Coptic SCRIPtorIuM

May 14, 2013 Workshop

Caroline T. Schroeder

Amir Zeldes

Research Objectives

Online, open source digital platform:

- For Corpus linguistics in Coptic
- Searchable text for historical and literary research
- Visualizations of Coptic texts and manuscripts
- Interoperable with current XML standards
- Collaborations with and contributions from other scholars

Sahidic **C**orpus **R**esearch: **I**nternet **P**la**t**f**o**rm for **I**nterdisciplinary m**u**ltilayer **M**ethods

Coptic Scriptorium is…

A corpus of Coptic texts:

- a richly-annotated database using ANNIS database technology
- TEI-XML versions of the texts for download
- Visualizations (including diplomatic editions) in HTML and inside the database

Tools for digitally processing texts in Coptic

Documentation

A platform where scholars and students will able to contribute texts and/or annotations and conduct research

Completely open source & open access

Texts

Proof-of-concept corpora:

Shenoute, Abraham Our Father (Canons vol. 3)

Select Coptic Apophthegmata Patrum

They contain:

- Diplomatic & normalized editions
- Translation
- Annotation

All of this is available online

http://go.pacific.edu/copticscriptorium

Tools: Encoding Converters

What you give: ASCII

1) Basic LaserCoptic or CMCL ASCII font
2) Exhaustive “Coptic” font (by G. Wurst & D. Van Damme)

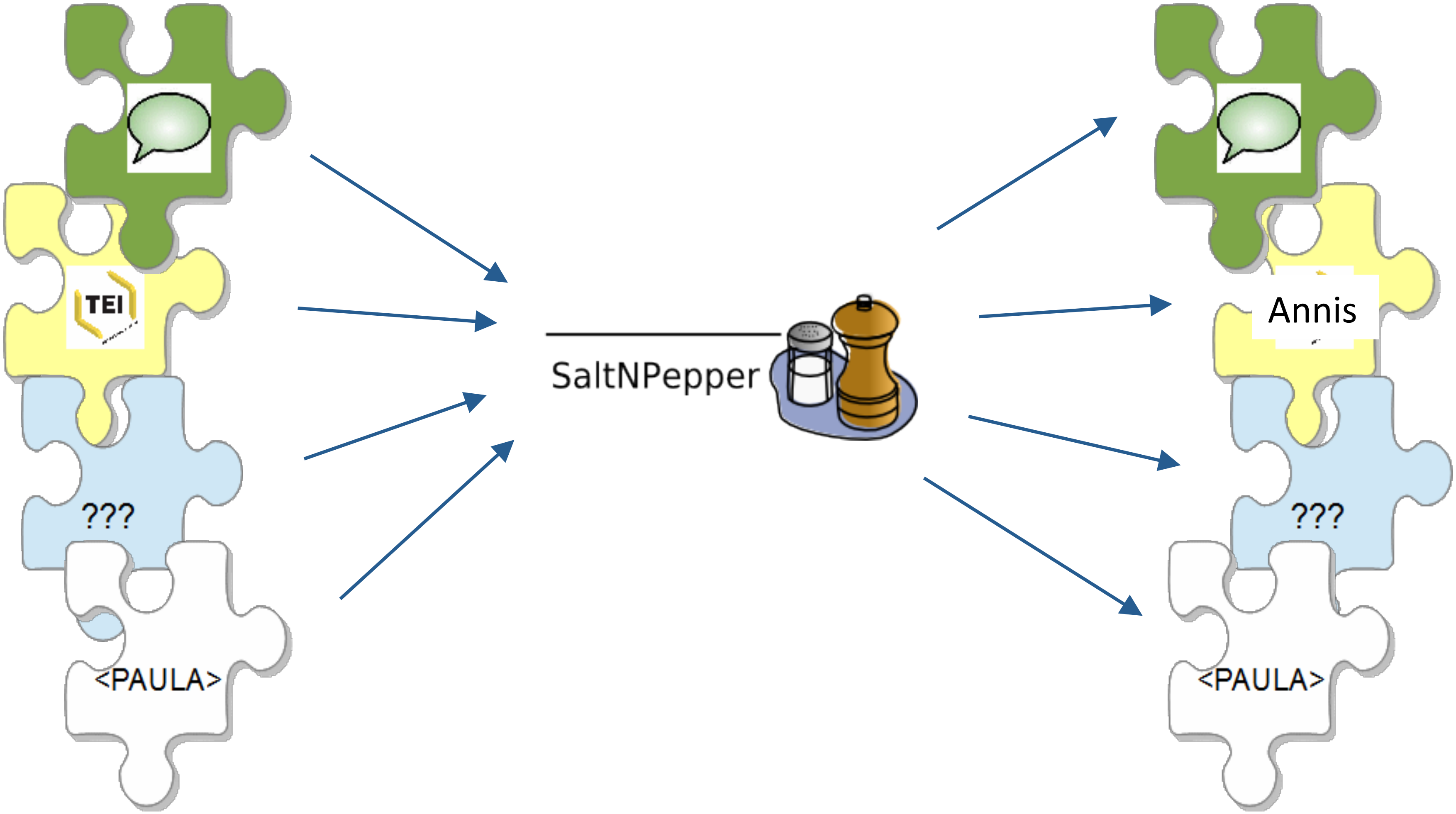
What you get: UTF-8

Fully compliant Unicode Coptic

etbe néeere de éhm
ntansvtm etbhhtoy

ⲉⲧⲃⲉ ⲛϣⲉⲉⲣⲉ ⲇⲉ ϣⲏⲙ
ⲛⲧⲁⲛⲥⲱⲧⲙ ⲉⲧⲃⲏⲏⲧⲟⲩ

File Format Converter (SaltNPepper)



Tokenization Scripts & Lexicon

Scripts process files of Coptic text to break the words into morphs

Based on CMCL Lexicon provided by T. Orlandi

Tokenization scheme follows morphemes established in B. Layton's grammar

Tokenization Script

What you give:

text as words

ⲁⲩϫⲟⲟⲥ ⲉⲧⲃⲉ ⲁⲡⲁ ⲡⲁⲡⲛⲟⲩⲧⲉ ϫⲉ ⲙⲉϥϭⲉⲡⲏ ⲉⲥⲉⲏⲣⲡ̄ .
ⲉϥⲙⲟⲟϣⲉ ⲇⲉ ⲛ̄ⲟⲩⲟⲉⲓϣ ⲁϥⲉⲓ ⲉϫⲛ̄ ⲟⲩⲕⲟⲗⲗⲩⲕⲓⲟⲛ ⲛ̄ⲗⲩⲥⲧⲏⲥ
ⲁⲩⲱ ⲁϥϭⲛ̄ⲧⲟⲩ ⲉⲩⲥⲱ . ⲁϥⲥⲟⲩⲱⲛϥ̄ ⲇⲉ ⲛ̄ϭⲓ ⲡⲉⲧⲟ ⲛ̄ⲁⲛϫⲱϫ
ⲁⲩⲱ ⲉϥⲥⲟⲟⲩⲛ ϫⲉ ⲙⲉϥⲥⲉ ⲏⲣⲡ̄ ⲁϥⲛⲁⲩ ⲉⲣⲟϥ ϩⲛ̄ ⲟⲩⲛⲟϭ ⲛ̄ϩⲓⲥⲉ
ⲁϥⲙⲟⲩϩ ⲛ̄ⲟⲩⲁⲡⲟⲧ ⲛ̄ⲏⲣⲡ̄ ⲉⲣⲉ ⲧⲉϥⲥⲏϥⲉ ϩⲛ̄ ⲧⲉϥϭⲓϫ ⲁⲩⲱ
ⲡⲉϫⲁϥ ⲙ̄ⲡϩⲗ̄ⲗⲟ ϫⲉ ⲉⲕⲧⲙ̄ⲥⲱ ϯⲛⲁϩⲟⲧⲃⲉⲕ . ⲁϥⲉⲓⲙⲉ ⲛ̄ϭⲓ
ⲡϩⲗ̄ⲗⲟ ϫⲉ ⲉϥⲟⲩⲱϣ ⲉⲉⲓⲣⲉ ⲛ̄ⲟⲩⲛ̄ⲧⲟⲗⲏ ⲛ̄ⲧⲉ ⲡⲛⲟⲩⲧⲉ ⲉϥⲟⲩⲱϣ
ⲉϯϩⲏⲩ ⲙ̄ⲙⲟϥ ⲁϥϫⲓⲧϥ̄ ⲁϥⲥⲟⲟϥ . ⲡⲁⲛϫⲱϫ ⲇⲉ ⲛ̄ⲗⲩⲥⲧⲏⲥ
ⲁϥⲙⲉⲧⲁⲛⲟⲓ ⲉϥϫⲱ ⲙ̄ⲙⲟⲥ ϫⲉ ⲕⲱ ⲛⲁⲓ ⲉⲃⲟⲗ ⲡⲁⲉⲓⲱⲧ ϫⲉ
ⲁⲓϯϩⲓⲥⲉ ⲛⲁⲕ . ⲡⲉϫⲁϥ ⲛⲁϥ ⲛ̄ϭⲓ ⲡϩⲗ̄ⲗⲟ ϫⲉ ϯⲡⲓⲥⲧⲉⲩⲉ
ⲉⲡⲛⲟⲩⲧⲉ ϫⲉ ϥⲛⲁⲉⲓⲣⲉ ⲛⲁⲕ ⲛ̄ⲟⲩⲛⲁ ⲉⲧⲃⲉ ⲡⲉⲓⲁⲡⲟⲧ ⲛ̄ⲏⲣⲡ̄ ϩⲙ̄
ⲡⲉⲓⲱⲛϩ̄ ⲙⲛ̄ ⲡⲉⲧⲛⲏⲩ . ⲡⲉϫⲁϥ ⲛⲁϥ ⲛ̄ϭⲓ ⲡⲁⲣⲭⲏⲗⲏⲥⲧⲏⲥ ϫⲉ
ϯⲡⲓⲥⲧⲉⲩⲉ ⲉⲡⲛⲟⲩⲧⲉ ϫⲉ ϫⲓⲛ ⲧⲉⲛⲟⲩ ⲛ̄ϯⲛⲁⲣ̄ ⲡⲉⲑⲟⲟⲩ ⲁⲛ
ⲛ̄ⲗⲁⲁⲩ . ⲡϩⲗ̄ⲗⲟ ⲇⲉ ⲁϥϫⲡⲟ ⲛⲁϥ ⲛ̄ⲧⲕⲟⲗⲗⲩϭⲉ ⲧⲏⲣⲥ̄ ϫⲉ
ⲁϥⲕⲱ ⲛ̄ⲥⲱϥ ⲙ̄ⲡⲉϥⲟⲩⲱϣ ⲉⲧⲃⲉ ⲡⲛⲟⲩⲧⲉ .

What you get:

morphemes ready for tagging

ⲁⲩϫⲟⲟⲥ	ⲁ
	ⲩ
	ϫⲟⲟ
	ⲥ
ⲉⲧⲃⲉ	ⲉⲧⲃⲉ
ⲁⲡⲁ	ⲁⲡⲁ
ⲡⲁⲡⲛⲟⲩⲧⲉ	ⲡⲁⲡⲛⲟⲩⲧⲉ
ϫⲉ	ϫⲉ
ⲙⲉϥϭⲉⲡⲏ	ⲙⲉ
	ϥ
	ϭⲉⲡⲏ
ⲉⲥⲉⲏⲣⲡ̄	ⲉ
	ⲥⲉ
	ⲏⲣⲡ
.	.
ⲉϥⲙⲟⲟϣⲉ	ⲉ
	ϥ
	ⲙⲟⲟϣⲉ
ⲇⲉ	ⲇⲉ

Tokenization Script

Give: formatted diplomatic text

```
ⲓ̈ⲁⲕⲱⲃ_ⲇⲉ_ⲛⲉ|ϥ|
ⲟⲩⲱϣ_ⲡⲉ`_ⲉ|
ⲧⲣⲉ|ϩⲉⲛ|ϣⲏ</cb></pb>
<pb xml:id="YA 520">
<cb>ⲣⲉ`_ϣⲱⲡⲉ`_ⲛⲁ|ϥ_
ⲉⲃⲟⲗ_ϩ̄ⲛ|ϩⲣⲁⲭ̇ⲏⲗ_·_
ⲛ̄|ⲑ|ⲉ_ⲛ̄|ⲛ|ⲉⲛⲧⲁ|ⲩ|
ϣⲱⲡⲉ_ⲛⲁ|ϥ_ⲉ
ⲃⲟⲗ_ϩ̄ⲛ_ⲗⲓⲁ_·_
```

Get: morphemes for tagging

Page, column, line breaks, word breaks aligned

cb					cb			
ⲡⲉ`	ⲉ	ⲧⲣⲉ	ϩⲉⲛ	ϣⲏⲣⲉ`		ϣⲱⲡⲉ`	ⲛⲁ	ϥ
ⲡⲉ`	ⲉⲧⲣⲉϩⲉⲛϣⲏⲣⲉ`					ϣⲱⲡⲉ`	ⲛⲁϥ	
lb		lb			lb			
[redacted]								
p								
YA519					YA520			
[redacted]								
ⲡⲉ`	ⲉ	ⲧⲣⲉ	ϩⲉⲛ	ϣⲏ	ⲣⲉ`	ϣⲱⲡⲉ`	ⲛⲁ	ϥ

Part of Speech Tagger

Give: tokenized text

ⲡ

ⲛⲟⲩⲧⲉ

ⲡⲉ

.

Get: POS-tagged text, following Layton

ⲡ	ART
ⲛⲟⲩⲧⲉ	N
ⲡⲉ	COP
.	PUNCT

Coarse grained tagger: 22 tags

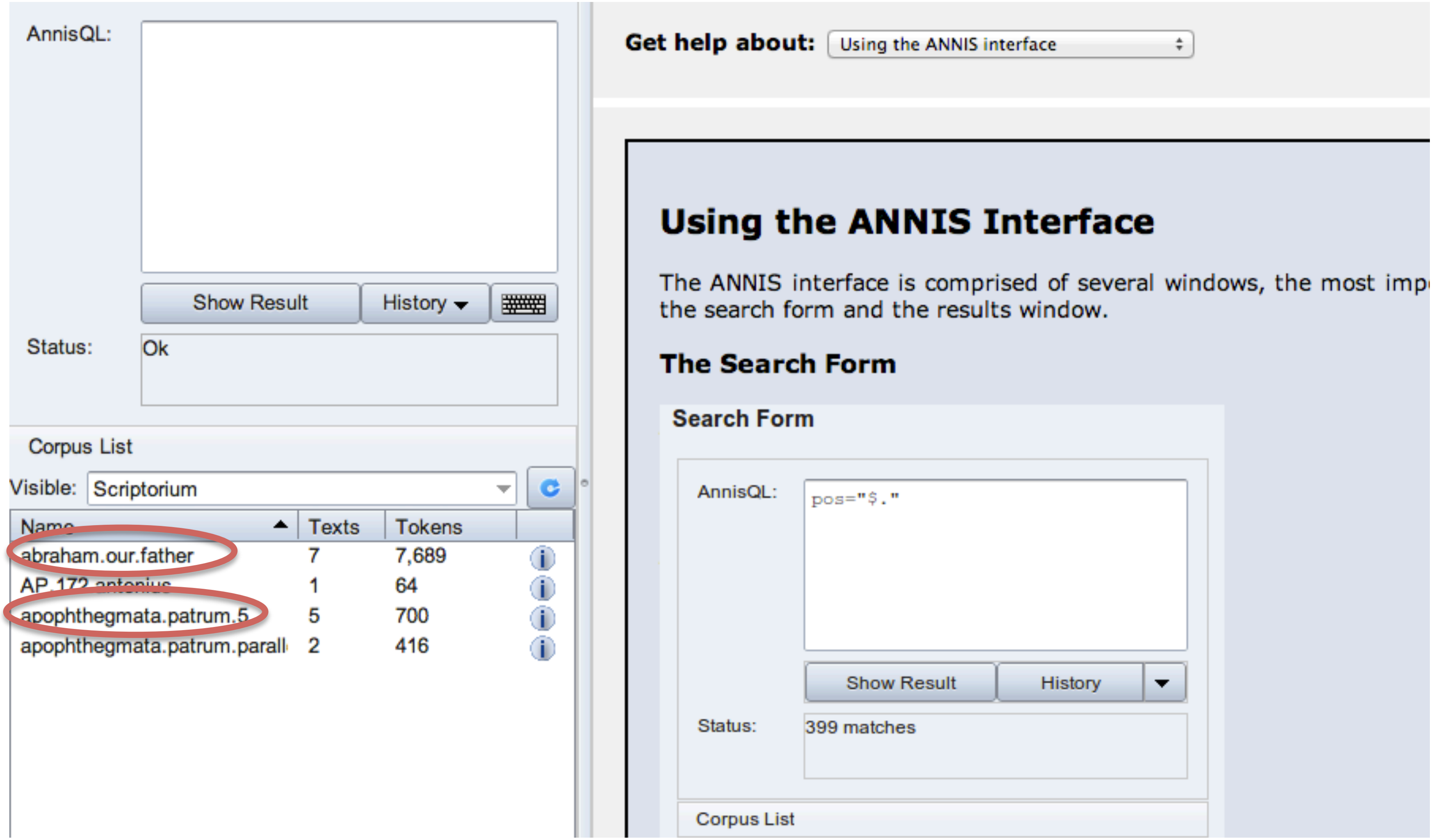
Fine grained tagger: 44 tags

All of these layers feed into the ANNIS database

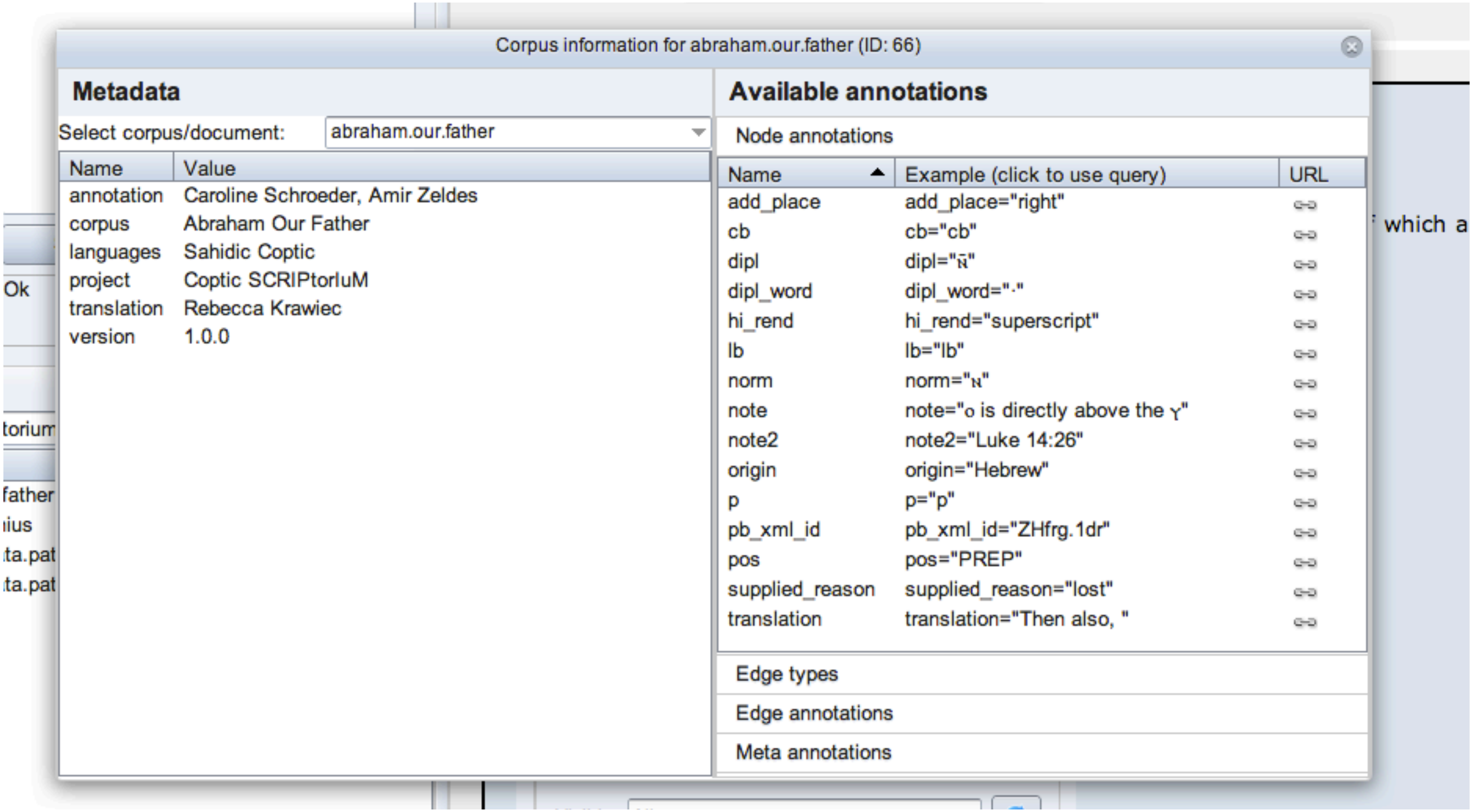
Open source (Apache 2.0)

Developed here at the Institute for German Language and Linguistics / SFB632

ANNIS & Search

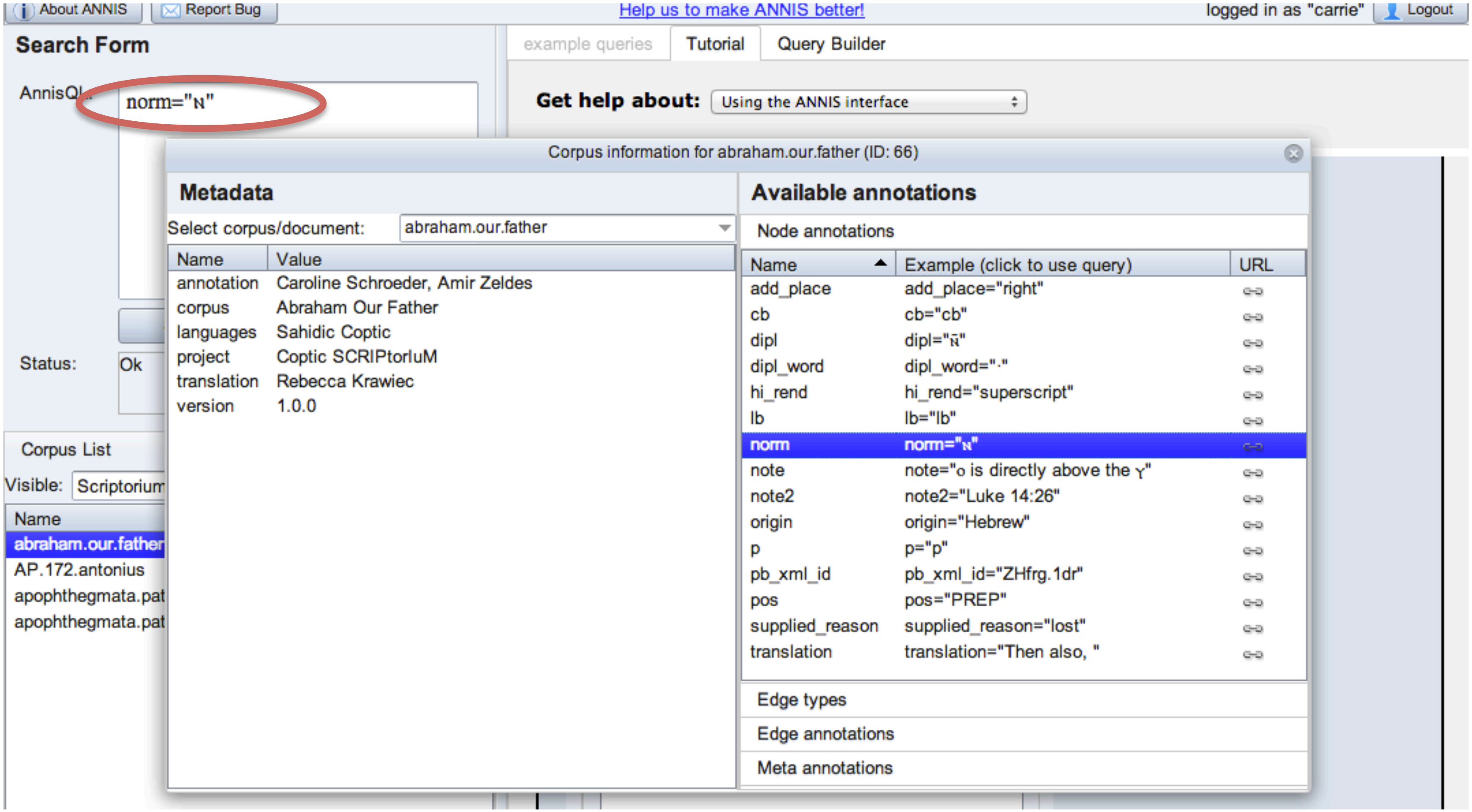


Corpus Information

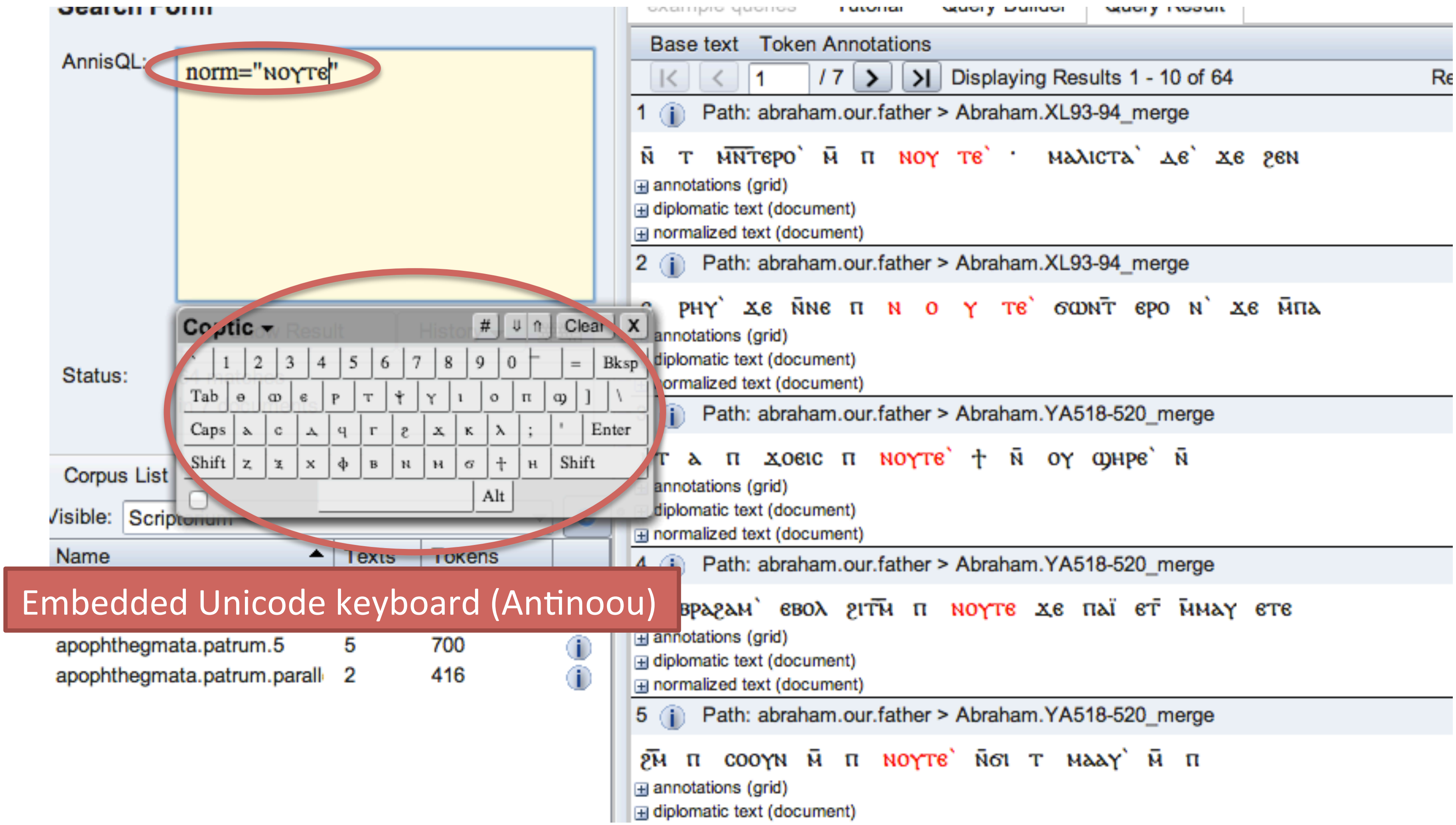


Nodes = the layers or fields of available information, e.g. words, line breaks in manuscripts, normalized tex.

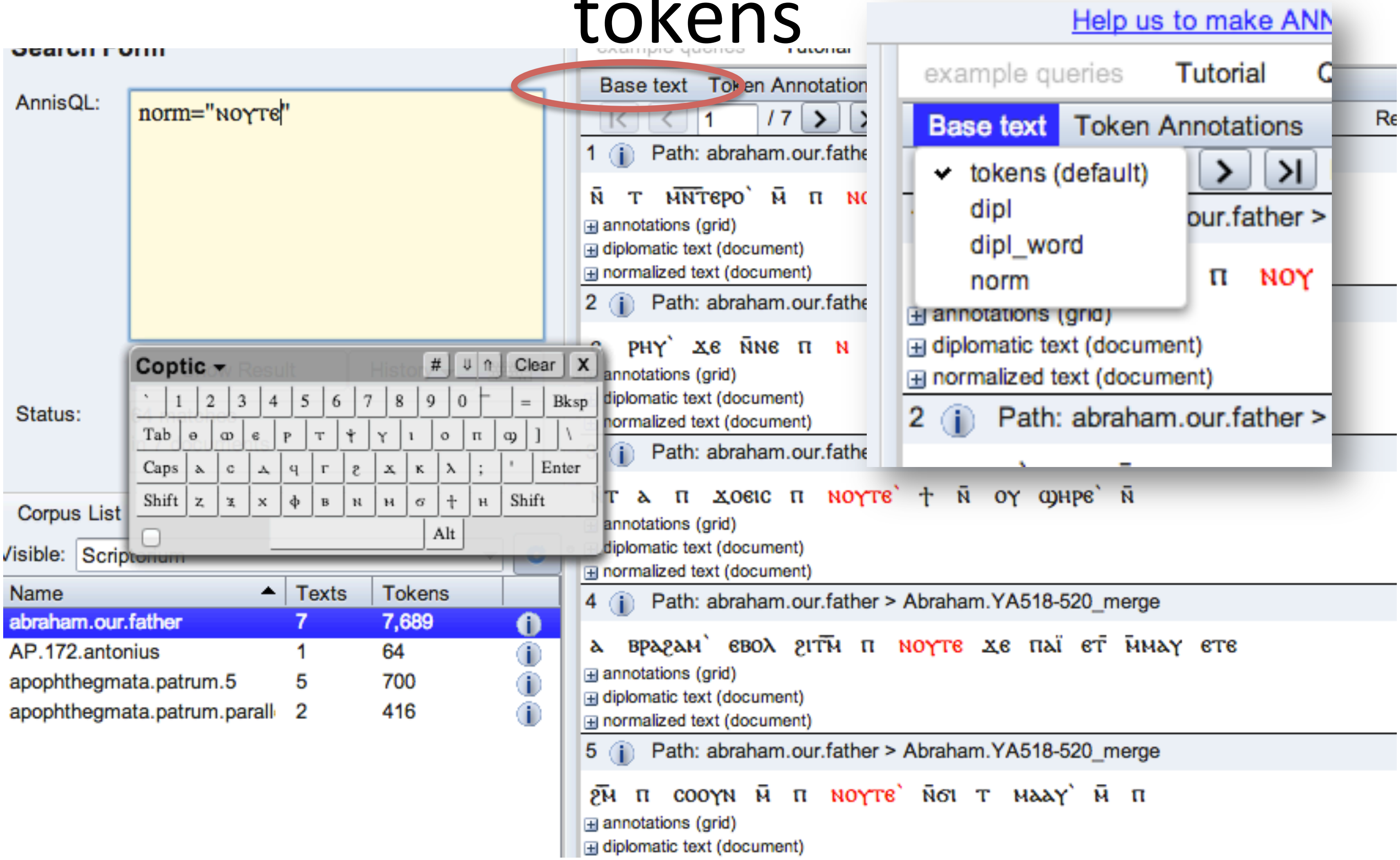
Example queries



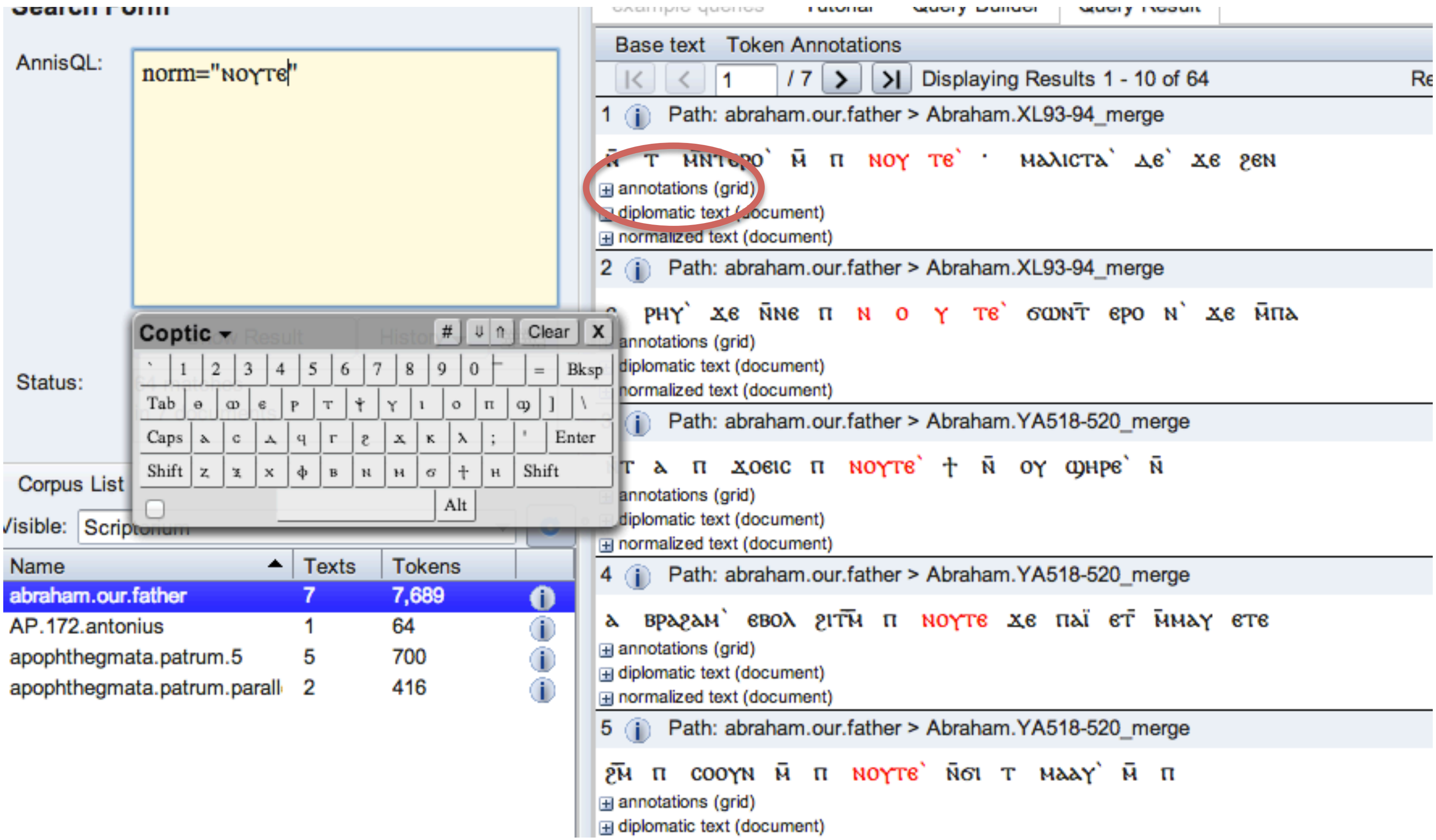
Edited previous query to search for ⲛⲟⲩⲧⲉ in the normalized text



Click on "Base text" to see search results as diplomatic, normalized, or tokens



Click on “annotations grid” to see all the annotations of the text



Grid of all annotations of text

1 / 7 Displaying Results 1 - 10 of 64 Result for query "norm="ⲛⲟⲩⲧⲉ""

3 Path: abraham.our.father > Abraham.YA518-520_merge

ⲛⲧ ⲁ ⲡ ϫⲟⲉⲓⲥ ⲡ ⲛⲟⲩⲧⲉ ϯ ⲛ ⲟⲩ ϣⲏⲣⲉ ⲛ

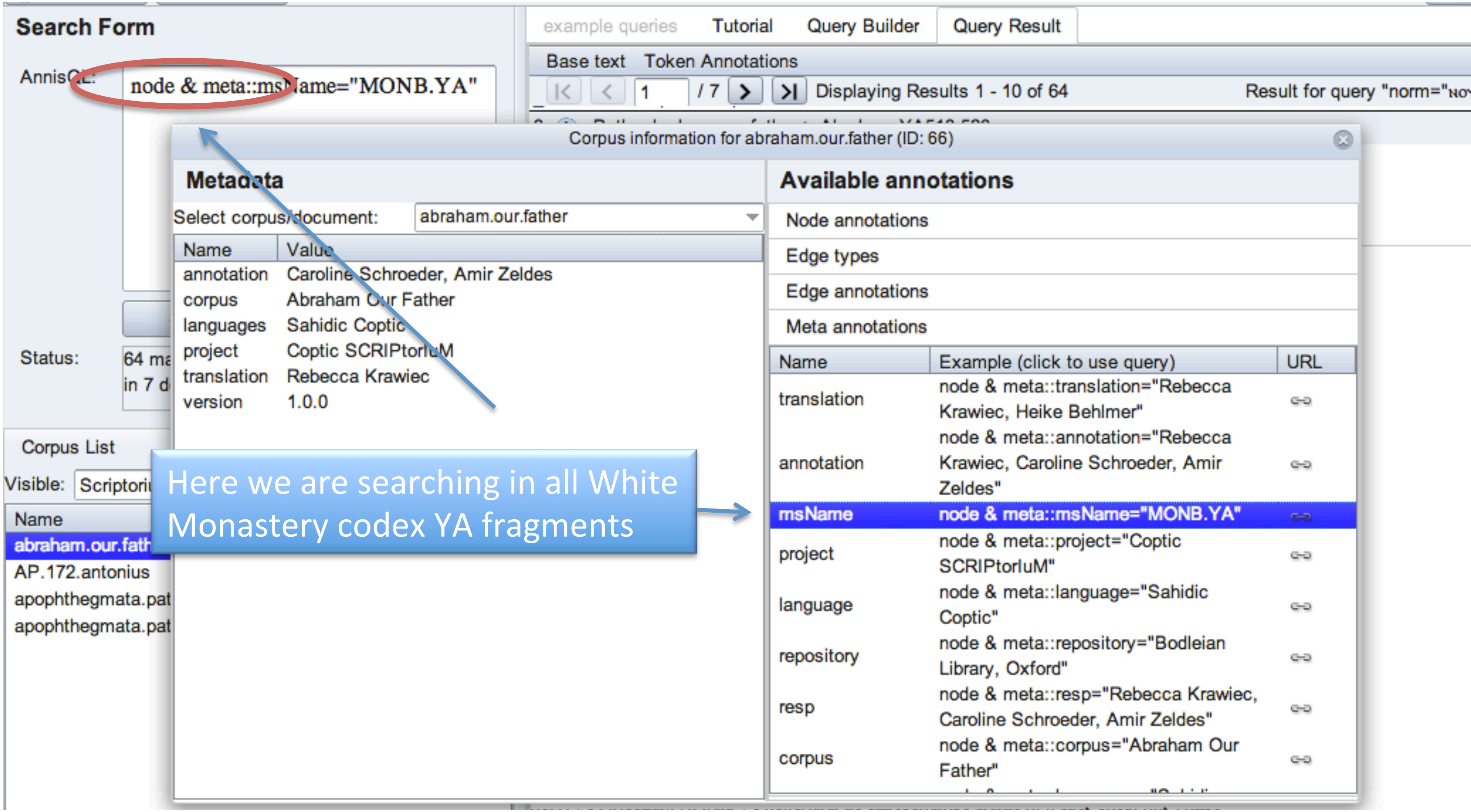
annotations (grid)

cb	cb										
dipl	ⲛ̄ⲧ	ⲁ	ⲡ	ϫⲟⲉⲓⲥ	ⲡ	ⲛⲟⲩⲧⲉ`	ϯ	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ`	ⲛ̄
dipl_word	ⲛ̄ⲧⲁⲡϫⲟⲉⲓⲥ				ⲡⲛⲟⲩⲧⲉ`		ϯ	ⲛ̄ⲟⲩϣⲏⲣⲉ`			ⲛ̄ⲁⲃⲣⲁ…
lb	lb				lb				lb		
norm	ⲛⲧ	ⲁ	ⲡ	ϫⲟⲉⲓⲥ	ⲡ	ⲛⲟⲩⲧⲉ	ϯ	ⲛ	ⲟⲩ	ϣⲏⲣⲉ	ⲛ
p	p										
pb_xml_id	YA518										
pos	CREL	APST	ART	N	ART	N	V	PREP	ART	N	PREP
translation	And thus the Lord God gave child(ren) to Abraham in his old age.										
tok	ⲛ̄ⲧ	ⲁ	ⲡ	ϫⲟⲉⲓⲥ	ⲡ	ⲛⲟⲩⲧⲉ`	ϯ	ⲛ̄	ⲟⲩ	ϣⲏⲣⲉ`	ⲛ̄

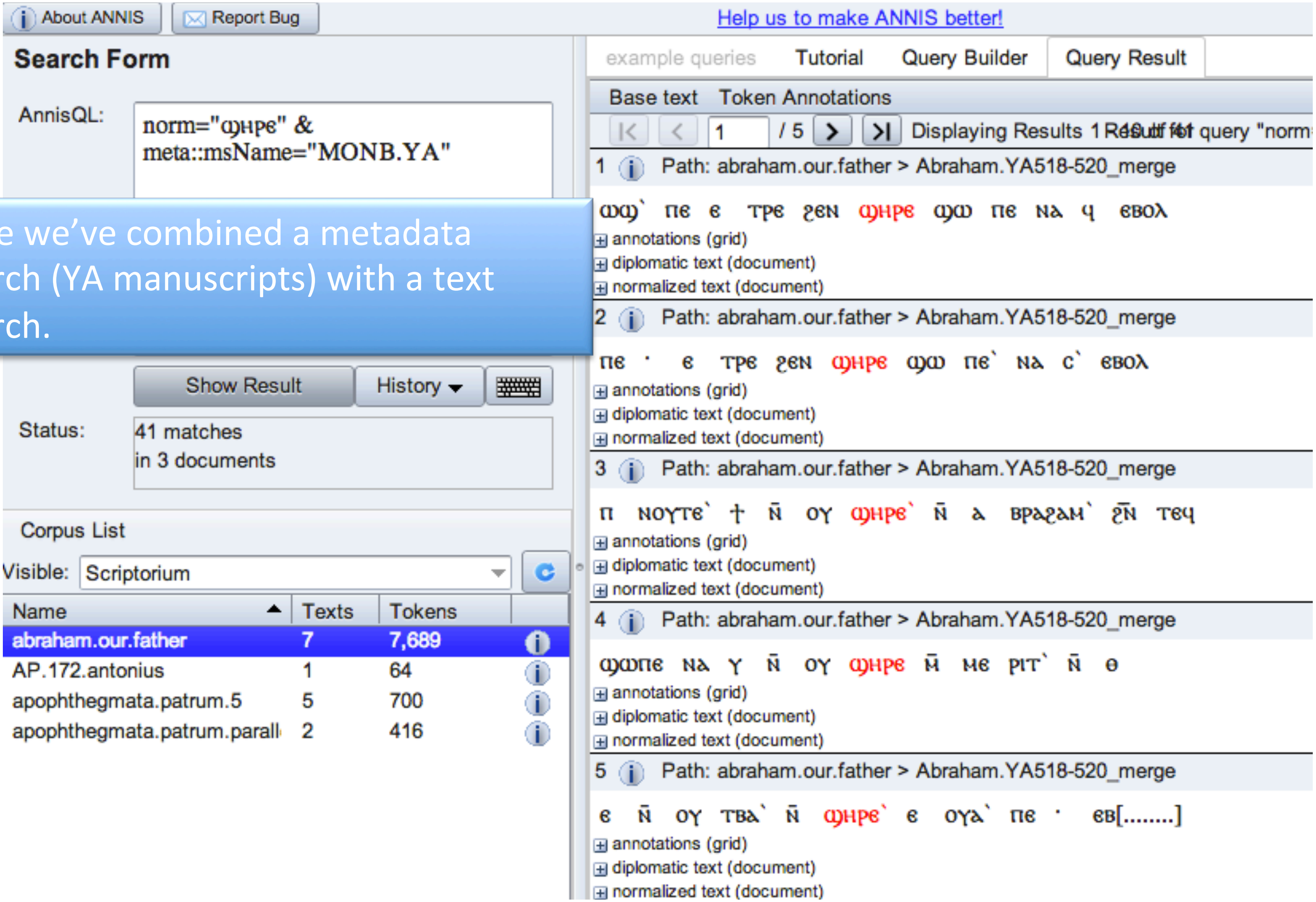
diplomatic text (document)

normalized text (document)

Click on sample searches to search metadata

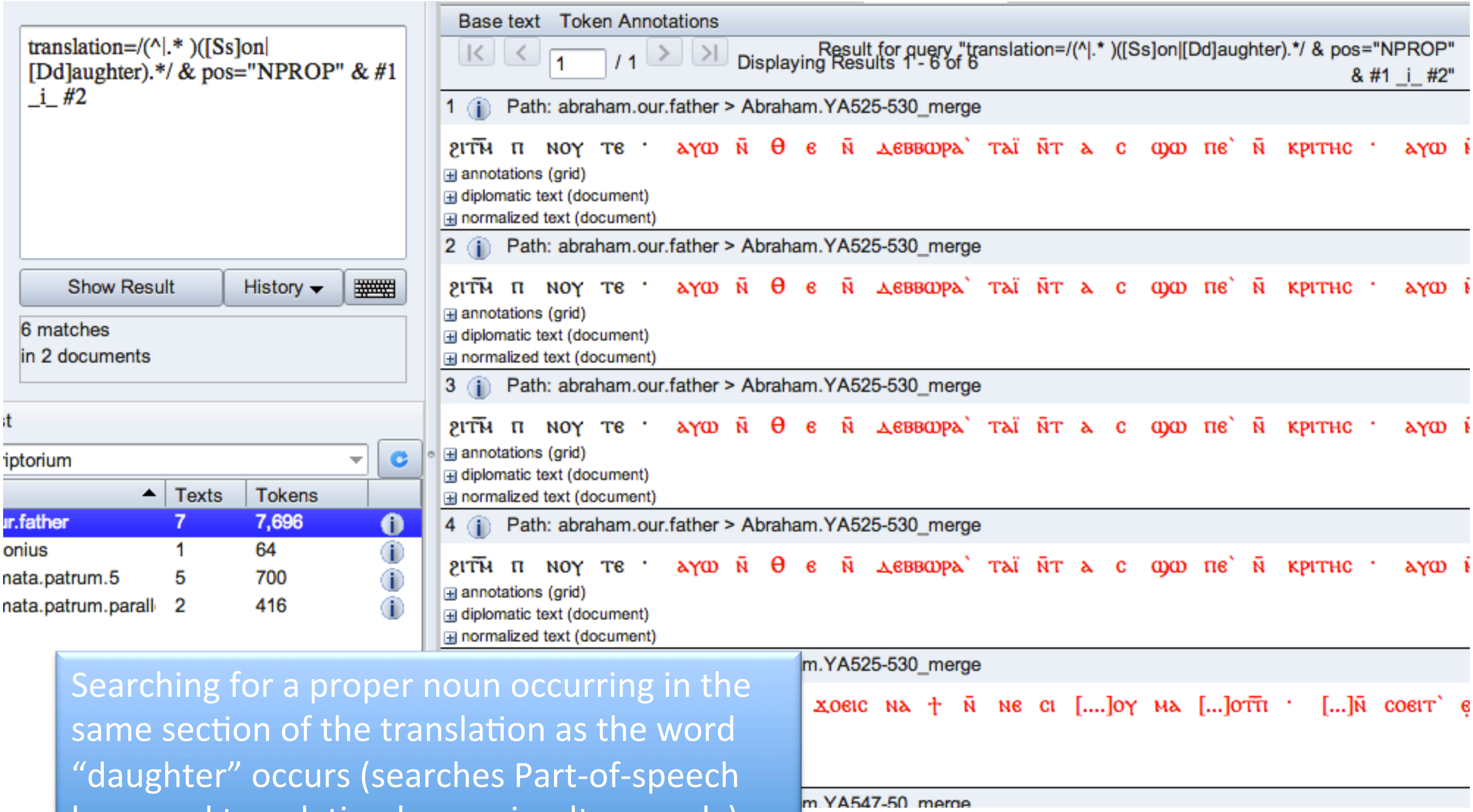


All instances of ϣнρε in YA



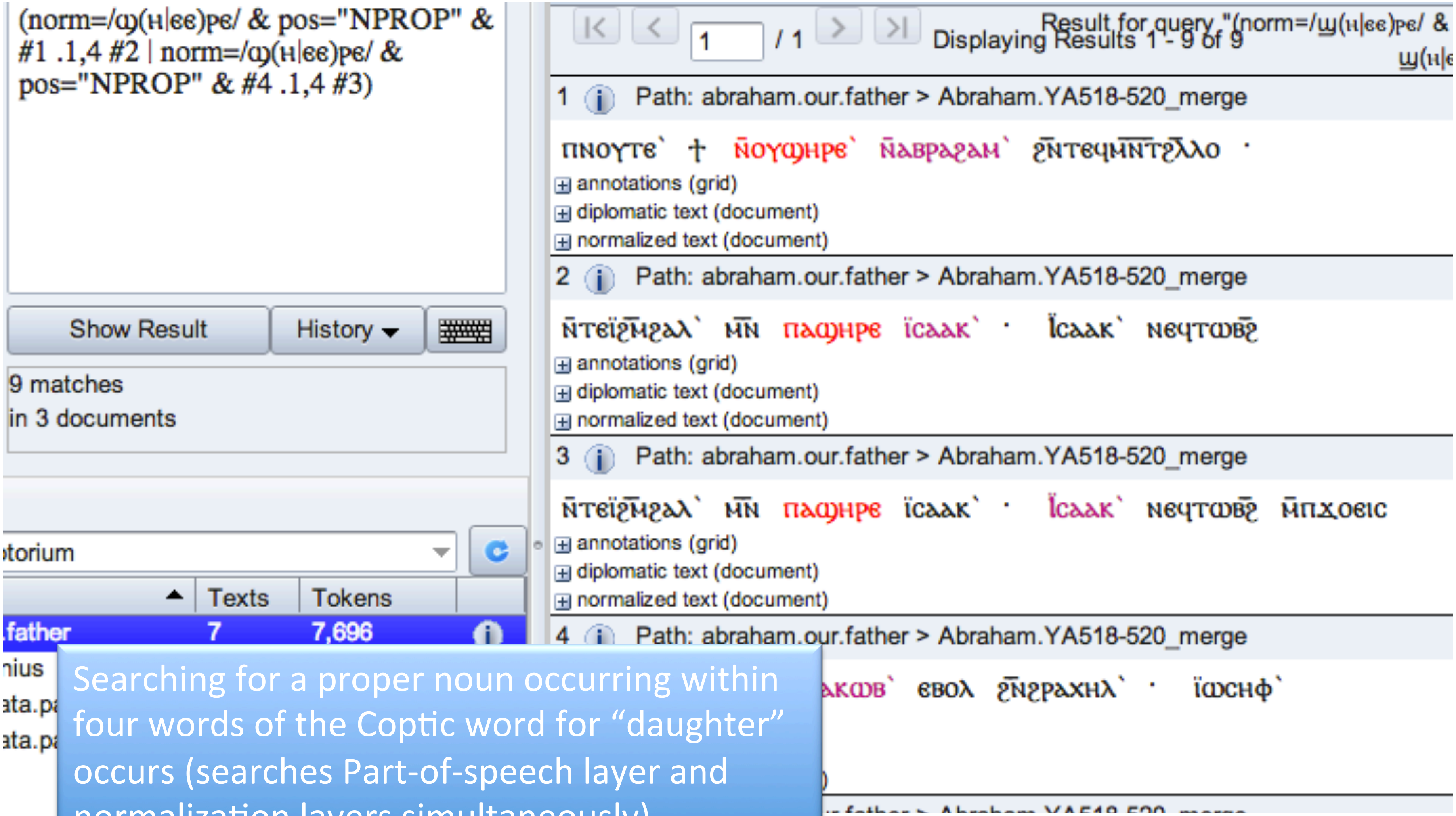
Here we've combined a metadata search (YA manuscripts) with a text search.

A little more complicated….



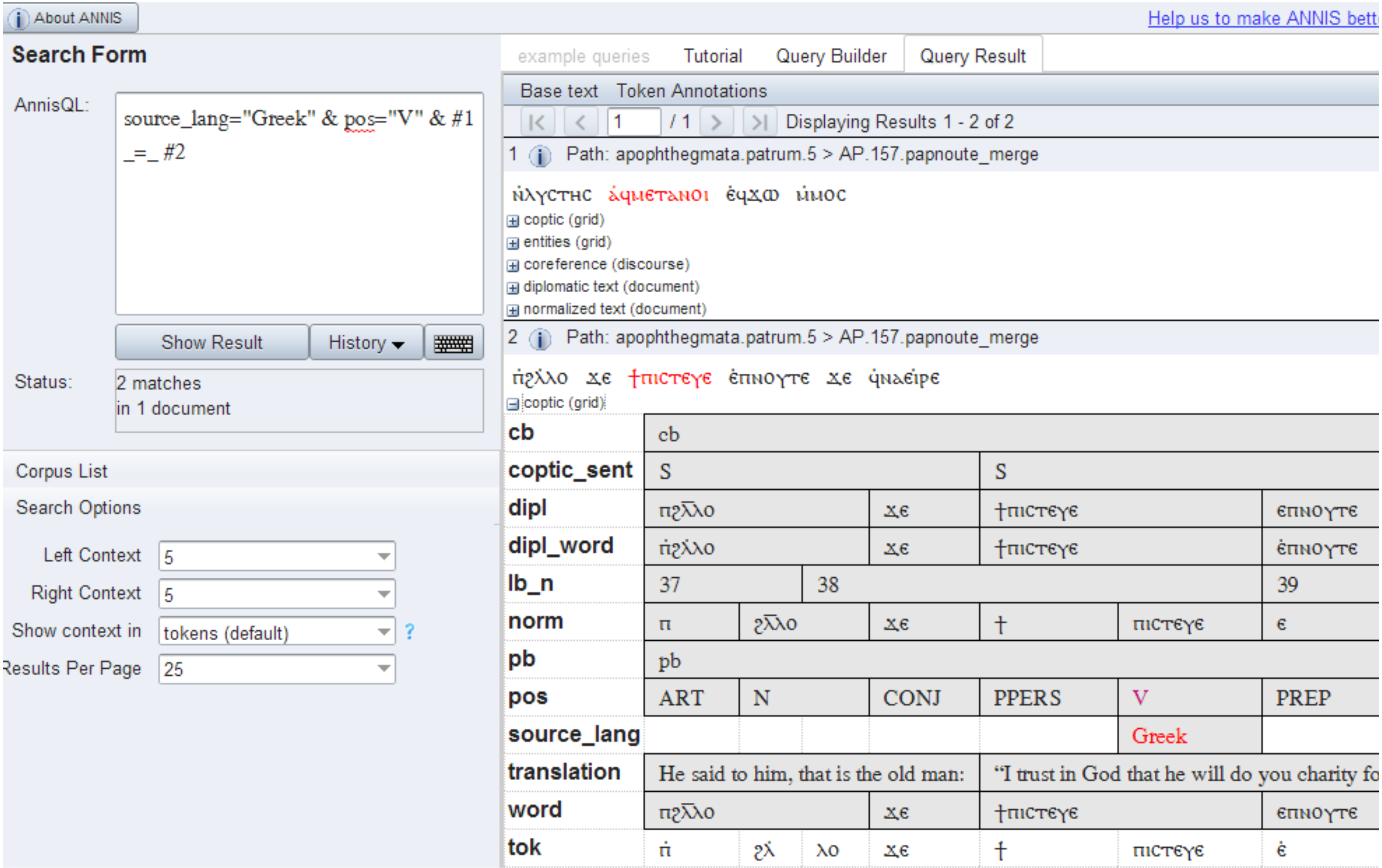
Searching for a proper noun occurring in the same section of the translation as the word "daughter" occurs (searches Part-of-speech layer and translation layers simultaneously)

Similar query searching Coptic normalized text

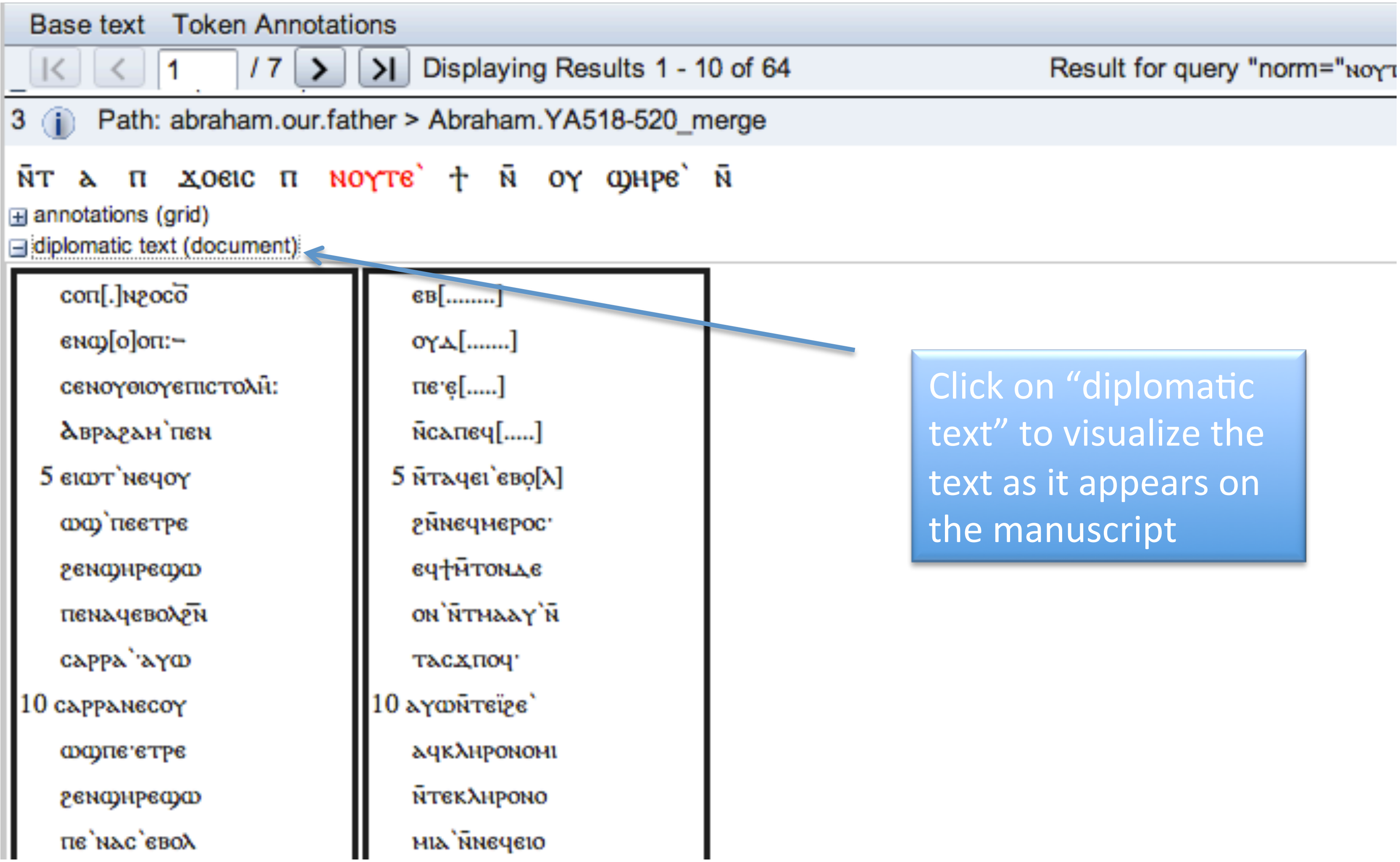


Searching for a proper noun occurring within four words of the Coptic word for "daughter" occurs (searches Part-of-speech layer and normalization layers simultaneously)

Search for Greek verbs



Visualization in ANNIS: diplomatic



Click on “diplomatic text” to visualize the text as it appears on the manuscript

Normalized text

ⲛ̄ⲧ ⲁ ⲡ ϫⲟⲉⲓⲥ ⲡ ⲛⲟⲩⲧⲉ` ϯ ⲛ̄ ⲟⲩ ϣⲏⲣⲉ` ⲛ̄

⊞ annotations (grid)

⊞ diplomatic text (document)

⊟ normalized text (document)

ⲥⲟⲡ[.]ⲛ ⲛϩⲟⲥⲟⲛ ⲉⲛϣⲟⲟⲡ : –

ⲥⲉⲛⲟⲩⲑⲓⲟⲩ ⲉⲡⲓⲥⲧⲟⲗⲏ :

ⲁⲃⲣⲁϩⲁⲙ ⲡⲉⲛⲉⲓⲱⲧ ⲛⲉϥⲟⲩⲱϣ ⲡⲉ ⲉⲧⲣⲉϩⲉⲛϣⲏⲣⲉϣⲱⲡⲉ ⲛⲁϥ ⲉⲃⲟⲗ ϩⲛⲥⲁⲣⲣⲁ · ⲁⲩⲱ ⲥⲁⲣⲣⲁ ⲛⲉⲥⲟⲩⲱϣ
ⲡⲉ · ⲉⲧⲣⲉϩⲉⲛϣⲏⲣⲉ ϣⲱⲡⲉ ⲛ ϯ

Abraham our father wished to have children with Sarah.

ⲛⲟⲩϣⲏⲣⲉ ⲛⲁⲃⲣⲁϩⲁⲙ ϩⲛⲧⲉϥ · ⲁϥϣⲱⲡⲉ
ⲛⲁⲩ ⲛⲟⲩϣⲏⲣⲉ ⲙⲙⲉⲣⲓⲧ ⲛⲧϩⲉ ⲛⲟⲩⲧⲃⲁ ⲛϣⲏⲣⲉ ⲉⲟⲩⲁ ⲡⲉ ·

ⲉⲃ[........]ⲟⲩⲇ[.......]ⲡⲉ · ⲉ[.....]ⲛⲥⲁ ⲡⲉϥ[.....] ⲛⲧⲁϥⲉⲓⲉⲃⲟⲗ ϩⲛ ⲛⲉϥⲙⲉⲣⲟⲥ · ⲉϥϯ ⲙⲧⲟⲛ ⲇⲉ ⲟⲛ ⲛⲧⲙⲁⲁⲩ
ⲛⲧⲁⲥϫⲡⲟϥ · ⲁⲩⲱ ⲛⲧⲉⲓϩⲉ ⲁϥⲕⲗⲏⲣⲟⲛⲟⲙⲓ ⲛⲧⲉⲕⲗⲏⲣⲟⲛⲟⲙⲓⲁ ⲛⲛⲉϥⲉⲓⲟⲧⲉ , ⲕⲁⲧⲁ ⲡⲉⲛⲧⲁⲩϫⲟⲟϥ
ⲛⲁⲃⲣⲁϩⲁⲙ ⲉⲃⲟⲗ ϩⲓⲧⲙⲡⲛⲟⲩⲧⲉ ϫⲉ ⲡⲁⲓ ⲉⲧⲙⲙⲁⲩ ⲉⲧⲉⲓⲥⲙⲁⲏⲗ ⲡⲉ ⲛϥⲛⲁⲕⲗⲏⲣⲟⲛⲟⲙⲉⲓ ⲙⲙⲟⲕ ⲁⲛ · ⲡⲗⲏⲛ
ϯⲛⲁⲥⲙⲟⲩ ⲇⲉ ⲟⲛ ⲉⲣⲟϥ ⲉⲧⲃⲏⲏⲧⲕ ⲧⲁⲁⲩⲝⲁⲛⲉ ⲙⲙⲟϥ ⲉⲙⲁⲧⲉ ⲛⲧⲁⲁⲁϥ ⲛⲟⲩⲛⲟϭ ⲛϩⲉⲑⲛⲟⲥ ⲉⲛⲁϣⲱϥ ·

Coreference

ⲁ ⲩ ϫⲟⲟ̇ ⲥ ⲉ̇ⲧⲃⲉ ⲁ̇ⲡⲁ ⲡⲁ̇ⲡ ⲛⲟⲩⲧⲉ ϫⲉ ⲙⲉ ϥ ϭⲉ ⲡⲏ ⲉ̇ ⲥⲉ ⲏ̇ⲣⲡ̇ . ⲻ ⲉ ϥ ⲙⲟⲟ̇ϣⲉ ⲇⲉ ⲛ̄ ⲟⲩ ⲟⲩⲟ̇ ⲉ̇ⲓϣ ⲁ ϥ ⲉ̇ⲓ ⲉ̇ϫⲛ ⲟⲩ ⲕⲟⲗⲗⲩⲕⲓ̇ⲟⲛ ⲛ̇ ⲗⲩⲥ ⲧⲏⲥ ⲁⲩⲱ̇ ⲁ̇ ϥ ϭⲛ̇ⲧ ⲟⲩ ⲉ̇ ⲩ ⲥⲱ . ⲁ̇ ϥ ⲥⲟⲩⲱⲛ ϥ̇ ⲇⲉ̇ ⲛ̇ϭⲓ̇ ⲡ ⲉⲧ ⲟ ⲛ ⲁ̇ⲛϫⲱϫ ⲁⲩⲱ̇ ⲉ̇ ϥ̇ ⲥⲟⲟ̇ⲩⲛ̇ ϫⲉ̇ ⲙⲉ ϥ ⲥⲉ ⲏ̇ⲣⲡ̇ · ⲻ ⲁ ϥ ⲛⲁⲩ ⲉ̇ⲣⲟ ϥ ϩⲛ̇ ⲟⲩ ⲛⲟϭ ⲛ̇ ϩⲓⲥⲉ ⲁ̇ ϥ ⲙⲟⲩϩ ⲛ ⲟⲩ ⲁ̇ⲡⲟⲧ ⲛ ⲏⲣ̇ⲡ̇ ⲉ̇ⲣⲉ ⲧⲉϥ ⲥⲏϥⲉ ϩⲛ̇ ⲧⲉϥ ϭⲓ̇ϫ · ⲁⲩⲱ̇ ⲡⲉϫⲁ ϥ ⲙ̇ ⲡ̇ ϩⲗ̇ ⲗⲟ ϫⲉ ⲉ̇ ⲕ ⲧⲙ̇ ⲥⲱ ϯ ⲛⲁ ϩⲟⲧⲃⲉ ⲕ .ⲻ ⲁ ϥ ⲉⲓⲙⲉ ⲛ̇ϭⲓ ⲡ̇ ϩⲗ̇ⲗⲟ ϫⲉ ⲉ̇ ϥ ⲟⲩⲱ̇ϣ ⲉ̇ ⲉ̇ⲓⲣⲉ ⲛ ⲟⲩ ⲛ̇ ⲧⲟⲗⲏ̇ ⲛ̇ⲧⲉ ⲡ̇ ⲛⲟⲩⲧⲉ ⲉ̇ ϥ ⲟⲩⲱ̇ϣ ⲉ̇ ϯ ϩⲏⲩ ⲙ ⲙⲟ ϥ ⲁ ϥ ϫⲓ̇ⲧ ϥ̇ ⲁ̇ ϥ ⲥⲟ̇ⲟ ϥ̇ . Ⲡ ⲁⲛϫⲱϫ ⲇⲉ ⲛ̇ ⲗⲩⲥ ⲧⲏⲥ ⲁ̇ ϥ ⲙⲉⲧⲁⲛⲟⲓ ⲉ̇ ϥ ϫⲱ ⲙ̇ⲙⲟ ⲥ ϫⲉ ⲕⲱ ⲛⲁ ⲓ̇ ⲉ̇ⲃⲟⲗ ⲡⲁ ⲉ̇ⲓⲱⲧ ϫⲉ ⲁ ⲓ̇ ϯ ϩⲓ̈ⲥⲉ ⲛⲁ ⲕ

Papnoute: referenced by name, the old man, “my father”

We also have the HTML visualizations you saw online

Wait, there's more…

Download TEI (Text Encoding Initiative) XML compatible (text and metadata)

Files validate to EpiDoc TEI schemas to ensure interoperability

Open Source & Open Access

Texts and Tools:

- Freely available
- Can be modified
- Attribution to sources required (cc-by)

Long term goals

Expand corpus (with collaborators)

Improve existing taggers and tokenizer

Expand tools:

- Named entity recognition
- Language of origin (Greek, Latin, Hebrew…)
- Lemmatizer

Collaborative platform

- Text contributions
- Annotation layers

What we want from you

Texts

Feedback: what do you want from a database?

Feedback: what visualizations do you want?

Feedback: comments on tools

Funding & Support has been provided by:

- A grant from the Pacific Fund, the University of the Pacific
- A grant from the Faculty Research Council, the University of the Pacific
- The Department of Religious and Classical Studies, The University of the Pacific
- The Institute for German Language and Linguistics, Humboldt University

C. Schroeder and A. Zeldes gratefully acknowledge the following individuals and organizations for their contributions to the project:

- The Corpus dei Manoscritti Copti Letterari (Coptic lexicon and other materials)
- Eric E. Johnson, Tibco Software
- Rebecca S. Krawiec, Ph.D., Canisius College (Abraham Our Father annotations and English translation)
- Tito Orlandi, Ph.D., Sapienza University, Rome; the University of Hamburg
- Janet Timbie, Ph.D., Catholic University of America (Apophthegmata Patrum annotations)
- The ANNIS team at Humboldt University
- Alex Dickerson, the University of the Pacific, student