Digital Developments for Dialects in Coptic

Amir Zeldes[1], Caroline T. Schroeder, Nicholas Wagner, Lydia Bremer-McCollum and Hany Takla

Presentation: https://www.youtube.com/watch?v=sXVF2m8j7JQ

In this paper we present ongoing work in expanding the coverage of Coptic materials online beyond the classical forms of the language to ones that have been less studied, but are no less important, for our understanding of Coptic cultural heritage. The Coptic Scriptorium project has spent the last decade building natural language processing (NLP) tools, an annotated corpus of texts, and an interactive research platform for the study of Coptic literature and language, especially in the classical dialect (Sahidic) (Schroeder & Zeldes 2020; Zeldes & Schroeder 2015). This paper describes the development of a suite of tools for the dialect of Bohairic as well as a pilot corpus of annotated Bohairic texts built using those tools.

Coptic is the heritage language of millions of people in Egypt and the diaspora. Coptic Christianity is the largest Christian community in the Middle East, with significant diasporic populations in the United States, Canada, and Australia. The history of the community goes back to the first and second centuries. The Coptic language is the last phase of the ancient Egyptian language family, having evolved ultimately from the hieroglyphs of pharaonic Egypt. It was used widely in the Roman and early Byzantine/Islamic periods of Egyptian history and consists of primarily Egyptian vocabulary (with substantial contributions from Greek terms and to a lesser extent Latin and Arabic) written in an alphabet comprised of the Greek letters with additional Egyptian characters; the grammar is Egyptian. (Allan 2020; Layton 2011; Müller 2021) Although Coptic declined as a spoken language with the increasing influence of Arabic, it remains a liturgical language in the Coptic Orthodox church, and there are movements in the Middle East and the United States to reinvigorate the spoken language. Researchers in linguistics, history, religious studies, biblical studies, Egyptology, papyrology, archaeology, classics, and art history all use the Coptic language, as well.

Important materials surviving in the classical dialect of Sahidic include letters, monastic rules, saints' lives, sermons, documentary sources (wills, receipts, etc.), magical texts, and biblical and other religious texts. Coptic Scriptorium has already produced a richly annotated corpus linked to an online dictionary with selections from all of these genres in the classical Sahidic dialect, but not yet in Bohairic. There is a need to expand Coptic digital resources to include Bohairic, since a substantial number of manuscripts survive in this dialect. Moreover, Bohairic is the liturgical language of the Coptic Orthodox Church and is still used in religious services in Egypt and the diaspora.

Although Sahidic and Bohairic are related, a range of differences make it impossible to analyze Bohairic texts using tools trained on Sahidic. On the most basic level of the alphabet, Bohairic has an additional letter. Both dialects contain the letter *hore* (Coptic ϩ, Unicode U+03E9 small/ U+03E8 capital). Bohairic, however, distinguishes between the *hore* /h/ and the *khei* ⳉ /x'/ (Unicode U+2CC9 small/U+2CC8 large). Compare the Sahidic word ⲉϩⲟⲩⲛ (*ehoun* "in") with Bohairic ⲉⳉⲟⲩⲛ (*ex'oun* "inward") and the two Bohairic words ⳉⲣⲏⲓ (*x'rēi* "lower part") and ϩⲣⲏⲓ (*hrēi* "upper part"). Moreover, unlike in Sahidic, Bohairic uses aspirated allophones (ⲑ, ⲫ, ⲭ / *th*, *ph*, *ch*) before sonorants (ⲃ, ⲗ, ⲙ, ⲛ, ⲣ, ⲟⲩ / *b*, *l*, *m*, *n*, *r*, ou as *w*). Compare the article+noun phrase "the god" in both dialects: in Sahidic it is ⲡⲛⲟⲩⲧⲉ (*pnoute*), but in Bohairic ⲫⲛⲟⲩϯ (*phnouti*). Also compare the term "on account of" in Sahidic ⲉⲧⲃⲉ (*etbe*) with Bohairic ⲉⲑⲃⲉ (*ethbe*). Other spelling differences between these dialects require individualized lemmata

[1] Corresponding author; Georgetown University, Department of Linguistics amir.zeldes@georgetown.edu

to be added to a comprehensive database and to be included in the lemmatizer in an NLP suite of tools. Other grammatical and linguistic differences illustrate why Sahidic NLP tools cannot be easily applied to the Bohairic language. The interrogative particle and negative particle in Bohairic, for example, are graphically identical (ⲁⲛ/*an*), while in Sahidic they are different (ϵⲛϵ/*ene* vs ⲁⲛ/*an*). Additionally words unique to Bohairic (*i.e.*, terms that do not appear in Sahidic) must receive their own identifiers as lemmata.

From a technical perspective, our work on expanding digital Coptic coverage to Bohairic consists of three iterative steps which feed into each other: 1. The establishment of guidelines for handling the analysis of Bohairic texts; 2. Creation of a core corpus of gold-standard annotations for training and evaluation of tools; and 3. Digitization of additional Bohairic works, to which we apply automatic analyses using tools trained on gold standard data, and which we can correct manually to expand 2., while refining the guidelines from step 1. Concretely, our work will represent the first fully segmented, part-of-speech tagged and dependency parsed treebank of Bohairic Coptic data, which we intend to release as part of the Universal Dependencies project (de Marneffe et al. 2021, https://universaldependencies.org/), a platform for the release of morpho-syntactically annotated data following a typological linguistic methodology.

In our paper, we will focus on the challenges of doing this work in the context of pre-existing work on the closely related Sahidic dialect. While that work has meant that we can leverage existing tools and guidelines as a starting point, it has also meant that decisions need to harmonize with those taken for Sahidic while also respecting the differences in Bohairic. Such considerations include harmonization of word segmentation decisions, part-of-speech tags, and syntactic annotation guidelines, but also considerations beyond single dialect analysis, such as the use of hyper-lemmatization (*i.e.*, grouping related words across dialects using common identifiers), in order to allow for linking to shared lexicographic resources, which cover multiple dialects (notably, the Coptic Dictionary Online, Feder et al. 2018). With our initial resources in hand, this paper will describe and evaluate our results using methods from NLP for closely related languages, which feed into the virtuous cycle of corpus expansion and automatic tool refinement.

Abbreviated Bibliography

Allen, James P. *Coptic: A Grammar of Its Six Major Dialects*. 1st edition. University Park, PA: Eisenbrauns, 2020.

Feder, Frank, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. “A Linked Coptic Dictionary Online.” In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 12–21. Santa Fe, New Mexico: Association for Computational Linguistics, 2018. https://www.aclweb.org/anthology/W18-4502.

Layton, Bentley. *A Coptic Grammar*. 3rd Edition, Rev. Porta Linguarum Orientalium Neue Serie 20. Wiesbaden: Harrassowitz, 2011.

Müller, Matthias. *Grammatik des Bohairischen*. Annotated edition. Hamburg: Widmaier Verlag, 2021.

Schroeder, Caroline T., and Amir Zeldes. “A Collaborative Ecosystem for Digital Coptic Studies.” *Journal of Data Mining & Digital Humanities* Special Issue on Collecting, Preserving, and Disseminating Endangered Cultural Heritage for New Understandings through Multilingual Approaches (September 23, 2020). https://doi.org/10.46298/jdmdh.5969.

Schroeder, Caroline T., and Amir Zeldes. “Coptic SCRIPTORIUM.” Coptic SCRIPTORIUM, 2013-2023. https://copticscriptorium.org/.

Zeldes, Amir, and Caroline T. Schroeder. “Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities.” *Digital Scholarship in the Humanities* 30, no. suppl 1 (December 1, 2015): i164–76. https://doi.org/10.1093/llc/fqv043.