Corpus Linguistics Tools for Sahidic Coptic

Amir Zeldes,
Humboldt-Universität zu Berlin
amir.zeldes@rz.hu-berlin.de

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

Leipzig eHumanities Seminar, 18.12.2013

Plan

- Introduction: Coptic and Corpus Linguistics
- Tools for annotating Coptic
  - Normalization
  - Tokenization
  - **POS Tagging**
- Tentative applications
- Conclusion and outlook

Who are these people?

- Dr. Amir Zeldes –
  Korpuslinguistik /
  SFB 632 Information Structure
  Humboldt-Universität zu Berlin

- Prof. Caroline T. Schroeder –
  Religious and Classical Studies /
  Humanities Center Director
  University of the Pacific

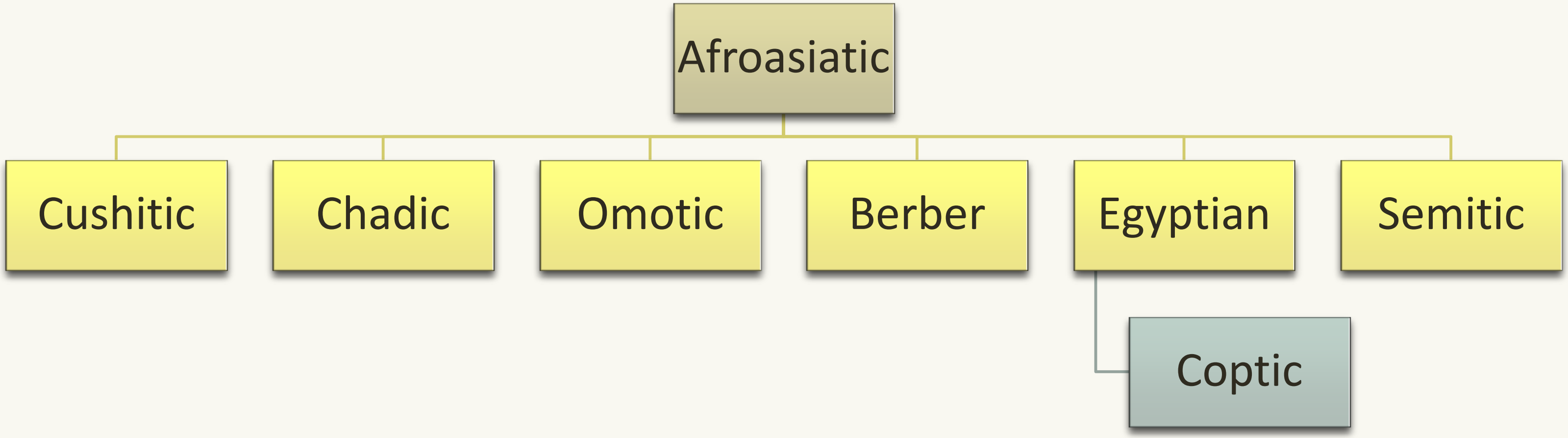
- Cooperation ***Coptic SCRIPTORIUM*** established at 2012 NEH summer institute on "Text in a Digital Age" (Tufts): http://coptic.pacific.edu/

What is Coptic?

- Last stage of the Ancient Egyptian Language
  (Longest continuous documentation of any language)
- Spoken in Hellenistic Egypt, primarily in 1st Millennium
- Heavy influence from Greek – a contact language
- Massive amounts of text preserved
  (Egyptian climate + papyrus = happy philologists ☺)
- ... but also pillaged, ripped up, sold to many different libraries, lost ...

Why study Coptic?

- Linguistically unique:
  - Documents transition: agglutinative < isolating < synthetic
  - Crucial for reconstructing Egyptian vowels, Proto-Afroasiatic
  - Comparative insights for Semitic, African languages



Afroasiatic

Cushitic

Chadic

Omotic

Berber

Egyptian

Semitic

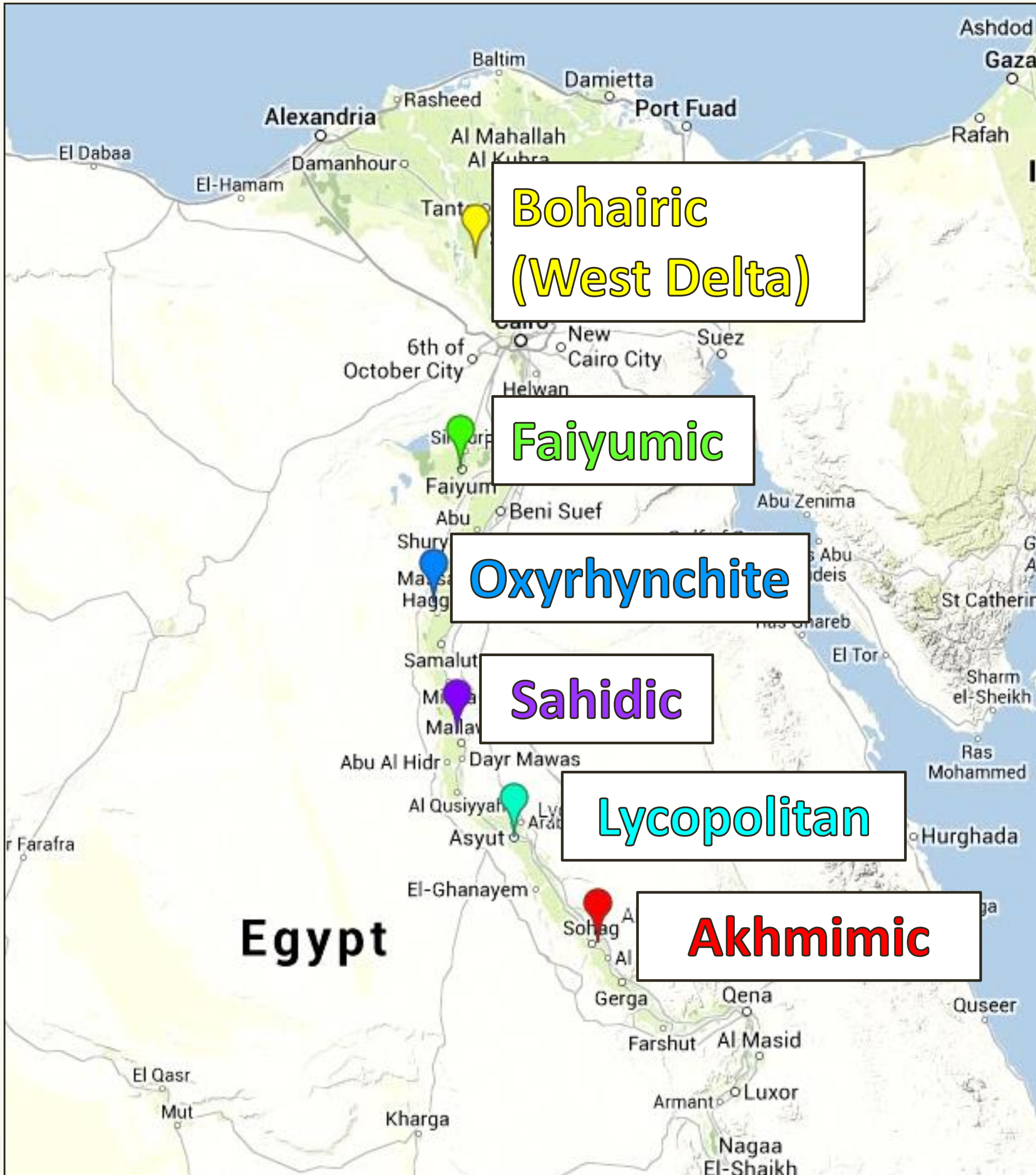
Coptic

Why study Coptic?

- Invaluable for the study of early Christianity
  - Rise of monasticism (Pachomius, the Desert Fathers)
  - Largest collection of Gnostic texts (Nag Hammadi library), unique hagiographies
  - Some of the most controversial texts, non-canonical gospels (e.g. Thomas, Mary, and most recently "Jesus's Wife")
- Much work to be done:
  - Only a fraction of texts are published
  - Extremely little online (compare Greek and Latin!)

Sahidic Coptic

- Coptic in use almost 2000 years
- Multiple dialects, periods
- Classical form: Sahidic (2nd-14th C.)
- Starting point for this project



What we would like to see

- Similar advances and availability to Greek and Latin
- As much text as possible online and free (CC-BY)
- Linguistically informed analyses
  - Segmentation (non-trivial as we will see)
  - Normalization (to find variants, abbreviations...)
  - Part-of-speech tagging (needed for linguistic analysis, vocabulary, identifying reuse; NB much homography!)
  - Search & visualization, corpus architecture, all respecting paleographic and text-linguistic interests, e.g. line breaks in words, but whole words... (→ talk in Berlin next month)

A word about the texts in this talk

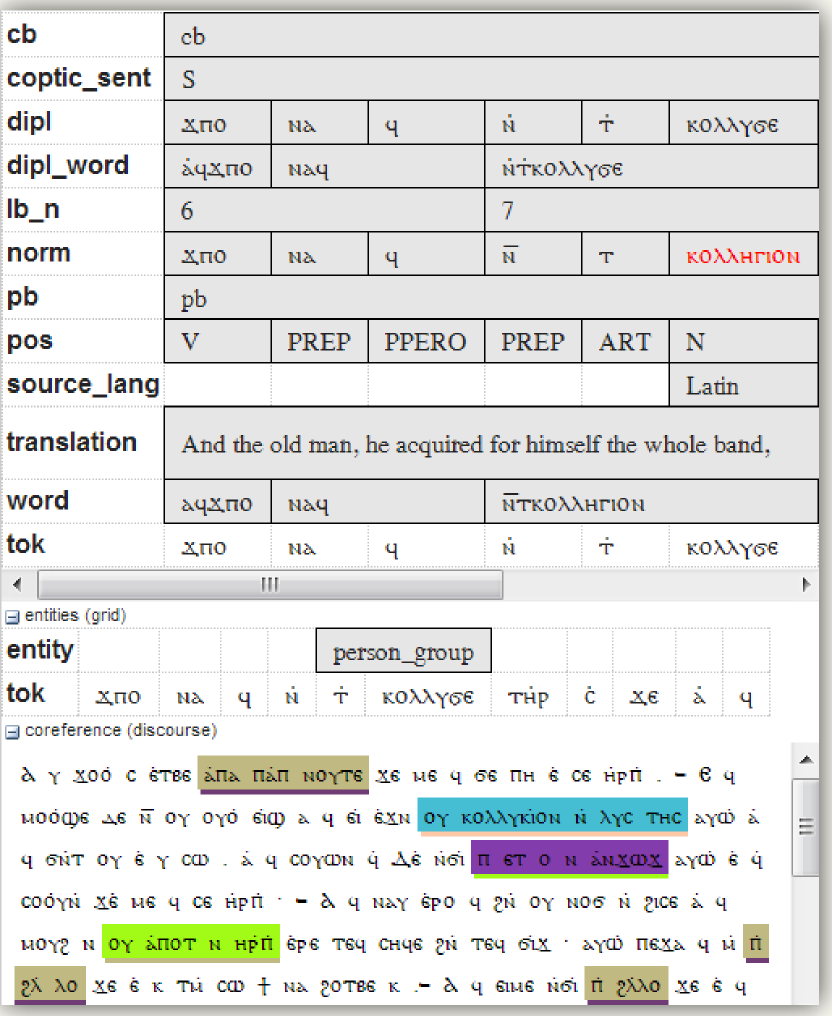
- So far we've concentrated on Shenoute's sermon Abraham our Father:
  - *"As for us, brethren, let us live by the truth so that we are upstanding in all our works, and so that the prophets, apostles and all the saints might dwell among us, ..."*
- Apophthegmata Patrum:
  - *"They said about the blessed Sarah the virgin that she spent sixty years living at the top of the river and she never set foot outside to see the river."*
- New Testament, esp. Gospel of Mark

Corpus linguistics

- Years of experience dealing with linguistic annotation (some examples in the next slides)
- Encoding, search, retrieval and visualization
- Mantras for re-usable, trainable, open source tools:
  - *Don't write your own POS-tagger – try training one first*
  - *Don't write a search webpage – use off the shelf software*
  - ....
  - *And put everything online for others to use/develop further!*

Some stuff we've been working on

- From running text to tokenized, segmented and tagged data (this talk)
- Representing diplomatic MSS, corpus architecture, metadata (talk at Berlin Digital Classicist Seminar next month)
- Language of origin (manual)
- Coreference and named entities (manual)



cb	cb					
coptic_sent	S					
dipl	ϫⲡⲟ	ⲛⲁ	ϥ	ⲛ̇	ⲧ̇	ⲕⲟⲗⲗⲩϭⲉ
dipl_word	ⲁ̀ϥϫⲡⲟ	ⲛⲁϥ		ⲛ̇ⲧ̇ⲕⲟⲗⲗⲩϭⲉ		
lb_n	6			7		
norm	ϫⲡⲟ	ⲛⲁ	ϥ	ⲛ̄	ⲧ	ⲕⲟⲗⲗⲏⲅⲓⲟⲛ
pb	pb					
pos	V	PREP	PPERO	PREP	ART	N
source_lang						Latin
translation	And the old man, he acquired for himself the whole band,					
word	ⲁϥϫⲡⲟ	ⲛⲁϥ		ⲛ̄ⲧⲕⲟⲗⲗⲏⲅⲓⲟⲛ		
tok	ϫⲡⲟ	ⲛⲁ	ϥ	ⲛ̇	ⲧ̇	ⲕⲟⲗⲗⲩϭⲉ

⊟ entities (grid)

entity					person_group						
tok	ϫⲡⲟ	ⲛⲁ	ϥ	ⲛ̇	ⲧ̇	ⲕⲟⲗⲗⲩϭⲉ	ⲧⲏ̇ⲣ	ⲥ̇	ϫⲉ	ⲁ̀	ϥ

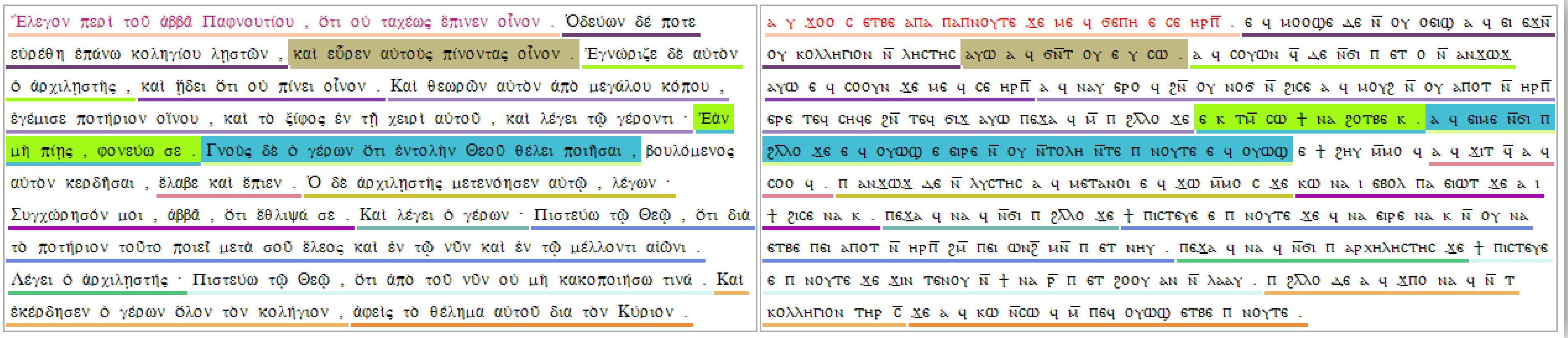
⊟ coreference (discourse)

ⲁ̀ ⲩ ϫⲟⲟ́ ⲥ ⲉ̀ⲧⲃⲉ ⲁ̀ⲡⲁ ⲡⲁ̀ⲡ ⲛⲟⲩⲧⲉ ϫⲉ ⲙⲉ ϥ ϭⲉ ⲡⲏ ⲉ̀ ⲥⲉ ⲏ̇ⲣⲡ̄ . – ⲉ̀ ϥ
ⲙⲟⲟ́ϣⲉ ⲇⲉ ⲛ̄ ⲟⲩ ⲟⲩⲟ́ ⲉ̀ⲓϣ ⲁ ϥ ⲉ̀ⲓ ⲉ̀ϫⲛ ⲟⲩ ⲕⲟⲗⲗⲩⲕⲓⲟⲛ ⲛ̇ ⲗⲩⲥ ⲧⲏⲥ ⲁⲩⲱ̀ ⲁ̀
ϥ ϭⲛ̇ⲧ ⲟⲩ ⲉ̀ ⲩ ⲥⲱ . ⲁ̀ ϥ ⲥⲟⲩⲱⲛ ϥ̀ ⲇⲉ̀ ⲛ̇ϭⲓ ⲡ ⲉⲧ ⲟ ⲛ ⲁ̀ⲛϫⲱϫ ⲁⲩⲱ̀ ⲉ̀ ϥ̀
ⲥⲟⲟ́ⲩⲛ̇ ϫⲉ̀ ⲙⲉ ϥ ⲥⲉ ⲏ̇ⲣⲡ̄ · – ⲁ̀ ϥ ⲛⲁⲩ ⲉ̀ⲣⲟ ϥ ϩⲛ̇ ⲟⲩ ⲛⲟϭ ⲛ̇ ϩⲓⲥⲉ ⲁ̀ ϥ
ⲙⲟⲩϩ ⲛ ⲟⲩ ⲁ̀ⲡⲟⲧ ⲛ ⲏⲣ̇ⲡ̄ ⲉ̀ⲣⲉ ⲧⲉϥ ⲥⲏϥⲉ ϩⲛ̇ ⲧⲉϥ ϭⲓϫ · ⲁⲩⲱ̀ ⲡⲉϫⲁ ϥ ⲙ̇ ⲡ̄
ϩⲗ ⲗⲟ ϫⲉ ⲉ̀ ⲕ ⲧⲙ̇ ⲥⲱ ϯ ⲛⲁ ϩⲟⲧⲃⲉ ⲕ .– ⲁ̀ ϥ ⲉⲓⲙⲉ ⲛ̇ϭⲓ ⲡ̇ ϩⲗ̄ⲗⲟ ϫⲉ ⲉ̀ ϥ

ANNIS search interface: https://korpling.german.hu-berlin.de/annis3/scriptorium

Some stuff we've been working on

- Parallel alignment Greek <> Coptic
  - Apophthegmata Patrum:



Ἔλεγον περὶ τοῦ ἀββᾶ Παφνουτίου , ὅτι οὐ ταχέως ἔπινεν οἶνον . Ὁδεύων δέ ποτε
εὑρέθη ἐπάνω κοληγίου λῃστῶν , καὶ εὗρεν αὐτοὺς πίνοντας οἶνον . Ἐγνώριζε δὲ αὐτὸν
ὁ ἀρχιλῃστὴς , καὶ ᾔδει ὅτι οὐ πίνει οἶνον . Καὶ θεωρῶν αὐτὸν ἀπὸ μεγάλου κόπου ,
ἐγέμισε ποτήριον οἴνου , καὶ τὸ ξίφος ἐν τῇ χειρὶ αὐτοῦ , καὶ λέγει τῷ γέροντι · Ἐὰν
μὴ πίῃς , φονεύω σε . Γνοὺς δὲ ὁ γέρων ὅτι ἐντολὴν Θεοῦ θέλει ποιῆσαι , βουλόμενος
αὐτὸν κερδῆσαι , ἔλαβε καὶ ἔπιεν . Ὁ δὲ ἀρχιλῃστὴς μετενόησεν αὐτῷ , λέγων ·
Συγχώρησόν μοι , ἀββᾶ , ὅτι ἔθλιψά σε . Καὶ λέγει ὁ γέρων · Πιστεύω τῷ Θεῷ , ὅτι διὰ
τὸ ποτήριον τοῦτο ποιεῖ μετὰ σοῦ ἔλεος καὶ ἐν τῷ νῦν καὶ ἐν τῷ μέλλοντι αἰῶνι .
Λέγει ὁ ἀρχιλῃστής · Πιστεύω τῷ Θεῷ , ὅτι ἀπὸ τοῦ νῦν οὐ μὴ κακοποιήσω τινά . Καὶ
ἐκέρδησεν ὁ γέρων ὅλον τὸν κολήγιον , ἀφεὶς τὸ θέλημα αὐτοῦ δια τὸν Κύριον .

ⲁ ⲩ ϫⲟⲟ ⲥ ⲉⲧⲃⲉ ⲁⲡⲁ ⲡⲁⲡⲛⲟⲩⲧⲉ ϫⲉ ⲙⲉ ϥ ϭⲉⲡⲏ ⲉ ⲥⲉ ⲏⲣⲡ̅ . ⲉ ϥ ⲙⲟⲟϣⲉ ⲇⲉ ⲛ̅ ⲟⲩ ⲟⲉⲓϣ ⲁ ϥ ⲉⲓ ⲉϫⲛ̅
ⲟⲩ ⲕⲟⲗⲗⲏⲅⲓⲟⲛ ⲛ̅ ⲗⲏⲥⲧⲏⲥ ⲁⲩⲱ ⲁ ϥ ϭⲛ̅ⲧ ⲟⲩ ⲉ ⲩ ⲥⲱ . ⲁ ϥ ⲥⲟⲩⲱⲛ ϥ̅ ⲇⲉ ⲛ̅ϭⲓ ⲡ ⲉⲧ ⲟ ⲛ̅ ⲁⲛϫⲱϫ
ⲁⲩⲱ ⲉ ϥ ⲥⲟⲟⲩⲛ ϫⲉ ⲙⲉ ϥ ⲥⲉ ⲏⲣⲡ̅ ⲁ ϥ ⲛⲁⲩ ⲉⲣⲟ ϥ ϩⲛ̅ ⲟⲩ ⲛⲟϭ ⲛ̅ ϩⲓⲥⲉ ⲁ ϥ ⲙⲟⲩϩ ⲛ̅ ⲟⲩ ⲁⲡⲟⲧ ⲛ̅ ⲏⲣⲡ̅
ⲉⲣⲉ ⲧⲉϥ ⲥⲏϥⲉ ϩⲛ̅ ⲧⲉϥ ϭⲓϫ ⲁⲩⲱ ⲡⲉϫⲁ ϥ ⲙ̅ ⲡ ϩⲗ̅ⲗⲟ ϫⲉ ⲉ ⲕ ⲧⲙ̅ ⲥⲱ ϯ ⲛⲁ ϩⲟⲧⲃⲉ ⲕ . ⲁ ϥ ⲉⲓⲙⲉ ⲛ̅ϭⲓ ⲡ
ϩⲗ̅ⲗⲟ ϫⲉ ⲉ ϥ ⲟⲩⲱϣ ⲉ ⲉⲓⲣⲉ ⲛ̅ ⲟⲩ ⲛ̅ⲧⲟⲗⲏ ⲛ̅ⲧⲉ ⲡ ⲛⲟⲩⲧⲉ ⲉ ϥ ⲟⲩⲱϣ ⲉ ϯ ϩⲏⲩ ⲙ̅ⲙⲟ ϥ ⲁ ϥ ϫⲓⲧ ϥ̅ ⲁ ϥ
ⲥⲟⲟ ϥ . ⲡ ⲁⲛϫⲱϫ ⲇⲉ ⲛ̅ ⲗⲩⲥⲧⲏⲥ ⲁ ϥ ⲙⲉⲧⲁⲛⲟⲓ ⲉ ϥ ϫⲱ ⲙ̅ⲙⲟ ⲥ ϫⲉ ⲕⲱ ⲛⲁ ⲓ ⲉⲃⲟⲗ ⲡⲁ ⲉⲓⲱⲧ ϫⲉ ⲁ ⲓ
ϯ ϩⲓⲥⲉ ⲛⲁ ⲕ . ⲡⲉϫⲁ ϥ ⲛⲁ ϥ ⲛ̅ϭⲓ ⲡ ϩⲗ̅ⲗⲟ ϫⲉ ϯ ⲡⲓⲥⲧⲉⲩⲉ ⲉ ⲡ ⲛⲟⲩⲧⲉ ϫⲉ ϥ ⲛⲁ ⲉⲓⲣⲉ ⲛⲁ ⲕ ⲛ̅ ⲟⲩ ⲛⲁ
ⲉⲧⲃⲉ ⲡⲉⲓ ⲁⲡⲟⲧ ⲛ̅ ⲏⲣⲡ̅ ϩⲙ̅ ⲡⲉⲓ ⲱⲛϩ̅ ⲙⲛ̅ ⲡ ⲉⲧ ⲛⲏⲩ . ⲡⲉϫⲁ ϥ ⲛⲁ ϥ ⲛ̅ϭⲓ ⲡ ⲁⲣⲭⲏⲗⲏⲥⲧⲏⲥ ϫⲉ ϯ ⲡⲓⲥⲧⲉⲩⲉ
ⲉ ⲡ ⲛⲟⲩⲧⲉ ϫⲉ ϫⲓⲛ ⲧⲉⲛⲟⲩ ⲛ̅ ϯ ⲛⲁ ⲣ̅ ⲡ ⲉⲧ ϩⲟⲟⲩ ⲁⲛ ⲛ̅ ⲗⲁⲁⲩ . ⲡ ϩⲗ̅ⲗⲟ ⲇⲉ ⲁ ϥ ϫⲡⲟ ⲛⲁ ϥ ⲛ̅ ⲧ
ⲕⲟⲗⲗⲏⲅⲓⲟⲛ ⲧⲏⲣ ⲥ̅ ϫⲉ ⲁ ϥ ⲕⲱ ⲛ̅ⲥⲱ ϥ ⲙ̅ ⲡⲉϥ ⲟⲩⲱϣ ⲉⲧⲃⲉ ⲡ ⲛⲟⲩⲧⲉ .

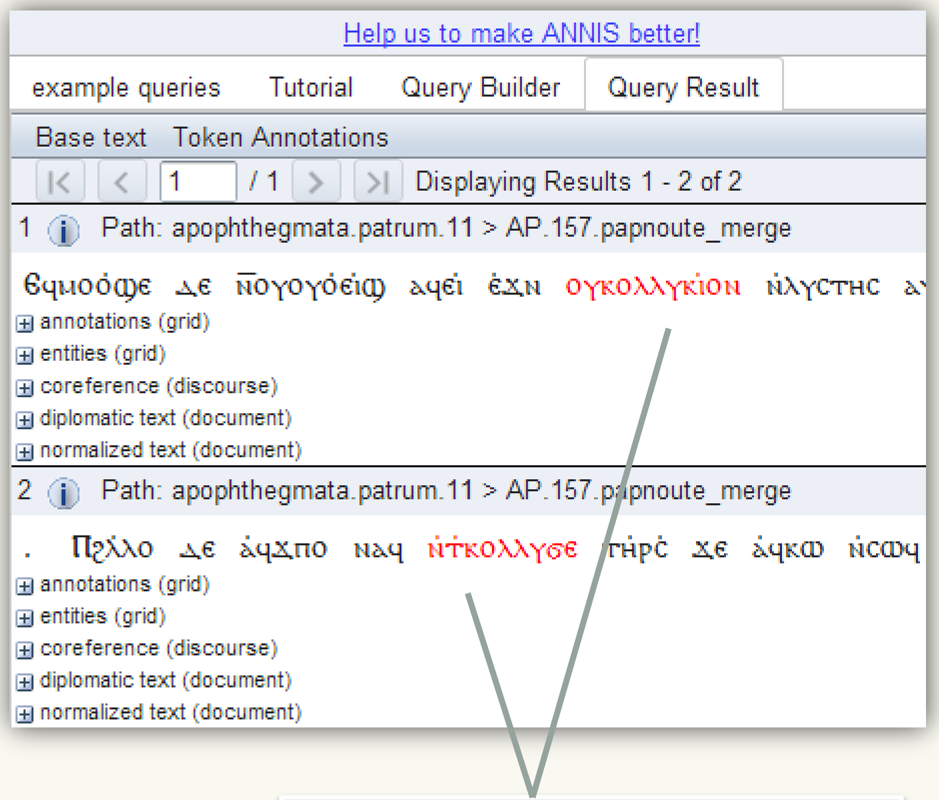
- Most of the corpus linguistics paradigm relies on **normalized, tokenized, consistently tagged data**
- How do we get there for Coptic?

Normalization

- Coptic uses a variant of the Greek alphabet
  - 24 + 6 letters adapted from Hieratic Egyptian: ϥ ϣ ϩ ϯ ϫ ϭ
    f sh h ti ch k^j
  - Many diacritics in MSS, e.g. superlinear strokes can signify: (but are often omitted)
    - Syllabic consonants: ⲙⲛ̄ⲧⲣⲙ̄ⲛ̄ⲕⲏⲙⲉ 'Coptic' (~ Egypt-man-ness)
    - Whole syllables containing these ⲙ̅ⲛ̅ⲧ̅
    - Omitted nasals: ⲥⲟⲟⲩ̅ for ⲥⲟⲟⲩⲛ 'to know'
    - Abbreviations (esp. nomina sacra, proper names): ⲓⲏ̄ⲗ = ⲓⲥⲣⲁⲏⲗ

Normalization

- Many other diacritics, potentially marking 'word' borders, potentially 'meaningless'
- Spelling can vary substantially, even for foreign words and even in the same manuscript



1 Path: apophthegmata.patrum.11 > AP.157.papnoute_merge

Ⲉϥⲙⲟⲟ́ϣⲉ ⲇⲉ ⲛ̄ⲟⲩⲟⲩⲟ́ⲉ̇ⲓϣ ⲁϥⲉ̇ⲓ ⲉ̇ϫⲛ ⲟⲩⲕⲟⲗⲗⲩⲕⲓ̇ⲟⲛ ⲛ̇ⲗⲩⲥⲧⲏⲥ ⲁ

2 Path: apophthegmata.patrum.11 > AP.157.papnoute_merge

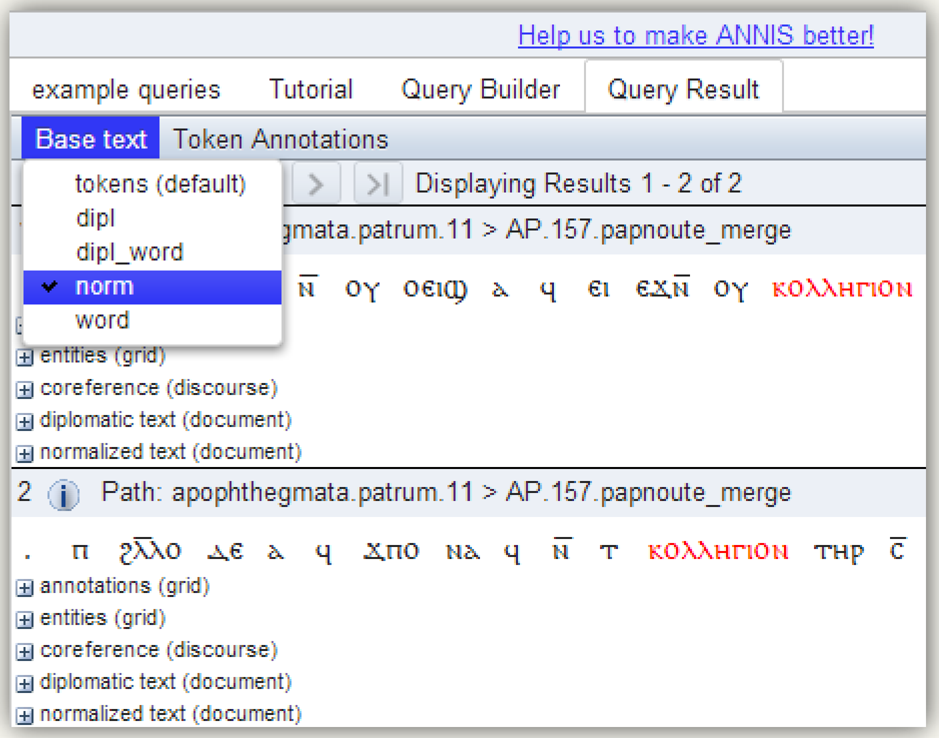
. Ⲡϩⲗ̀ⲗⲟ ⲇⲉ ⲁ̀ϥϫⲡⲟ ⲛⲁϥ ⲛ̇ⲧ̇ⲕⲟⲗⲗⲩϭⲉ ⲧⲏⲣⲥ̇ ϫⲉ ⲁ̀ϥⲕⲱ ⲛ̇ⲥⲱϥ

Can you guess the word?

Solution: *Collegium*

Normalization

- Current approach:
  - Keep diplomatic form and add normalization
  - Auto-normalization for diacritics
  - List of known abbreviations, growing
  - Switch freely between views in interface (ANNIS, Zeldes et al. 2009)



Help us to make ANNIS better!

example queries | Tutorial | Query Builder | Query Result

Base text | Token Annotations

tokens (default)
dipl
dipl_word
✓ norm
word

Displaying Results 1 - 2 of 2

gmata.patrum.11 > AP.157.papnoute_merge

ⲛ̄ ⲟⲩ ⲟⲉⲓϣ ⲁ ϥ ⲉⲓ ⲉϫⲛ̄ ⲟⲩ ⲕⲟⲗⲗⲏⲅⲓⲟⲛ

⊞ entities (grid)
⊞ coreference (discourse)
⊞ diplomatic text (document)
⊞ normalized text (document)

2 Path: apophthegmata.patrum.11 > AP.157.papnoute_merge

. ⲡ ϩⲗ︤ⲗ︥ⲟ ⲇⲉ ⲁ ϥ ϫⲡⲟ ⲛⲁ ϥ ⲛ̄ ⲧ ⲕⲟⲗⲗⲏⲅⲓⲟⲛ ⲧⲏⲣ ⲥ̄

⊞ annotations (grid)
⊞ entities (grid)
⊞ coreference (discourse)
⊞ diplomatic text (document)
⊞ normalized text (document)

Tokenization

- Coptic is an agglutinative language:
  - ϫⲓⲛⲧⲁⲓⲣ̄ⲙⲟⲛⲁⲭⲟⲥ *'Since I became a monk'*
    since-that-PAST-1sg-do-monk
  - ⲉⲛⲧⲁϥⲧⲣⲉⲛⲣⲡϣⲁ *'he who made us keep the ceremony'*
    REL-PAST-3sgM-CAUS-1pl-do-the-observance
- Impossible to analyze grammatically without segmenting
- But documents are written in *scriptio continua*(!)
- Different conventions on how to segment "words" (Layton 2004), some hints from "meaningless diacritics"

Tokenization – Step 1/2

- Word segmentation: (manual + re-segmentation script)

...........ⲛ̄

ⲟⲩϣⲏⲣⲉ`ⲛ̄ⲁ → ⲛ̄ⲟⲩϣⲏⲣⲉ` ⲛ̄ⲁⲃⲣⲁϩⲁⲙ`

ⲃⲣⲁϩⲁⲙ`... 'of-a-son of-Abraham'

most texts 'come like this' from researchers – phew!
(e.g. in EpiDoc XML, text files, MS Word etc.)

- The "apostrophes" in these examples correspond to our idea of **word forms** but this is only *sometimes* so

Tokenization – Step 2/2

- Morpheme segmentation: (automatic)

ⲛ̄ⲟⲩϣⲏⲣⲉ`	ⲛ̄ⲁⲃⲣⲁϩⲁⲙ`	→	ⲛ ⲟⲩ ϣⲏⲣⲉ	ⲛ	ⲁⲃⲣⲁϩⲁⲙ
of-a-son	of-Abraham		of a son	of	Abraham

- Automatic script operates on normalized text
- Lexicon and rule based (full-form lexicon supplied by CMCL, courtesy of Prof. Tito Orlandi)
- Ideally followed by manual correction
(possible for smaller MSS, less so for the whole Bible)

Examples and challenges

- Rules formulated as cascade of regular expressions, e.g.: Indefinite durative present/future:
  - `...`
  - `/^($exist)($nounlist)($verblist|$vstatlist|$advlist)$/`
  - `/^($exist)($nounlist)(ⲛⲁ)($verblist)$/`
  - `/^($exist)($nounlist)(ⲛⲁ)($verblist)($ppero)$/`
  - `...`
- Biggest problem – handling of out-of-lexicon items
- Secondary problem – rule order occasionally causes errors

Examples and challenges

- A further problem comes from letters belonging to two tokens: ⲧ /p/ + ϩ /h/ > ⲑ /th/ (aspirated pronunciation of ⲑ, ⲫ, ⲭ)
  - ⲑⲉ = ⲧ + ϩⲉ 'the way'
  - similarly: ⲑⲁⲗⲁⲥⲥⲁ = ⲧ + ϩⲁⲗⲁⲥⲥⲁ 'the sea' ☺
- digraph ϯ /ti/ also a problem (e.g. ⲛϯⲟⲩⲇⲁⲓⲁ 'of Judea')
- Lexicon must be consulted even before tokenization!
- In practice: two step process with and without trying to split the word form
- Current accuracy: 84.29% (Bible) – 94.44% (Apophthegmata)

Part-of-speech tagging

- With segmented text, computational linguistics methods become more easily applicable
- Two part-of-speech tag sets developed:
  (based on Layton 2004)
  - Fine-grained: 45 tags (all different auxiliaries, converters, proper and common nouns, imperative and stative verbs, different types of pronouns)
  - Coarse-grained: 22 tags (APST → A, ... C, N, V, PPER)

Tagset overview

A[*]	Auxiliary base	ⲁ[ϥ], ⲙⲉ[ϥ], ⲧⲣⲉ[ϥ]
ADV	Adverb	ⲉⲃⲟⲗ, ⲟⲛ, ⲡⲱⲥ
ART	Article	ⲡ(ⲉ), ⲧ(ⲉ), ⲛ(ⲉ), ϩⲉⲛ
C[*]	Converter	ⲉ, ⲉⲧⲉ, ⲛⲉ, ...
CONJ	Conjunction	ⲁⲩⲱ, ⲏ, ⲙⲏ, ⲕⲁⲓ, ⲉⲓⲧⲉ
COP	Copula	ⲡⲉ/ⲧⲉ/ⲛⲉ
EXIST	Existential	ⲟⲩⲛ/ⲙⲛ
FUT	Future	ⲛⲁ
IMOD	Inflected modifier	ⲧⲏⲣ[ϥ], ϩⲱⲱ[ⲧ], …
N[*]	Noun	ⲁⲑⲏⲧ, ⲣⲱⲙⲉ, ⲁⲣⲭⲏ, ...
NEG	Negation	ⲛ, ⲁⲛ, ⲧⲙ[ⲥⲱⲧⲙ]
NUM	Numeral	ⲟⲩⲁ, ⲥⲛⲁⲩ, …

PDEM	Pronoun, demonst.	ⲡⲉⲓ/ⲡⲁⲓ, ⲧⲉⲓ/ⲧⲁⲓ
PINT	Pronoun, interrog.	ⲟⲩ, ⲛⲓⲙ
PPER[*]	Pronoun, personal	ϥ,ⲥ,ϯ,ⲛ,ⲁⲛⲟⲕ,ⲁⲛⲅ̄
PPOS[*]	Pronoun, possess.	ⲡⲉϥ,ⲧⲉⲧⲛ̄,ⲡⲟⲩ,ⲡⲁ
PREP	Preposition	ⲉⲧⲃⲉ, ϩⲛ̄, ⲛ, ⲙ̄ⲙⲟ[ϥ]
PTC	Particle	ⲇⲉ, ⲛ̄ϭⲓ, ϫⲉ, …
PUNCT	Punctuation	. , · …
UNKNOWN	Unknown, lacuna	ⲃ_ _ _, _ _ⲟⲥ, _ _ _
V[*]	Verb	ⲥⲱⲧⲡ, ⲥⲟⲧⲡ, ⲟ, ⲁⲣⲓ
VBD	Verboid	ⲛⲁⲛⲟⲩ[ϥ], ⲡⲉϫⲁ[ϥ]

Interannotator agreement

- The quality of a tag set is only as good as a **human's** ability to tag text correctly
- Guidelines must be provided to decide each case – SCRIPTORIUM guidelines (Zeldes & Schroeder 2013)
- Agreement experiment Schroeder / Zeldes
  - 1500 tokens (minus some invalidated cases)
  - Identical pos tags: 1396 / 1482 = **94.19%** (coarse: 96.15%)
  - Cohen's Kappa: **κ = 93.67**
    (considers chance agreement, cf. Artstein & Poesio 2008)

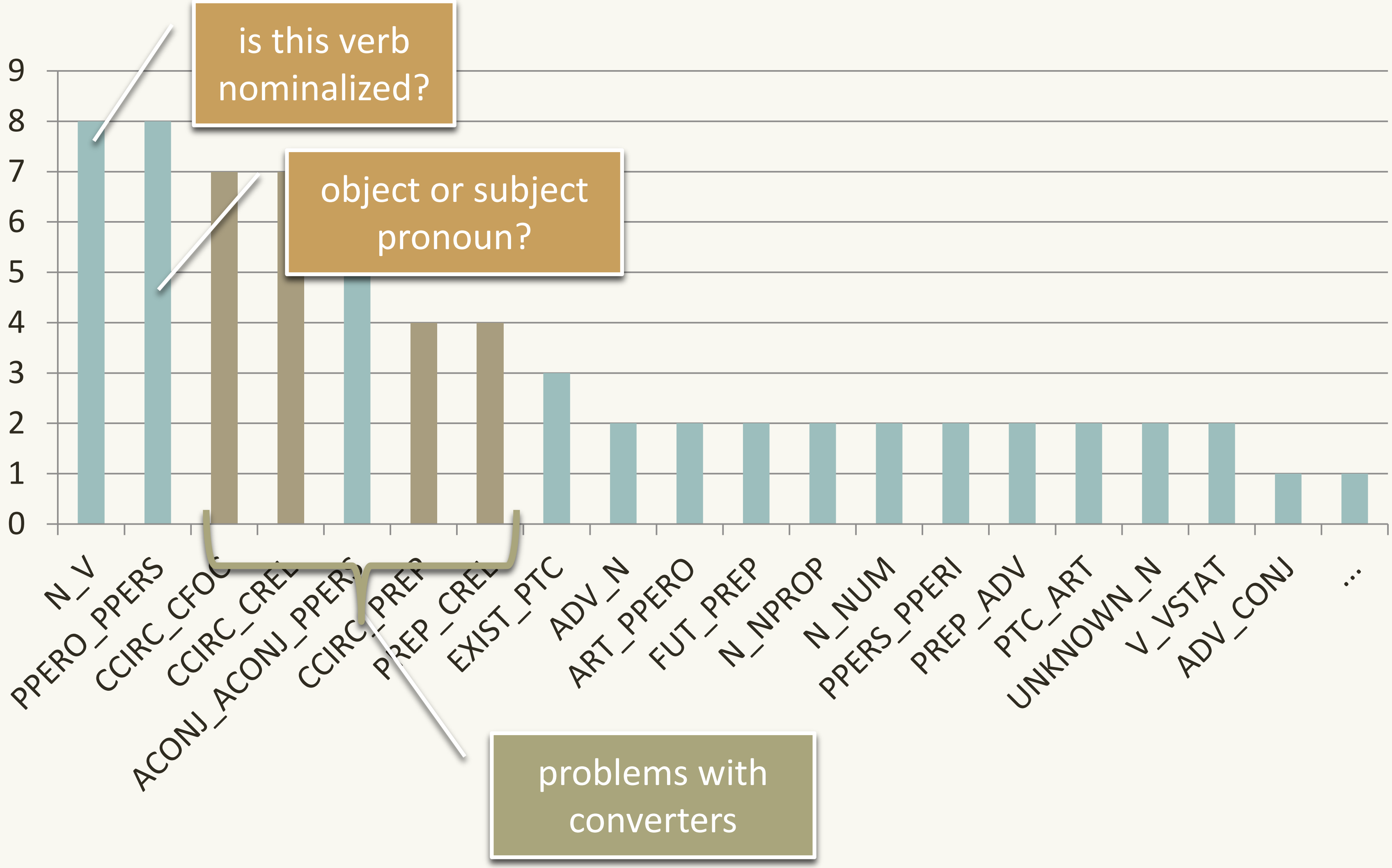
Where are the problems?

- Agreement similar across genres:

Shenoute – Abraham our Father:	854/906=94.26%
Apophthegmata Patrum:	542/576=94.09%

- Some problems can be solved by refining guidelines, continuing training, etc.
- Other problems are not so easy

Where are the problems?



is this verb nominalized?

object or subject pronoun?

problems with converters

Example: Converters

- Coptic has morphemes called "converters"
- Three in particular share the same form in **some** environments: **ⲉ**
- Decision often based on interpretation:

ⲉϥϯⲙ̄ⲧⲟⲛ ⲇⲉ ⲟⲛ` ⲛ̄ⲧⲙⲁⲁⲩ` ⲛ̄ⲧⲁⲥϫⲡⲟϥ
*(...**?** And thus he gives rest to the mother who bore him...)*

  - Focalizing (CFOC): *It is **to the mother** that he gives rest...*
  - Circumstantial (CCIRC): ***while** he gives rest to the mother...*
  - Relative (CREL): ***who** gives rest to the mother...*

Training a tagger for Coptic

- Tag set is brand new
- No training data available
- How do we get the most out of a small sample?
  - Diversify genres
  - Carefully craft the tag set
- Work in progress:
  - Select "best" data to include in training set
  - Extrapolate additional training data

Different genres in manual training set

Corpus	manual morphs	+auto morphs	total tokens
Abraham our Father (Shenoute of Atripe)	1908	7111	7688
Apophthegmata Patrum	1395	1395	1501
Sahidica NT	1229	209,633	209,633
	4532	**218,139**	**218,822**

Crafting the tag set

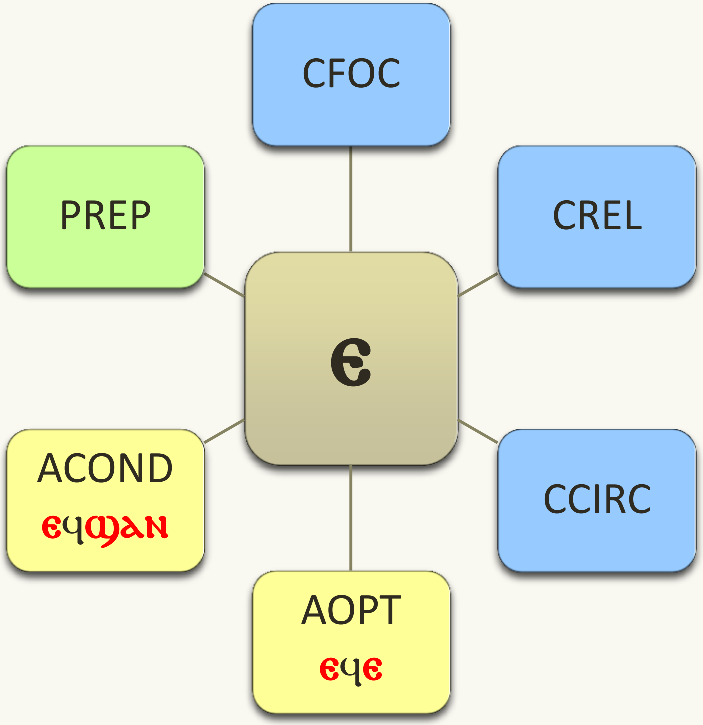
- Tag sets should be informative – we want to know if something is a noun or a verb
- But don't bite off more than you can chew:
  - Example: Should an English tagger try to identify subjunctive verbs?
  - probably not (none do!)
  - usually indistinguishable from indicatives:
    - *I demand that John go / goes* (distinguishable case – how to identify?)
    - *I demand that you go* (indistinguishable!)

Crafting the tag set

- Some Coptic compromises:
  - Tag the visibly different verb forms: stative, morphological imperatives
  - Don't try for other imperatives, plural vs. singular nouns...
  - Don't tag the internal structure of words:
    Coptic = mnt-rm-n-kēme < Egypt + man + ness
    *tempting to break down but obscures this being a **noun***
  - Annotating morphemes below the POS level is still possible on a separate annotation layer!
- Try to make things uniform: a sentence has a subject (noun with article etc. or pronoun), predicate, objects, prepositional phrases...

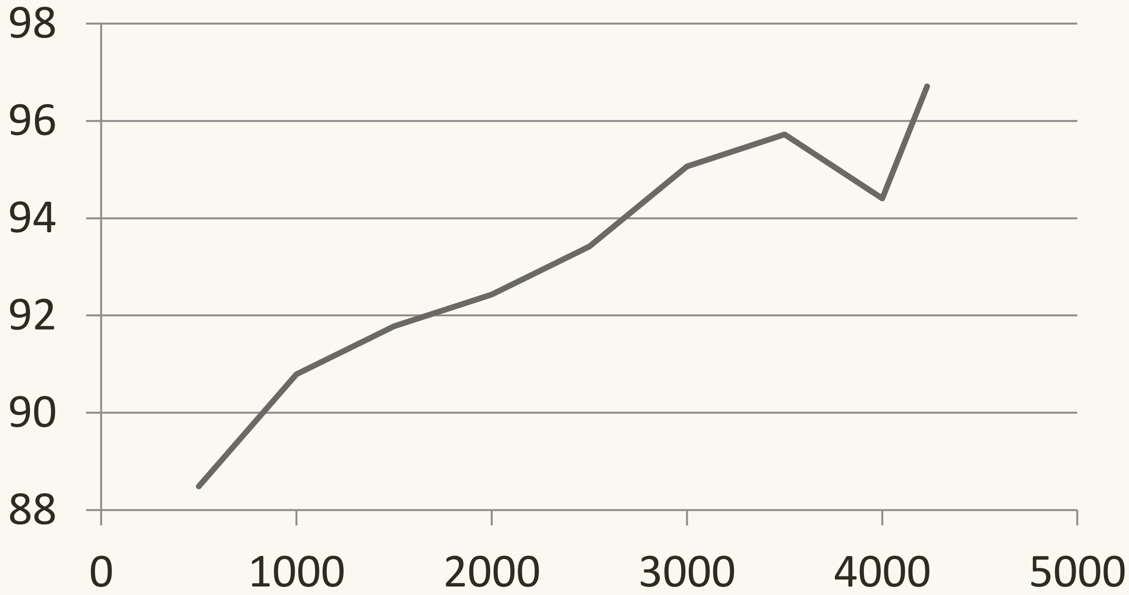
Closed vs. open classes

- Open classes are hard, unknown items are **allowed**
  - Nouns, verbs; no attempt to identify adjectives in Coptic
  - In fine grained tag set also proper nouns (hard, but important!)
- Closed classes are no problem when unambiguous...



Evaluation

- Performance on 10% held out data (500 tokens) with almost full lexicon coverage, using TreeTagger (Schmid 1994)



- A little too good to be true – easy dataset?

Evaluation

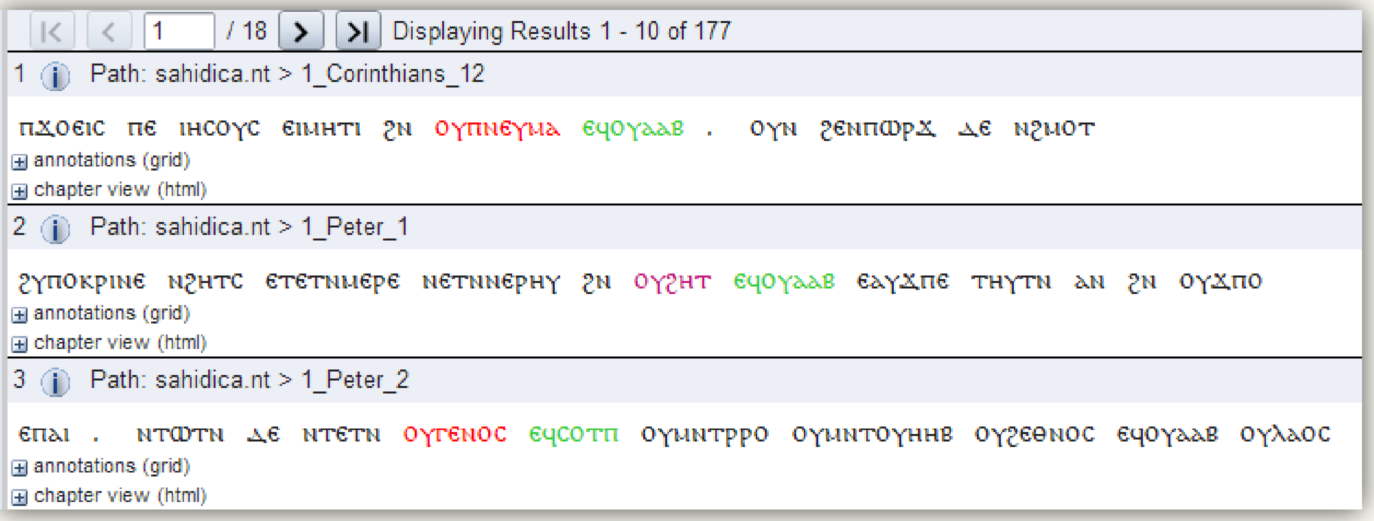
- 10-fold cross validation (each 10th is held out):
  - Average slice accuracy: 94.04%
  - More realistic
  - Sounds good, but remember: every 20th token is wrong!
- Results still very good for such a small training set
- Primary reason: lexicon coverage
  (even with 10% missing, Shenoute is Shenoute...)

Evaluation

- Out of domain toy evaluation: randomly selected text from papyri.info
  - First 50 tokens as a sanity check
  - Contract for delivering honey – completely different genre
  - Many open class items out-of-vocabulary, proper names
  - Accuracy: 79.6% (fine) / 87.7% (coarse)
- Work on robustness still needed
- Some ideas in the work, current WIP: "extrapolated data" (thanks to Ines Rehbein for this idea)

Getting more of the "best" data

- How to teach a stochastic tagger difficult distinctions?
  - ⲟⲩⲣⲱⲙⲉ ⲉϥⲥⲱⲧⲙ "a man who hears" or "while hearing"?
  - Some patterns exist: e.g. definite noun → CCIRC
- Idea: find unambiguous cases from the Bible in ANNIS



1 Path: sahidica.nt > 1_Corinthians_12

ⲡϫⲟⲉⲓⲥ ⲡⲉ ⲓⲏⲥⲟⲩⲥ ⲉⲓⲙⲏⲧⲓ ϩⲛ ⲟⲩⲡⲛⲉⲩⲙⲁ ⲉϥⲟⲩⲁⲁⲃ . ⲟⲩⲛ ϩⲉⲛⲡⲱⲣϫ ⲇⲉ ⲛϩⲙⲟⲧ

2 Path: sahidica.nt > 1_Peter_1

ϩⲩⲡⲟⲕⲣⲓⲛⲉ ⲛϩⲏⲧⲥ ⲉⲧⲉⲧⲛⲙⲉⲣⲉ ⲛⲉⲧⲛⲛⲉⲣⲏⲩ ϩⲛ ⲟⲩϩⲏⲧ ⲉϥⲟⲩⲁⲁⲃ ⲉⲁⲩϫⲡⲉ ⲧⲏⲩⲧⲛ ⲁⲛ ϩⲛ ⲟⲩϫⲡⲟ

3 Path: sahidica.nt > 1_Peter_2

ⲉⲡⲁⲓ . ⲛⲧⲱⲧⲛ ⲇⲉ ⲛⲧⲉⲧⲛ ⲟⲩⲅⲉⲛⲟⲥ ⲉϥⲥⲟⲧⲡ ⲟⲩⲙⲛⲧⲣⲣⲟ ⲟⲩⲙⲛⲧⲟⲩⲏⲏⲃ ⲟⲩϩⲉⲑⲛⲟⲥ ⲉϥⲟⲩⲁⲁⲃ ⲟⲩⲗⲁⲟⲥ

Extrapolation – making up more good data! ☺

- Data covers usage of some lexemes and inflections
- We have a lexicon with more words and paradigms
- Why not make up sentences by swapping out open class words like nouns and verbs?
- Let's try this for English

Not so easy

IN TRAINING DATA

- *the man ate a sandwich*
- *a boy sees a tree*
- …

EXTRAPOLATION VIA LEXICON

- *a sandwich drank the man*
- *a computer sees a people*
- …

- Need to consider morphosyntax (gender, number)
- Semantic compatibility
- Need to get appropriate *combinations* from the Bible

Automatic generation: some examples

Disambiguating ⲉ-

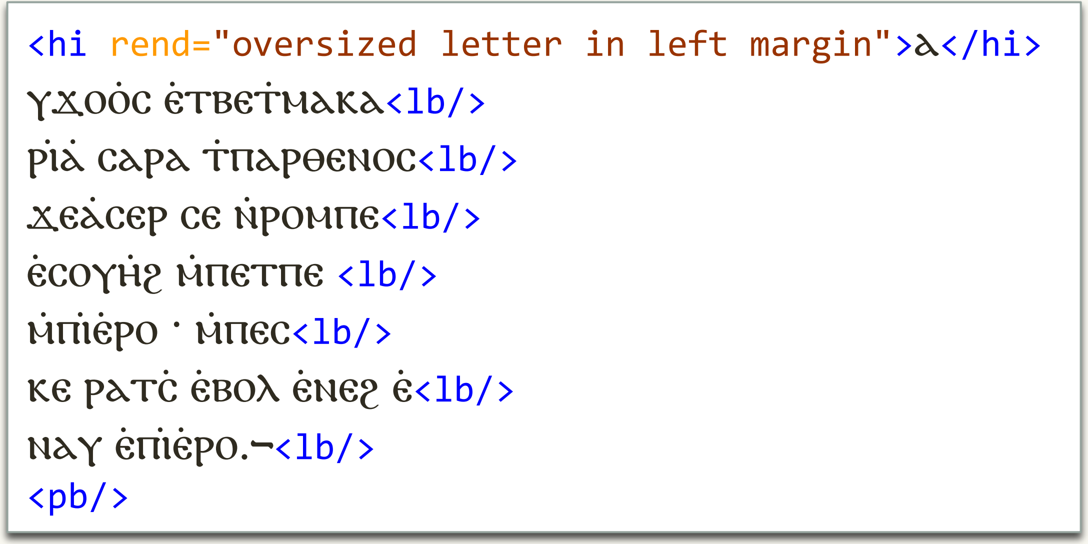
- ⲁⲩⲱ ⲉ|ⲥ|ⲛⲁⲩ . ⲁ|ⲛ|ⲥⲱⲧⲙ ⲉ|ⲡ|ⲛⲟⲩⲧⲉ .
  auō esnau ansōtm epnoute
  And while she saw, we listened to God.

- ϣⲁ|ⲕ|ⲉⲓ ⲉ|ⲡ|ϣⲏⲣⲉ · ϣⲁⲛⲧ|ⲩ|ϩⲁⲣⲉϩ ⲉ|ⲡ|ϫⲟⲉⲓⲥ .
  šakei epšēre šantuhareh epjoeis
  You always go to the son, until they observe ø the Lord

Stay tuned for how this turns out!

Why do all this corpus linguistics stuff?

- A lot of projects are digitizing manuscripts in TEI
- Huge advances over print editions in many ways
- Do we need more than plain text and fuzzy search, considering the effort?



```
<hi rend="oversized letter in left margin">ⲁ</hi>
ⲩϫⲟⲟ̇ⲥ ⲉ̇ⲧⲃⲉⲧ̇ⲙⲁⲕⲁ<lb/>
ⲣⲓ̇ⲁ̇ ⲥⲁⲣⲁ ⲧ̇ⲡⲁⲣⲑⲉⲛⲟⲥ<lb/>
ϫⲉⲁ̇ⲥⲉⲣ ⲥⲉ ⲛ̇ⲣⲟⲙⲡⲉ<lb/>
ⲉ̇ⲥⲟⲩⲏ̇ϩ ⲙ̇ⲡⲉⲧⲡⲉ <lb/>
ⲙ̇ⲡⲓ̇ⲉ̇ⲣⲟ · ⲙ̇ⲡⲉⲥ<lb/>
ⲕⲉ ⲣⲁⲧⲥ̇ ⲉ̇ⲃⲟⲗ ⲉ̇ⲛⲉϩ ⲉ̇<lb/>
ⲛⲁⲩ ⲉ̇ⲡⲓ̇ⲉ̇ⲣⲟ.¬<lb/>
<pb/>
```

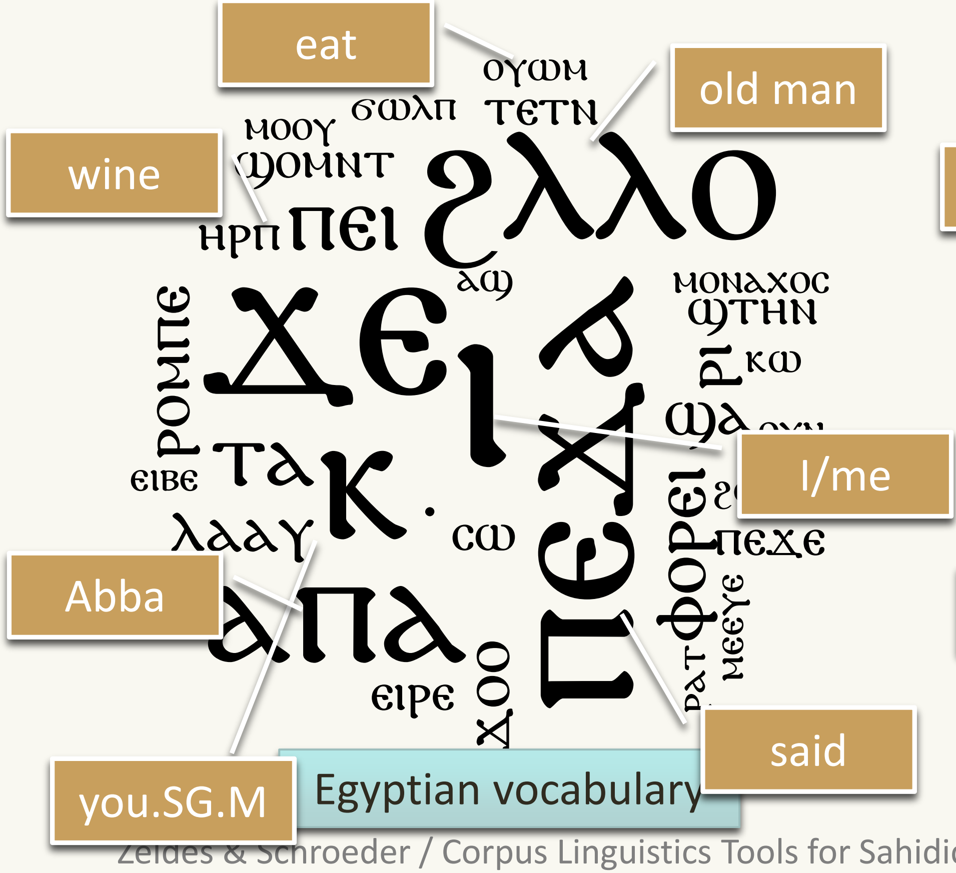
Why do all this corpus linguistics stuff?

- We need normalization, segmentation and tagging to run informed statistics and gain new insights:
  - What style is a text written in?
  - What is the most similar text to it?
  - What entities / kinds of entities is a text about?
  - Authorship?
  - Intertextuality?
  - POS tags for entry level quantitative work on grammar
- "Premium" machine readability – preaching to the choir?

What is a text about?

Run of the mill word clouds...

11 APOPHTHEGMATA PATRUM



eat

old man

wine

I/me

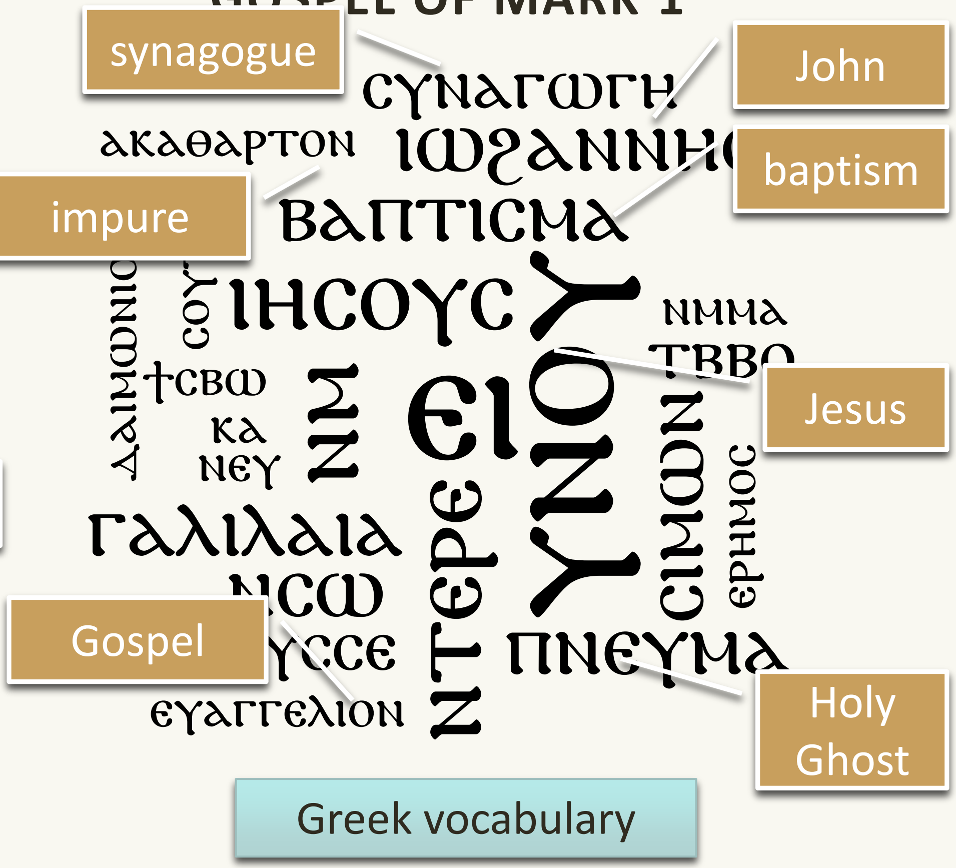
Abba

said

you.SG.M

Egyptian vocabulary

GOSPEL OF MARK 1



synagogue

John

baptism

impure

Jesus

Gospel

Holy Ghost

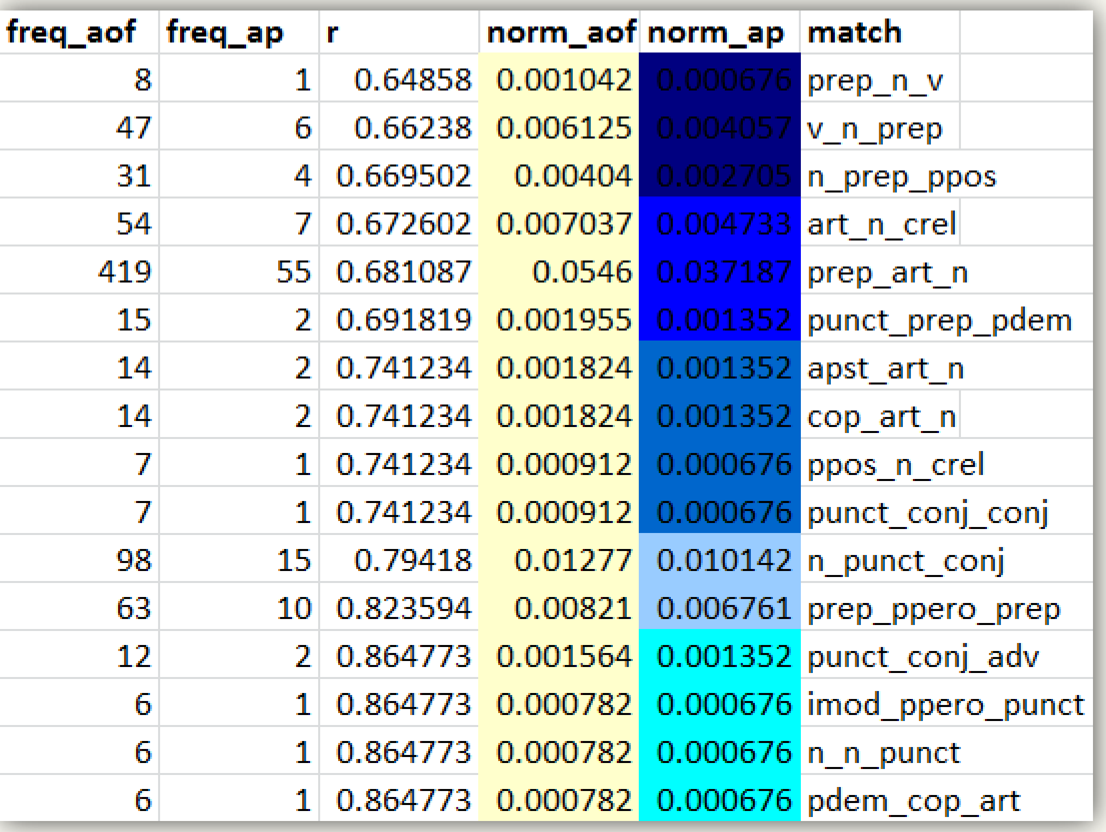
Greek vocabulary

What is a text about?

- Can't analyze vocabulary on complex word forms like ϫⲓⲛⲧⲁⲓⲣ̄ⲙⲟⲛⲁⲭⲟⲥ *'since I became a monk'*
- Can't deal with non-normalized text like ⲓ̅ⲏ̅ⲗ̅ = ⲓⲥⲣⲁⲏⲗ
- For many purposes we need more
  - Plots of just the verbs? Proper names? → POS tagging
  - Highlight, search and link place-names? → Entity tagging
  - Collapse inflected variants? → Lemmatization
  - Collapse prominent referents? → Coreference annotation
  - Dispersion of any of the above, alignment ... and much more

Grammatical characteristics

- Underuse/overuse analysis on POS n-grams in AP versus AOF:



freq_aof	freq_ap	r	norm_aof	norm_ap	match
8	1	0.64858	0.001042	0.000676	prep_n_v
47	6	0.66238	0.006125	0.004057	v_n_prep
31	4	0.669502	0.00404	0.002705	n_prep_ppos
54	7	0.672602	0.007037	0.004733	art_n_crel
419	55	0.681087	0.0546	0.037187	prep_art_n
15	2	0.691819	0.001955	0.001352	punct_prep_pdem
14	2	0.741234	0.001824	0.001352	apst_art_n
14	2	0.741234	0.001824	0.001352	cop_art_n
7	1	0.741234	0.000912	0.000676	ppos_n_crel
7	1	0.741234	0.000912	0.000676	punct_conj_conj
98	15	0.79418	0.01277	0.010142	n_punct_conj
63	10	0.823594	0.00821	0.006761	prep_ppero_prep
12	2	0.864773	0.001564	0.001352	punct_conj_adv
6	1	0.864773	0.000782	0.000676	imod_ppero_punct
6	1	0.864773	0.000782	0.000676	n_n_punct
6	1	0.864773	0.000782	0.000676	pdem_cop_art

Excel Plug-in: http://korpling.german.hu-berlin.de/~amir/uoaddin.htm

Grammatical characteristics

- Examples from cursory eyeballing:
  - Apopthegmata patrum:
    - **PTC_APST_PPERS**: particles preceding past tense verbs – Greek enclitics (especially ⲇⲉ, but others too)
    - **VBD_PPERS_PREP**: dialog, doubtless from '*he said to (him)*'
  - Abraham our Father:
    - **N_CREL_VSTAT**: a noun which is in a state → explicative – *marriage which is legitimate*, *brethren which are superior to them*, *thoughts of alienation which exist in our hearts*...)
    - Lots of **CFOC** n-grams: focalization as argumentative device
- Much more interesting: syntax trees... not yet there!

Conclusion

- Corpus linguistics tools are out there, ready to be used on historical texts in any language
- Worth the effort to (re-)train existing tools, adapt standards while not re-inventing the wheel
- The case of Coptic:
  - Promising early results on tagging and segmentation (need better handling of out of vocabulary items)
  - Disseminate tag set and tools, revise and retrain as needed

Outlook

- More data:
    - Test version of Acephalous 22 (Shenoute)
    - New Testament corpus
        - Gospel of Mark subset (manual)
        - Entire NT (automatic)
    - Letters by Besa (Shenoute's successor)
- More annotations:
    - Lemmatization
    - More work on entities
    - Syntax?

Outlook

- Next year – BMBF funded young researcher group on eHumanities at HU Berlin
- **KOMeT**:
  KOrpuslinguistische Methoden für ePhilologie mit TEI
  - Focus on marrying TEI resources with computational linguistics methods and formats
  - Developing NLP tools, search and visualization for ancient world textual resources
  - Pilot phase (2014, approved): Coptic
  - Main phase (2015-2019, pending): Other languages as well

Ⲙⲓⲱⲧⲛ ⲧⲱⲛⲟⲩ!

well-being+your.PL   greatly

=>

Thanks!

References

- Artstein, Ron & Massimo Poesio (2008), Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4), 556–596.
- Layton, Bentley (2004), *A Coptic Grammar*. Second Edition, Revised and Expanded. (Porta linguarum orientalium 20.) Wiesbaden: Harrassowitz.
- Schmid, Helmut (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, 44–49. Available at: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.
- Zeldes, Amir, Julia Ritz, Anke Lüdeling & Christian Chiarcos (2009), ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.
- Zeldes, Amir & Caroline Schroeder (2013), *SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic. Version: 1.0.1_2013.7.6a*. Available at: http://coptic.pacific.edu/download/tools/scriptorium_tagset_documentation.pdf.

Links

- Coptic SCRIPTORIUM: http://coptic.pacific.edu/
- ANNIS: http://www.sfb632.uni-potsdam.de/annis/
- Search engine for our corpora: https://korpling.german.hu-berlin.de/annis3/scriptorium
- Papyri.info: http://papyri.info/
- CMCL: http://cmcl.let.uniroma1.it/