Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities

Amir Zeldes and Caroline T. Schroeder

Abstract

This paper motivates and details the first implementation of a freely available part of speech tag set and tagger for Coptic. Coptic is the last phase of the Egyptian language family and a descendent of the hieroglyphs of ancient Egypt. Unlike classical Greek and Latin, few resources for digital and computational work have existed for ancient Egyptian language and literature until now. We evaluate our tag set in an inter-annotator agreement experiment and examine some of the difficulties in tagging Coptic data. Using an existing digital lexicon and a small training corpus taken from several genres of literary Sahidic Coptic in the first half of the first millennium, we evaluate the performance of a stochastic tagger applying a fine grained and coarse grained set of tags within and outside the domain of literary texts. Our results show that a relatively high accuracy of 94-95% correct automatic tag assignment can be reached for literary texts, with substantially worse performance on documentary papyrus data. We also present some preliminary applications of natural language processing to the study of genre, style and authorship attribution in Coptic and discuss future directions in applying computational linguistics methods to the analysis of Coptic texts.

1. Introduction

Despite widespread illiteracy, late antique Egypt increasingly became a land that loved the book. The codex as an information technology device came to prominence in the Roman era. The earliest Christian monasteries in Egypt contributed to book production, book ownership, and book learning. (Bagnall, 2009) The rules of the monastery of Pachomius (4th century) required that all monks learn to read, even against their will:

> Whoever enters the monastery uninstructed shall be taught first what he must observe; and when, so taught, he has consented to it all, they shall give him 20 Psalms or two of the Apostle's epistles, or some other part of the Scripture. And if he is illiterate, he shall go at the first, third and sixth hours to someone who can teach him and who has been appointed for this. He shall stand before him and learn very studiously with all gratitude. Then the fundamentals of a syllable, the verbs, and nouns shall be written for him, and he shall be forced to read, even if he refuses. (Veilleux, 1981: 166–67)

Pachomius's rules, written in the 300s, were translated into Latin by the church father Jerome, who also created the Latin biblical translation known as the Vulgate. In this form, the rules would go on to influence the monasteries of medieval Europe, where monks imagined the desert fathers and mothers of Egypt to be their ascetic forbears. (Lefort, 1956; Boon, 1932)

Pachomius, like many other literate late antique Egyptians, originally composed his letters and rules in Coptic, the last phase of the ancient Egyptian language family. In use during the Roman and early Islamic periods of Egyptian history, it evolved ultimately

from the language of the hieroglyphs, and together with them, forms the longest chain of historical documentation of any language in the world.

In the computer age of the 21st century, scholars face a new challenge: deciphering, studying and documenting the vast library of Coptic texts in digital formats. This means adapting accepted standards, software and best practices to a domain that has seen little attention from computational quarters. At the same time, advances in computational and corpus linguistics offer great promise for unraveling new ways to study Coptic language and literature. This article aims to contribute to the new wave of Digital Coptic studies by presenting and evaluating a comprehensive part of speech (POS) tagging schema for Sahidic Coptic, the classical dialect of the language, as well as discussing some applications. We begin by outlining the importance of some of the Coptic works that can be accessed using computational methods (section 2). Section 3 briefly outlines Coptic 'words' and smaller morphological units, and section 4 describes the proposed tag set for evaluation. Section 5 presents an inter-annotator agreement study to determine how well humans can apply the tag set, and section 6 evaluates tagging performance on test data. Section 7 showcases some applications for tagged text, and section 8 concludes with lessons from this work and suggestions for future development.

2. Why Tag Coptic?

Coptic emerged during the Roman era when Greek was the "lingua franca" in the Eastern Empire. The alphabet is primarily Greek, though it includes a handful of Egyptian characters, taken from the previous phase of the Egyptian language, known as Demotic. But the language's structure derives from Egyptian grammar and syntax. . Coptic absorbed some Greek vocabulary, as well as to a lesser extent Latin and, later, Arabic

loan words. These factors make Coptic a rich environment for studies of culture and language.

Sources in Coptic are pivotal for a range of humanistic disciplines, such as linguistics, biblical studies, the history of Christianity, Egyptology, and ancient history. For example, some of the most important extra-canonical Christian texts (such as the Gnostic scriptures) survive in Coptic. Pachomius, his successors, and others among the earliest Christian monks also documented their history, theology, and ways of life in this language. In many cases, the correspondences between Coptic texts, indigenous religious works in adjacent traditions, and translations in the area cannot be studied based on one to one lexical correspondence, but necessitate reference to quantitative studies at more abstract levels, including signature author styles and grammar, which help scholars to study transmission histories, textual re-use and authorship attribution. A richly annotated Coptic corpus, tagged for POS as well as language of origin for foreign words, enables research into questions of bilingualism, education, and translation practices in multilingual environments as well as fundamental questions about the Coptic language and Egyptian language family. Computational analyses of the corpus may allow us to identify Coptic texts in translation (e.g. original Greek) vs. natively authored Coptic sources, preliminarily classify texts into genres, or identify authorship styles. Tagged data may ultimately be able to help scholars understand the large number of untranslated or understudied Coptic texts at an early stage in the digitization, editorial, and research process. The annotated corpus has also recently been used for digital pedagogy in Coptic at the Humboldt University in Berlin.

The realization of the potential in interdisciplinary work on computational methods for Coptic led us to establish Coptic SCRIPTORIUM (http://www.copticscriptorium.org), an open access, open source and collaborative project on the study of Coptic in the digital age. No open source corpus in the Egyptian language family, tagged for both POS and language of origin, exists aside from the project's developing body of work, which meant that creating it required the adaptation of standards and software. To produce this corpus and address these research questions, we have developed the first fully implemented part-of-speech tagging schema for the Coptic language (though see below for some previous pioneering work). We next explain how Coptic morphology presents some challenges in applying POS tagging to the correct units of analysis.

3. Language as Lego: Picking apart and Piecing together Coptic

Coptic is an agglutinative language, in which several morphemes can join together to create complex noun and verb forms. Example (1) shows a sentence comprised of three such forms known as "bound groups" in standardized Sahidic Coptic script as well as a transliteration, a glossed form, and a translation. We follow Layton (2011: 12-18) for transliteration conventions. PST indicates the auxiliary indicating past tense, 3sgM the third person singular masculine pronoun, and 3sgF the equivalent feminine pronoun:

(1) ⲁⲫⲥⲱⲧⲙ̄ ⲉⲣⲟⲥ ⲛ̄ⲥⲓⲡⲣⲱⲙⲉ

a.f.sōtṃ ero.s ṇk^{y}i.p.rōme

PST.3sgM.hear to.3sgF namely.the.man

'He heard her, that is the man'

Though original Coptic literary texts are generally written on manuscripts in *scriptio continua*, without spaces, there are conventions for transcribing groups of morphemes together, leaving spaces between larger 'word forms' or 'bound groups.' This is motivated both by some phonological considerations (clitics are written together with stressed stems) and by orthographic hints in manuscripts, such as the symbol '̀' in the diplomatic text from a manuscript in example (2). Such symbols often correspond to modern notions of 'words' or 'bound groups' in Coptic (but not always, see Layton 2011:19-20). Note that in the manuscript there is no space between the two large graphemic units. Additionally, lines in a manuscript may break in the middle of a word. In the 19th and 20th centuries, scholars have segmented words according to different standards, which means that published editions of Coptic do not show uniformity in the divisions between words. The emerging scholarly standard is now Layton's, but even he notes that word segmentation in Coptic is a modern convenience.

(2) ⲛ̄ⲟⲩⲱⲛⲣⲉ̀ ⲛ̄ⲁⲃⲣⲁϩⲁⲙ̀

ṇ.u.šēre ṇ.Abraham

of.a.son of.Abraham

'of a son of Abraham'

For the purposes of tagging, however, the relevant unit is not the units delimited by spaces in (1–2), but rather the constituent morphemes, separated by dots in the transcription and glosses (e.g. a noun like 'son', not the sequence 'of a son'). Therefore, we must first segment the text into such units before submitting it to the tagger. While some texts may already be segmented in this way at digitization, many are not. We

therefore prepared a simple segmentation script based on a lexicon lookup using the lexicon described in the next section. While correct segmentation is important for accurate tagging, the issue of segmentation is not within the scope of this article. We therefore assume correctly segmented text for the discussion below..

4. Tag sets

Designing a POS tag set for Coptic is complicated, since there has been almost no previous work on tagging for the language, and grammatical traditions and terminology vary. A notable exception can be found in Orlandi (2004), which included the development of an electronic full form lexicon with some useful categories and an attempt at a dedicated rule based system for Sahidic Coptic, but did not culminate in robust, publically available tagging software. Our work builds on Orlandi (2004) by reusing the same lexicon, recoded for the tag set described below, but using the freely available TreeTagger (Schmid 1994), a trainable stochastic tagger, instead of a dedicated rule based approach.[1]

Creating Coptic training data for a stochastic tagger is work-intensive. Digitized text in multiple genres (for robustness) is hard to find and usually not normalized, meaning extensive manual work, and texts are often difficult to translate and understand, making tagging more laborious. Thus in order to achieve high accuracy with comparatively little data, we opted to create a coarse tag set (SCRIPTORIUM Coarse, or SC), making only minimal distinctions, and a more fine-grained tag set (SCRIPTORIUM Fine, SF) that could be less robust. We begin with the coarse tag set.

4.1 Coarse tag set (SC)

SC comprises 22 distinct tags, only five of which are open ended (i.e. allow items not found in the lexicon). The open classes are:

Tag	Description
N	noun
V	verb
ADV	adverb
NUM	numeral (including letter combinations standing for numbers)
UNKNOWN	missing or illegible/unknown words[2]

Table 1. Open classes in the coarse tag set (SC).

Assigning an open class to items not in the lexicon is one of the main challenges facing the tagger. However Coptic syntax makes it often possible to tell nouns from verbs based on syntactic environment alone, and unknown adverbs are rare (primarily loans from Greek, which can also be recognized based on their suffixes).[3] Unlisted numerals (large numbers, letter combinations) and unknown items are also rare, the latter also being recognizable by some notations for lacunae (e.g. ⲃ[....]), which can also be treated as distinctive 'affixes' by the tagger. A greater challenge is posed by disambiguating the closed classes, since many Coptic morphemes are homographs in certain environments (though they are distinguishable in others). The closed classes are:

Tag	Description
A	auxiliary (any Coptic conjugation base, see also next section)
ART	article
C	converter (several subordinators, e.g. relativizers; cf. Layton 2011:319-366 and the next section)
CONJ	conjunctions (e.g. ⲁⲩⲱ *awō* 'and', ⲏ *ē* 'or')
COP	copula
EXIST	existential predicates (ⲟⲩⲛ *wn*/ⲙⲛ *mn* 'there is/isn't')
FUT	future marker (ⲛⲁ *na*)
IMOD	inflected modifier (ⲧⲏⲣ- *tēr*- 'all of', ϩⲱⲱ- *hō'* 'also, for one's part')
NEG	negations
PDEM	pronoun, demonstrative

PINT	pronoun, interrogative
PPER	pronoun, personal
PPOS	pronoun, possessive
PREP	preposition
PTC	particle (e.g. ⲇⲉ *de* ‘but’, ⲛⲥⲓ *nkʲi* ‘namely’)
PUNCT	punctuation
VBD	verboid (a closed class of suffixally conjugated predicates, e.g. ⲛⲁⲛⲟⲩ- ‘be good’)

Table 2. Closed classes in the coarse tag set (SC).

Disambiguating even the coarse closed classes can be difficult, as some forms can belong to multiple classes. For example, the letter ⲛ *n* can stand for a preposition (‘of’), an auxiliary (conjunctive, somewhat similar to an English *-ing* form or a Latin *ablativus absolutus*), a negation, a plural definite article, or a personal pronoun (1st person plural, ‘we’). These are not generally difficult for humans to distinguish in context (see Section 5 below), but nevertheless mean a substantial challenge to the tagger.

4.2 Fine tag set (SF)

SF comprises 44 distinct tags, which add to and expand on SC in the following ways. Firstly, an additional open class of proper nouns NPROP is distinguished from common nouns N. This distinction is primarily recognizable for unknown words by checking for the presence of an article, as proper nouns generally don’t carry an article. However this rule is not absolute, as some place names take articles, and at the same time common nouns occasionally occur without articles, especially in generic readings (e.g. ‘man’ to mean mankind, or any man in general).

Secondly, 15 different auxiliaries are distinguished, which have multiple, partly overlapping spellings but otherwise form closed classes. These are:

Tag	Name	Example	Approx. translation
AAOR	Aorist	ⲱⲗ	he always/generally does
ACAUS	Causative	ⲧⲣⲉ	he causes to do

ACOND	Conditional	ⲉⲣϣⲁⲛ	if he does
ACONJ	Conjunctive	ⲛⲧⲉ	doing
AFUTCONJ	Future Conjunctive	ⲧⲁⲣⲉ	he shall do
AJUS	Jussive	ⲙⲁⲣⲉ	let him do
ALIM	Limitative	ϣⲁⲛⲧⲉ	until he does
ANEGAOR	Negative Aorist	ⲙⲉ	he never does
ANEGJUS	Negative Jussive	ⲙⲡⲣⲧⲣⲉ	let him not do
ANEGOPT	Negative Optative	ⲛⲛⲉ	may he not do
ANEGPST	Negative Past	ⲙⲡⲉ	he did not do
ANY	Not Yet	ⲙⲡⲁⲧⲉ	he has not yet done
AOPT	Optative	ⲉⲣⲉ	may he do
APREC	Precursive	ⲛⲧⲉⲣⲉ	after he does
APST	Past	ⲁ	he did

Table 3. Auxiliary tags in the fine tag set (SF).

The remaining added tags specify subtypes of verbs, personal pronouns, and the aforementioned converters. Verbs distinguish morphological imperative (VIMP) and stative forms (VSTAT), where they are distinguishable. The former exist for only a handful of verbs (e.g. ⲁⲣⲓ *ari* 'do', ⲁϫⲓ *ači* 'say'), and no attempt is made to tag other verbs used in the imperative as VIMP (cf. Schiller et al. 1999 for a similar decision in the standard tag set for German, STTS). The latter exist for most verbs and signify a state in the case of intransitive verbs (e.g. ϩⲟⲗϭ *holk^j* 'be sweet') or a passive for transitive verbs (ⲕⲏⲧ *kēt* 'be built').

For pronouns, subject, object, and independent forms are distinguished as PPERS, PPERO and PPERI respectively. The latter are used for emphatic purposes ('As for me, I…') or in nominal sentences ('It is I'). Converters (the tag C in the coarse set) include: CREL for the relative converter ('which'), CCIRC for the circumstantial ('while'), CFOC for the focalizing converter ('it is X!', see below) and the preterit conversion CPRET, which signifies an anterior past (imperfect and pluperfect readings, depending on tenses it combines with). Though they have rather different semantics, the converters share morphosyntactic characteristics, including partly identical forms depending on their

environment, an initial position before fully inflected sentences (which they 'convert') and fusional morphology together with adjacent pronouns.

Thus the primary differences between the fine and coarse grained tag sets revolve around more detailed distinctions in the closed classes, as well as the addition of proper names. How challenging the decision between closed classes is can best be illustrated using the example of the form ϵ *e*, which can have as many as five different tags in SF:

- PREP – a preposition meaning 'to'
- CREL – a form of the relative converter in some environments '(…) which'
- CCIRC – a form of the circumstantial converter '(…) while'
- CFOC – a form of the focalizing converter 'it's that (…)', stressing some element in the following sentence.
- PPERO – an object pronoun (2nd person feminine singular)

In some cases, especially when the text is fragmentary, even a human annotator cannot disambiguate these with absolute certainty, as in the following example, for which the preceding context is lost:

(3) ⲉϥϯⲙ̄ⲧⲟⲛ ⲇⲉ ⲟⲛ ⲛ̄ⲧⲙⲁⲁⲩ ⲛ̄ⲧⲁⲥⲭⲡⲟϥ

e.f.ti.mton de on n.t.ma'u nt.a.s.čpo.f

?.3sgM.give.rest but still of.the.mother that.PST.3sgF.bore.3sgM

The first e- in the sentence is definitely a converter, but in this environment three converters share the same form, and a translation with any of the three is possible:

- Relative: ‘which however still gives rest to the mother that bore him’
- Circumstantial: ‘while he still however gives rest...’
- Focalizing: ‘But it is TO THE MOTHER WHO BORE HIM that he gives rest!’

These ambiguities are also a major source of disagreement between human annotators, see Section 5.

4.3 Portmanteau tags

In some comparatively infrequent cases, a single orthographic form can contain two categories. For example the verb ⲉⲓⲛⲉ *eine* ‘bring’ takes the form ⲛⲧ *nt-* before personal pronouns objects (e.g. ⲛⲧⲟ *nt.f* ‘bring him’). However if it takes the first person object form ⲧ *-t* ‘me’, then this is not written separately, leading to a plain ⲛⲧ *nt* ‘bring me’. In these cases we assign a portmanteau tag consisting of both relevant categories separated by an underscore: V_PPERO (a verb and its object in one; cf. Schiller et al. 1999 for a similar decision for German).

The same can occur in many forms of the 2nd person singular feminine subject, which is often realized as a ‘zero,’ as in the case of the preterit conversion:

(4) **ⲛⲉⲣⲉ**-ⲡⲣⲱⲙⲉ ⲥⲱⲧⲙ̄ **ⲛⲉ**ⲕⲥⲱⲧⲙ̄ **ⲛⲉⲣⲉ**ⲥⲱⲧⲙ̄

nere.p.rōme sōtm **ne**.k.sōtm **nere**.sōtm

CPRET.the.man hear CPRET.2sgM.hear CPRET+2sgF.hear

the man used to hear you (m.) used to hear you (f.) used to hear

In the third case in (4), the converter takes the same form as in the first case (*nere*), but there is no overt realization of the word 'you (fem.)'. For a masculine 2nd person subject, the converter is *ne*, and the word 'you (masc.)' is realized as *k*. Thus the tag for *nere* 'you used to (fem.)' is CPRET_PPERS, a converter form which also contains a personal pronoun marking.

We have so far assigned 12 combination tags (mostly 2nd person singular feminine subjects connected to various auxiliaries), but these form only 57 tokens within our test corpus of over 12,000 tokens, i.e. less than 0.4%.

5. Inter-annotator agreement

Automatic POS tagging is only useful, and can only be evaluated for accuracy, if human annotators can agree on the 'gold standard' tag for every word (or more realistically for most words) in a text. We therefore conducted a small experiment to evaluate our tag set. Both authors independently annotated the same two subcorpora using the maximally granular SF tag set. The data was taken from two different texts in order to give a first indication whether agreement robustness might be affected by text type or genre. We selected a section from the letter *Abraham Our Father* by the classical monastic author Shenoute and a collection of short narrative anecdotes from the *Sayings of the Desert Fathers*, which both have good orthography and few lacunae, but have rather different styles. The two text types contained 906 and 576 Coptic morphs respectively. Our agreement on tagging is presented in Table 4.

Text	Identical SF	Identical SC
Abraham our Father	854/906 (94.26%)	872/906 (96.24%)
Apophthegmata Patrum	542/576 (94.09%)	553/576 (96.01%)
Total	1396/1482 (94.19%)	1425/1482 (96.15%)

Table 4. Percentage of agreement between two annotators by text.

The figures in Table 4 are quite positive, with absolute agreement accuracy of around 94% for the fine-grained tag set and 96% for the coarse-grained one. However, these figures don't give us an idea of how likely this agreement is to arise by chance (e.g. if most words are nouns, it is easier to just guess that something is a noun whenever in doubt). For this reason, the Kappa metric is commonly used to evaluate annotation schemes, which takes into account the difficulty of the annotation task in terms of reaching agreement by chance. Kappa ranges from 1 (perfect agreement) to 0 (absolutely random, but not zero agreement). For our experiment we calculated a Kappa value of 93.96 for SF and 95.69 for SC, which can be considered very high (see Artstein & Poesio, 2008).

he primary disagreements occurred in telling apart the open classes of nouns and verbs, and disambiguating closed classes, particularly the converters. Figure 1 shows the most frequent confusion categories in SF.

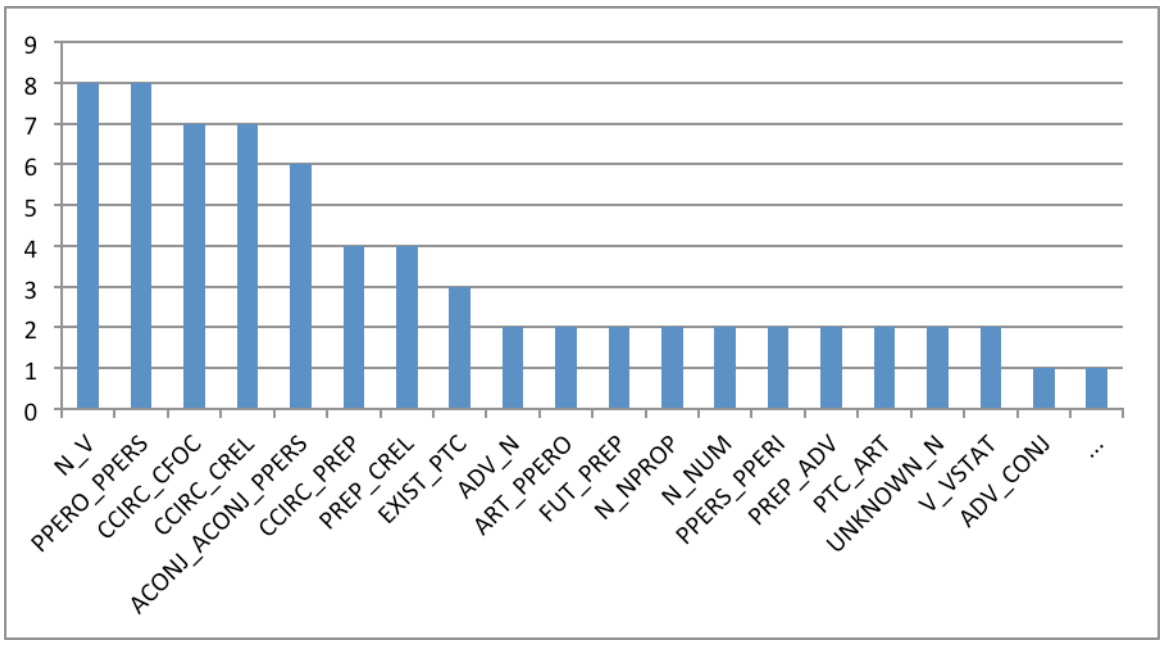

Figure 1. Confused tag pairs by disagreement frequency.

The confusion of nouns and verbs may seem surprising given the linear, agglutinative nature of Coptic grammar. However cases of confusion primarily arose in the context of nominalized verbs, as in example (5).

(5) ⲛⲥⲉⲁⲡⲟⲧⲁⲥⲥⲉ ϩⲛⲟⲩ†ⲧⲱⲛ

ṇ.se.apotasse hṇ.u.**ti**.tōn

and.they.renounce in.a.**give**.quarrel

‘and they renounce (them) argumentatively’

The morph under disagreement in this example is the verb *ti* ‘give’, which is part of the complex expression *ti-tōn* ‘quarrel (lit. give quarrel)’. The entire combination *ti-tōn* has been nominalized in the presence of an indefinite article *u*, so that the second bound group in (5) literally reads ‘in a give quarrel’, roughly meaning ‘argumentatively’ (or ‘in argument, while arguing’). The morph *ti* is morphologically a verb, but syntactically

converted to a noun, which leads to disagreement. This type of issue can probably be resolved by refining guidelines.

A different class of problem occurs primarily in disagreements about converters, which can stem from subtle translation differences, as in (6).

(6) ⲛⲉⲟⲩⲛ ⲟⲩϩⲗⲗⲟ ϩⲛⲛⲣⲓ … **ⲉ**ⲫⲟⲣⲉⲓ ⲛⲟⲩⲧⲙⲏ

ne.wn̩ u.hl̩lo hn̩.n̩.ri **e**.f.phori n̩.u.tmē

PRET.was an.old in.the.cells **C**.3sgM.carry ACC.a.mat

'There was an old man in Kellia … (who carried/carrying) a reed mat'

In this case, it is difficult to make a certain decision about the converter *e* in bold in (6). Coptic relative clauses modifying an indefinite noun take the same converter form *e* as circumstantial clauses meaning roughly 'while'. Therefore in (6) the text could mean that there was a man who was in the habit of carrying a reed mat (relative), or that there was man there carrying a reed mat at that point in time (circumstantial). Ambiguities like this are not likely to be answered completely consistently by human annotators, and an automatic tagger is likely to vary, as well, though generally preferring the option that is more frequent in training data.

6. Automated Tagging Accuracy

To train the tagger and evaluate accuracy we tagged the texts in Table 5. Texts were selected for scholarly interest (linguistic and philological), and in order to offer a breadth of genres in literary Coptic, including religious discourse, letters, and Biblical and non-Biblical narrative (see Section 7 for more details on the texts and authors).

Text	**Morphs**
Shenoute / Abraham Our Father	2061

Shenoute / Acephalous 22	229
Shenoute / Not Because a Fox Barks	1767
Besa / Letter to Aphthonia	1123
Besa / Letter to Thieving Nuns	785
New Testament / Mark 1	1229
Apophthegmata Patrum (11 texts)	1388
Artificial sentences	62
Total	8582

Table 5. Breakdown of texts used in the gold standard training corpus.

The inclusion of some artificial sentences at the bottom of the table was motivated by the need to generate examples for the tagger of some particularly infrequent combinations not otherwise attested in the corpus, in particular cases of portmanteau tags (Section 4.3) which we had foreseen based on combinatoric possibilities in Coptic grammar, e.g. possible 2nd person singular feminine forms that were not attested in our corpus. This need was minimized, particularly in the context of rare 2sgF forms, by including Besa's Letter to Aphthonia (5th century), written to address a female nun in the 2nd person.

To evaluate the tagger's performance we take a 10 fold cross-validation approach, dividing the data into 10 portions of which each portion is held out once as test data while the remaining 9 are used as training data (excluding the artificial sentences, which are never in the test set). The training data was fed to the freely available and trainable TreeTagger (Schmid 1994), which was also given a list of the open and closed tags and a POS tagged lexicon containing 5265 entries (derived from CMCL's database mentioned above). The best model used trigram context (looking at probabilities in sequences of three morphs). Table 6 gives the accuracy per slice as well as the percentage of unknown words encountered by the tagger which were missing from the lexicon.

Slice	% correct SF	% correct SC	% out of lexicon
1	89.66	91.10	0.46
2	95.16	96.82	0.00

3	95.07	95.28	0.96
4	94.03	95.76	1.65
5	92.92	96.55	2.85
6	94.96	96.90	1.76
7	94.93	93.62	1.38
8	96.26	96.48	0.70
9	94.17	93.64	2.74
10	94.44	95.00	3.67
average	**94.16**	**95.12**	**1.62**

Table 6. Tagger accuracy in 10-fold cross validation.

A relatively high total accuracy of over 94% for SF and 95% for SC was reached, meaning that on average, about every 20th morph receives an incorrect tag (slightly more often for SF). While climbing even fractions of a percent higher will become exponentially more difficult, it should be noted that these figures are not very far below tagging performance for languages with much larger training sets, which is due in large part to the high coverage of the lexicon: on average, only 1.62% of morphs in the test texts had no lexicon information, meaning that at least for open classes, the tagger could usually rely on dictionary information to establish whether a word was known to be a noun or a verb. The large part of tagging errors was due to incorrect disambiguation, the major cause of human disagreements in Section 5. For example, the top 5 tagging errors, making up 16.7% of all errors, were due to different confusions of the correct tag for the morph ϵ *e*, which has 5 different readings (cf. Section 4.2).

These results suggest that the tagger may be vulnerable in texts with a higher proportion of out-of-data vocabulary, and possibly also different genres. While we cannot offer a full exploration of this issue in this paper, we give a toy evaluation on a much more 'unruly' text type, documentary papyri. We tested both models on two documentary papyri taken from papyri.info, together comprising 137 tokens. For SF, tagging accuracy degraded to 80.29%, while SC remained more robust at 87.59%. Of the 137 tokens, 23 were not found in the lexicon (16.78%). The difference between the two models'

performance is due in large part (but not only) to the distinction between proper and common nouns, as proper nouns are often out of lexicon items and difficult to distinguish from common nouns. At the same time it is highly likely that the worse performance on papyrus data is due not only to out-of-data items and the frequency of proper nouns, but also because the tagger has been trained on completely different text types and language domains, all coming from literary Coptic. We therefore feel that there is room for much work on expanding the domains and text types on which the tagger is trained, as well as for obtaining more lexicon data, including lists of proper names and toponyms.

7. Applications in Research and Pedagogy

A Coptic corpus tagged for POS can enable research projects in a variety of disciplines. Coptic literature (hagiography, sermons, epistles) can be analyzed for knowledge about the rhetorical structure of diverse texts. Traditional scholarship on genre and literary formulae, (e.g., Choat, 2007 on epistolary formulae), can be enhanced by the ability to query and analyze large corpora in terms of grammar and syntax as well as vocabulary. A statistical analysis of a corpus that spans several centuries of Coptic history can yield information about the evolution of the language over time, especially as Arabic enters Egypt.

We present here some preliminary research on genre and style. These results come from a subset of the corpora used in Section 6, but they illustrate the potential for research with a larger corpus. The documents include the letter *Abraham Our Father* by the monastic leader Shenoute writing in the late 300s or early 400s, portions of an untitled fragmentary text by Shenoute (*Acephalous Work 22*), two letters by Shenoute's

successor Besa, a selection of sayings from the Coptic *Sayings of the Desert Fathers* (the *Apophthegmata Patrum*) and the first six chapters from the Coptic Gospel of Mark.

Table 7 shows the frequencies for the most prevalent parts of speech in each of the works above as deviations from the expected norm using a chi-squared test. The redder a cell is, the more the frequency for that POS in that particular document (or set of documents) is *above the norm*; the bluer the cell, the more it is below norm. The data was taken from our corpora in June 2014 and annotated either manually (parts from the gold standard data) or automatically using the tagger.[4]

ID	total_freq	pos	Besa	A22	AOF	AP	Mark1_6
51	3240	PREP	-0.9223	2.245374	3.720749	-1.83133	-3.60583
40	2925	N	0.080235	2.600809	3.098838	-0.51138	-4.27406
21	1949	ART	-1.18379	1.800207	4.449135	-3.38187	-3.27729
57	1903	V	-1.16191	-2.07814	-3.87829	0.551977	5.31527
49	1406	PPERS	-1.98882	-2.5304	-8.07701	3.180776	9.002659
47	1173	PPERO	-1.0792	0.041699	-3.04763	-0.49954	3.790747
30	1007	CONJ	0.570794	-1.54252	-2.43828	0.739032	2.621228
54	971	PUNCT	5.768879	-3.21255	5.669246	2.864461	-8.19385
19	604	APST	-1.82084	-4.9142	-4.00445	1.51988	6.850586
34	523	CREL	1.668746	2.013423	2.620324	-3.31901	-3.06952
7	514	ADV	-1.58016	1.466869	0.979143	-2.08521	-0.01307
50	387	PPOS	4.262987	0.407573	-0.71813	0.272237	-1.82168
22	339	CCIRC	-2.62532	0.757333	-1.08926	-0.03072	2.054625
53	310	PTC	-1.14017	-0.99507	-5.51929	2.960991	5.304131
42	254	NPROP	-2.57099	-3.90854	1.538779	-0.96275	2.282299
61	205	VSTAT	2.454121	1.992609	-1.78869	0.626819	-0.81402
44	172	PDEM	0.04957	0.829026	3.35968	1.07025	-4.28088
37	136	FUT	1.843619	0.574767	-0.62318	0.099953	-0.67892
38	124	IMOD	-0.03142	0.102049	2.679328	-1.6264	-1.9884
31	123	COP	0.28516	1.825912	2.145972	-1.60893	-2.53799
5	112	ACONJ	2.740252	0.765349	-3.71106	-0.34661	2.029526
41	109	NEG	0.10896	0.267472	0.862211	2.955049	-2.35958

Table 7. Tag frequency deviation heat map for five subcorpora.

The table contains two narrative texts: the *Sayings of the Desert Fathers* (*Apophthegmata Patrum*, AP) and chapters from the Gospel of Mark. These works are similar in many respects. They both have a low proportion of articles (ART), indicating pronominal noun phrases and/or longer predicates. Conversely they both contain high

frequencies of personal subject pronouns (PPERS), a possible source of the lower proportion of articles. This likely results from the use of "you" and "I" in dialog, or possibly narration chains about human protagonists ("s/he"). These results lead to various hypotheses about genre that need further testing, including whether Coptic narrative texts as a category show high frequencies of pronouns and low frequencies of articles.

There are also some differences within each genre group: APST (the auxiliary indicating past tense) is much more overwhelming in the Gospel of Mark than in the Apophthegmata. Both, however, use the APST much more frequently than the other document sets. The monastic letter known as *Abraham Our Father* (AOF) groups with the Mark selections in terms of frequency of proper nouns. This is likely due to Biblical narrative integration: The Gospel of Mark is a biblical text (much like an ancient biography) about Jesus and his disciples. Originally written in Greek in the first century, it was later translated into Coptic. *Abraham Our Father* is an entirely different genre: a letter from a monastic leader (Shenoute) in the late fourth/early fifth century to the female monks in the women's residence of the monastery. However *Abraham Our Father* contains extensive biblical citations and exegesis; Shenoute interprets of the lives of various biblical characters (beginning with the patriarch Abraham and his wife Sarah) and applies them to the monastic life.

We can also compare the syntactic structure of texts authored originally in Coptic with translations from Greek to gain new insights into translation practices in a multilingual environment. (Late antique Egypt was home to native writers and speakers of Greek, Latin, Demotic, various dialects of Coptic, and eventually Arabic; in a monastic setting even more languages could be encountered, including Syriac.) The

syntactic analysis may also lead to a better understanding of the textual history of documents that survive only in Coptic but are theorized to have been written originally in Greek. In our sample, the two texts written originally in Greek contain more particles (which are overwhelmingly of Greek origin) relative to the rest of the corpus. The language of origin analysis, which was annotated in a separate annotation layer, confirms this. The high frequency of particles may be a sign of a translated text. Future research on tagged corpora may lead to the discovery of other indicators that can identify translated texts.

The authorship of many Coptic texts also remains in question, because they are anonymous, fragmentary, or pseudonymously attributed to earlier historical figures. Many Coptic manuscript fragments lie unattributed in museums, libraries, or private collections. The best metrics for authorship attribution of Coptic texts still need to be investigated. Given the more limited vocabulary of Coptic literature, syntactic analysis may be useful instead of or in combination with stylometry based on vocabulary alone. Our preliminary data show some stylistic distinctions between two authors with extremely similar demographic profiles. Shenoute and Besa were both leaders (abbots) of the White Monastery federation in Egypt. Besa was Shenoute's successor, knew Shenoute personally, likely heard many of his sermons in person, and read the letters, texts, and rules in Shenoute's literary corpus. Within our preliminary dataset, Besa prefers conjunctions over Shenoute. And while both utilize the relative converter (CREL) to create relative clauses more often than the narrative texts, the verbose CREL is more frequent in Shenoute's corpora than Besa's. Prepositions (PREP), nouns (N), and articles (ART) are also significantly more prevalent in Shenoute's writings than in the other

samples. Besa, by contrast, strongly favors the conjunctive auxiliary, which is a special form used to tightly chain predications together with relatively little phonological material.

These conclusions are necessarily preliminary at this point. The datasets are small, and the analysis has not fully accounted for portmanteau tags (e.g., including CPRET_PPERS in the aggregate counts for both CPRET and PPERS). We are also only beginning to look at multiword n-grams and characteristic POS sequences. Yet the indicators are suggestive. Many Coptic texts remain to be digitized, or even edited and translated. The automatic annotation of texts at an early stage of the digitization and editorial process may help us identify texts in translation, posit authorship of anonymous or unidentified texts, or find named entities for prosopograpical or geospatial analyses.

A tagged corpus may also be used pedagogically to study Coptic language. The corpora described here were recently used as the basis of a beginners' Coptic course at Humboldt University in Berlin, with students' final assignments leading to new contributions to the corpus. Aligned with a translation, the tagged Coptic text provides a sample set for students seeking to translate and understand the use of unique Egyptian grammar. Advanced undergraduates and graduates s can improve their knowledge of the language by applying the tag set to untagged and/or untranslated texts, or using the automatically tagged corpus to produce more rapid translations. Non-academics in the general population with interests in Egypt, the history of Christianity, or linguistics can read the English translations aligned with the tagged Coptic to get a better appreciation of the sources in the original language and to advance their understanding of it outside of formal academic instruction.

8. Conclusion and Outlook

The work described in this paper has shown that it is possible to reach promising results in natural language processing (NLP) for an under-resourced, dead language such as Coptic with a relatively small set of training data. Using off-the-shelf freely available tools, a usable POS tagging model can be trained by crafting a tag set that is informative but does not make unrealistic demands on the tagger. That said, there are many distinctions that the tagger does not make in the interest of higher consistency, and in some cases it will make sense to use a more coarse grained tag set. This is particularly evident in the case of out of domain data such as the small evaluation of documentary papyri, for which the fine-grained tagging efforts degraded much more than the course grained model.

The applications discussed throughout this article suggest exciting directions for work with linguistically annotated corpora for Coptic. Little quantitative work on Coptic exists, and using tagged data will allow us to abstract away from particular texts and vocabulary and look at underlying linguistic structures from a grammatical point of view. We also see great potential in extending our model to other dialects and periods of Coptic besides classical Sahidic writing, a comparison which will be greatly facilitated across spelling and pronunciation variants if abstract POS tags are used.

The resources presented here are freely available to researchers from the SCRIPORIUM website, and we are very hopeful that they will be taken up and extended by the community of scholars working on Coptic texts. The next tasks for advancing computational methods for Coptic depend in large part on contributions from Coptologists, who can extend the manually tagged gold standard as well as make use of

the existing corpus of digitized texts with automatic tagging or even develop new forms of annotation. Some examples of new levels of analysis include detailed parallel alignment with Greek originals and other translations, syntax tree analyses, named entity recognition and more. All of these forms of annotation have extensive tools and standards for other languages that can be adapted to Coptic. However many second-tier NLP tools rely on tagged data, and their quality subsequently depends on the quality of this preliminary task. The present work therefore lays a foundation for further types of corpus annotation, to allow additional types of research. We are only beginning to evaluate the part of speech analysis of Coptic texts across authors, genres, periods, and subject matter. We have no doubt that the digital age holds many advances, but also challenges, as we negotiate the representation and accessibility of Coptic texts for the coming decades.

Notes

1. We thank Prof. Orlandi and the Corpus dei Manoscritti Copti Letterari project (CMCL) for making the resources available to us.

2. Including such a tag is not just necessary for cases in which the tagger cannot determine the tag assignment: there are many cases in which human annotators also cannot determine the correct tag for a specific word because of damaged manuscripts. This situation is expected to recur often in texts to be tagged by the tagger.

3. We also annotate language of origin for loanwords from Greek and Latin, or Biblical Hebrew terms, on a separate annotation layer using a lexicon and a list of prefix and suffix patterns, but these are not given separate POS tags.

4. Previous releases of Coptic SCRIPTORIUM's corpora are available on GitHub at https://github.com/CopticScriptorium/corpora-legacy-releases.

Funding

This work was supported by the National Endowment for the Humanities [PW-51672-14, HD-51907-14]; the German Federal Ministry of Education and Research [01UG1406]; Humboldt University; and the University of the Pacific.

9. References

Artstein, R., and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4): 556–596.

Bagnall, R.S., 2009. *Early Christian Books in Egypt*, Princeton: Princeton University Press.

Boon, A., 1932. *Pachomiana latina: Règle et épîtres de S. Pachôme, épître de S. Théodore et "liber" de S. Orsiesius. Texte latin de S. Jérôme*, Louvain: Bureaux de la Revue.

Choat, M., 2007. Epistolary Formulae in Early Coptic Letters. In N. Bosson and A. Boud'hors (eds.), *Actes du huitième congrès international d'études Coptes: Paris, 28 juin - 3 juillet 2004*. (Orientalia Lovaniensia Analecta 163.) Leuven: Peeters Publishers, pp. 667–77.

Layton, B. (2011). *A Coptic Grammar*. Third Edition, Revised and Expanded. (Porta linguarum orientalium 20.) Wiesbaden: Harrassowitz.

Lefort, L.T. ed., 1956. *Oeuvres de S. Pachôme et de ses disciples*, Louvain: Imprimerie orientaliste L. Durbecq.

Orlandi, T. (2004). Towards a Computational Grammar of Sahidic Coptic. In M. Immerzeel, and J. van der Vliet (eds), *Coptic Studies on the Threshold of a New Millennium. Proceedings of the Seventh International Congress of Coptic Studies*. Vol. 1. Leiden: Peeters, pp. 125–130.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung and Universität Tübingen, Seminar für Sprachwissenschaft, Technical Report.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf (accessed 10.9.2013).Veilleux, A. ed., 1981. *Pachomian Koinonia, Volume Two: Pachomian Chronicles and Rules*, Kalamazoo: Cistercian Publications.