paris BN 130/4 ff. 110-11

Naples IB16

ff. 1-3 Cyril of Alexandria (frag.)

ff. 4-5 Ps-Macarius homily (frag.)

ff.8-11 unidentified

f. 16 Shenoute *Abraham (codex XL)*

ff. 20-23 John Chrysostom Homily on Hebrews (frag.)

<msPart> feature request submitted to the TEI

Current definition:

> **<msPart>** (manuscript part) contains information about an originally distinct manuscript or part of a manuscript, now forming part of a composite manuscript.
> http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-msPart.html

Request:

> change the description so that the element can be used for encoding from either perspective

Accepted 1 July 2014

Coptic word segmentation

'Since I became a monk'

"since" + that + past tense marker + subject pronoun "I" + verb "do" + noun "monk"

6 items (morphemes)=1 word or "bound group"

'he who made us keep the ceremony'

8 morphemes

Different scholars use different conventions for segmenting words.

(We follow Bentley Layton, *Coptic Grammar*, 1st-3rd editions.)

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Data-model in standoff markup

Morphemes	ⲛ̄	ⲟⲩ	ϣⲛⲣⲉ̀	ⲛ̄	ⲁⲃⲣⲁϩⲁⲙ̀
Bound groups	ⲛ̄ⲟⲩϣⲛⲣⲉ̀			ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀	

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Normalize morphemes and bound groups (semi-automated)

Morphemes	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃⲣⲁϩⲁⲙ̀
Bound groups	ⲛ̄ⲟⲩϣⲏⲣⲉ̀			ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀	
Normalized morphemes	ⲛ	ⲟⲩ	ϣⲏⲣⲉ	ⲛ	ⲁⲃⲣⲁϩⲁⲙ
Normalized groups	ⲛⲟⲩϣⲏⲣⲉ			ⲛⲁⲃⲣⲁϩⲁⲙ	

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Manuscript architecture (line, column, page)

also letter renderings (color, size, super/subscript, etc.)

Morphemes	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃⲣⲁϩⲁⲙ̀	
Bound groups	ⲛ̄ⲟⲩϣⲏⲣⲉ̀			ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀		
Normalized morphemes	ⲛ	ⲟⲩ	ϣⲏⲣⲉ	ⲛ	ⲁⲃⲣⲁϩⲁⲙ	
Normalized groups	ⲛⲟⲩϣⲏⲣⲉ			ⲛⲁⲃⲣⲁϩⲁⲙ		
Line breaks	5					6
Column breaks	cb					
Page breaks	pb@xml_id					

Token layer is therefore smaller than the morphemes, to reflect page/line/column breaks between words

Token	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃ	ⲣⲁϩⲁⲙ̀
Morphemes	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃⲣⲁϩⲁⲙ̀	
Bound groups	ⲛ̄ⲟⲩϣⲏⲣⲉ̀			ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀		
Normalized morphemes	ⲛ	ⲟⲩ	ϣⲏⲣⲉ	ⲛ	ⲁⲃⲣⲁϩⲁⲙ	
Normalized groups	ⲛⲟⲩϣⲏⲣⲉ			ⲛⲁⲃⲣⲁϩⲁⲙ		
Line breaks	5					6
Column breaks	cb					
Page breaks	pb@xml_id					

Fuller view of data-model in standoff markup

Token	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃ	ⲣⲁϩⲁⲙ̀
Morphemes	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ̀	ⲛ̄	ⲁⲃⲣⲁϩⲁⲙ̀	
Bound groups	ⲛ̄ⲟⲩϣⲏⲣⲉ̀			ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀		
Normalized morphemes	ⲛ	ⲟⲩ	ϣⲏⲣⲉ	ⲛ	ⲁⲃⲣⲁϩⲁⲙ	
Normalized groups	ⲛⲟⲩϣⲏⲣⲉ			ⲛⲁⲃⲣⲁϩⲁⲙ		
Line breaks	5					6
Column breaks	cb					
Page breaks	pb@xml_id					
Part of speech	PREP	ART	N	PREP	NPROP	
Lang of origin					Hebrew	

Project KOMeT

Coptic SCRIPTORIUM presents this data open source in a variety of models, including the EpiDoc subset of TEI XML.

http://coptic.pacific.edu

http://sourceforge.net/p/epidoc/wiki/Home/

For research questions that need annotations not covered by TEI XML, Project KOMeT offers a standard for annotating TEI documents from the outside, using separate external XML files in standoff annotation. TEI XML texts can be annotated without editing or changing the TEI XML.

http://korpling.german.hu-berlin.de/komet/

4. Linguistic Annotations and Research Applications

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Part of Speech Tagging

TreeTagger (Schmid 1994, natural language processing) and a lexicon from Prof. Tito Orlandi and the Corpus dei Manoscritti Copti Letterari project http://cmcl.aai.uni-hamburg.de/

Two tag sets: fine grained (45 tags) and coarse (22 tags)

See http://coptic.pacific.edu for full documentation

Accuracy:

- In domain, 10-fold cross-validation: 94.04% (fine)
- Out of domain (test with papyri.info): 79.6% (fine) / 87.7% (coarse)

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

A primary difficulty is disambiguating homonyms

ɴ and ϵ each can have 6 different tags!

Word cloud of the part of speech tags for ɴ in the SCRIPTORIUM corpora.

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Examining Style in Coptic Literary Texts

Part of Speech Tags: Residuals from Chi Squared Test Frequencies
Red=more frequent than expected; blue=less

pos	Besa	A22	AOF	AP	Mark1_6
PREP	-0.9223	2.245374	3.720749	-1.83133	-3.60583
N	0.080235	2.600809	3.098838	-0.51138	-4.27406
ART	-1.18379	1.800207	4.449135	-3.38187	-3.27729
V	-1.16191	-2.07814	-3.87829	0.551977	5.31527
PPERS	-1.98882	-2.5304	-8.07701	3.180776	9.002659
PPERO	-1.0792	0.041699	-3.04763	-0.49954	3.790747
CONJ	0.570794	-1.54252	-2.43828	0.739032	2.621228
PUNCT	5.768879	-3.21255	5.669246	2.864461	-8.19385
APST	-1.82084	-4.9142	-4.00445	1.51988	6.850586
CREL	1.668746	2.013423	2.620324	-3.31901	-3.06952
ADV	-1.58016	1.466869	0.979143	-2.08521	-0.01307
PPOS	4.262987	0.407573	-0.71813	0.272237	-1.82168
CCIRC	-2.62532	0.757333	-1.08926	-0.03072	2.054625
PTC	-1.14017	-0.99507	-5.51929	2.960995	5.304131
NPROP	-2.57099	-3.90854	1.538779	-0.96275	2.282299
VSTAT	2.454121	1.992609	-1.78869	0.626819	-0.81402

Narrative texts:

- few articles
- more personal subject pronouns

G Mark

- more past tense

Abraham and G Mark

- about biblical figures & biblical text > more proper names

The translated texts have more particles (e.g., Grk *de*)

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Language of Origin Tagging

Significant amounts of Greek vocabulary in Coptic literature, plus some other languages:

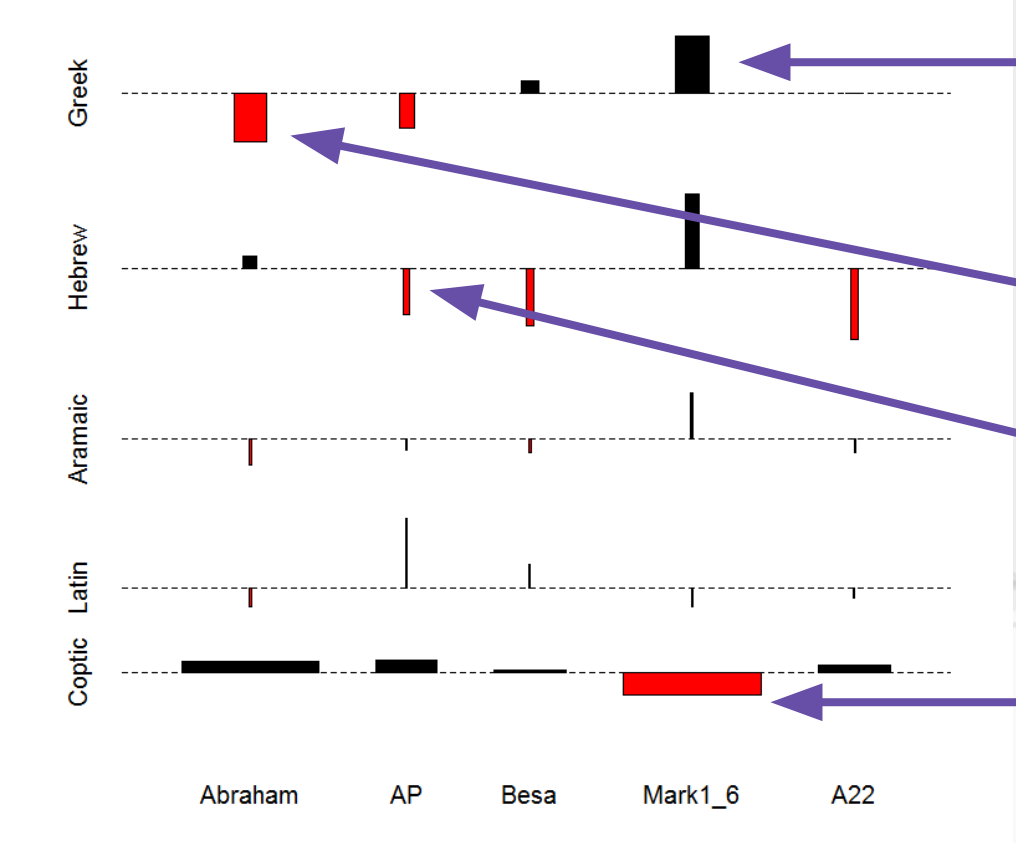
- Hebrew (mostly names)
- Aramaic (mostly names)
- Latin
- eventually Arabic

Annotations can enable research in

- translation practices
- loanword borrowing across languages
- multilingualism

Association plot for language of origin

idth corresponds to amount of data, height to deviation from the expected value



Gospel of Mark has most Greek vocabulary

Abraham Our Father by Shenoute has the least Greek

Surprise: Sayings of the Desert Fathers have less Greek than expected (a translated text)

Gospel of Mark has less than expected native Coptic vocabulary

Much more remains to be done

Abba Abraham told of a man of Scetis who was a scribe and did not eat bread. A brother came to beg him to copy a book. The old man whose spirit was engaged in contemplation, wrote, omitting some phrases and with no punctuation. The brother, taking the book and wishing to punctuate it, noticed that words were missing. So he said to the old man, 'Abba, there are some phrases missing.' The old man said to him, 'Go, and practise first that which is written, then come back and I will write the rest.'

Abraham 3
Sayings of the Desert Fathers
trans. Benedicta Ward

Acknowledgments and Contact Information

We thank the University of the Pacific and Humboldt University for ongoing support.

Funding for 2014-15 provided by the National Endowment for the Humanities Office of Digital Humanities and Div. of Preservation & Access (USA).

Funding for 2014 provided by the BMBF/Federal Ministry of Education and Research (Germany).

Caroline T. Schroeder

University of the Pacific, Stockton, California
carrie@carrieschroeder.com

Amir Zeldes

Humboldt University, Berlin
beg. August 2014: Georgetown, Washington, DC
amir.zeldes@rz.hu-berlin.de

http://coptic.pacific.edu
http://korpling.german.hu-berlin.de/komet/
All project participants are listed on the websites.

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Image & source credits

Full bibliography at the Coptic SCRIPTORIUM Zotero site.

Background and global Shenoute manuscript distribution map: visualization from palladio.designhumanities.org with data from Stephen Emmel, *Shenoute's Literary Corpus* and personal correspondence wiht Alin Suciu; google map at http://goo.gl/c3uaM6.

2. A. Veilleux, *Pachomian Koinonia* 2:166-67.
3. "Le_menu_de_Tepemankh_Louvre_d2.jpg" http://upload.wikimedia.org/wikipedia/commons/9/9c/Le_menu_de_Tepemankh_Louvre_d2.jpg. 3 July 2014.
"Demotic_Ostrakon.jpg" http://upload.wikimedia.org/wikipedia/commons/c/c6/Demotic_Ostrakon.jpg. 3 July 2014.
RPM_Ägypten_284.jpg http://upload.wikimedia.org/wikipedia/commons/7/78/RPM_%C3%84gypten_284.jpg. 3 July 2014.
4. "Champollion_table.jpg" http://upload.wikimedia.org/wikipedia/commons/b/b3/Champollion_table.jpg. 3 July 2014.
6-7. White Monastery manuscript MONB.YA page 535/Biblioteca Nazionale Naples, Borgia Collection, IB2 f. 31r
8. "NagHammadi_1.jpg" http://upload.wikimedia.org/wikipedia/en/f/f7/NagHammadi_1.jpg. 3 July 2014.

NATIONAL ENDOWMENT FOR THE Humanities
Federal Ministry of Education and Research

Image & source credits continued...

10. Photograph by Schroeder, White Monastery near Sohag, December 2012. Licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
11. Visualization using http://palladio.designhumanities.org with data from Emmel, *Shenoute's Literary Corpus* and personal correspondence with Alin Suciu; google map at http://goo.gl/c3uaM6.
12. Visualization using Google fusion tables with data from Emmel, *Shenoute's Literary Corpus*.
13. Visualization using http://palladio.designhumanities.org with data from Emmel, *Shenoute's Literary Corpus*, and Schroeder, Zeldes, et al (2013-2014), *Coptic SCRIPTORIUM*, http://coptic.pacific.edu.
14. Visualization using Google fusion tables with data from Emmel, *Shenoute's Literary Corpus*, and Schroeder, Zeldes, et al (2013-2014), *Coptic SCRIPTORIUM*, http://coptic.pacific.edu.
15. Visualization using Google fusion tables with data from Emmel, *Shenoute's Literary Corpus*.
18. Emmel, *Shenoute's Literary Corpus*; Alin Suciu, "The Borgian Coptic Manuscripts in Naples."
19. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-msPart.html
20. Layton, *Coptic Grammar*, 3rd ed.
28. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees."
29. Visualization using http://www.wordle.net with data from Schroeder, Zeldes, et al (2013-2014), *Coptic SCRIPTORIUM*, http://coptic.pacific.edu.
30. Data from Schroeder, Zeldes, et al (2013-2014), *Coptic SCRIPTORIUM*, http://coptic.pacific.edu.
32. Visualization using R with data from Schroeder, Zeldes, et al (2013-2014), *Coptic SCRIPTORIUM*, http://coptic.pacific.edu.
33. Ward, *Sayings of the Desert Fathers*, 29.