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
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."

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)

ANNIS search interface: https://korpling.german.hu-berlin.de/annis3/scriptorium
Some stuff we've been working on

- Parallel alignment Greek <-> Coptic
  - Apophthegmata Patrum:

  ![Parallel alignment example]

- 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: ϒ ω ε Τ ι χ ο
     f sh h ti ch k j
  - Many diacritics in MSS, e.g. superlinear strokes can signify:
    (but are often omitted)
    - Syllabic consonants: μετρική 'Coptic' (~ Egypt-man-ness)
    - Whole syllables containing these μντ
    - Omitted nasals: ςονγ for ςονγ 'to know'
    - Abbreviations (esp. nomina sacra, proper names): μηλ = μεριμναλ
Normalization

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

- Coptic is an agglutinative language:
  - \(\text{ΔΙΝΤΑΙΦΜΟΝΑΧΟΣ}\) 'Since I became a monk'
    since-that-PAST-1sg-do-monk
  - \(\text{ΕΝΤΑΨΗΡΙΝΙΩΝ}\) '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)

  
  
  ⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀⠀莉

  ϋⲏⲗⲓⲁⲣⲓⲗ`ⲓⲏⲣⲉ ⎯→ ⌀ⲑⲓⲣⲓⲏⲣⲓⲗ` ⌀ⲏⲣⲓⲏⲣⲓⲗ` ⌀ⲗⲅⲡⲣⲓⲏ㎝` ⌀_aspect of-Abraham'

  ϋⲏⲗⲓⲣⲓⲏⲣⲓⲗ` ⌀_aspect of-a-son

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)

\[
\text{ⲛⲟⲩⲓⲣⲉ} \text{ⲛⲁⲃⲣⲁ} \Rightarrow \text{n oy whre n abraam}
\]

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: \( \tau /p/ + \varepsilon /h/ > \theta /th/ \) (aspirated pronunciation of \( \theta, \phi, \chi \))
  - \( \theta \varepsilon = \tau + \varepsilon \varepsilon \) 'the way'
  - similarly: \( \theta \alpha \lambda \alpha \varsigma \varsigma \alpha = \tau + \varepsilon \alpha \lambda \alpha \varsigma \varsigma \alpha \) 'the sea'

- digraph \( \dot{\tau} /ti/ \) also a problem (e.g. \( \nu \tau o y \delta \alpha \alpha \alpha \) '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

<table>
<thead>
<tr>
<th>A[*]</th>
<th>Auxiliary base</th>
<th>ω[ŋ], me[ŋ], tpe[ŋ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV</td>
<td>Adverb</td>
<td>εβολ, ὀν, πως</td>
</tr>
<tr>
<td>ART</td>
<td>Article</td>
<td>π(ε), τ(ε), ν(ε), γεν</td>
</tr>
<tr>
<td>C[*]</td>
<td>Converter</td>
<td>e, ete, ne, ...</td>
</tr>
<tr>
<td>CONJ</td>
<td>Conjunction</td>
<td>αυω, ὧ, μν, καί, είτε</td>
</tr>
<tr>
<td>COP</td>
<td>Copula</td>
<td>πε/τε/νε</td>
</tr>
<tr>
<td>EXIST</td>
<td>Existential</td>
<td>οΥΝ/ΜΝ</td>
</tr>
<tr>
<td>FUT</td>
<td>Future</td>
<td>ηα</td>
</tr>
<tr>
<td>IMOD</td>
<td>Inflected modifier</td>
<td>τρη[ŋ], ςωω[τ], ...</td>
</tr>
<tr>
<td>N[*]</td>
<td>Noun</td>
<td>αωητ, ρωμε, αρΧη, ...</td>
</tr>
<tr>
<td>NEG</td>
<td>Negation</td>
<td>η, ΑΝ, ΤΗ[ΣΩΤΗ]</td>
</tr>
<tr>
<td>NUM</td>
<td>Numerical</td>
<td>οΥΔ, ΚΝΑΥ, ...</td>
</tr>
<tr>
<td>PDEM</td>
<td>Pronoun, demonst.</td>
<td>πει/παι, τει/ται</td>
</tr>
<tr>
<td>PINT</td>
<td>Pronoun, interrog.</td>
<td>ογ, ΝΗ</td>
</tr>
<tr>
<td>PPER[*]</td>
<td>Pronoun, personal</td>
<td>ι, ū, η, ηοκ, ηογ</td>
</tr>
<tr>
<td>PPOS[*]</td>
<td>Pronoun, possess.</td>
<td>ηο, ηετή, ιού, ΠΑ</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition</td>
<td>ετβε, ηη, η, ηΗΗ[Ο]</td>
</tr>
<tr>
<td>PTC</td>
<td>Particle</td>
<td>δε, ηης, ΧΕ, ...</td>
</tr>
<tr>
<td>PUNCT</td>
<td>Punctuation</td>
<td>. , · ...</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>Unknown, lacuna</td>
<td>Β_ _ _, _ _ος, _ _</td>
</tr>
<tr>
<td>V[*]</td>
<td>Verb</td>
<td>σωτπ, σοτπ, ο, ΑΡΙ</td>
</tr>
<tr>
<td>VBD</td>
<td>Verboid</td>
<td>ΝΑΝΟΥ[Ν], ΝΕξΑ[Ν]</td>
</tr>
</tbody>
</table>
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: $\kappa = 93.67$

  (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: €
- Decision often based on interpretation:

€ⲉⲙⲧⲟⲛ ⲉⲥⲛ ⲉⲛⲥⲥ

(...? 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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>manual morphs</th>
<th>+auto morphs</th>
<th>total tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abraham our Father (Shenoute of Atrie)</td>
<td>1908</td>
<td>7111</td>
<td>7688</td>
</tr>
<tr>
<td>Apophthegmata Patrum</td>
<td>1395</td>
<td>1395</td>
<td>1501</td>
</tr>
<tr>
<td>Sahidica NT</td>
<td>1229</td>
<td>209,633</td>
<td>209,633</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4532</strong></td>
<td><strong>218,139</strong></td>
<td><strong>218,822</strong></td>
</tr>
</tbody>
</table>
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 10\textsuperscript{th} is held out):
  - Average slice accuracy: 94.04%
  - More realistic
  - Sounds good, but remember: every 20\textsuperscript{th} token is wrong!
- Results still very good for such a small training set
- Primary reason: lexicon coverage
  (even with 10\% missing, Shenoute is Shenoute...)

Leipzig, 18.12.2013
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?
  - οὑρωμε εὑρωτι "a man who hears" or "while hearing"?
  - Some patterns exist: e.g. definite noun → CCIRC

- Idea: find unambiguous cases from the Bible in ANNIS
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 e-

- αγω ϵε|ς|ναγ. α|ν|εωτμ ε|π|νογτε. 
aud esnau ansōtm epnoute

  And while she saw, we listened to God.

- ωα|κ|ει ε|π|ῳηρε. ωαντ|γ|γαρεζ ε|π|ξοεις.
  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?
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
- l/me

GOSPEL OF MARK 1

- synagogue
- impure
- l/me
- Gospel

- John
- baptism
- Jesus
- Holy Ghost

Egyptian vocabulary

Greek vocabulary

you.SG.M

Abba

eat

said

Greek vocabulary

Abba
What is a text about?

- Can't analyze vocabulary on complex word forms like ΔΙΝΤΑΙΠΜΟΝΑΧΟΣ 'since I became a monk'
- Can't deal with non-normalized text like ιηλ = ιεραηλ
- 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:

Excel Plug-in: [http://korpling.german.hu-berlin.de/~amir/uoaddin.htm](http://korpling.german.hu-berlin.de/~amir/uoaddin.htm)
Grammatical characteristics

- Examples from cursory eyeballing:
  - Apopthegmata patrum:
    - **PTC_APST_PERS**: particles preceding past tense verbs – Greek enclitics (especially ἢ, but others too)
    - **VBD_PERS_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
Miwtn twnoy!

well-being+your.PL greatly
=>
Thanks!
References

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/ascriptorium
- Papyri.info: http://papyri.info/
- CMCL: http://cmcl.let.uniroma1.it/