Morphological Analysis of Sahidic Coptic for Automatic Glossing

Daniel E. Smith, Mans Hulden
University of Colorado
daniel.e.smith@colorado.edu, mans.hulden@colorado.edu

Abstract

We report on the implementation of a morphological analyzer for the Sahidic dialect of Coptic, a now extinct Afro-Asiatic language. The system is developed in the finite-state paradigm. The main purpose of the project is provide a method by which scholars and linguists can semi-automatically gloss extant texts written in Sahidic. Since a complete lexicon containing all attested forms in different manuscripts requires significant expertise in Coptic spanning almost 1,000 years, we have equipped the analyzer with a core lexicon and extended it with a 'guesser' ability to capture out-of-vocabulary items in any inflection. We also suggest an ASCII transliteration for the language. A brief evaluation is provided.

Keywords: Sahidic Coptic, Egyptian, finite-state, foma, morphological analysis, glossing

1. Introduction

The aim of this paper is to present the results of a project to create a computational analyzer/lemmatizer/glosser that processes phonological, morphological and semantic information about the Coptic language. This analyzer is capable of processing a section of raw text which has been properly transliterated and producing a series of glosses based on what is known of the structure of the Coptic language, specifically the Sahidic dialect.

This paper is structured as follows. We begin by introducing the Coptic language and giving a brief outline of its word structure. Following this section will be a brief summary of some of the specific complications that must be overcome working with Coptic in a computational setting. Then, the analyzer will be presented with examples of the input and output given. Finally, the utility of the analyzer as a tool for both teaching and research is outlined. This paper aims to show that computational methods and the Coptic language present a promising synergy for furthering research into both fields.

2. Overview of Coptic

The Coptic language (ISO 639-1: cop) represents the last stage of Egyptian. The diagnostic traits of Coptic last from roughly 300 AD until the final movement of the spoken language into a basically dead liturgical language in 1300 AD (Loprieno, 1995; Layton, 2004). Coptic is generally divided into five dialects: Sahidic, Bohairic, Fayyumic, Achmimic and Subachmimic (Lambdin, 1983). Of these, this paper will focus on the Sahidic dialect which was the standard dialect of the early Coptic Church[1] and seems to be descended from the primary dialect of the Pharaonic political elite in Memphis and Thebes (Lambdin, 1983). Sahidic is also the only dialect with a substantial body of original literature.

2.1. Writing system

Written Coptic used a phonemic system based on the Greek script which incorporated elements of the earlier Demotic writing system native to Egypt. For the most part there is a one-to-one correspondence between the phonemes of the spoken language and the graphemes of the written language. Superliner strokes are thought to have represented syllabic consonants and doubled vowels are thought to have represented a glottal stop interjected during the vowel (Depuydt, 1993). Table 1 shows (a) the phonemes of Sahidic Coptic, (b) their graphical rendering and (c) our one-to-one ASCII transliteration, which we employ in the analyzer being presented.[2] Except as noted to be otherwise in table 1, a capital letter in the transliteration is considered to be syllabic. Coptic is written in *scriptio continua*, i.e. does not contain punctuation or inter-word spacing; see Figure 1 for an illustration of the typical nature of extant manuscripts.

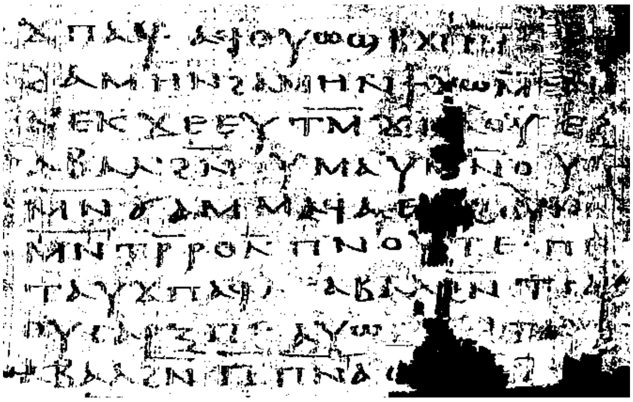

Figure 1: Fragment of Coptic manuscript (The Gospel of St. John, from Thompson (1924)).

2.2. Morphology

In what follows, we shall be concerned mainly with modeling the nominal, pronominal, verbal, and prefix system of Coptic. Morphologically, Coptic is a highly flexional language. It tends towards prefixes and preclitics, but has a small class of suffixes. Person marking is usually achieved

[1] Bohairic is the current lithurgical language of the Coptic Orthodox Church.

[2] Though a Coptic Unicode block exists as of the Unicode standard 4.1, proper rendering is difficult to guarantee across applications which is why we have chosen to use a transliteration scheme. This scheme does not follow the standard Coptic transliteration scheme in order to facilitate transcription of the macrons used in Coptic writing.

Grapheme	Phoneme	Transliteration
ⲁ	/a/	a
ⲃ	/b/	b
ⲅ	/g/	g
ⲇ	/d/	d
ⲉ	/e/	e
ⲍ	/z/	%
ⲏ	/e:/	E
ⲑ	/th/	t h
ⲓ	/j/	j
ⲉⲓ	/i:/	i
ⲕ	/k/	k
ⲗ	/l/	l
ⲙ	/m/	m
ⲛ	/n/	n
ⲝ	/ks/	k s
ⲟ	/o/	o
ⲡ	/p/	p
ⲣ	/r/	r
ⲥ	/s/	s
ⲧ	/t/	t
ⲩ	/w/	w
ⲟⲩ	/u:/	u
ⲫ	/ph/	p h
ⲭ	/kh/	k h
ⲯ	/ps/	p s
ⲱ	/o:/	O
ϣ	/ʃ/	z
ϥ	/f/	f
ϩ	/h/	h
ϫ	/c/	x
ϭ	/k^j/	c
ϯ	/ti/	t i

Table 1: The writing system of Sahidic Coptic and our transliteration, adapted from information in (Depuydt, 1993; Lambdin, 1983; Loprieno, 1995)
.

through a small closed class of prefix personal pronouns which also code for possession, although suffix and free pronouns exist as well. Number and gender marking is achieved with a definite prefix that is marked for either singular with gender or a common plural. Verbal tense is achieved with a series of combinatory prefixes. Both nouns and verbs can receive much of the same morphology with different meanings being intended by each. Thus, a prefix pronoun on a verb indicates subject but on a noun indicates possession (Layton, 2007).

Most of Coptic morphology is prefixing; the prefixes attach to both nouns and verbs. For the most part these prefixes are quite stable across forms and uses, but in some highly fused instances they can mutate. For example, the possessive construction is easily derivable as the combination of **owNt+** syllabic pronominal prefix. But, for example, this is not applicable in the second singular feminine form where one would expect **owNtR** based on the above prefix by adding *2SF* pronoun **r**, but instead one gets **owNte**. All such irregularities must be simply modeled as such in the lexicon.

All Coptic nouns are marked for gender, usually with a prefixed article. However, morphological irregularity is also found in nouns, and is more complex to handle elegantly. For example, while for the most part the only gender marking on Coptic nouns comes from this definite article, a small class of nouns retain artifacts of the older Egyptian gender system. Unfortunately for computational analysis (although fortunately for those reconstructing older stages) these alternations are highly irregular and cannot be predicted, such as in the table below.

Gloss	Masc	Fem
the sibling	pson	tsOne
the child	pzEre	tzeere
the old person	phLlo	thLlO
the dog	puhor	tuhOre

The same effects can be seen within the Egyptian number system which, like the gender system, does not usually mark directly on the noun. Still, there is a small class of highly common nouns that seem to retain these remnants of the old system; a broken plural common to Semitic languages:

Gloss	Singular	Plural
the father	piOt	niote
the brother	pson	nesnEw
the ship	pxoj	nexEw

The greater difficulty in handling the morphological irregularity in Coptic comes from the verbal system which is rife with root-and-pattern alternations left over from a templatic past.[3] This results in a system with four different verbal roots which sometimes still follow traces of the old templatic pattern, but are still by and large hard to predict. A set of example verbs in this paradigm are presented in the table below.

Gloss	Infinitive	Suffix	Preverbal	Stative
build	kOt	ket-	kotˆ	kEt
dry up	OzM	ezM-	ozMˆ	ozM
bend	rjke	rek-	rakˆ	roke
order	tsano	tsane-	tsanoˆ	tsanEw

These pose a greater problem than the nominal irregularity because (a) this necessitates four separate entries for each verb in the lexicon and (b) different roots have different morphological possibilities for prefixes or suffixes. Implementing root-and-pattern morphology as transducers is generally well understood and there are relatively simple mechanisms for doing so: multi-tape automata (Kay,

[3] In Coptic, these behave very much like ablauting verbs in Indo-European languages where a single vowel alternates, as for tense in some English lexemes **run** ∼ **ran**.

1987; Hulden, 2009b), intersection of roots and patterns (Beesley, 1998), perhaps by specialized regular expression operations (Beesley and Karttunen, 2003), or through composition of transducers that directly modify vowels (Jaber and Delmonte, 2008). In the case of Coptic, however, we have decided to simply hard-code all the different verbal grades in the lexicon, because of the unpredictability. Some of the alternations are illustrated below.

The final hurdle presented by Coptic is the enormous number of homophonous forms. Just the simple form **N** has an exceptional number of possible meanings in the lexicon including: genitive, dative, object marker, negative marker, and so on. Furthermore, **N** can also surface as **ne**, **M**, and **me** creating even more possibilities for homophony. Alternations like this produce a large number of potential parses for some word forms. Any disambiguation—although usually obvious at a glance for an expert—must then be performed by the user.

2.3. Previous computational work

The only computational effort we know of regarding Sahidic Coptic is that of Orlandi (2004), which is largely a implementationless sketch concerning the possibility of automatically analyzing Coptic word forms (POS tagging). Ashton (2012) also presents a formal analysis of the so-called 'second position clitics' in Coptic through monadic second-order transductions. No implementation is given. A ongoing larger-scale project is Zeldes and Schroeder (2015) who have developed a coarse-grained tagset for Sahidic Coptic and trained a statistical tagger on small amounts of labeled text, producing reasonable accuracy on held-out data. No morphological analysis module is included and the authors rely on pre-tokenized text.

coptic.lexc

```
...

LEXICON Noun

GUESSNOUNSTEM[*GUESS*]+N+C:GUESSNOUNSTEM  Inf;
GUESSNOUNSTEM[*GUESS*]+N+M:GUESSNOUNSTEM  NM;
[book]+N+M:xOOme                          NM;
[man]+N+M:rOme                            NM;
[mountain]+N+M:toow                       NM;
[old_man]+N+M:hLlo                        NM;
[stone]+N+M:One                           NM;
...

GUESSNOUNSTEM[*GUESS*]+N+F:GUESSNOUNSTEM  NF;
[city]+N+F:poljs                          NF;
[girl]+N+F:zeepe                          NF;
[letter]+N+F:epjstolE                     NF;
[old_woman]+N+F:hLlO                      NF;
[queen]+N+F:RrO                           NF;
...

LEXICON Names

[Zecharia]:sakharias                      #;
[Elizabeth]:eljsabet                      #;
[Herod]:hErOdEs                           #;
...

LEXICON Verb

GUESSVERBSTEM[*GUESS*]+V:GUESSVERBSTEM  Inf;
[give]+V+INF+:ti                      Infix;
[go]+V+INF+:bOk                       Infix;
[situated]+V+INF+:kO                  Infix;
[write]+V+INF+:shaj                   Infix;
[walk]+V+INF+:mooze                   Infix;
[come]+V+INF+:ej                      Infix;
```

Table 2: Selected fragment with example entries from the lexicon in *lexc* format illustrating the gloss-tag combination that the analyzer yields. Also show are the single-symbol entries that will be expanded into a 'guesser' such as GUESSVERBSTEM.

3. Implementation

The analyzer is implemented using the *foma* toolkit (Hulden, 2009a). We use the *lexc* (Beesley and Karttunen, 2003) formalism to specify a lexicon transducer in a standard fashion that handles morphotactics and maps tagged and glossed lemmas to an intermediate representation. This intermediate representation is then modified to yield the actual surface forms by composing the lexicon transducer with a set of transducers that handle morphophonological alternation. Table 2 shows a snippet of our lexicon model and illustrates how glosses are implemented together with the morphological parse.

Long-distance agreement patterns are modeled with *flag diacritics* (Beesley and Karttunen, 2003). These are special unification symbols that can be introduced into lexical entries. In doing so, one creates a grammar that overgenerates wildly, but where the overgeneration is curbed at runtime by a compatibility evaluation of these special symbols. For example, nominal-attaching morphemes in our lexicon are decorated with symbols for gender such as @U.GENDER.M@ and @U.GENDER.F@, which cannot be combined in the same word.

3.1. Evaluation

An initial implementation of the analyzer includes lexical entries for 95 verbs, 50 nouns, 65 productive prefixes, 36 closed-class words (demonstratives, conjunctions), and several proper names. Additionally, a guessing facility is provided—that is, the lexicon is augmented with special lexicon entries GUESSNOUNSTEM, GUESSVERBSTEM that are replaced with any phonotactically plausible sequence before application of phonological alternations. This yields a grammar that will generally be able to analyze any word into its morphological constituents, however, one that produces glosses only where the constituents are found in the lexicon. We additionally combine the guessing grammar with the original grammar through a finite-state operation called *priority union* (Beesley and Karttunen, 2003), which produces a combined transducer that will output guesses only in the event that an analysis based on the included lexicon is not possible. We expect the lexicon to be straightforward to augment with new entries as desired.

As a small evaluation corpus, we have performed a manual word separation and annotation of 111 content word forms from the first passages of *The Gospel of St. Luke*. In this evaluation, many words receive multiple analyses (2.9 on average). The number of word forms containing the correct analysis was 105, yielding a recall of 94.6%. Of the six unanalyzed forms, we were unable to provide a plausible manual analysis to five of them, given the documented sources of the language. See Table 3 for an illustration of the type of analyses produced by the system.

```
Example Analysis
INPUT
Mpesnaw
OUTPUT
Mpesnaw NEG+past+[two]+N+M
Mpesnaw NEG+past+PRF+CIRC+[two]+N+M
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[see]+V
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[see]+V2SF+DEP+
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[have_pity]+V+INF2SF+DEP+[what]
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[have_pity]+V+INF[what]
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[go]+V+INF2SF+DEP+[what]
Mpesnaw NEG+past+PRF+CIRC+3SF+DEP+[go]+V+INF[what]
```

Table 3: Example analysis of the word **Mpesnaw** illustrating the types of ambiguity produced by the analyzer.

The system can also be used for automatic spacing of Coptic texts—we can extract the output projection of the resulting transducer, converting it to an automaton that accepts all and only well-formed Coptic words, which can subsequently be re-converted into a transducer which inserts spaces between legal words in all possible ways.

4. Conclusion & Future work

We expect the analyzer to be useful in a number of contexts. From a teaching perspective, this is a useful tool for helping students understand Coptic grammar, especially those early on in their studies. One of the seemingly difficult tasks for students new to the language appears to be the segmentation of texts, especially in cases of homophony or large units. This system offers a tool whereby students struggling with a text can automatically produce all of the possible readings for a unit. Additionally, the analyzer provides the possibility of automatically spacing and glossing larger texts in Coptic. Some manual disambiguation remains to be done by the user. However, as future work, the analyzer could profit from being combined with a disambiguator. Such a tool could be coupled with a hand-written constraint grammar (Karlsson, 1990), such as is done in Bick (2000), and other developers in different domains (Forcada et al., 2011). As Zeldes and Schroeder (2015) reports encouraging POS-tagging accuracy results using two different less fine-grained annotation schemes than are assumed in this paper, we expect that a statistical disambiguator could also be trained as more labeled data becomes available.

Increasing coverage of the analyzer can in most cases be done without major additions to the grammar as we expect to have captured the major morphophonological alternations and morphotactic constraints. The majority of work to expand the system thus falls in the domain of lexicography. Training a statistical word-divider of Coptic is also possible using the system if a larger corpus of Coptic were to be made available.

Acknowlegdements

We thank the anonymous reviewers and Joshua Roberson at the University of Memphis for their helpful comments.

5. Bibliographic References

Ashton, N. (2012). Second position clitics and monadic second-order transduction. In *Proceedings of the Workshop on Applications of Tree Automata Techniques in Natural Language Processing*, pages 31–41. Association for Computational Linguistics.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Publications, Stanford, CA.

Beesley, K. R. (1998). Arabic morphology using only finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic languages*, pages 50–57. Association for Computational Linguistics.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

Depuydt, L. (1993). On Coptic sounds. *Orientalia*, pages 338–375.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, jul.

Hulden, M. (2009a). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Hulden, M. (2009b). Revisiting multi-tape automata for Semitic morphological analysis and generation. *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, pages 19–26.

Jaber, S. and Delmonte, R. (2008). Arabic morphology parsing revisited. In *Computational Linguistics and Intelligent Text Processing*, pages 96–105. Springer.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Papers presented to the 13th International Conference on Computational Linguistics*, volume 3, pages 168–173, Helsinki, Finland.

Kay, M. (1987). Nonconcatenative finite-state morphology. In *Proceedings of EACL 1987.*

Lambdin, T. O. (1983). *Introduction to Sahidic Coptic*. Mercer University Press.

Layton, B. (2004). *A Coptic grammar with chrestomathy and glossary: Sahidic dialect*. Otto Harrassowitz Verlag.

Layton, B. (2007). *Coptic in 20 lessons: introduction to Sahidic Coptic with exercises & vocabularies*. Peeters Publishers.

Loprieno, A. (1995). *Ancient Egyptian: a linguistic introduction*. Cambridge University Press.

Orlandi, T. (2004). Towards a computational grammar of Sahidic Coptic. In Jacques van der Vliet et al., editors, *Coptic studies on the threshold of a new millennium*, pages 125–130.

Thompson, H. (1924). *The Gospel of St. John according to the earliest Coptic manuscript*. British School of Archaeology in Egypt.

Zeldes, A. and 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*, 31:164–176.