Building Linguistically and Intertextually-Tagged Coptic Corpora with Open Source Tools

So Miyagawa[1], Amir Zeldes[2], Marco Büchler[3], Heike Behlmer[1], Troy Griffitts[4]

1: University of Göttingen, Seminar for Egyptology and Coptic Studies, and Deutsche Forschungsgemeinschaft Collaborative Research Centre 1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam" Project Area B 05 "Scriptural Interpretation and Educational Tradition in Coptic-Speaking Egyptian Christianity of the Late Antiquity: Shenoute, Canon 6" 2: Georgetown University, Department of Linguistics, 3: University of Göttingen, Institute for Computer Science, eTRAP Research Group, 4: Göttingen Academy of Sciences and Humanities

* This work has been supported by joint funding from the National Endowment for the Humanities (NEH grant HG-229371) and Deutsche Forschungsgemeinschaft (DFG project 273503199) and funded by Deutsche Forschungsgemeinschaft's Collaborative Research Centre 1136 "Education and Religion in Cultures of the Mediterranean and Its Environment from Ancient to Medieval Times and to the Classical Islam" Project Area B05 "Scriptural Interpretation and Educational Tradition in Coptic-Speaking Egyptian Christianity of the Late Antiquity: Shenoute, Canon 6."

Coptic is the last stage of the Egyptian language. Before Coptic, Ancient Egyptian was written in Hieroglyphs, Hieratic, and Demotic scripts. Starting in the third century CE (excluding "Old Coptic"), Coptic used an alphabet based on the Greek and several added Demotic letters. A large but understudied corpus of literary texts exists in Coptic, including important Gnostic, monastic and Manichaean texts, as well as early Biblical translations. Efforts to build a digital Coptic corpus are still in their initial phases. In this paper, we present the most recent work in a partnership of Digital Humanities projects. Coptic SCRIPTORIUM (Schroeder and Zeldes, 2016) is a major initiative endeavoring to put corpora online which are linguistically and philologically annotated (i.e. supporting grammatical, paleographical and literary annotations), while projects in Göttingen are producing digital editions of Coptic texts focusing on philological standards and critical editions: A project at the Göttingen Academy of Sciences and Humanities is preparing a complete digital edition of the Coptic Old Testament (Behlmer and Feder, 2017), and in a project of Collaborative Research Centre 1136 "Education and Religion" digital diplomatic editions of selected works of Shenoute and Besa, 4th-5th century abbots of the White Monastery in Upper Egypt, are being prepared for text reuse research (see below). Based on our experiences, we have schematized workflows for building Coptic corpora with linguistic and literary information by using open source programs, merging data from OCR (Optical Character Recognition) and transcription sources, Natural Language Processing (NLP) tools, and manual annotation interfaces allowing for the correction of automatic tool output.

Digital transcriptions of Coptic texts are acquired in several ways, taking care to target either out-of-copyright editions or diplomatic transcriptions of manuscripts, both of which can be made freely available under Creative Commons licenses. For data not yet available in digital transcription, we adopted OCRopus, an open-source, language-independent neural network-based OCR program first developed by Thomas Breuel (Bulert et al., 2017). Our OCRopus model, trained for Coptic print editions, achieves a high accuracy rate close to 97%. The open-source data is available at the GitHub repository of the KELLIA project (https://github.com/KELLIA/CopticOCR, accessed 6 July 2018).

In addition to OCR data, we are working to offer consistent representations of already digitized texts. Pioneers of Coptic DH such as Tito Orlandi have accumulated digital transcriptions using old ASCII-based fonts that display the Latin alphabet in a Coptic font. Since 2013, Coptic SCRIPTORIUM has undertaken to convert and re-publish such data, as well as digitizing new texts in Unicode. A Unicode converter for old ASCII font encodings is available on the SCRIPTORIUM website. Using converted texts, OCR data and new transcriptions, a broad collection of digital Coptic texts has been produced.

To validate the results of both conversions and of new digital transcriptions, we use two freely available annotation interfaces: the Virtual Manuscript Room developed by Troy Griffitts (VMR, https://vmrcre.org/, accessed 6 July 2018, see Griffitts, 2017), and GitDox (Zhang and Zeldes, 2017, https://corpling.uis.georgetown.edu/gitdox/, accessed 6 July 2018), an annotation environment optimized for correcting linguistic annotations.

The VMR editor enables a team to produce online diplomatic and critical editions, using digital manuscripts images. In this phase, we can correct the errors of the OCR, the digital or the original transcriber. Moreover, one can tag philological information appropriate for Coptic and Greek manuscripts and export the data in a TEI XML format. The GitDox interface offers XML validation options as well, but also includes a spreadsheet-based interface, which makes it easy to view aligned annotations at the levels of word forms, phrases, and sentences. It is being used to correct automatic part-of-speech (POS) tagging, lemmatization, and morphological analysis, among other things.

Coptic Scriptorium's NLP pipeline (Zeldes and Schroeder, 2016) provides automatic linguistic analyses, including a morphological tokenizer for the highly agglutinative complex word forms used in Coptic, a lemmatizer linked to the Coptic Dictionary Online (https://corpling.uis.georgetown.edu/coptic-dictionary/, accessed 6 July 2018), automatic POS tagging, language of origin detection for Greek loan words, and a syntactic dependency parser which outputs annotations in the Universal Dependencies scheme (http://universaldependencies.org/, accessed 6 July 2018). All of these tools are trainable, meaning they could also be used to benefit automatic analysis in other languages with similar challenges. The Coptic texts of the manuscripts are in *scriptio continua*. Many modern editions, however, insert spaces between phrases known as bound groups, similarly to the analysis of complex space-delimited word forms in modern Arabic and Hebrew, or related ancient language varieties, such as Biblical Hebrew, Classical Arabic, or Syriac. Thus, the tokenization and "word"-segmentation are a key component for Coptic NLP. Coptic SCRIPTORIUM's tools currently achieve an average of 98.82% correct boundary detection, or 94.87% perfectly segmented bound groups in tokenization. Based on this tokenization, we tag the linguistic categories mentioned above, i.e. POS tagging, syntactic parsing, lemmatization and language of origin, which can then undergo manual correction. The data is exported in a number of formats, including EpiDoc XML (Bodard and Stoyanova, 2016), TreeTagger SGML (Schmid, 1994) and Paula XML (Dipper, 2005), all of which are well documented open standards. Finally, one can visualize the corpus with linguistic and philological tags using ANNIS (Krause and Zeldes, 2016), a search and

visualization platform for richly annotated corpora which is currently in use for a variety of Digital Humanities corpora.

With annotated corpora at hand, we have focused on applications such as text reuse detection and visualizing intertextuality among Coptic monastic texts. The latter incorporates quotations from other texts considered authoritative, especially from the Coptic translation of the Bible. The eTRAP research group of the University of Göttingen has developed TRACER (Büchler et al., 2018 forthcoming), a program to detect text reuse, especially in classical or historical languages. Using TRACER we have found previously undetected quotations of the Bible in selected works of the abbots Shenoute and Besa. The information gained is visualized using the TRAViz program (Jänicke et al., 2015) and can be represented in ANNIS.

The means to build a Coptic corpus using only open data and tools are currently only available for Sahidic, the main literary dialect of Coptic in Late Antiquity. Extending this work to other dialects, we ultimately hope to provide standards of Natural Language Processing for the entire Coptic literary corpus.

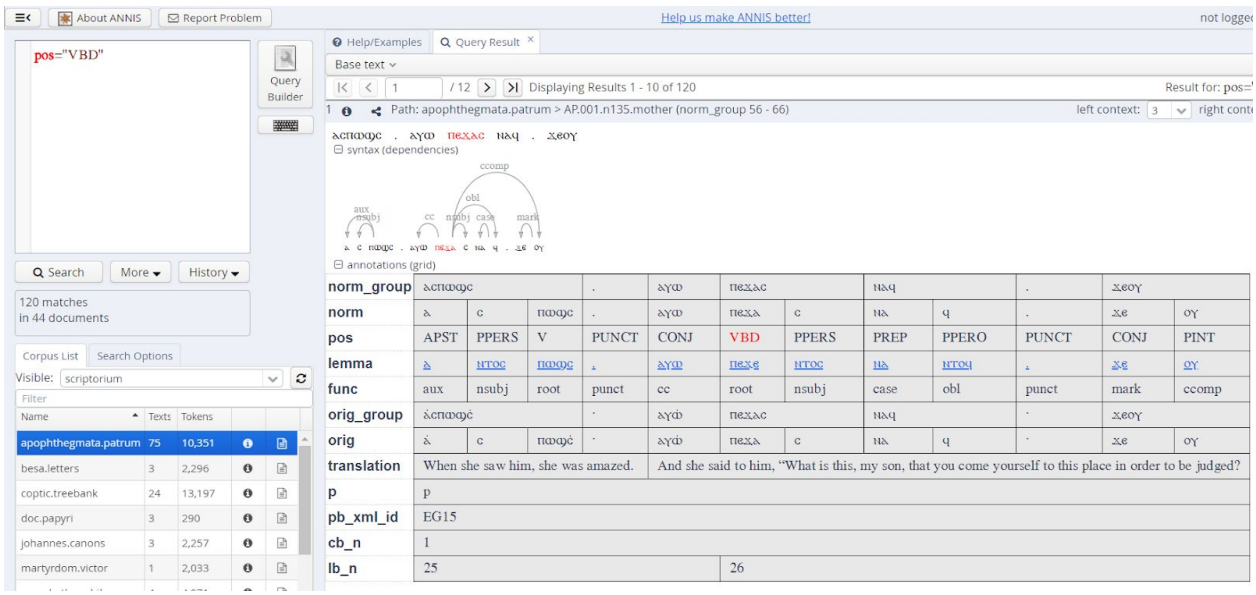

Figure 1: Coptic XML corpora visualized by ANNIS from Coptic SCRIPTORIUM

References

Behlmer, H. and Feder, F. (2017). "The Complete Digital Edition and Translation of the Coptic Sahidic Old Testament. A New Research Project at the Göttingen Academy of Sciences and Humanities." *Early Christianity*, 8: 97–107.

Bodard, G. and Stoyanova, S. (2016). "Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy." In Bodard, G. and Romanello, M. (eds) *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagemen*t. London: Ubiquity Press, pp. 51–68.

Büchler, M., Franzini, G., Franzini, E., Moritz, M. and Bulert, K. (2018 forthcoming). "TRACER - a Multilevel Framework for Historical Text Reuse Detection." *Journal of Data Mining and Digital Humanities* (Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages).

Bulert, K., Miyagawa, S. and Büchler, M. (2017). "Optical Character Recognition with a Neural Network Model for Printed Coptic Texts." In Rhian L. et al. (eds) *Digital Humanities 2017 Conference Abstracts*. pp. 657–9.

Dipper, S. (2005). "XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation." In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, pp. 39-50.

Griffitts, T. (2017). *Software for the Collaborative Editing of the Greek New Testament*. Ph.D. Thesis, University of Birmingham.

Jänicke, S, Geßner, A, Franzini, G., Terras, M., Mahony, S. and Scheuermann, G. (2015). "TRAViz: A Visualization for Variant Graphs." *Digital Scholarship in the Humanities* (Digital Humanities 2014 Special Issue).

Krause, T. and Zeldes, A. (2016). "ANNIS3: A new architecture for generic corpus query and visualization." *Digital Scholarship in the Humanities 2016*, 31. http://dsh.oxfordjournals.org/ content/31/1/118 (accessed 6 July 2018).

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing.* Manchester.

Schroeder, C. T. and Zeldes, A. (2016). "Raiders of the Lost Corpus." *Digital Humanities Quarterly* 10(2). http://www.digitalhumanities.org/ dhq/vol/10/2/000247/000247.html (accessed 6 July 2018).

Zeldes, A. and Schroeder, C. T. (2016). "An NLP Pipeline for Coptic." In *Proceedings of LaTeCH 2016 - The 10th SIGHUM Workshop at the Annual Meeting of the ACL*. Berlin, pp. 146–55.

Zhang, S. and Zeldes, A. (2017). "GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription." In *Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages*. Marco Island, Florida, pp. 619–23.