GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

BILDUNG
UND
RELIGION
SFB 1136

# Digital Text Reuse Detection in Coptic Literature
## A Progress Report

So Miyagawa, Marco Büchler, Heike Behlmer
University of Göttingen

MONB.XV, pag. 13, K 939, recto, Österreichische Nationalbibliothek, Vienna

# Aspects of study

Digital humanities and corpus linguistics study on Shenoute and Besa's text reuses (quotations, allusions, etc.) using a digital text reuse detection tool

- Formal characteristics of text reuses of Shenoute and Besa
  - Alternation in quotations
  - Quotation signals

- More unknown text reuses
  - Evaluation of text reuse detection tool (TRACER)
  - Contribution to Coptic Linguistics and Septuagint Studies

- Functions of the text reuses

# SFB 1136 (2015–2020)

- German Research Foundation (DFG) Collaborative Research Centre (SFB) 1136 “Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam”
- Project Area B 05 "Biblical Interpretation and Educational Traditions in the Coptic-speaking Egyptian Christianity of Late Antiquity: Shenoute, Canon 6" (Heike Behlmer, Julien Delhez, and So Miyagawa)

BILDUNG
UND
RELIGION
SFB 1136

Deutsche
Forschungsgemeinschaft
DFG

# eTRAP (2015–2019)

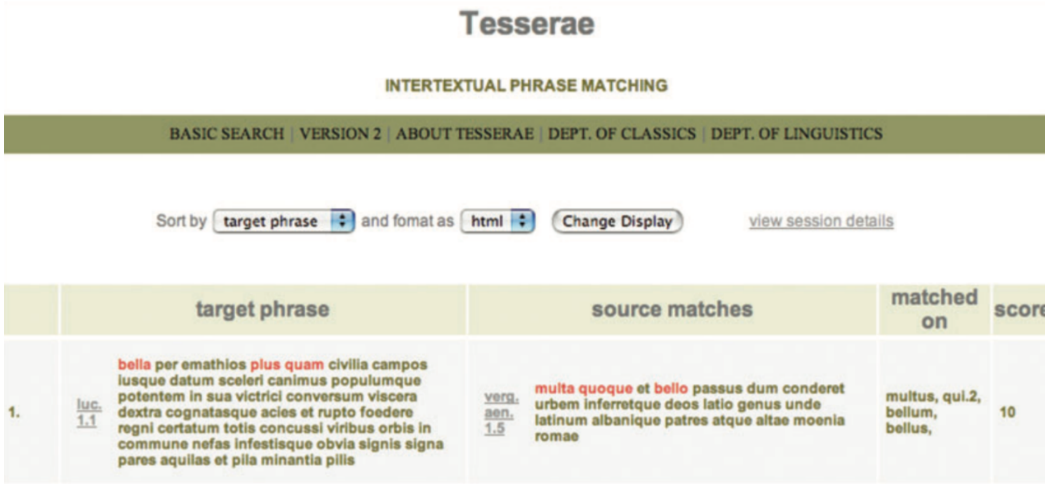
- electronic Text Reuse Acquisition Project (PI: Marco Büchler)
- An Early Career Research Group funded by the German Federal Ministry of Education and Research (BMBF)
- Producing the TRACER software for digital text reuse detection

eTRAP
Electronic Text Reuse Acquisition Project

Bundesministerium
für Bildung
und Forschung

# History of digital text reuse detection

- Data mining: 1980’s–
- Text mining and NLP: late 1980’s–
- Plagiarism checker: 1990’s–



Tesserae

INTERTEXTUAL PHRASE MATCHING

BASIC SEARCH | VERSION 2 | ABOUT TESSERAE | DEPT. OF CLASSICS | DEPT. OF LINGUISTICS
Sort by
target phrase
and fomat as
html
Change Display
view session details

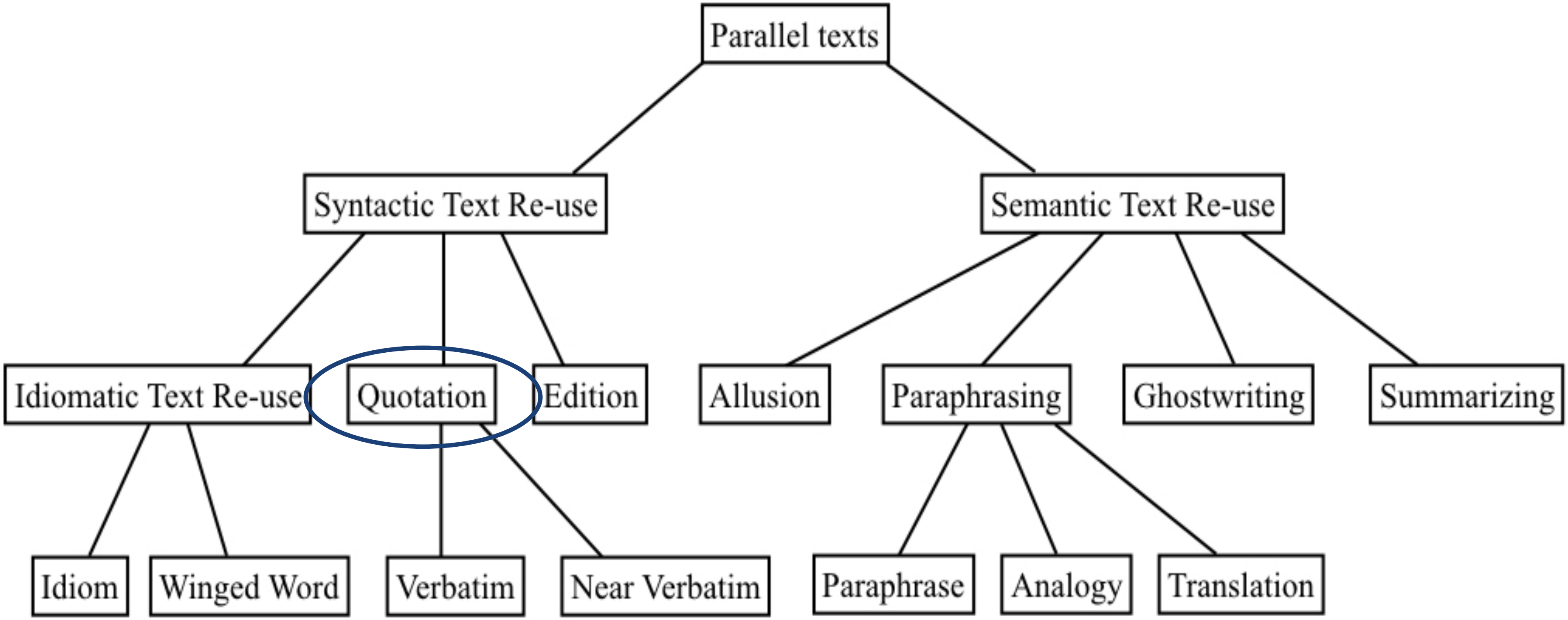
| | | target phrase | | source matches | matched on | score |
|---|---|---|---|---|---|---|
| 1. | luc. 1.1 | bella per emathios plus quam civilia campos iusque datum sceleri canimus populumque potentem in sua victrici conversum viscera dextra cognatasque acies et rupto foedere regni certatum totis concussi viribus orbis in commune nefas infestisque obvia signis signa pares aquilas et pila minantia pilis | verg. aen. 1.5 | multa quoque et bello passus dum conderet urbem inferretque deos latio genus unde latinum albanique patres atque altae moenia romae | multus, qui.2, bellum, bellus, | 10 |

Coffee et al. (2012:3)

- Tesserae (since 2006)
  - Detects text reuses in Latin literature
  - E.g., found 93 text reuses from Vergil’s Aeneid in Lucan's Civil War, including 46 unknown in previous studies (Coffee et al. 2012:401)

# Text reuse studies

A field of computational linguistics



Parallel texts

Syntactic Text Re-use

Semantic Text Re-use

Idiomatic Text Re-use

Quotation

Edition

Allusion

Paraphrasing

Ghostwriting

Summarizing

Idiom

Winged Word

Verbatim

Near Verbatim

Paraphrase

Analogy

Translation

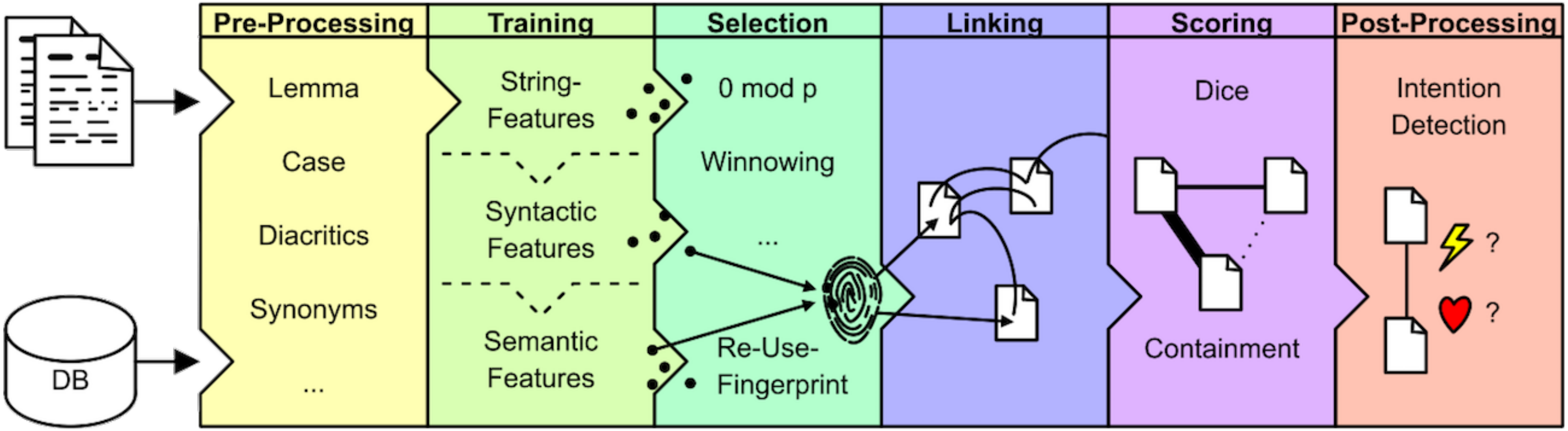
Büchler (2013), Franzini et al. (2016), Miyagawa et al. (2020)

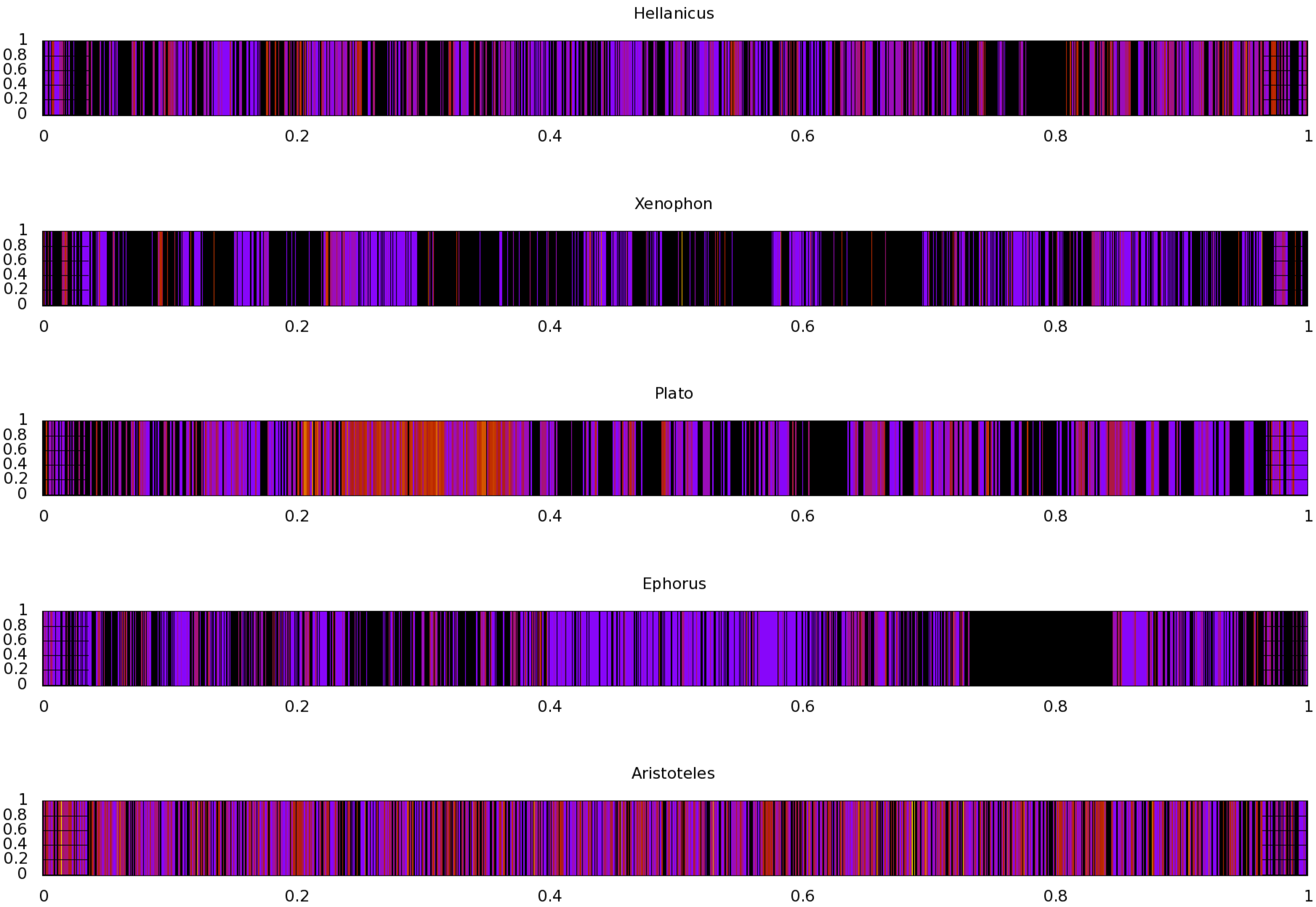
# TRACER

Bundesministerium
für Bildung
und Forschung

eTRAP
Electronic Text Reuse Acquisition Project

- The latest computer program to detect text reuses
- Developed by the eTRAP team, directed by Marco Büchler, and belonging to the Institute of Computer Science (University of Göttingen)
- Can automatically detect text reuses irrespective of how big the text corpus is


DB
Pre-Processing
Lemma
Case
Diacritics
Synonyms
...
Training
String-Features
Syntactic Features
Semantic Features
Selection
0 mod p
Winnowing
...
Re-Use-Fingerprint
Linking
Scoring
Dice
Containment
Post-Processing
Intention Detection
?
?


Hellanicus
Xenophon
Plato
Ephorus
Aristoteles
0
0.2
0.4
0.6
0.8
1

# Shenoute, *Canon* 6

Attestation:

- 6 codices (siglum: MONB = Monastero Bianco/White Monastery)
- MONB.XF, XL, XM, XV, YJ, YK (none complete)
- appr. 325 existing pages

Works (minimum of 5), incipits:

- He Who Sits upon His Throne
- Remember, O Brethren
- Is It Not Written
- Then Am I Not Obliged
- People Have Not Understood

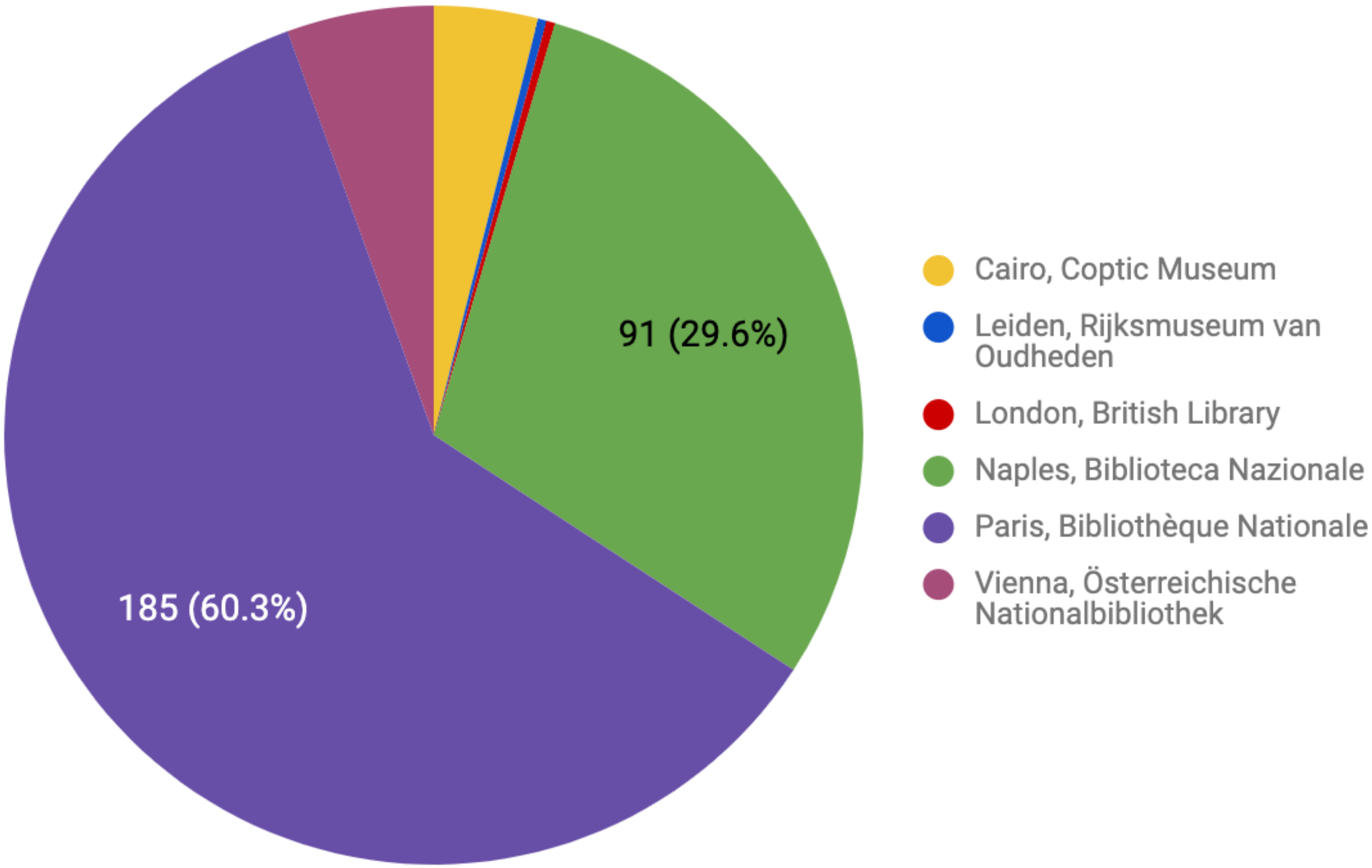
Themes:

- Accusations against Shenoute of excessive violence (and his defense)
- Shenoute’s illness
- Problems of the female community
- Monastic rules

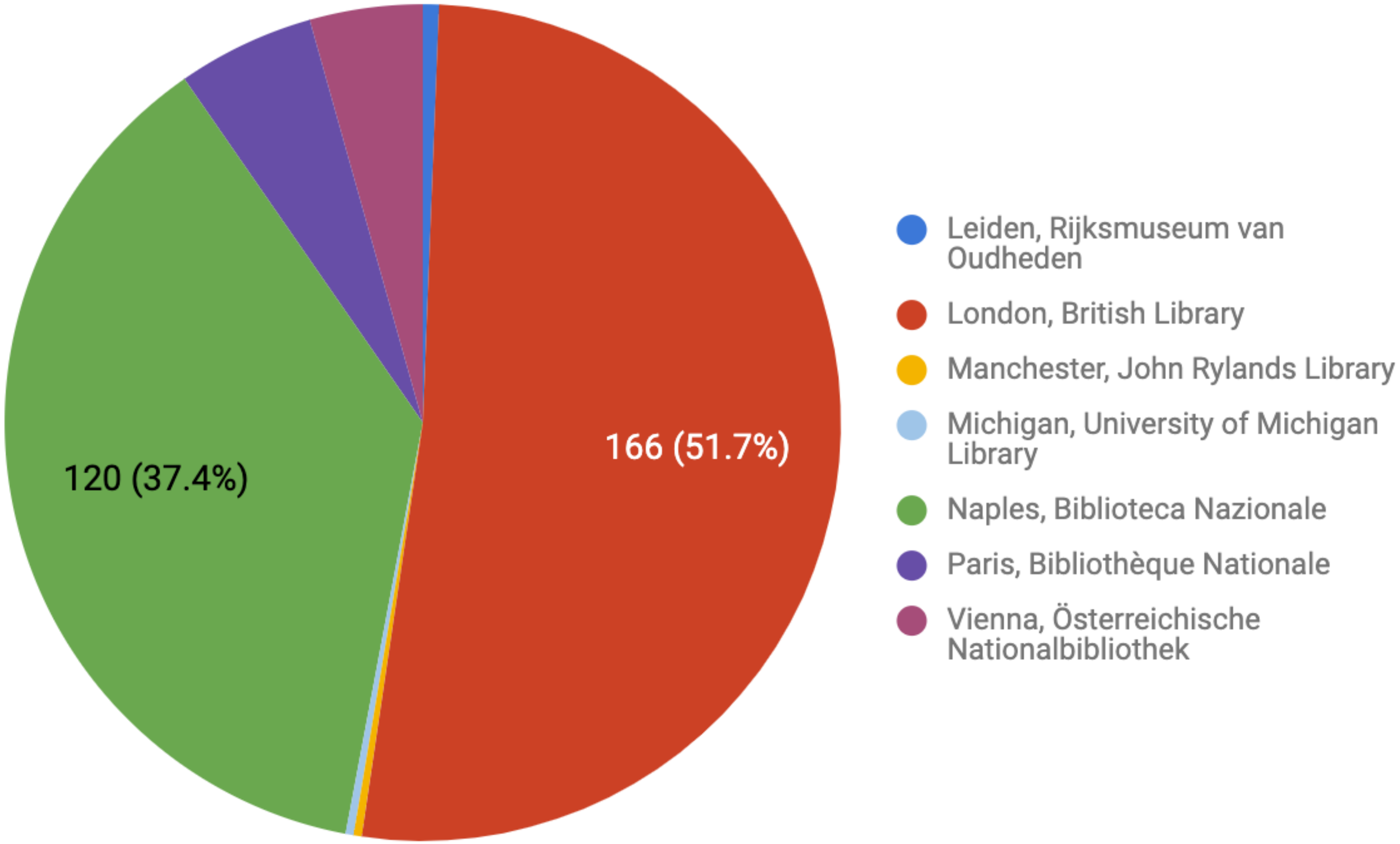
ⲀⲠⲀ ϢⲈⲚⲞⲨⲦⲈ

Shenoute
Relief, Bode Museum, Berlin

# Shenoute, *Canon* 6

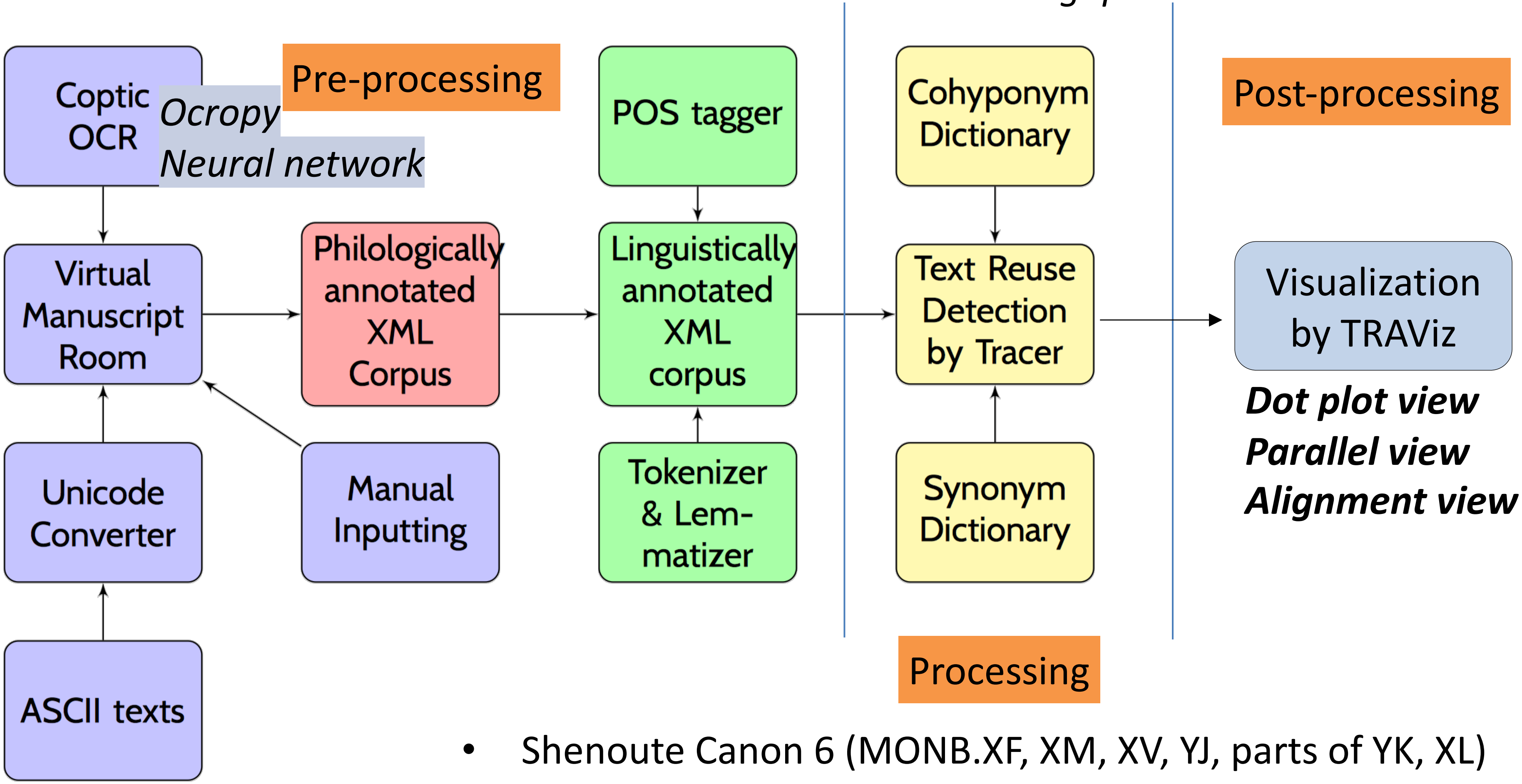


# Besa, *Letters and Sermons* (Codices A and B)



# Workflow of the project

***Coptic WordNet***
*Göttingen,*
*Oslo and Singapore*



Coptic OCR

*Ocropy*
*Neural network*

Pre-processing

Virtual Manuscript Room

Unicode Converter

ASCII texts

Philologically annotated XML Corpus

Manual Inputting

POS tagger

Linguistically annotated XML corpus

Tokenizer & Lemmatizer

Cohyponym Dictionary

Text Reuse Detection by Tracer

Synonym Dictionary

Processing

Post-processing

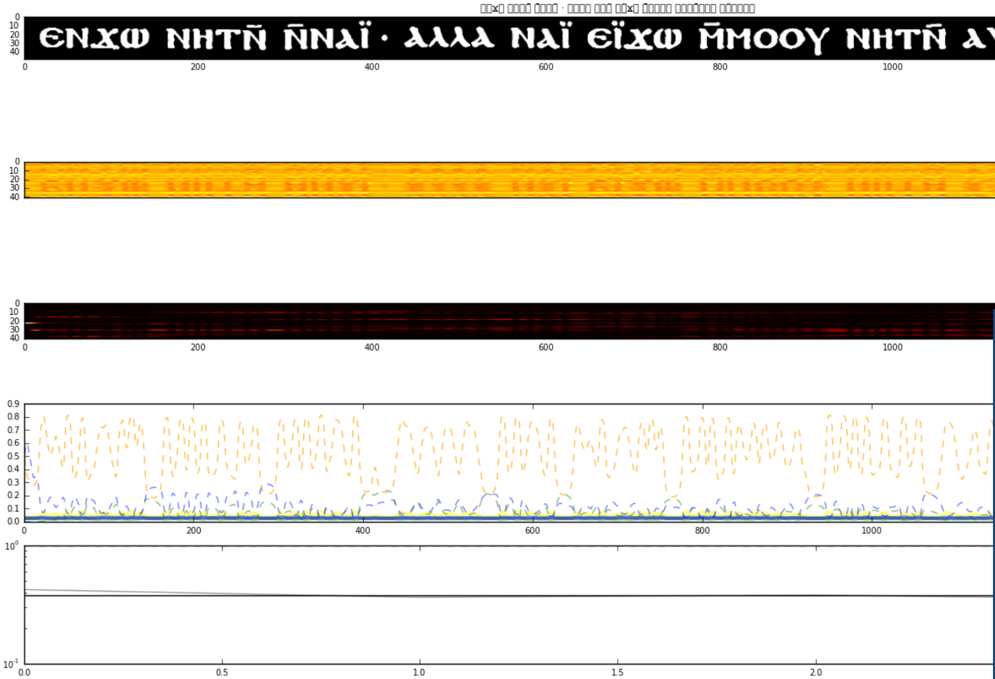
Visualization by TRAViz

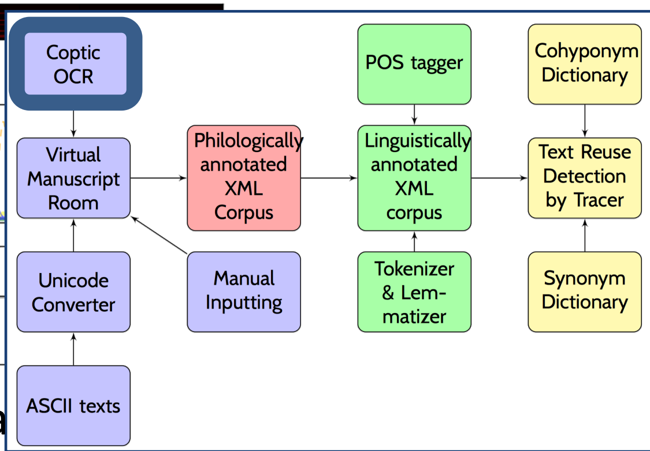
***Dot plot view***
***Parallel view***
***Alignment view***

- Shenoute Canon 6 (MONB.XF, XM, XV, YJ, parts of YK, XL)
- Besa (MONB.BA, BB, some other codices)

# OCR with neural networks

- Lincke et al. (2019). “Optical Character Recognition for Coptic fonts: A multi-source approach for scholarly editions.” In: Proceedings of DATeCH 2019.
- Miyagawa et al. (2019). “Optical character recognition of typeset Coptic text with neural networks.” *Digital Scholarship in the Humanities*, Volume 34.

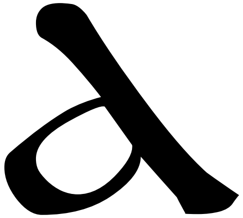



Coptic OCR
Virtual Manuscript Room
Unicode Converter
ASCII texts
Philologically annotated XML Corpus
Manual Inputting
POS tagger
Linguistically annotated XML corpus
Tokenizer & Lem-matizer
Cohyponym Dictionary
Text Reuse Detection by Tracer
Synonym Dictionary

Team: Kirill Bulert, Eliese Sophia-Lincke, Marco Büchler, So Miyagawa

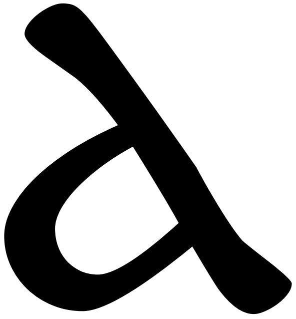
# Unicode converter

- Zeldes, Amir, Caroline Schroeder and So Miyagawa (2013-2016) “recode_coptic.pl” (https://github.com/CopticScriptorium/converters/blob/master/recode_coptic.pl)
- Can convert old ASCII representation fonts of Coptic such as “Coptic,” “CopticLS", "CMCL", "NagHammadi", "avva_shenouda", "low" fonts into the proper Coptic Unicode.
- Written in Perl Script



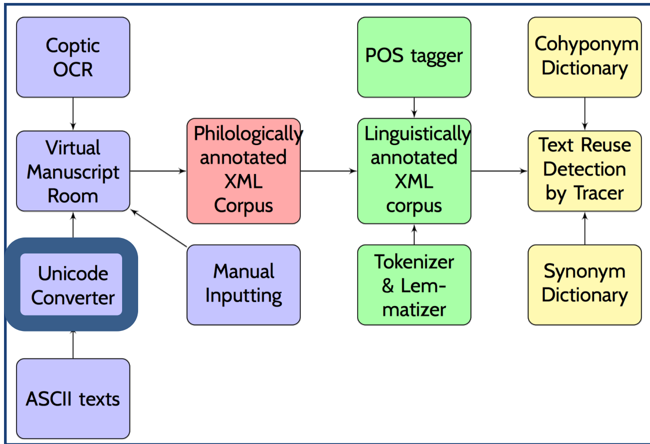
U+0041

(actually ASCII A)

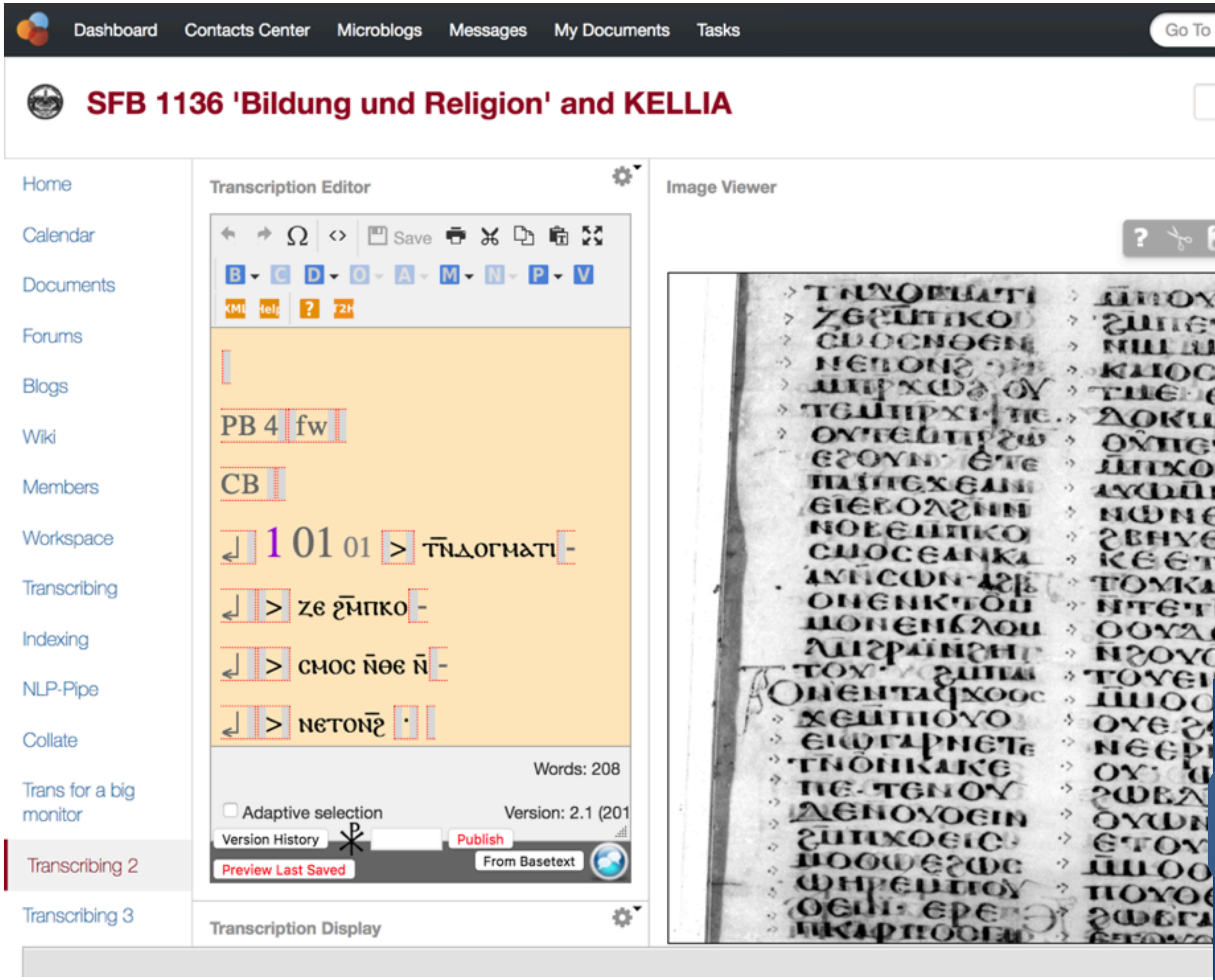


U+2C80

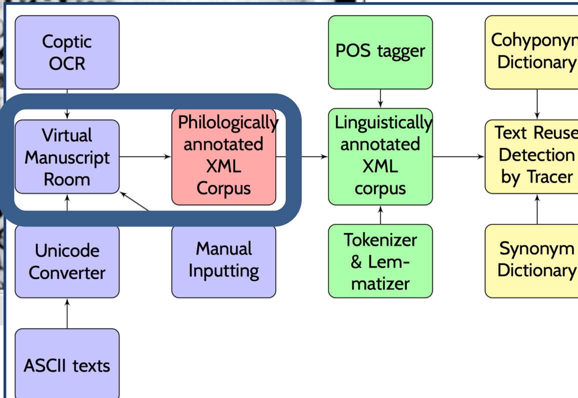
(Coptic Unicode)


Coptic OCR
POS tagger
Cohyponym Dictionary
Virtual Manuscript Room
Philologically annotated XML Corpus
Linguistically annotated XML corpus
Text Reuse Detection by Tracer
Unicode Converter
Manual Inputting
Tokenizer & Lem-matizer
Synonym Dictionary
ASCII texts

# Virtual Manuscript Room (VMR)


Dashboard
Contacts Center
Microblogs
Messages
My Documents
Tasks
Go To
Admin
0
So Miyagawa
SFB 1136 'Bildung und Religion' and KELLIA
Home
Calendar
Documents
Forums
Blogs
Wiki
Members
Workspace
Transcribing
Indexing
NLP-Pipe
Collate
Trans for a big monitor
Transcribing 2
Transcribing 3
Transcription Editor
Save
Image Viewer
Words: 208
Adaptive selection
Version: 2.1 (201
Version History
Publish
Preview Last Saved
From Basetext
Transcription Display

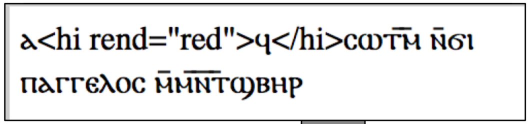
PB 4 fw
CB
1 01 01 > ⲧ̄ⲛⲁⲟⲣⲙⲁⲧⲓ -
> ⲍⲉ ϩ̄ⲙⲡⲕⲟ -
> ⲥⲙⲟⲥ ⲛ̄ⲑⲉ ⲛ̄ -
> ⲛⲉⲧⲟⲛ̄ϩ ·


Coptic OCR
Virtual Manuscript Room
Philologically annotated XML Corpus
Unicode Converter
Manual Inputting
ASCII texts
POS tagger
Linguistically annotated XML corpus
Tokenizer & Lem-matizer
Cohyponyr Dictionary
Text Reuse Detection by Tracer
Synonym Dictionary

# Coptic NLP Service

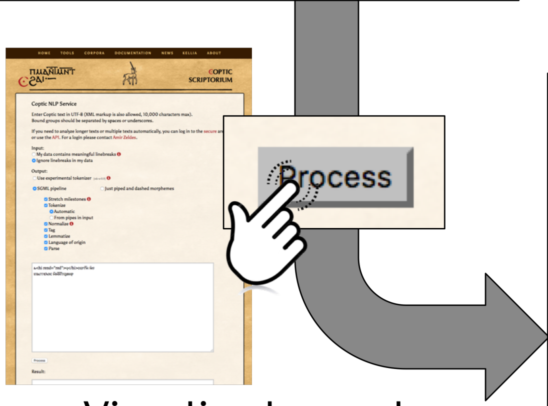
*Everything is done by one click!*

Input Coptic texts:



```
ⲁ<hi rend="red">ϥ</hi>ⲥⲱⲧⲙ̄ ⲛ̄ϭⲓ
ⲡⲁⲅⲅⲉⲗⲟⲥ ⲙ̄ⲙⲛ̄ⲧϣⲃⲏⲣ
```

- Part of speech tagging
- Syntactic roles annotation
- Lemmatization
- Tokenization (Word segmentation for Coptic)



HOME
TOOLS
CORPORA
DOCUMENTATION
NEWS
KELLIA
ABOUT
COPTIC SCRIPTORIUM

Coptic NLP Service

Enter Coptic text in UTF-8 (XML markup is also allowed, 10,000 characters max).
Bound groups should be separated by spaces or underscores.

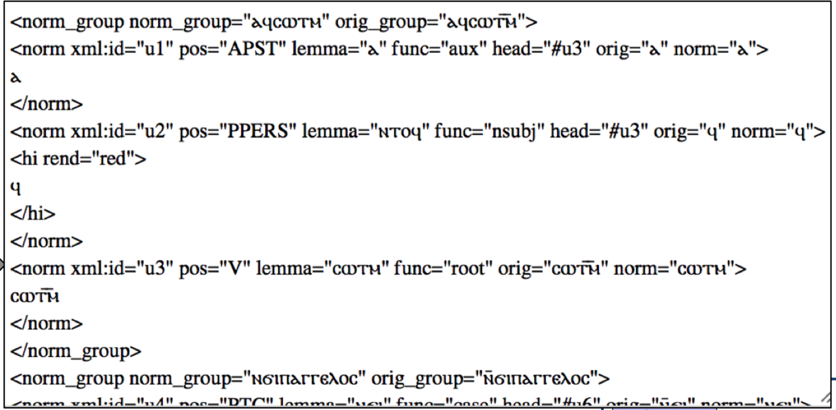
If you need to analyze longer texts or multiple texts automatically, you can log in to the secure ar… or use the API. For a login please contact Amir Zeldes.

Input:
○ My data contains meaningful linebreaks
● Ignore linebreaks in my data

Output:
☐ Use experimental tokenizer
● SGML pipeline ○ Just piped and dashed morphemes
- ☑ Stretch milestones
- ☑ Tokenize
  - ● Automatic
  - ○ From pipes in input
- ☑ Normalize
- ☑ Tag
- ☑ Lemmatize
- ☑ Language of origin
- ☑ Parse

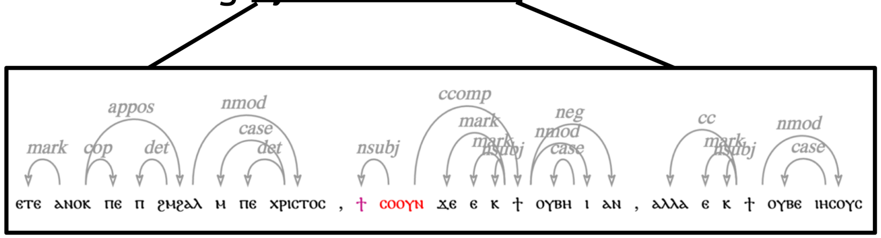
Process

Result:



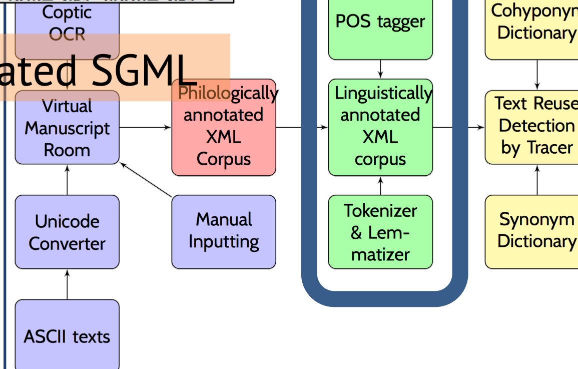
```
<norm_group norm_group="ⲁϥⲥⲱⲧⲙ" orig_group="ⲁϥⲥⲱⲧⲙ̄">
<norm xml:id="u1" pos="APST" lemma="ⲁ" func="aux" head="#u3" orig="ⲁ" norm="ⲁ">
ⲁ
</norm>
<norm xml:id="u2" pos="PPERS" lemma="ⲛⲧⲟϥ" func="nsubj" head="#u3" orig="ϥ" norm="ϥ">
<hi rend="red">
ϥ
</hi>
</norm>
<norm xml:id="u3" pos="V" lemma="ⲥⲱⲧⲙ" func="root" orig="ⲥⲱⲧⲙ̄" norm="ⲥⲱⲧⲙ">
ⲥⲱⲧⲙ̄
</norm>
</norm_group>
<norm_group norm_group="ⲛϭⲓⲡⲁⲅⲅⲉⲗⲟⲥ" orig_group="ⲛ̄ϭⲓⲡⲁⲅⲅⲉⲗⲟⲥ">
<norm xml:id="u4" pos="PTC" lemma="ⲛϭⲓ" func="case" head="#u6" orig="ⲛ̄ϭⲓ" norm="ⲛϭⲓ">
```

- Visualized as web corpus using ANNIS
- Drawing syntactic trees



appos
nmod
ccomp
mark
cop
det
case
det
nsubj
mark
mark
nsubj
neg
nmod
case
cc
mark
nsubj
nmod
case
ⲉⲧⲉ ⲁⲛⲟⲕ ⲡⲉ ⲡ ϩⲙϩⲁⲗ ⲙ ⲡⲉ ⲭⲣⲓⲥⲧⲟⲥ , ϯ ⲥⲟⲟⲩⲛ ϫⲉ ⲉ ⲕ ϯ ⲟⲩⲃⲏ ⲓ ⲁⲛ , ⲁⲗⲗⲁ ⲉ ⲕ ϯ ⲟⲩⲃⲉ ⲓⲏⲥⲟⲩⲥ

See Zeldes & Schroeder (2015, 2016) and Miyagawa et al. (2018)

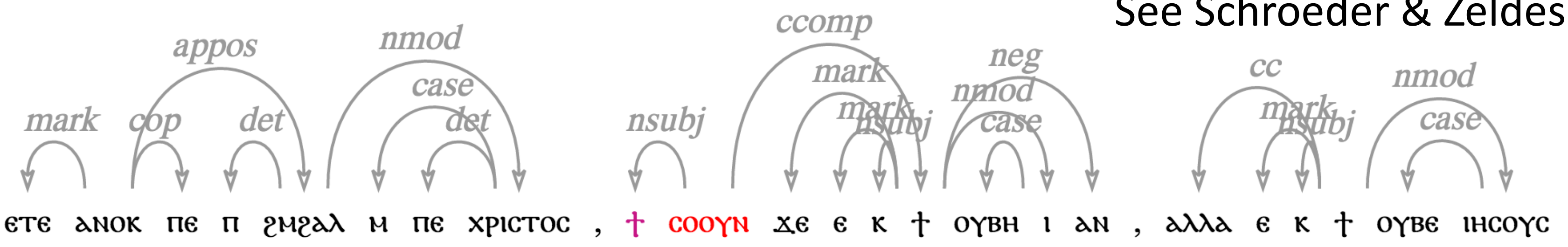


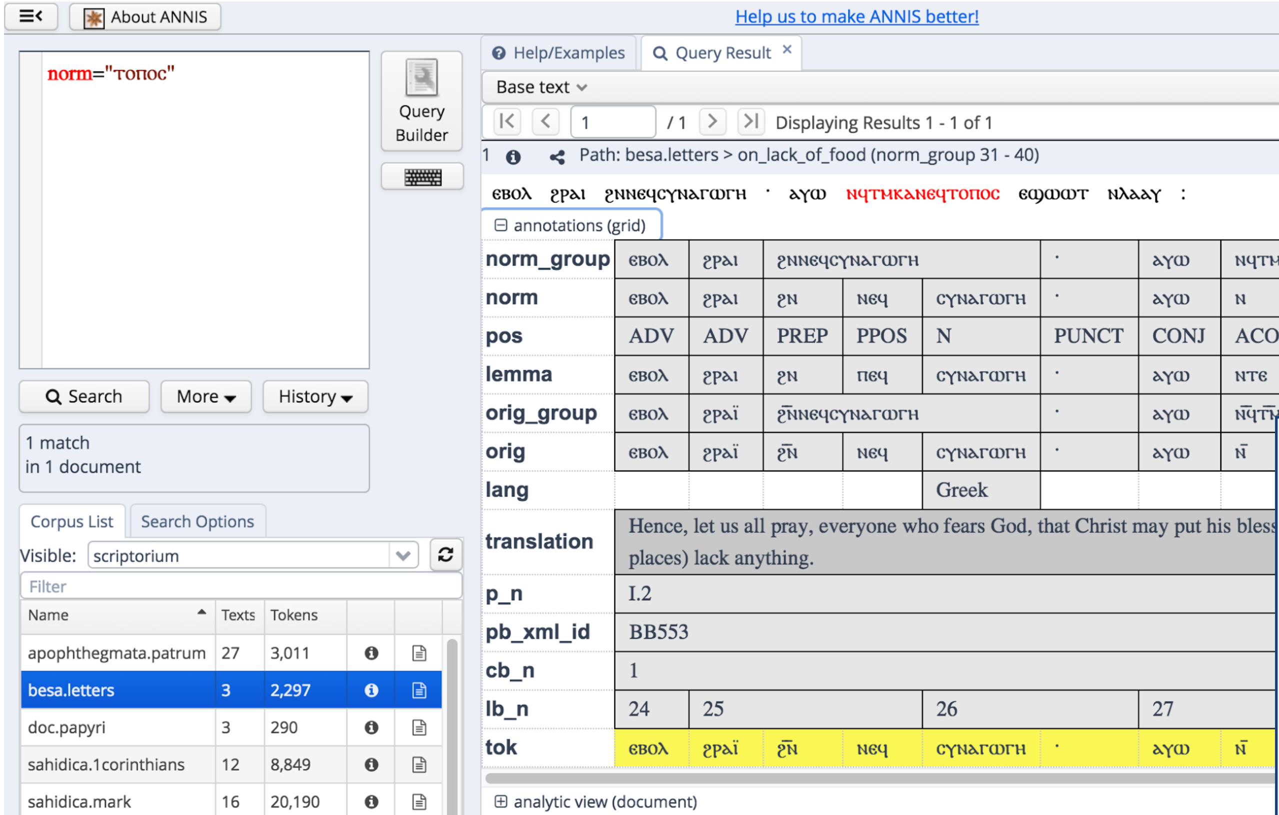
Annotated SGML

Coptic OCR
Virtual Manuscript Room
Unicode Converter
ASCII texts
Manual Inputting
Philologically annotated XML Corpus
POS tagger
Linguistically annotated XML corpus
Tokenizer & Lemmatizer
Cohyponym Dictionary
Text Reuse Detection by Tracer
Synonym Dictionary

# Coptic SCRIPTORIUM

See Schroeder & Zeldes (2016)


appos
nmod
ccomp
mark
neg
nmod
cc
nmod
mark
cop
det
case
det
nsubj
mark
nsubj
case
mark
nsubj
case
ⲉⲧⲉ ⲁⲛⲟⲕ ⲡⲉ ⲡ ϩⲙϩⲁⲗ ⲙ ⲡⲉ ⲭⲣⲓⲥⲧⲟⲥ , ϯ ⲥⲟⲟⲩⲛ ϫⲉ ⲉ ⲕ ϯ ⲟⲩⲃⲏ ⲓ ⲁⲛ , ⲁⲗⲗⲁ ⲉ ⲕ ϯ ⲟⲩⲃⲉ ⲓⲏⲥⲟⲩⲥ


About ANNIS
Help us to make ANNIS better!
not logged in
Login
Query Builder
Search
More
History
Help/Examples
Query Result
Base text
Corpus List
Search Options
Filter
annotations (grid)
analytic view (document)

norm="ⲧⲟⲡⲟⲥ"

1 match
in 1 document

Visible: scriptorium

| Name | Texts | Tokens |
|---|---|---|
| apophthegmata.patrum | 27 | 3,011 |
| besa.letters | 3 | 2,297 |
| doc.papyri | 3 | 290 |
| sahidica.1corinthians | 12 | 8,849 |
| sahidica.mark | 16 | 20,190 |

1 / 1 Displaying Results 1 - 1 of 1

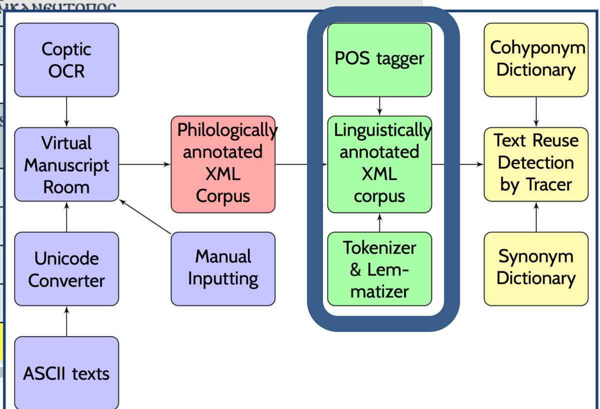
Result for: norm="ⲧⲟⲡⲟⲥ"

1 Path: besa.letters > on_lack_of_food (norm_group 31 - 40)

left context: 5 right context: 5

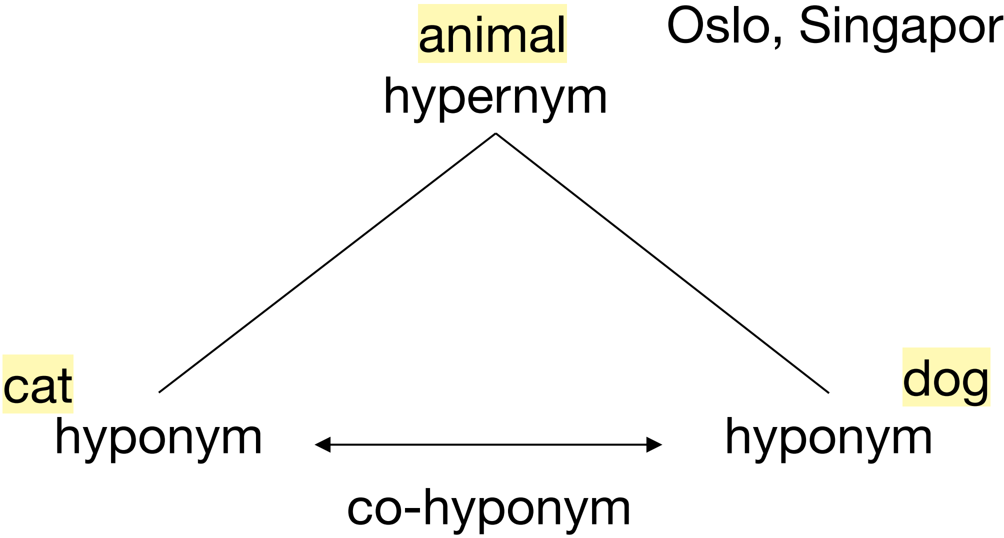
ⲉⲃⲟⲗ ϩⲣⲁⲓ ϩⲛⲛⲉϥⲥⲩⲛⲁⲅⲱⲅⲏ · ⲁⲩⲱ ⲛϥⲧⲙⲕⲁⲛⲉϥⲧⲟⲡⲟⲥ ⲉϣⲱⲱⲧ ⲛⲗⲁⲁⲩ :

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **norm_group** | ⲉⲃⲟⲗ | ϩⲣⲁⲓ | ϩⲛⲛⲉϥⲥⲩⲛⲁⲅⲱⲅⲏ | | | · | ⲁⲩⲱ | ⲛϥⲧⲙⲕⲁⲛⲉϥⲧⲟⲡⲟⲥ | | | | | |
| **norm** | ⲉⲃⲟⲗ | ϩⲣⲁⲓ | ϩⲛ | ⲛⲉϥ | ⲥⲩⲛⲁⲅⲱⲅⲏ | · | ⲁⲩⲱ | ⲛ | ϥ | ⲧⲙ | ⲕⲁ | ⲛⲉϥ | ⲧⲟⲡⲟⲥ |
| **pos** | ADV | ADV | PREP | PPOS | N | PUNCT | CONJ | ACONJ | PPERS | NEG | V | PPOS | N |
| **lemma** | ⲉⲃⲟⲗ | ϩⲣⲁⲓ | ϩⲛ | ⲡⲉϥ | ⲥⲩⲛⲁⲅⲱⲅⲏ | · | ⲁⲩⲱ | ⲛⲧⲉ | ⲛⲧⲟϥ | ⲧⲙ | ⲕⲱ | ⲡⲉϥ | ⲧⲟⲡⲟⲥ |
| **orig_group** | ⲉⲃⲟⲗ | ϩⲣⲁϊ | ϩ̄ⲛⲛⲉϥⲥⲩⲛⲁⲅⲱⲅⲏ | | | · | ⲁⲩⲱ | | | | | | |
| **orig** | ⲉⲃⲟⲗ | ϩⲣⲁϊ | ϩ̄ⲛ | ⲛⲉϥ | ⲥⲩⲛⲁⲅⲱⲅⲏ | · | ⲁⲩⲱ | ⲛ̄ | | | | | |
| **lang** | | | | | Greek | | | | | | | | |
| **translation** | Hence, let us all pray, everyone who fears God, that Christ may put his bless<br>places) lack anything. | | | | | | | | | | | | |
| **p_n** | I.2 | | | | | | | | | | | | |
| **pb_xml_id** | BB553 | | | | | | | | | | | | |
| **cb_n** | 1 | | | | | | | | | | | | |
| **lb_n** | 24 | 25 | | | 26 | | 27 | | | | | | |
| **tok** | ⲉⲃⲟⲗ | ϩⲣⲁϊ | ϩ̄ⲛ | ⲛⲉϥ | ⲥⲩⲛⲁⲅⲱⲅⲏ | · | ⲁⲩⲱ | ⲛ̄ | | | | | |

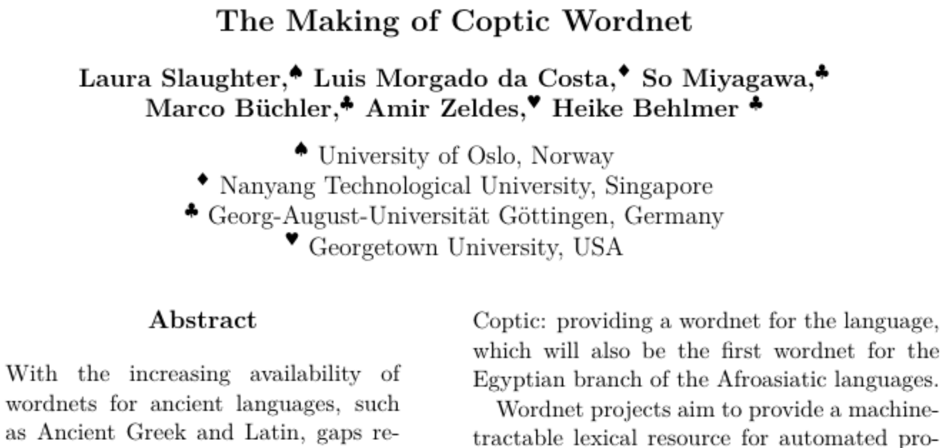

Coptic OCR
Virtual Manuscript Room
Unicode Converter
ASCII texts
Philologically annotated XML Corpus
Manual Inputting
POS tagger
Linguistically annotated XML corpus
Tokenizer & Lem-matizer
Cohyponym Dictionary
Text Reuse Detection by Tracer
Synonym Dictionary

# Coptic Wordnet

Oslo, Singapore and Göttingen


animal
hypernym
cat
hyponym
dog
hyponym
co-hyponym

Slaughter et al. (2019)


### The Making of Coptic Wordnet

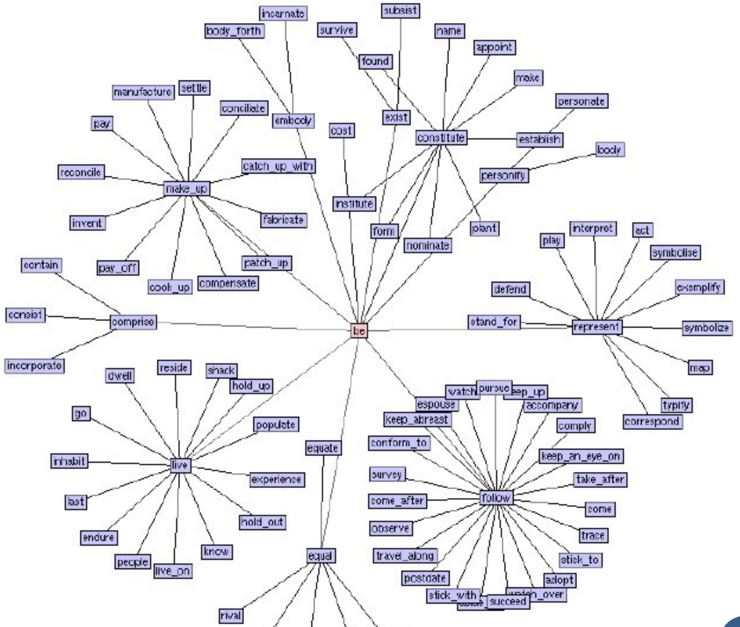
Laura Slaughter,♠ Luis Morgado da Costa,♦ So Miyagawa,♣
Marco Büchler,♣ Amir Zeldes,♥ Heike Behlmer ♣

♠ University of Oslo, Norway
♦ Nanyang Technological University, Singapore
♣ Georg-August-Universität Göttingen, Germany
♥ Georgetown University, USA

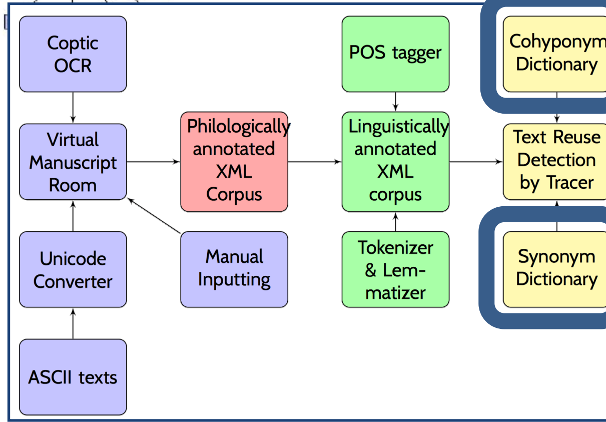
**Abstract**

With the increasing availability of wordnets for ancient languages, such as Ancient Greek and Latin, gaps re-

Coptic: providing a wordnet for the language, which will also be the first wordnet for the Egyptian branch of the Afroasiatic languages.
Wordnet projects aim to provide a machine-tractable lexical resource for automated pro-


incarnate
body_forth
survive
subsist
name
appoint
found
make
manufacture
settle
conciliate
embody
exist
personate
pay
cost
constitute
establish
body
catch_up_with
reconcile
make_up
personify
institute
fabricate
invent
form
plant
nominate
interpret
act
play
symbolise
contain
pay_off
patch_up
cook_up
compensate
defend
exemplify
consist
comprise
be
stand_for
represent
symbolize
incorporate
dwell
reside
shack
hold_up
map
watch
pursue
keep_up
espouse
accompany
typify
go
keep_abreast
correspond
populate
comply
equate
conform_to
inhabit
live
keep_an_eye_on
experience
survey
take_after
come_after
follow
come
last
hold_out
observe
trace
endure
people
know
equal
travel_along
stick_to
live_on
postdate
adopt
stick_with
watch_over
succeed
rival

Kamps (2002)


Coptic OCR

Virtual Manuscript Room

Unicode Converter

ASCII texts

Philologically annotated XML Corpus

Manual Inputting

POS tagger

Linguistically annotated XML corpus

Tokenizer & Lemmatizer

Cohyponym Dictionary

Text Reuse Detection by Tracer

Synonym Dictionary

# TRACER processing

- Remote access to ROEDEL, a high-performance computer

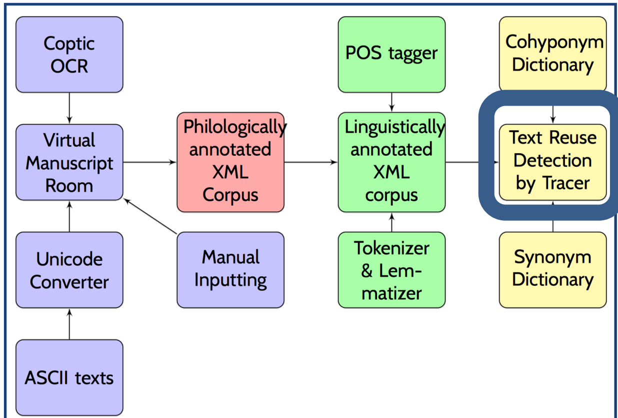
TRACER …

- is a Java program
- has a 700 algorithms
- uses an XML config file
- has no graphical user-interface

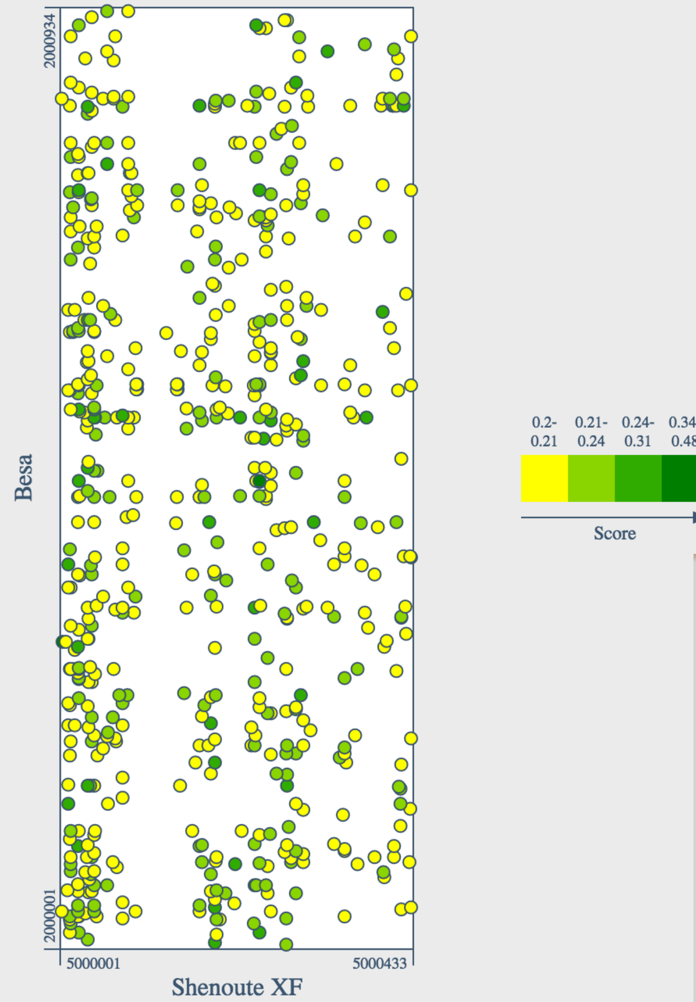
TRACER
TEXT REUSE DETECTION MACHINE

eTRAP
Electronic Text Reuse Acquisition Project

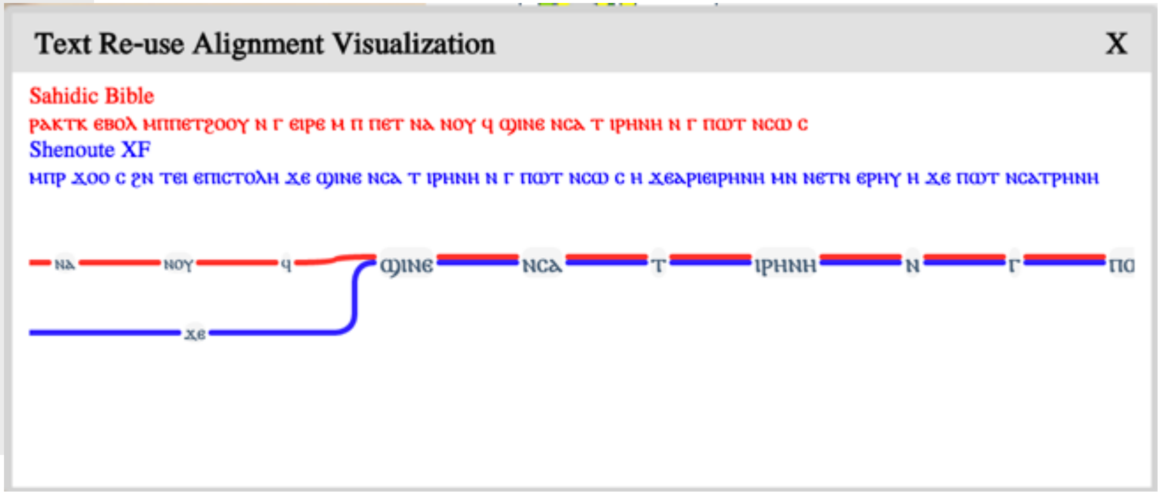
Bundesministerium
für Bildung
und Forschung


Coptic OCR
POS tagger
Cohyponym Dictionary
Virtual Manuscript Room
Philologically annotated XML Corpus
Linguistically annotated XML corpus
Text Reuse Detection by Tracer
Unicode Converter
Manual Inputting
Tokenizer & Lemmatizer
Synonym Dictionary
ASCII texts

# TRAViz: Visualization of TRACER's results



- x-axis = Shenoute XF
- y-axis = Besa
- left and bottom = the start of the texts
- right and top = the end of the texts


Text Re-use Alignment Visualization
X

Sahidic Bible

ⲣⲁⲕⲧⲕ ⲉⲃⲟⲗ ⲙⲡⲡⲉⲧϩⲟⲟⲩ ⲛ ⲅ ⲉⲓⲣⲉ ⲙ ⲡ ⲡⲉⲧ ⲛⲁ ⲛⲟⲩ ϥ ϣⲓⲛⲉ ⲛⲥⲁ ⲧ ⲓⲣⲏⲛⲏ ⲛ ⲅ ⲡⲱⲧ ⲛⲥⲱ ⲥ

Shenoute XF

ⲙⲡⲣ ϫⲟⲟ ⲥ ϩⲛ ⲧⲉⲓ ⲉⲡⲓⲥⲧⲟⲗⲏ ϫⲉ ϣⲓⲛⲉ ⲛⲥⲁ ⲧ ⲓⲣⲏⲛⲏ ⲛ ⲅ ⲡⲱⲧ ⲛⲥⲱ ⲥ ⲏ ϫⲉⲁⲣⲓⲉⲓⲣⲏⲛⲏ ⲙⲛ ⲛⲉⲧⲛ ⲉⲣⲏⲩ ⲏ ϫⲉ ⲡⲱⲧ ⲛⲥⲁⲧⲣⲏⲛⲏ

ⲛⲁ
ⲛⲟⲩ
ϥ
ϣⲓⲛⲉ
ⲛⲥⲁ
ⲧ
ⲓⲣⲏⲛⲏ
ⲛ
ⲅ
ⲡⲱ
ϫⲉ

# TRACER's first results

| Source Text | Target Text | TR Cand. |
|---|---|---|
| Sahidic Bible | Besa | 13,835 |
| Sahidic Bible | Canon 6 | 8414 |

Too many text reuse candidates!

From the entire Sahidic Bible to the Sahidic Psalms

# Psalms in Coptic monasticism

- Desert Fathers and Mothers
    - Both in solitude and with others
- Rituals and liturgies at coenobitic monasteries
- Pachomian Rules

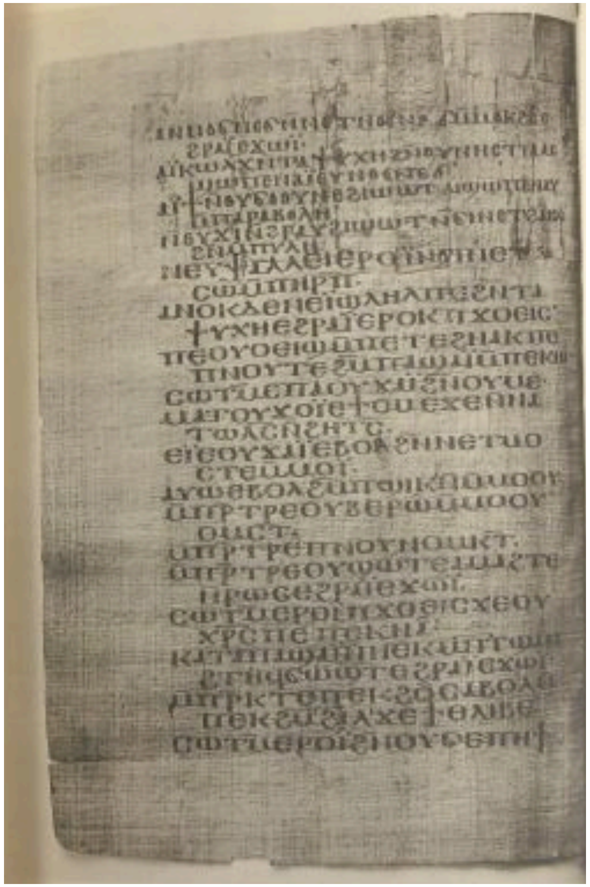
Jürgen Horn (2000)’s typology of Coptic Psalm mss.

(A) Manuscripts of the Psalms (Psalters)

(B) Manuscripts for Rituals (e.g. lectionaries)

(C) Quotations (e.g. by Shenoute)

(D) Occasional uses (magic, school exercises etc.)



British Library, Or. 5000
“London Psalter"

# TRACER's second results

| Source | Target | TR Cand. |
|---|---|---|
| Psalms | Besa | 629 |
| Psalms | Canon 6 | 423 |

| Canon 6 *Incipit* | TR Cand. |
|---|---|
| *He Who Sits Upon His Throne* | 84 |
| *Remember, O Brethren* | 31 |
| *I Am Not Obliged* | 207 |
| *Is It Not Written* | 98 |
| *People Have Not Understood* | 3 |

- Newly found quotations unknown to previous research :

6 (Psalms–Besa), 14 (Psalms–Shenoute *Canon* 6)

See Miyagawa, Büchler and Behlmer (forthcoming 2020)

# Many idiomatic text reuses…

| Translation | Idiom | Frequency in Psalms | Frequency in Besa | Shared morphs |
|---|---|---|---|---|
| Fear the Lord | ⲣ-ϩⲟⲧⲉ ϩⲏⲧ-ϥ ⲙ-ⲡ-ϫⲟⲉⲓⲥ | 7 | 7 | 7 |
| Evil in the place of good | ϩⲉⲛ-ⲡ-ⲉⲑ-ⲟⲟⲩ ⲉ-ⲡ-ⲙⲁ ⲛ-ϩⲉⲛ-ⲡ-ⲉⲧ-ⲛⲁⲛⲟⲩ-ⲟⲩ | 5 | 2 | 13 |
| Fear God | ⲣ-ϩⲟⲧⲉ ϩⲏⲧ-ϥ ⲙ-ⲡ-ⲛⲟⲩⲧⲉ | 4 | 18 | 7 |
| Before God | ⲙ-ⲡⲉ-ⲙⲧⲟ ⲉⲃⲟⲗ ⲙ-ⲡ-ⲛⲟⲩⲧⲉ | 4 | 10 | 7 |
| Everyone who works for the lawlessness | ⲟⲩⲟⲛ ⲛⲓⲙ ⲉⲧ-ⲣ-ϩⲱⲃ ⲉ-ⲧ-ⲁⲛⲟⲙⲓⲁ | 4 | 1 | 8 |
| Before the Lord | ⲙ-ⲡⲉ-ⲙⲧⲟ ⲉⲃⲟⲗ ⲙ-ⲡ-ϫⲟⲉⲓⲥ | 2 | 8 | 7 |

# Unknown quotation found by TRACER: 1 (Ps → Besa)

ⲉ⸗ⲕ-ⲉ-ⲡⲱⲧ ⲛⲥⲱ⸗ⲟⲩ ⲛ-ⲧⲉⲓ-ϩⲉ ϩⲛ-ⲧⲉ⸗ⲕ-ϩⲁⲧⲏⲩ ⲁⲩⲱ ⲉ⸗ⲕ-ⲉ-ϣⲧⲣⲧⲱⲣ⸗ⲟⲩ ϩⲛ-ⲧⲉ⸗ⲕ-ⲟⲣⲅⲏ
(Ps. 82:16, BL)
“you will pursue them in this way with your whirlwind and vex them with your wrath.”

ⲁⲩⲱ ⲟⲛ ⲣⲱⲙⲉ ⲛⲓⲙ ⲉⲧ⸗ⲕ-ⲛⲁ-ϯ-ϩⲏⲩ ⲉⲣⲟ⸗ⲟⲩ ⲁⲛ ϩⲣⲁⲓ ϩⲙ-ⲡⲓ-ⲙⲁ ⲉ⸗ⲕ-ⲉ-ⲡⲱⲧ ⲛⲥⲱ⸗ⲟⲩ ϩⲛ-ⲧⲉ⸗ⲕ-ϩⲁⲧⲏⲩ ⲛ⸗ⲅ-ϣⲧⲣⲧⲱⲣ-ⲟⲩ ϩⲛ-ⲧⲉⲕ-ⲟⲣⲅⲏ ⲛⲧⲉ-ⲟⲩ-ⲥⲁⲧⲉ ⲟⲩⲟⲙ⸗ⲟⲩ (Besa 4:1:2)
“And all men from whom you will get no gain in this place, you will pursue with your whirlwind and vex with your wrath and a fire shall devour them.” (Kuhn)

Deletion 1: ⲛ-ⲧⲉⲓ-ϩⲉ “in this way”
Deletion 2: ⲁⲩⲱ “and”
Auxiliary change: Future III (ⲉ-ⲕ-ⲉ-) → conjunctive (ⲛ-ⲅ-)

# Unknown quotation found by TRACER: 2 (Ps → Besa)

ⲉ-ⲕ-ⲉ-ⲁⲁ-ⲩ ⲛ-ⲑⲉ ⲛ-ⲟⲩ-ⲧⲣⲓⲣ ⲛ-ⲕⲱϩⲧ ⲉ-ⲩ-ⲟⲩⲟⲉⲓϣ ⲙ-ⲡⲉⲕ-ⲙⲧⲟ ⲉⲃⲟⲗ ⲡ-ϫⲟⲉⲓⲥ ⲡ-ϫⲟⲉⲓⲥ ⲕ-ⲛⲁ-ϣⲧⲣⲧⲱⲣ-ⲟⲩ ϩⲛ-ⲧⲉⲕ-ⲟⲣⲅⲏ ⲛⲧⲉ-ⲟⲩ-ⲥⲁⲧⲉ ⲟⲩⲟⲙ-ⲟⲩ (Ps. 20.10, BL)
"You will make them like an oven of fire at the time of your presence. Lord, Lord, you will vex them with your wrath and a fire shall devour them."

↓

ⲁⲩⲱ ⲟⲛ ⲣⲱⲙⲉ ⲛⲓⲙ ⲉⲧ-ⲕ-ⲛⲁ-ϯ-ϩⲏⲩ ⲉⲣⲟ-ⲟⲩ ⲁⲛ ϩⲣⲁⲓ ϩⲙ-ⲡⲓ-ⲙⲁ ⲉ-ⲕ-ⲉ-ⲡⲱⲧ ⲛⲥⲱ-ⲟⲩ ϩⲛ-ⲧⲉⲕ-ϩⲁⲧⲏⲩ ⲛ-ⲅ-ϣⲧⲣⲧⲱⲣ-ⲟⲩ ϩⲛ-ⲧⲉⲕ-ⲟⲣⲅⲏ ⲛⲧⲉ-ⲟⲩ-ⲥⲁⲧⲉ ⲟⲩⲟⲙ-ⲟⲩ (Besa 4/I/2)
"And all men from whom you will get no gain in this place, you will pursue with your whirlwind and vex with your wrath and a fire shall devour them." (Kuhn)

Auxiliary change: Future I → conjunctive

# Unknown quotation found by TRACER: 3 (Ps → Besa)

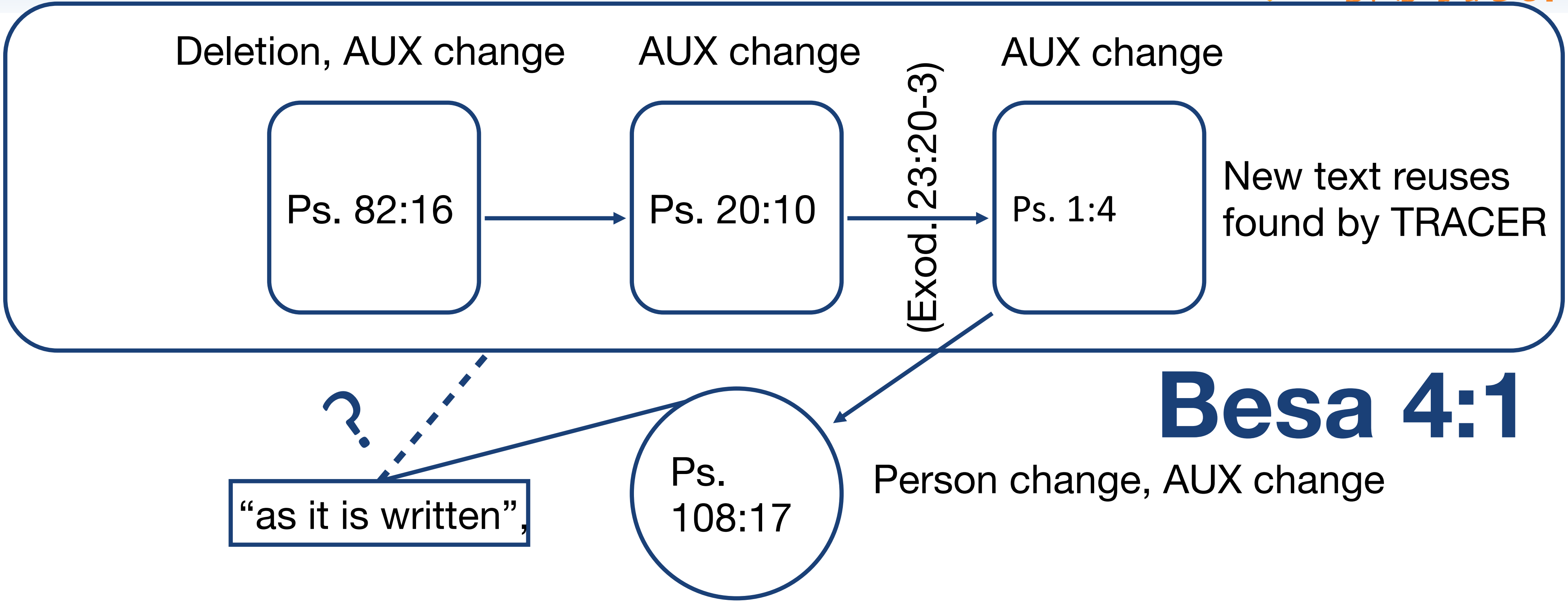
ⲛ-ⲧⲁⲓ ⲁⲛ ⲧⲉ ⲑⲉ ⲛ-ⲛ-ⲁⲥⲉⲃⲏⲥ ⲛ-ⲧⲁⲓ ⲁⲛ ⲧⲉ ⲑⲉ ⲁⲗⲗⲁ ⲉ-ⲩ-ⲛⲁ-ⲣ-ⲑⲉ ⲙ-ⲡ-ϣⲟⲉⲓϣ ⲉ-ϣⲁⲣⲉ-ⲡ-ⲧⲏⲩ ⲑⲗⲟ-ϥ ⲉⲃⲟⲗ ϩⲓϫⲙ-ⲡ-ϩⲟ ⲙ-ⲡ-ⲕⲁϩ (Ps. 1:4, BL)
“This is not like the ungodly. It is not so but they will be like the dust which the wind scatters upon the face of the earth. ”

↓

ⲛⲧⲉ-ⲡ-ⲁⲅⲅⲉⲗⲟⲥ ⲛ-ⲧ-ⲇⲓⲁⲑⲏⲕⲏ ϥⲓⲧ-ⲟⲩ ⲛ-ϥ-ⲟⲩⲟⲧⲃ-ⲟⲩ ⲙ-ⲡ-ⲃⲟⲗ ⲙ-ⲡ-ⲥⲟⲃⲧ ⲛ- ⲧ-ⲥⲩⲛⲁⲅⲱⲅⲏ ⲛ-ϥ-ϫⲟⲟⲣ-ⲟⲩ ⲇⲉ ⲟⲛ ⲉⲃⲟⲗ ⲛ-ⲑⲉ ⲙ-ⲡ-ϣⲟⲉⲓϣ ⲉ-ϣⲁⲣⲉ ⲡ-ⲧⲏⲩ ⲑⲗⲟ-ϥ ⲉⲃⲟⲗ ϩⲓϫⲙ-ⲡ-ϩⲟ ⲙ-ⲡ-ⲕⲁϩ (Besa 4:1:2)

"The angel of the covenant shall take them and remove them outside the wall of the community, and shall moreover disperse them, even as the dust which the wind scatters upon the face of the earth” (Kuhn)

De-verbalization: ⲉ-ⲩ-ⲛⲁ-ⲣ-ⲑⲉ… “while they will do like…” → ⲛ-ⲑⲉ… “like…”



Deletion, AUX change

AUX change

AUX change

Ps. 82:16

Ps. 20:10

(Exod. 23:20-3)

Ps. 1:4

New text reuses found by TRACER

?

“as it is written”,

Ps. 108:17

# Besa 4:1

Person change, AUX change

And all men from whom yon will get no gain in this place, **you will pursue with your whirlwind and vex** **with your wrath** (Ps. 82:16) **and a fire shall devour them** (Ps. 20:10). *The angel of the covenant shall take them and remove them outside the wall of the community, and shall moreover disperse them* (cf. Exod. 23:20-3), even ***as the dust which the wind scatters upon the face of the earth*** (Ps.1:4), as it is written, ‘Forasmuch as (ⲉⲃⲟⲗ ϫⲉ) they loved the curse, it came upon them; they did not desire the blessing, it fled from them ’(Ps. 108:17) (Besa 4/I/2, translation by Kuhn)

# Quotation Index Phrase (QIP)

- Inspired by the term “Quotative Index” by Tom Güldemann for African languages (2008), but a different concept

For example,

| | | | |
|---|---|---|---|
| ⲛ-ⲑⲉ | ⲉⲧ-ⲥⲏϩ | (ϩⲛ-*X*) | (ϫⲉ-*QUOT*) |
| OBL-DEF.SG.F:way.SG.F | REL-write.STA | (in-*X*) | (COMP-*QUOT*) |

“as it is written (in *X*) (that *QUOT*)”

- Attested 49 times in Shenoute’s Canon 6 and 36 times in Besa
- 3 positions: preposed, inserted and postposed

# Preposed QIP construction

Ps. 32(33):15

ⲡ-ⲉⲛⲧ-ⲁ⸗ϥ-ⲡⲗⲁⲥⲥⲉ ⲙⲁⲩⲁⲁ⸗ϥ ⲛ-ⲛⲉ⸗ⲩ-ϩⲏⲧ ⲡ-ⲉⲧ-ⲉⲓⲙⲉ ⲉ-ⲛⲉ⸗ⲩ-ϩⲃⲏⲩⲉ ⲧⲏⲣ⸗ⲟⲩ
“The one who shaped himself their hearts, the one who knows all their works“

Shenoute, *He Who Sits Upon His Throne*

ⲁⲗⲗⲁ ⲉⲃⲟⲗ ⲣⲱ ϫⲉ-ⲛⲧⲟϥ ⲡ-ⲉⲛⲧ-ⲁ⸗ϥ-ⲡⲗⲁⲥⲥⲉ ⲛ-ⲛⲉ⸗ⲩ-ϩⲏ[ⲧ] ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ ϩⲛ̄-ⲛⲉ-ⲯⲁⲗⲙⲟⲥ ϫⲉ-ⲡ-ⲉⲛⲧ-ⲁ⸗ϥ-ⲡⲗⲁⲥⲥⲉ ⲙⲁⲩⲁⲁ⸗ϥ ⲛ-ⲛⲉ⸗ⲩ-ϩⲏⲧ (MONB.YJ p. 33)
“but because it is he who has shaped their hearts, as it is written in the Psalms: "the one who himself shaped their hearts”

# Inserted QIP construction

1 Sam 16:7

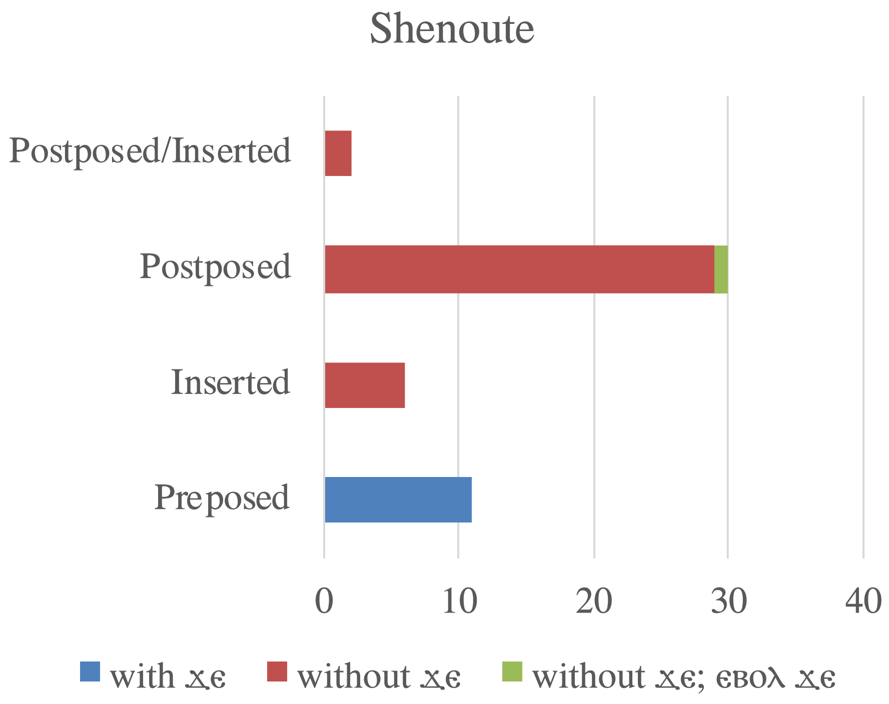
ⲡⲉϫⲉ-ⲡ-ϫⲟⲉⲓⲥ ⲛ-ⲥⲁⲙⲟⲩⲏⲗ ϫⲉ-ⲙⲡⲉⲣ-ϭⲱϣⲧ ⲉϫⲙ-ⲡⲉ⸗ϥ-ⲥⲁ ⲟⲩⲇⲉ ⲉϫⲛ-ⲧ-ϭⲟⲧ ⲛ-ⲧⲉ⸗ϥ-ϣⲓⲏ ϫⲉ⸗ϥ-ⲣ-ⲥⲟϣϥ ⲛⲁϩⲣⲁⲓ ϫⲉ-ⲛ-ⲑⲉ ⲁⲛ ⲉⲧⲉⲣⲉ-ⲛⲉ-ⲣⲱⲙⲉ ϭⲱϣⲧ ⲉ-ⲡ-ϩⲟ ⲉⲣⲉ-ⲡ-ⲛⲟⲩⲧⲉ ϩⲱⲱ⸗ϥ ϭⲱϣⲧ ϫⲉ-ⲉⲣⲉ-ⲛ-ⲣⲱⲙⲉ ϭⲱϣⲧ ⲉ-ⲡ-ϩⲟ ⲙⲙⲁⲧⲉ ⲡ-ⲛⲟⲩⲧⲉ ⲇⲉ ⲛⲧⲟϥ ⲉ⸗ϥ-ϭⲱϣⲧ ⲉ-ⲡ-ϩⲏⲧ
“And the Lord said to Samouêl, Look not upon his beauty nor upon his stature; for he is despised in my sight; for not as men look at the face, does God also look; for men look at the face only and God looks rather at the heart. ” (Drescher 1970:36)

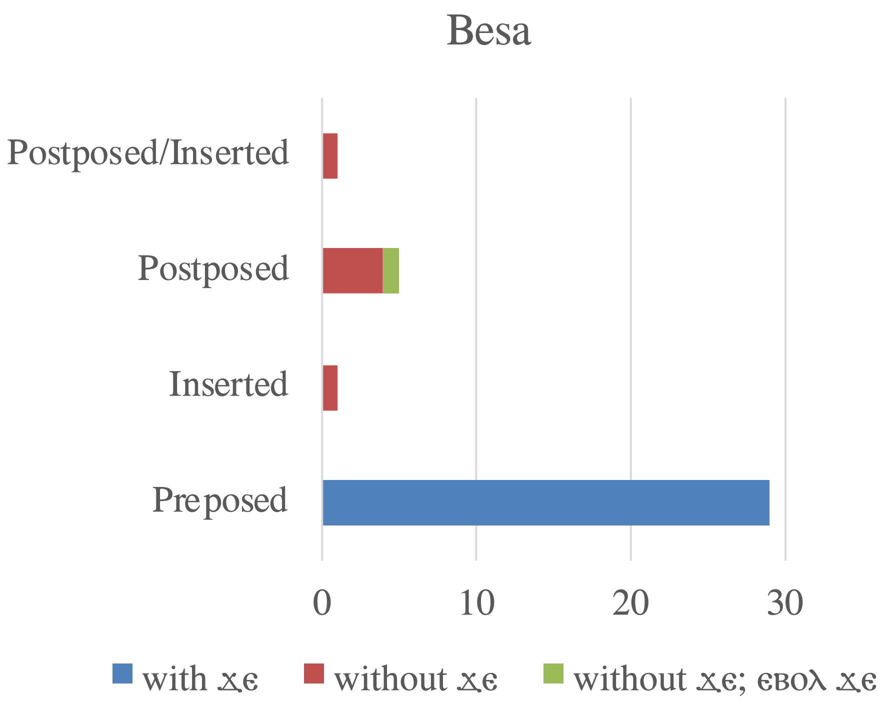
Shenoute, *He Who Sits Upon His Throne*

ⲡ-ⲣⲱⲙⲉ ⲉ⸗ϥ-ϭⲱϣⲧ ⲉ-ⲡ-ϩⲟ ⲙⲙⲁⲧⲉ ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ ⲡ-ⲛⲟⲩⲧⲉ ⲉ⸗ϥ-ϭⲱϣⲧ ⲉ-ⲡ-ϩⲏⲧ[...]
“the man looks at the face alone,” as it is written, “it is at the heart that God looks [...]” (MONB.YJ p. 33)

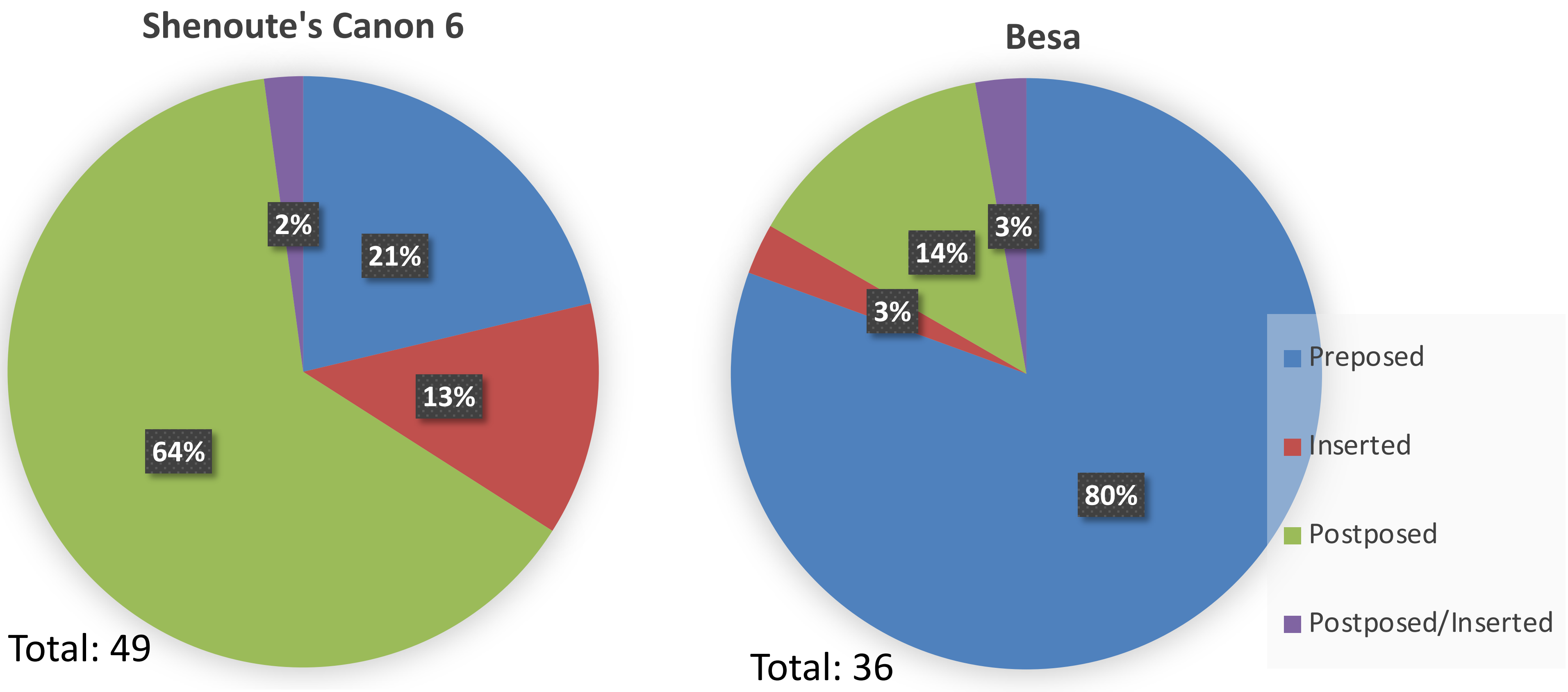
# Corpus analysis of location of ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ

- Only ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ (ϩⲛ-*X*) (ϫⲉ-*QUOT*)
- Shenoute’s Canon 6: 49,234 tokens vs Besa: 60,572 tokens
- Preposed QIP accompanied the complementizer always, but the others not


Shenoute


Besa

# Proportion of locations of ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ



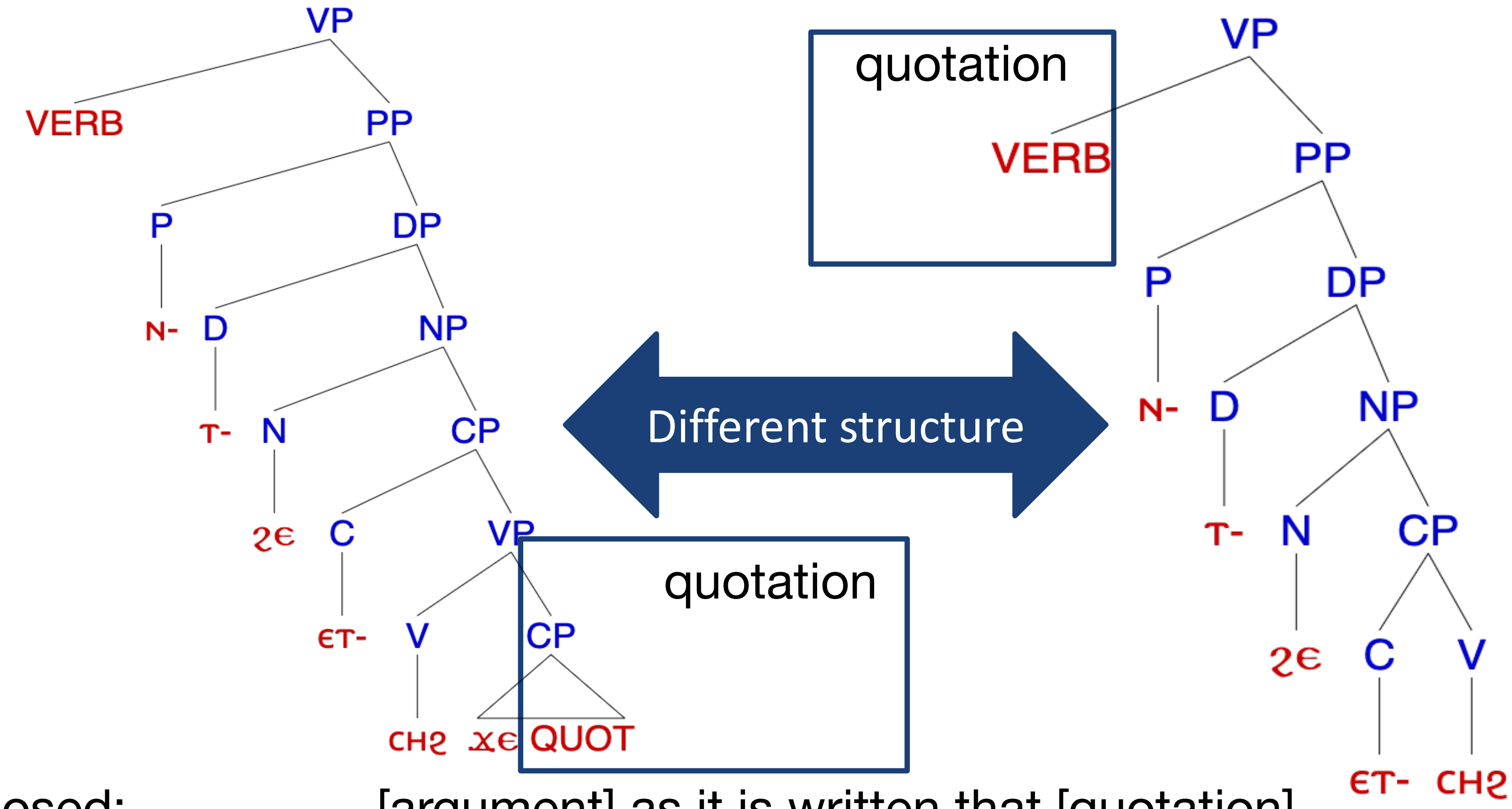
Total: 49

Total: 36

# Difference of QIP constructions

Pre-posed QIP construction

Post-posed QIP construction


VP
VERB
PP
P
DP
ⲛ-
D
NP
ⲧ-
N
CP
ϩⲉ
C
VP
ⲉⲧ-
V
CP
ⲥⲏϩ
ϫⲉ
QUOT
quotation
Different structure
quotation
VP
VERB
PP
P
DP
ⲛ-
D
NP
ⲧ-
N
CP
ϩⲉ
C
V
ⲉⲧ-
ⲥⲏϩ

Pre-posed: [argument] as it is written that [quotation]
Post-posed/inserted: [quotation = argument], as it is written

# [near verbatim quotation = argument] QIP [verbatim quotation] formula

- 4 times in Shenoute‘s Canon 6 and once in Besa

- Sort of a hybrid of pre-posed and post-posed QIPs.

Shenoute, *He Who Sits Upon His Throne*

ⲁⲗⲗⲁ ⲉⲃⲟⲗ ⲣⲱ ϫⲉ-ⲛⲧⲟϥ ⲡ-ⲉⲛⲧ-ⲁ⸗ϥ-ⲡⲗⲁⲥⲥⲉ ⲛ-ⲛⲉ⸗ⲩ-ϩⲏ[ⲧ] ⲛ-ⲑⲉ ⲉⲧ-ⲥⲏϩ ϩ̄ⲛ̄-ⲛⲉ-ⲯⲁⲗⲙⲟⲥ ϫⲉ-ⲡ-ⲉⲛⲧ-ⲁ⸗ϥ-ⲡⲗⲁⲥⲥⲉ ⲙⲁⲩⲁⲁ⸗ϥ ⲛ-ⲛⲉ⸗ⲩ-ϩⲏⲧ (MONB.YJ p. 33)
“but because it is he who has shaped their hearts, as it is written in the Psalms: "the one who himself shaped their hearts”

- Shenoute‘s tendency to use postposed QIPs → more skillful than Besa (?)

# Conclusions

- The project of SFB1136 and eTRAP have collaborated working on text reuses in Shenoute’s Canon 6 and Besa’s letters and sermons
- We found quotations unknown to the previous studies in both corpora
- With newly found quotations, we conducted a study on the location of ⲛⲑⲉ ⲉⲧⲥⲏϩ
- Besa uses more pre-posed ⲛⲑⲉ ⲉⲧⲥⲏϩ than Shenoute
- Perhaps this indicates that Shenoute embedded biblical quotations more skillfully than Besa in monastic educational discourse
- Full application of Coptic Wordnet to TRACER is the next challenge

# Selected Bibliography 1

- Büchler, M. (2013). Informationstechnische Aspekte des Historical Text Re-use. Doctoral dissertation, Universität Leipzig.
- Coffee, N., Koenig, J. P., Poornima, S., Forstall, C. W., Ossewaarde, R., & Jacobson, S. L. (2012). The Tesserae Project: intertextual analysis of Latin poetry. *Literary and linguistic computing*, *28*(2), 221-228.
- Feder, F., Kupreyev, M., Manning, E., Schroeder, C. T., & Zeldes, A. (2018). A linked Coptic dictionary online. In: *Proceedings of LaTeCH 2018 - The 11th SIGHUM Workshop at COLING2018 (Santa Fe, NM).*
- Franzini, G., Franzini, E., Büchler, M. (2016). Historical Text Reuse: What Is It?. http://www.etrap.eu/historical-text-re-use/.
- Horn, J. (2000). Die koptische (sahidische) Überlieferung des alttestamentlichen Psalmenbuches - Versuch einer Gruppierung der Textzeugen für die Herstellung des Texte. In: A. Aejmelaeus & U. Quast (eds.), *Der Septuaginta-Psalter und seine Tochterübersetzungen* (AAWG.PH Folge 3 320; MSU 24), Göttingen: Vandenhoeck & Ruprecht. 97-106.
- Kamps, J., & Marx, M. (2002). Visualizing wordnet structure. In: *Proceedings of the 1st International Conference on Global WordNet.* 182-186.
- Lincke, E.-S., Bulert, K., & Büchler, M. (2019). Optical Character Recognition for Coptic fonts: A multi-source approach for scholarly editions. In: *Proceedings of DATeCH 2019*.

# Selected Bibliography 2

- Miyagawa, S., Zeldes, A., Büchler, M., Behlmer, H., & Troy Griffitts. (2018). Building linguistically and intertextually tagged Coptic corpora with open source tools. In: C. Suzuki (ed.), *Proceedings of the 8th Conference of Japanese Association for Digital Humanities*. Tokyo: Center for Open Data in the Humanities. 139-141.
- Miyagawa, S., Bulert, K., Büchler, M. & Behlmer, H. (2019). Optical character recognition of typeset Coptic text with neural networks. *Digital Scholarship in the Humanities* 34 (Supplement 1). i135-141.
- Schroeder, C. T., & Zeldes, A. (2016). Raiders of the Lost Corpus. *Digital Humanities Quarterly* 10.2.
- Slaughter, L., Morgado da Costa, L., Miyagawa, S., Büchler, M., Zeldes, A. & Behlmer, H. (2019). The making of Coptic Wordnet. In: C. Fellbaum, P. Vossen, E. Rudnicka, M. Maziarz, and M. Piasecki (eds.), *Proceedings of 10th Global WordNet Conference (GWC 2019)*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej. 166-175.
- Zeldes, A. & Schroeder, C. T. (2016) An NLP Pipeline for Coptic. In: *Proceedings of LaTeCH 2016 - The 10th SIGHUM Workshop at the Annual Meeting of the ACL (Berlin).* 146–55.
- Zeldes, A & Schroeder, C. T. (2015) Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities, *Digital Scholarship in the Humanities* 30 (Supplement 1). i164–76.

# Thank you!

So Miyagawa (miyagawa.so.36u@kyoto-u.jp)
Marco Büchler (buechler@infai.org)
Heike Behlmer (hbehlme@uni-goettingen.de)

+ABBA
ϹΕ
ΝΟΥ
ΘΪΟΥ
ΑΡ
ΧΗ
ΜΑΝ