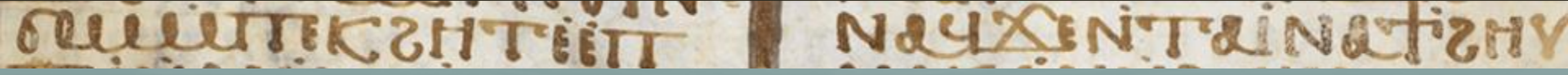


Digital Coptic 3
July 12-13, 2020

NATIONAL ENDOWMENT FOR THE HUMANITIES

Exposing Coptic entities

Automation, search and visualization

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Caroline T. Schroeder
University of the Pacific
cschroeder@pacific.edu

Lance Martin
Catholic University of America
71martin@cua.edu

GEORGETOWN UNIVERSITY

THE CATHOLIC UNIVERSITY OF AMERICA

COPTIC SCRIPTORIUM

copticscriptorium.org

Open source, open access research platform for Coptic language and literature

Tools, texts, collaborative environment

Interdisciplinary

Today: focus on **entities**

Thanks!

- This is joint work with project members and students:
  - Sichang Tu
  - Elizabeth Davidson
  - Mitchell Abrams

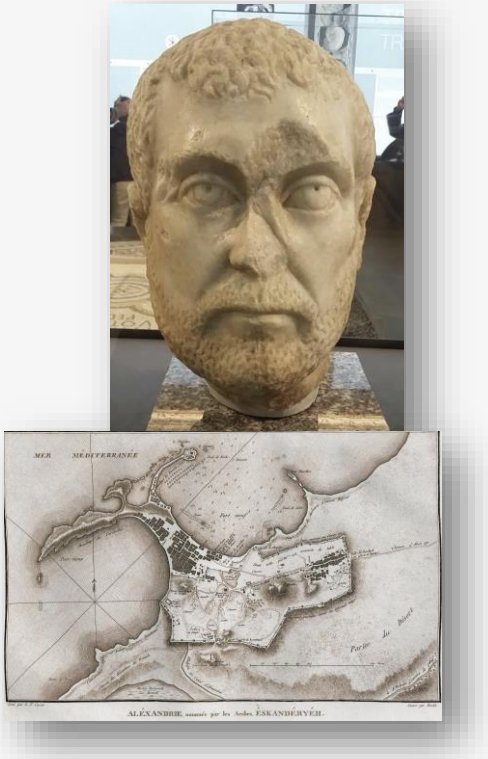
- Funding: NEH (HAA-261271-18, Digital Humanities Advancement Grant III *A Linked Digital Environment for Coptic Studies*)

NATIONAL ENDOWMENT FOR THE HUMANITIES

What are entities?

Texts talk about things: people, places…

- Some of these are **named:**
  - “Emperor Diocletian”, “Alexandria”, …



- But we also might care about **non-named** entities:
  - “A holy man”, “The desert”, “a different monastery”, …

What are entities?

- For this talk: entities are **anything you can refer to** in a text (e.g. refer back to with a pronoun)
- Entities can be **nested** and belong to many **types**:

[*also the faith of* [*the Catholic Church*]$_{\text{ORGANIZATION}}$]$_{\text{ABSTRACT}}$

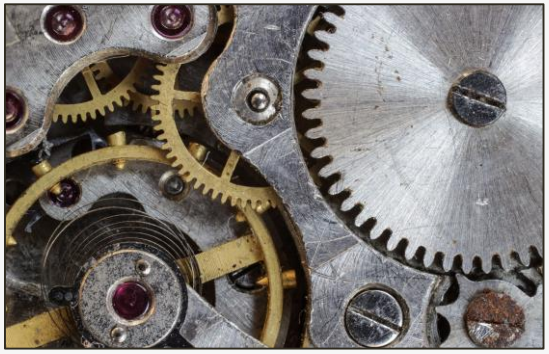
Main question: *how can we **expose** these entities to users?*

Goals

- Provide gold standard (non-)named entity **data**
- **Link** named entities to searchable stable identifiers
- Publish tools for **automatic** entity recognition

Data

Linking



Automation

Data

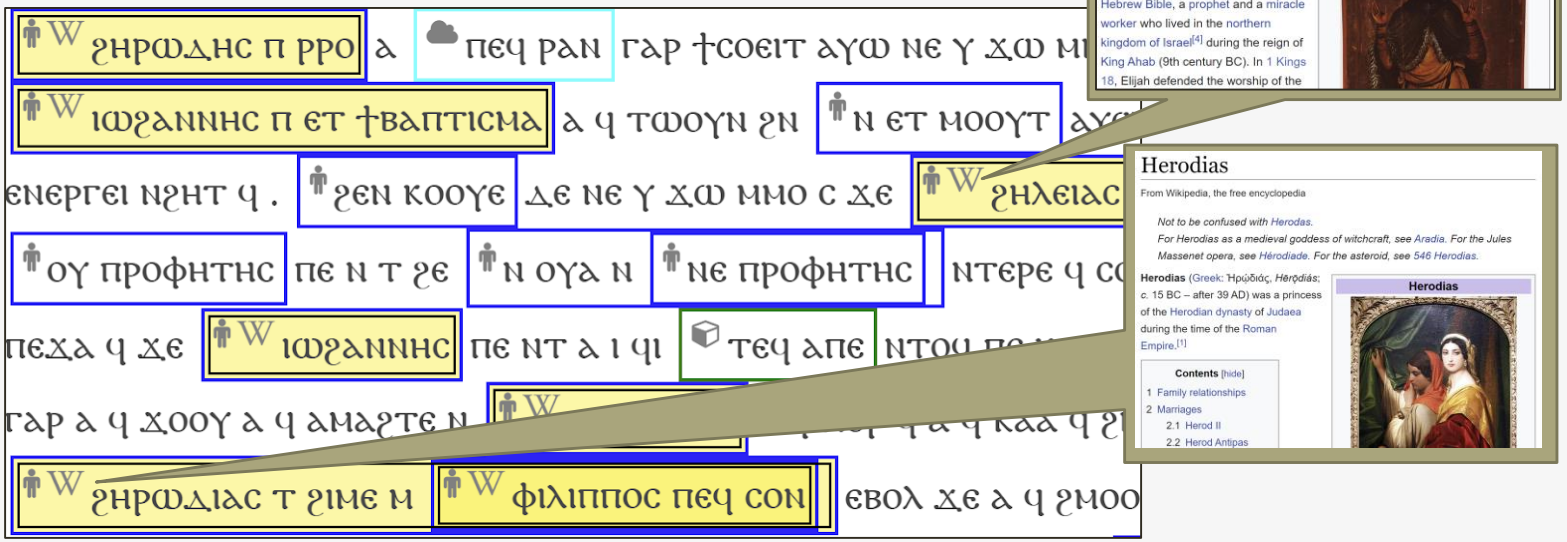
- As a pilot project, we have manually annotated the Coptic Treebank (Zeldes & Abrams 2018) for entities:
  - 46,000 words
  - 6,500 entity mentions
    - 610 named (~10%)
    - 441 identifiable, linked to 104 unique Wikipedia articles

Why Wikipedia?

- Common practice in Computational Linguistics – entity linking=“Wikification” (Shnayderman et al. 2019)
- Offers a table of authorities out-of-the-box
- Very stable, (rare) identifier changes are tracked
- Comparable and linkable to other language resources
- Automatically gives us Wiki articles about mentioned entities, GIS (geo-location) information, …

Wikification and links

- Users of Coptic Scriptorium data can immediately find out who someone/where some place is
- Get quantitative/distributional information



What's included?

- Currently: Any nominal referring expression, except pronouns
- Ten entity classes:

	ABSTRACT ('humility', 'thoughts')		PERSON ('Ruth the Moabite', 'all the angels')
	ANIMAL ('a dog', '200 horses')		PLACE ('a mountain', 'Alexandria')
	EVENT ('his death', 'war')		PLANT ('the tree', 'fruit', 'wheat')
	OBJECT ('bottles', 'her hand')		SUBSTANCE ('the water', 'blood')
	ORGANIZATION ('the king's army', 'Catholic Church')		TIME ('the month of Parmoute', 'ten years')

What can users search for?

- Entity types, identities
- Named/non-named
- Since our data is also 100% syntactically parsed:
  - Entity head noun ([the **army** of the emperor])
  - Part of speech
  - Grammatical function
  - Foreign/Greek loan words
  - …

What can users search for?

entity ->head pos="ART"	*entities headed by articles, ⲡⲉⲧ...*
text=/.*ⲡ ⲛⲟⲩⲧⲉ.*/	*full text contains ⲡ ⲛⲟⲩⲧⲉ*
entity="person" _i_ entity="place"	*PERSON containing a PLACE*
identity=/.*Herod.*/	*linked entity contains 'Herod'*
identity="Jesus" & meta::corpus=/.*mark.*/	*mentions of Jesus in Mark*
entity="person" ->head norm & pos="V" & #3 ->dep[func="obj"] #2	*verb with PERSON as object*
entity="place" _i_ lang="Greek"	*PLACE containing Greek loanword*

- Try it here:
  - https://corpling.uis.georgetown.edu/annis/scriptorium/
  - Select ***coptic.treebank***, enter a query from the above and search!
- Detailed ANNIS QL reference:
  - https://corpus-tools.org/annis/aql.html
  - https://copticscriptorium.org/ANNIS-tips.html

What can users search for?

ϩⲁⲡϩⲟ ⲛⲏⲥⲁⲩ · ⲡⲉϥⲥⲟⲛ ⲁϥⲃⲱⲕ ⲉⲧⲙⲉⲥⲟⲡⲟⲧⲁⲙⲓⲁ ⲛⲧⲥⲩⲣⲓⲁ ⲁϥⲉⲓ ⲉϫⲙⲡⲙⲁ ⲉⲧⲉⲣⲉⲛⲉⲥⲟⲟⲩ ⲛⲗⲁⲃⲁⲛ ⲛϩⲏⲧϥ

⊟ annotations (grid)

				Greek			Greek					
	ⲃⲱⲕ	ⲉ	ⲡ	ⲙⲉⲥⲟⲡⲟⲧⲁⲙⲓⲁ	ⲛ	ⲡ	ⲥⲩⲣⲓⲁ	ⲁ	ⲛⲧⲟϥ	ⲉⲓ	ⲉϫⲛ	ⲡ
	ⲃⲱⲕ ⲉ											
	ⲃⲱⲕ	ⲉ	ⲧ	ⲙⲉⲥⲟⲡⲟⲧⲁⲙⲓⲁ	ⲛ	ⲧ	ⲥⲩⲣⲓⲁ	ⲁ	ϥ	ⲉⲓ	ⲉϫⲙ	ⲡ
		ⲉⲧⲙⲉⲥⲟⲡⲟⲧⲁⲙⲓⲁ			ⲛⲧⲥⲩⲣⲓⲁ			ⲁϥⲉⲓ			ⲉϫⲙⲡⲙⲁ	
	ⲃⲱⲕ	ⲉ	ⲧ	ⲙⲉⲥⲟⲡⲟⲧⲁⲙⲏⲁ	ⲛ̄	ⲧ	ⲥⲩⲣⲓⲁ	ⲁ	ϥ	ⲉⲓ	ⲉϫⲙ̄	ⲡ
		ⲉⲧⲙⲉⲥⲟⲡⲟⲧⲁⲙⲏⲁ			ⲛ̄ⲧⲥⲩⲣⲓⲁ			ⲁϥⲉⲓ			ⲉϫⲙ̄ⲡⲙⲁ	

	V	PREP	ART	NPROP	PREP	ART	NPROP	APST	PPERS	V	PREP	A
	ccomp	case	det	obl	case	det	nmod	aux	nsubj	parataxis	case	d
			place									p
						place						
						Syria (region)						

s brother, went into Mesopotamia of Syria, and how when he had arrived at the place wherein were the sheep of Laban

Automatic processing

- The Treebank is limited in scale
- Our goal is to include **automatic** entity annotations for all Coptic Scriptorium corpora by the end of summer
- Use **Natural Language Processing** tools
- Approach:
  - Rely on syntax parser to identify entity mention boundaries (also tested: Recurrent Neural Network)
  - Use state-of-the-art machine learning classifiers and high coverage knowledge base to classify entities (using CRF classifier)
  - Cascaded entity link matching – full text, entity head, corpus and frequency biases

Entity detection: trees and neural networks

Tree based detection

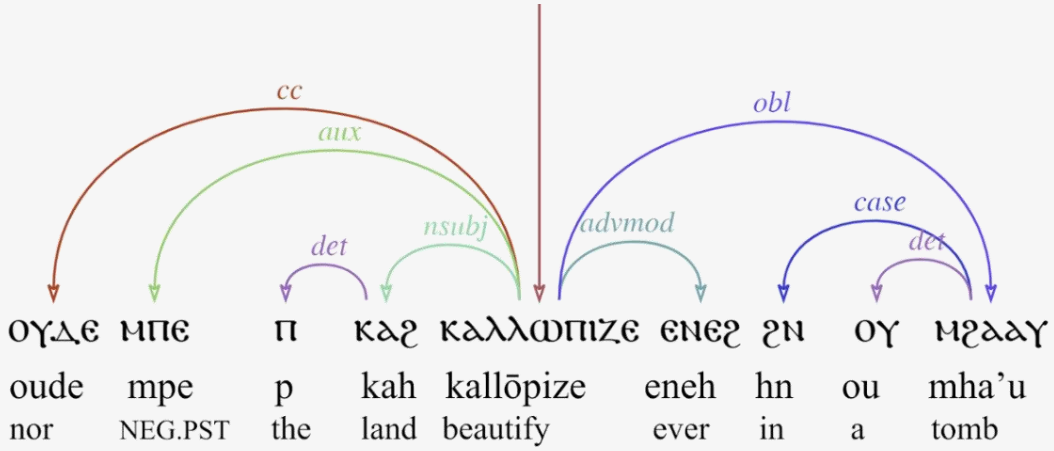
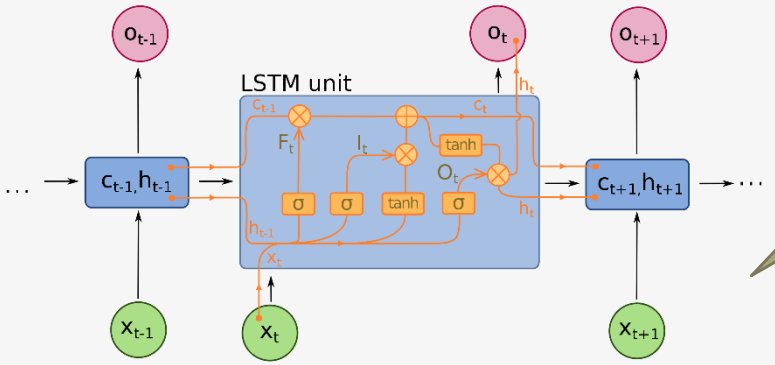


Figure 1. Universal Dependencies tree for a Coptic sentence: *"nor has the land ever adorned itself with a tomb"* (Proclus, homily 13 'On Easter', in Budge ed., urn:cts:copticLit:proclus.homily13.budge)

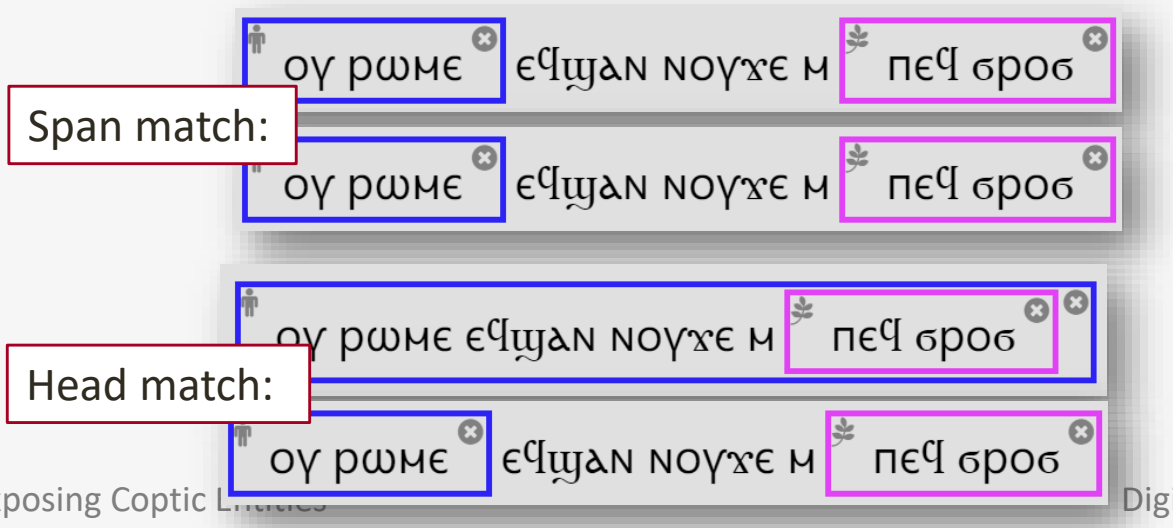
Recurrent Neural Network (RNN)



Bi-LSTM with biaffine attention, Yu et al. (2020)

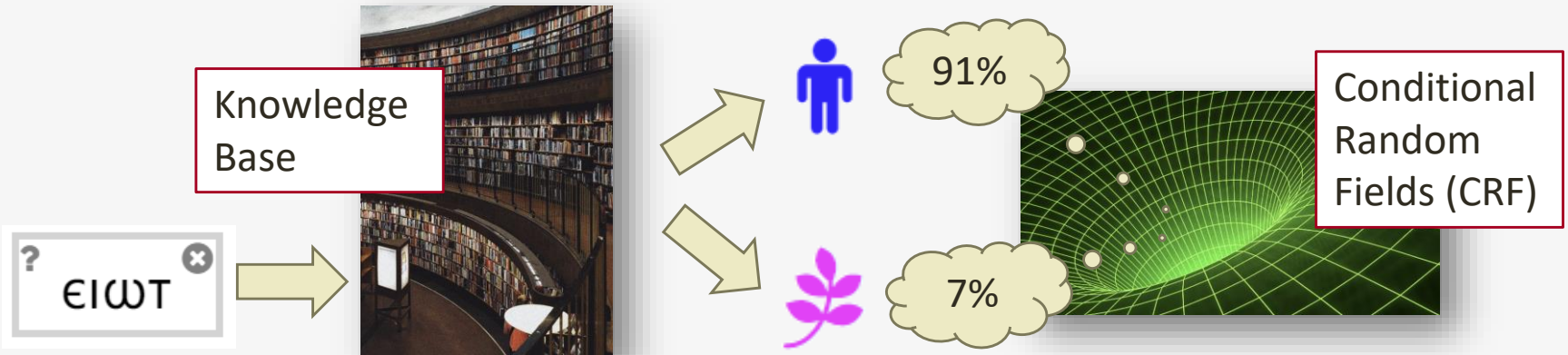
Evaluation – entity detection

	Span match			Head match		
method	**Recall**	**Precision**	**F1**	**Recall**	**Precision**	**F1**
LOOKUP (baseline)	0.386	0.555	0.455	0.591	0.850	0.697
NOUN (entities = nouns)	0.123	0.111	0.117	0.855	0.773	0.812
TREE (gold parse)	0.879	0.862	**0.870**	0.948	0.929	**0.938**
TREE (pred parse)	0.831	0.815	**0.823**	0.941	0.922	**0.931**
RNN (binary)	0.653	0.732	0.690	0.793	0.725	0.757



Evaluation – entity classification

	Span match			Head match		
method	**Recall**	**Precision**	**F1**	**Recall**	**Precision**	**F1**
MAJORITY	0.213	0.209	0.211	0.235	0.230	0.232
RNN	0.476	0.614	0.536	0.527	0.757	0.621
KB	0.681	0.660	0.670	0.728	0.705	0.717
CRF	0.805	0.778	0.791	0.861	0.831	0.846
CRF+KB	**0.827**	**0.810**	**0.818**	**0.889**	**0.869**	**0.879**



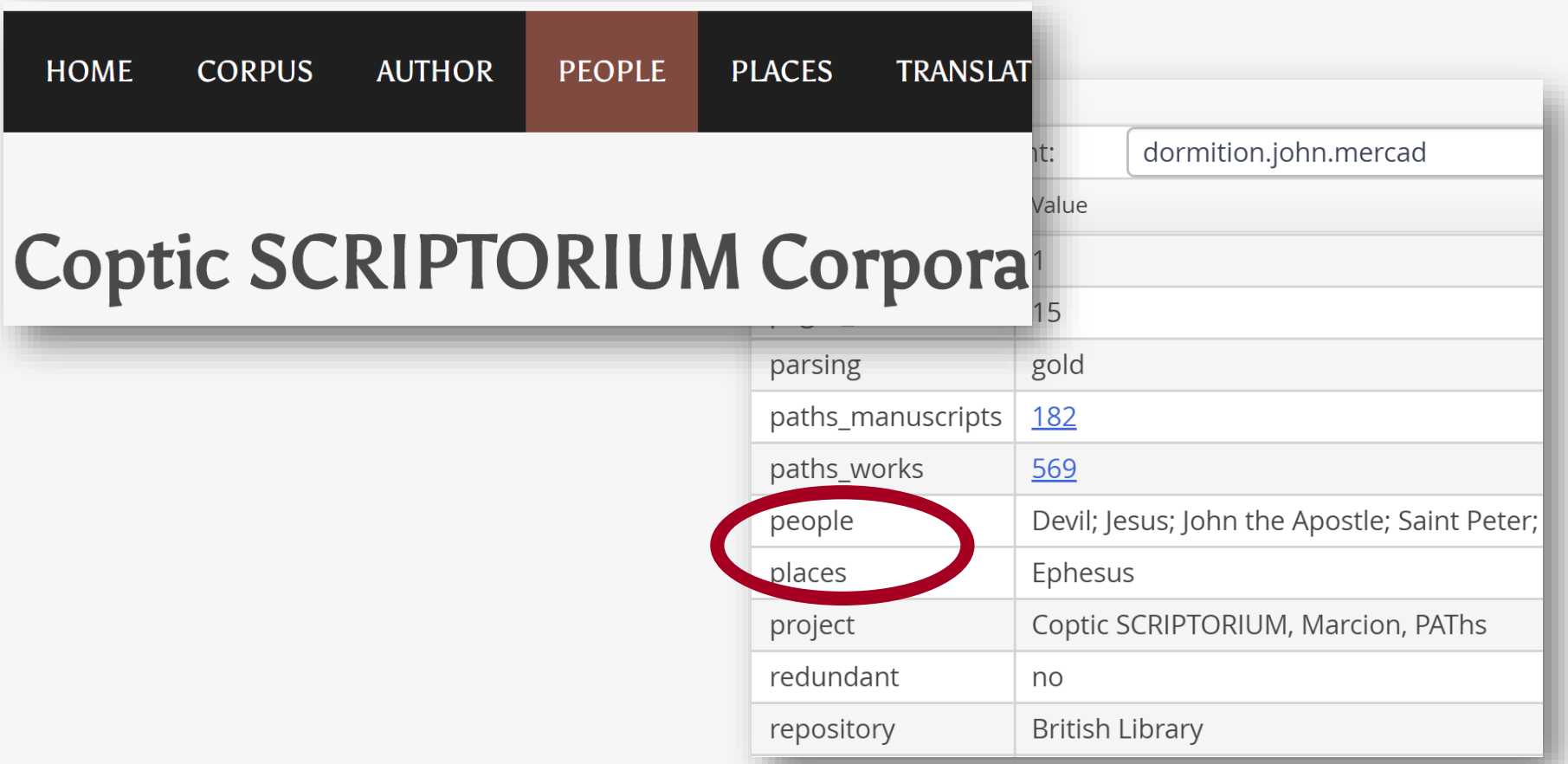
Evaluation – entity linking

- GitDox interface (Zeldes & Zhang 2016)
- (corpus) > best match > (corpus) > best head

method	accuracy	coverage	no error
exact	0.227	0.273	0.953
head	0.433	0.500	0.933
cascade	**0.460**	**0.500**	**0.960**

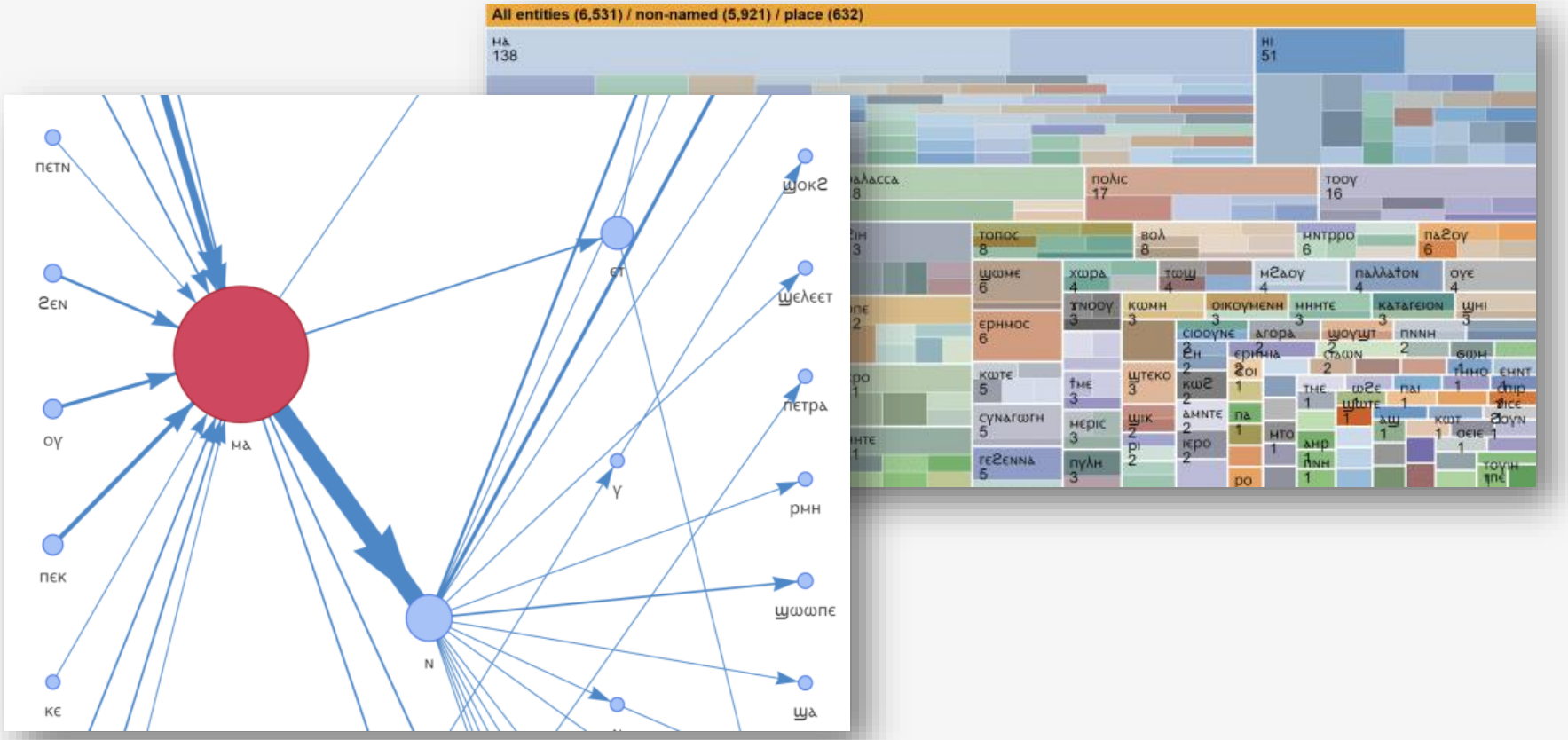
Metadata linking

- Enables search for documents by mentioned entities



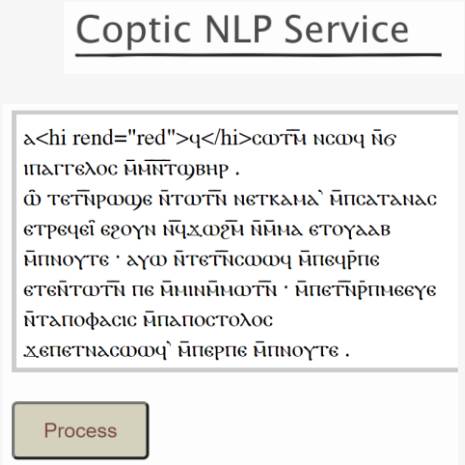
Explore the data

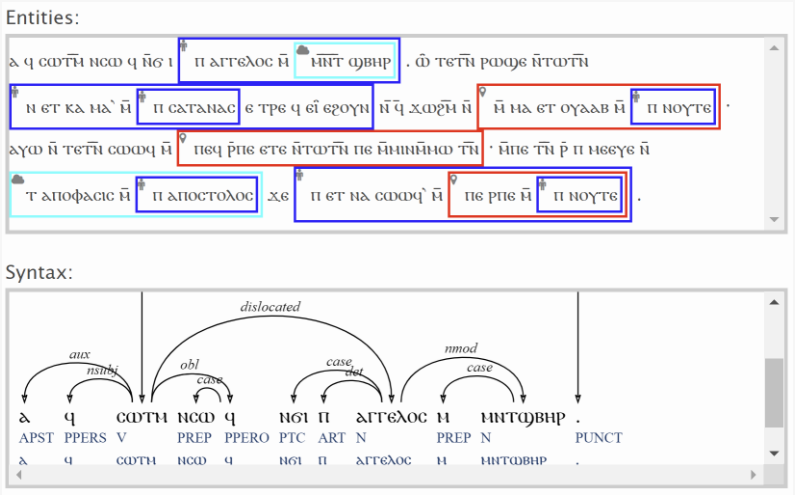
- ‘distant reading’ interactive visualizations of the entity data: https://copticscriptorium.org/entities/breakdown.html



Automatic tools demo

- You can now use the Coptic-NLP tool suite to do automatic entity recognition!
- Online demo here, code and tools on GitHub:
  - https://corpling.uis.georgetown.edu/coptic-nlp/





Future plans

- Add automatic entity recognition to all Coptic Scriptorium corpora
- Manually correct named entity linking for core corpora (excl. OT, NT)
- Disseminate automatic entity tagging tools and data online

Ⲙⲓⲱⲧⲛ̄ ⲧⲱⲛⲟⲩ!

Thank You

This work has been supported by

NEH Division of Preservation and Access (PW-51672-14)

NEH Office of Digital Humanities (HD-51907; HG-229371)

DFG

BMBF (German Federal Ministry of Education and Research)

Georgetown University

University of Oklahoma

The University of the Pacific

Humboldt University

Canisius College