Submitted on: 05.08.2019

Open Greek and Latin: Digital Humanities in an Open Collaboration with Pedagogy

Thomas Köntges
Alexander-von-Humboldt-Chair for Digital Humanities at the Institute for Computer Science and Mathematics, University of Leipzig, Leipzig, Germany.
E-mail address: thomas.koentges@gmail.com

Rhea Lesage
Americas, Europe, and Oceania Division, Widener Library, Harvard University, Cambridge, MA, United States.
E-mail address: karabel@fas.harvard.edu

Bruce Robertson
Classics, Mount Allison University, Sackville, New Brunswick, Canada.
E-mail address: bruce.g.robertson@gmail.com

Jeannie Sellick
Department of Religious Studies, University of Virginia, Charlottesville, VA, United States.
E-mail address: jms5fe@virginia.edu

Lucie Wall Stylianopoulos
Fiske Kimball Fine Arts Library, University of Virginia, Charlottesville, VA, United States.
E-mail address: lucie@virginia.edu

Abstract:

This paper outlines and describes the work flow used to create the First Thousand Years of Greek component of the Open Greek and Latin project. Open Greek and Latin (OGL) is an international collaborative consortium of librarians, faculty and researchers committed to creating an Open Educational Resource (OER) featuring a corpus of digital texts, deep-reading tools, and open-source software. The consortium is working to free researchers (and library resources) from dependence on commercialized data and addresses the growing need for open textual corpora to support new forms of born-digital annotation, advanced reading practices, and expanded audiences for pre-modern Greek and Latin. The authors include two use cases for the open access collection and suggest expanded research opportunities as it grows to include multilingual translations and editions. They describe the challenges and opportunities encountered in the process and propose ways in which this international collaboration can grow via distributed teams of librarians, faculty, students, and researchers.

Keywords:
Collaboration, Open Access, Digital Collections, OCR, Greek and Latin

Open Greek and Latin (OGL) is an international collaborative consortium of librarians, faculty and researchers committed to creating an Open Educational Resource (OER) featuring a corpus of digital texts, deep-reading tools, and open-source software. The consortium is working to free researchers (and library resources) from dependence on commercialized data and to address the growing need for open textual corpora to support new forms of born-digital annotation, advanced reading practices, and expanded audiences for pre-modern Greek and Latin. The OGL aims to go beyond text entry and allow for robust searching capabilities. Adding a generation of machine-actionable annotations will create new "smart" editions that can personalize themselves to the needs of audiences.

This paper aims to outline and describe the work flow used to create the First Thousand Years of Greek (First 1K Greek) component of the OGL. We will move from content creation using custom-designed Optical Character Recognition (OCR) for quality control and structural metadata followed by text correction and review according to international standards for encoding scholarly editions. Finally, we will discuss creating and encouraging student digital humanists through internship programs.

We include two use cases for the open corpus: one by a digital humanities scholar and the second by a graduate student and propose future research opportunities as the corpus expands to include multilingual translations and editions. The OGL corpus is growing thanks to a diverse group of institutional partners each of whom offer unique strengths. Using the experience gained from the First 1K Greek segment, the team proposes ways in which this international collaboration can grow via distributed teams of librarians, students, and other interested parties. OGL offers opportunities for students at every level, from the introductory student to advanced researchers who are proposing new analyses of complex sources, whether in formal classes or informal settings, to progress from the most basic to the advanced levels. Contributors to OGL can be citizens in a new, democratized republic of letters, and what better place to initiate the call than at IFLA WLIC 2019 in Athens, Greece?

Retrospective and Building the Corpus

OGL was established in 2013 and its chief architect is Dr. Gregory Crane, Editor-in-Chief of the Perseus Project at Tufts University and Professor of Digital Humanities at the University of Leipzig. Since its beginnings at the Humboldt Chair of Digital Humanities at the University of Leipzig, OGL has developed into an international collaborative consortium. Current partners include the Harvard Center for Hellenic Studies, the Harvard Library, the University of Virginia Library, Mount Allison University, the Perseus Digital Library at Tufts University, the Open Philology Project at the University of Leipzig, and most recently, the National Library of Greece.

In 2016, the Harvard Library and the Harvard Center for Hellenic Studies joined forces with Mount Allison University and the University of Virginia to help the OGL implement a proof of concept of the project, focusing on the first thousand years of Greek literature. Funding for the First 1K Greek component of the OGL came from the Harvard Library through a grant from the Arcadia Foundation and the generous support of the Center for Hellenic Studies.

The goal of the First1K Greek is to collect at least one edition of every Greek work composed between Homer and fourth century CE, with a focus on texts that do not already exist in the Perseus Digital Library (which is already part of OGL). The project's work flow reflects the international collaboration that defines OGL. The Perseus team created the metadata for the targeted First 1K Greek texts, identified the highest quality digital scans available and sent them to the Mount Allison University (Canada) Lace OCR Project, led by Professor Bruce Robertson. His group ran the scanned texts through Robertson's customized OCR open source software and outsourced these files to an outside contractor that corrected the OCR and converted texts to TEI XML. The completed texts were uploaded to the GitHub repository for additional TEI XML editing and correction for international standards

compliance, a process performed by the Leipzig team and interns from the Center for Hellenic Studies and University of Virginia.

Methods and Tools: An Iterative Process

Our task was to digitize within an 8-month period a volume of ancient Greek far larger than all that had been digitized during the previous decades. Clearly, we needed to approach the task differently; and new technologies and software were aligned at the beginning of our project to make this possible. Hitherto, ancient Greek texts were digitized through double-key entry, a slow, and therefore expensive, process that was made necessary by the poor quality of OCR on ancient Greek. While page images containing Latin-script languages had been reliably converted to digital texts through OCR, ancient Greek was among the many scripts that generated excessive errors in OCR; OCR correction was more time-consuming than manual transcription. Ancient Greek's use of diacritics -- small marks written above and below letters (especially vowels) to show changes in pitch, aspiration and silent letters --, as well as the variety of fonts and layouts used to render Greek in public-domain volumes, made the script's OCR a unique problem.

One of our group had, however, collaborated on an experimental OCR engine for ancient Greek, called Rigaudon (Robertson and Boschetti 2017). Based on the Gamera document recognition library, it produced high quality results; but it required labour-intensive manual re-training for each new font, excellent page images, and the computing power of 40 CPUs and 40 Gb of memory, which was made available in 2012-13 through a grant from Compute Canada. While Compute Canada continued to provide us with its facilities, a less resource-intensive OCR process was needed, ideally one that could be trained more easily and would be more tolerant of the poorer quality, lower resolution images that were available through services such as Google Books and Internet Archive.

This process was the Ocropus OCR engine (Thomas M. Breuel 2008), an innovative application of neural networks, which formed the basis of our project's ancient Greek OCR process. Ocropus does not require manual training for new classifiers; instead, it is trained with thousands of pairs of line images and corresponding ground truth text. Fortunately, for many fonts, we had access to this data from the results of Rigaudon, so initially we could train Ocropus rapidly. Called 'Ciaconna', our process integrated all the separate steps in the Ocropus suite and made them operate together within Compute Canada's HPC environment. As described in Robertson (Robertson 2019), it also added dehyphenation and spellcheck routines specifically designed for scholarly texts. Finally, it extended the Ocropus code so the image rectangle corresponding to each line and word of the output text was available as metadata, conforming to the hOCR standard (T. M. Breuel and Kaiserslautern 2007).

This, then provided the crucial middle step in our digitization workflow, which was conceptualized in the following way. First, our team's librarians identified the required texts and sourced their available page images from projects such as Internet Archive. If the text's font was one for which we already had a classifier, we processed these into hOCR text and metadata using Compute Canada's grid. The images and text were sent to offshore editors, who would correct these and add TEI encoding. Given the value of our grant and the labour cost of our offshore editing team, the volume of text that could be digitized depended therefore on two factors: the accuracy of the machine-generated OCR and the speed with which the editors -- none of whom were familiar with Greek -- could correct it. It was clear that speedier editing would mean more texts. There was one additional issue we faced: what to do if we needed to OCR a volume whose Greek font did not yet perform well in Ocropus/Ciaconna?

We endeavored to speed up the OCR correction and effectual training sets through an experimental editing Web app called 'Lace'. As shown in Figure 1, it presents the original page image on the left and the OCR text on the right. The text is fully editable within the browser and uses the metadata discussed above to speed up the correction process. Words which appear in a dictionary of valid forms are colored light green; those that differ from a dictionary word only by the changing of a diacritic are colored olive; and other colours indicate progressively less reliable forms of automatic correction. Finally, a light red

is used to note words that cannot be corrected and therefore likely erroneous. When an editor clicks on a word, its image pops up beside it, allowing the editor to verify it more rapidly than would be the case if she had to find the word within the page opposite. When the return key is pressed, the word is stored on the server and permanently tagged as verified by an editor. Thereafter it is highlighted with a light-blue colour.

This Web app not only sped up the editing process, it also provided a means of generating training data efficiently. A volume could be OCR'd initially with a less suitable classifier. Once a few pages of this output had been edited, the results could be saved as training material for a new, better classifier. Usually after three such steps, a very high-quality classifier resulted, with the initial pages serving both as corrected results and as training material for OCR'ing the rest of the volumes using this font.

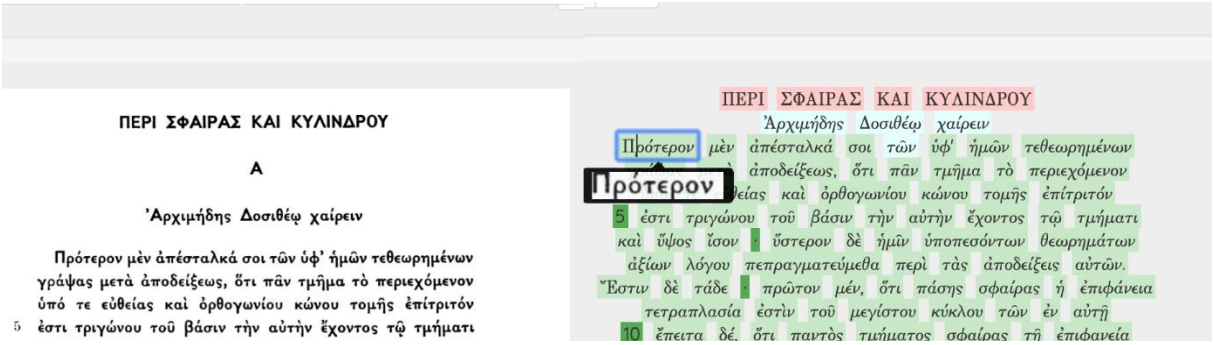

Figure 1: The Lace Web app for viewing and editing ancient Greek OCR results

Use Case 1: Using OGL's machine-actionable data in Historical Language Processing

Through a workflow involving OCR, XML editing, and the curation of library metadata, the Perseus Digital Library and OGL have made available close to 30 million words of literary Ancient Greek text under an open license. As shown above, every text in the OGL corpus starts in an hOCR-XML format and is then transformed through a number of steps into CTS-compatible EpiDoc TEI XML.[1] Aside from the effort involved in OCR corrections, the OGL team is transforming a book- and page-based citation format to a work- and section-based citation format; that is, so that a machine can identify not only the start of a work, but also the start of subsections in that work. This provides granularity and makes the OGL/Perseus corpus the biggest open machine-actionable collection of Classical Greek. While this paper already mentioned the benefits of the openness of the collection, this section will show examples of the use of this *machine-actionable* data. If philology is the art of reading slowly, this section shows a method of identifying passages to read slowly when faced with a 30-million-word corpus.

Historical Language Processing (HLP) extracts information from large amounts of historical text; a sub-discipline of Natural Language Processing, HLP has historically focused on English or other modern Indo-European languages written almost entirely in the Latin alphabet. HLP can be broken down into several analysis and synthesis processes. For instance, individual analyses of functional words, co-occurrence of words, topic modeling, metrical analysis, or morpho-syntactical analysis could be combined for clustering or classifying text. Additionally, hypotheses gained through one experiment can be evaluated using other independent data experiments or more traditional analysis. Here, we will showcase how topic modeling—that is, a statistical method to find recurring patterns of co-occurring words—can help us find passages of interest and can be used as a mathematical representation of a

[1] For the early beginnings of the CTS/CITE architecture see Smith (2009). For a more recent overview see the forthcoming chapter T. Koentges, C. Blackwell, J. Tauber, N. Smith, and G. Crane (2019, forthcoming). For CTS's independence from XML see C. Blackwell, T. Koentges, and N. Smith (2018).

complex topic.[2] All results have been generated using the Ancient Greek texts of OGL within the topic modeling app ToPān.[3]

In the first example, we used the CTS textual nodes for the Greek text of Herodotus' *Histories* and mapped them to their English translation. We then modeled them together. Mixing the translation with the original Ancient Greek has two advantages: first, as LDA topic modeling is a document-based approach, combining the two languages increases the information in the document and produces a more reliable model. Second, a person who does not know Greek very well, such as a student reading Herodotus in translation, can not only find all passages that are connected to a certain topic, for instance religious sacrifice (see *Figure 2*), but also identify the Greek terms that are related to this topic. One can easily see how this increases the accessibility of the corpus and how it can help teachers to build a source text collection.

In the second example, we produced a topic model for all Greek CTS-compatible text in the Perseus Digital Library and OGL for a larger analysis of the *Corpus Platonicum*.[1] The goal was to see whether it is possible for a machine to detect philosophical texts after only a few hours of modeling. If one looks at the most common topics for Plato and Aristotle, for example, they seem to differ at first glance. The reason for this, however, is that the most common topic in the *Corpus Platonicum* is a structural topic that can be associated with dialogue *(Figure 3)*. Once that topic is removed Aristotle and Plato score high in the same two topics: one topic related to scientific inquiry and one topic related to the nature of virtue. We can use those topics and trace them through the whole corpus; if we use the 'scientific inquiry' topic our search will lead us to passages by Aristoxenus of Tarentum, Nicomachus of Gerasa, Theon of Smyrna, and so on. In short: philosophical texts with mathematical or astronomical connections. If we search for the 'virtue' topic, we find passages by Plotinus and Maximus of Tyre among many other philosophers *(Figure 4)*. In this pilot study it seems that we can express some kind of "philosophical-ness" as numeric values of the dimension 'virtue' and the dimension 'scientific inquiry'. Thus, it is possible for a machine to extract passages based on a more complex topic from a large textual corpus with only a few hours of modeling and some human interpretation."

Neither of the two examples shown would be possible without the textual data being structured as outlined above; that is, in an open standard format that is easy to understand and supplemented by a robust citation framework. Having the data structured in such a way enables us to produce useful results within hours through Historical Language Processing. Given the swiftness of such processes, one can expect that we are just seeing the tip of the iceberg of what is possible with large, open educational resources like the OGL corpus.

[2] See Blei (2012) for LDA topic modelling. See Koentges (2016) for topic modelling ancient languages.
[3] See Koentges (2018a).

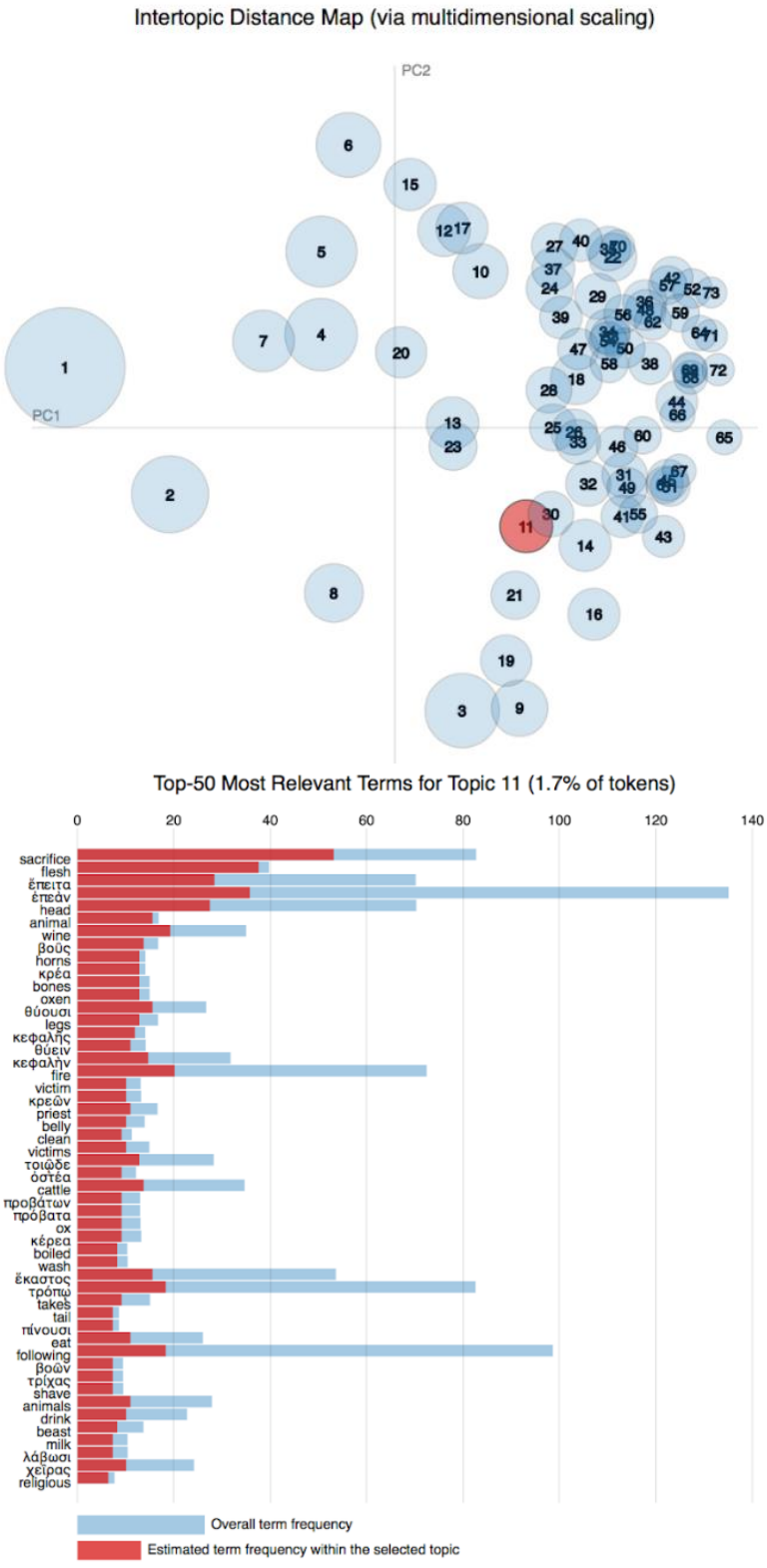

Figure 2: Principal Component Analysis of a bilingual LDA topic model of Herodotus (English and Greek). Topic 11 is selected, which shows Greek and English terms related to the topic of religious sacrifice.

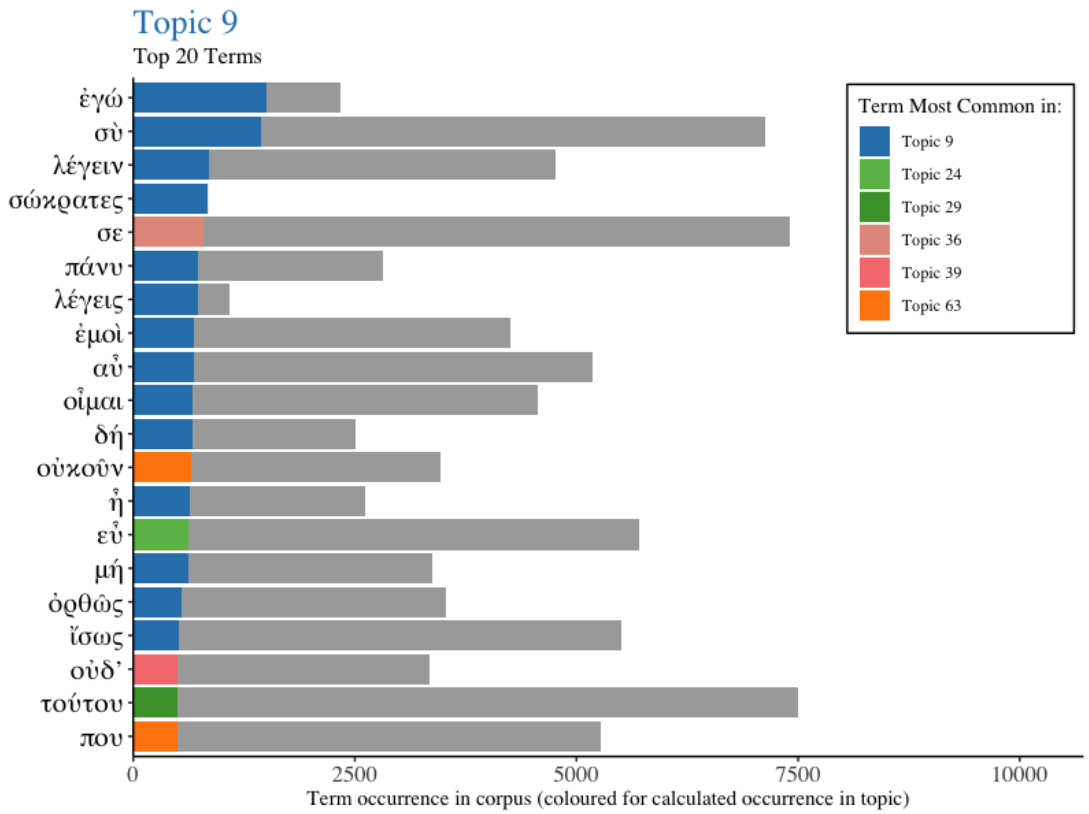

Figure 3: Top 20 Terms, Corpus Platonicum

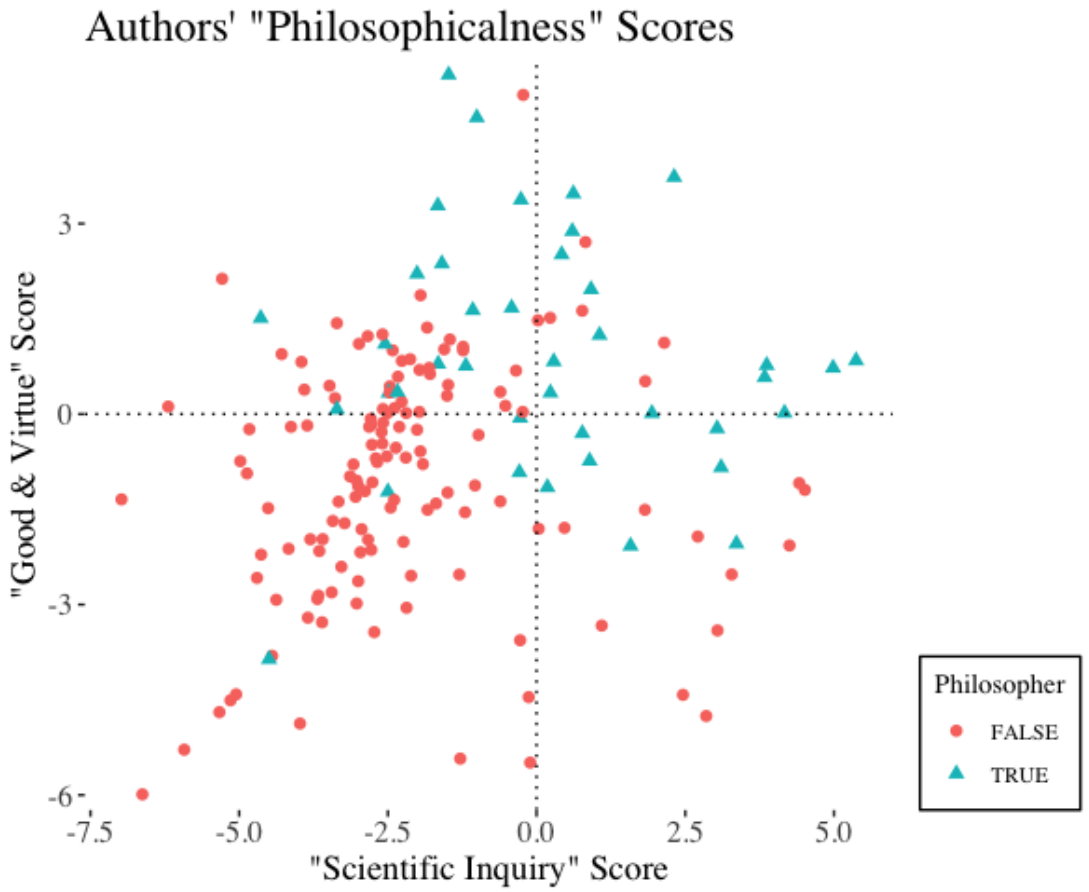

Figure 4: Authors' "Philosophicalness" Scores

Use Case 2: "Editing in your Pajamas" – The Acts of John and the Acts of Thomas

As noted earlier, the Lace environment has greatly sped up the editing process and made it more efficient, and even enjoyable. As a UVA graduate student intern, my task was to edit two early Christian apocryphal texts, the Acts of John and the Acts of Thomas.

For those unacquainted with the wonder of these texts, the acts of John and Thomas are part of a loosely related collection of stories from the 2nd and 3rd centuries known as the Apocryphal Acts of the Apostles.[4] Both the Acts of John and Acts of Thomas follow the respective journeys of the titular apostle "to the ends of the earth" (Acts 1:8). John, Thomas, and all of the Apocryphal Acts are an invaluable part of early Christian literature, yet they are woefully overlooked in many New Testament classes. While there are endless websites on which you can find the canonical New Testament in Greek, it's much more difficult to find the Apocryphal Acts in English, let alone in their original Greek.

Here is where the work of Open Greek and Latin becomes so significant. The work we do together – the scanning, correcting, Lace editing – makes possible the wider dissemination of stories like those of John and Thomas. OGL supports students, researchers, professors, and history nerds in resurrecting the study of the lesser known texts of antiquity.

Two years ago, I had the pleasure of editing Max Bonnet's 1898 critical edition of the Acts of John.[5] When I got to the project, the text had already undergone the initial step of digitalization by being run through the Lace software. My job then, became to use OCR to correct the mistakes of the software. The Lace software made many mistakes ranging from missing accents to misreading entire words. Even though the computer struggled to read the Greek, the software did do an excellent job demarcating where it struggled. Open any unedited page on the OCR and you will find a cornucopia of colors indicating the software's insecurity over individual words.

One of the most glaring mistakes the computer makes is by reading Greek characters as non-Greek letters. For example, in the Acts of John in Rome, the Emperor Domitian hears of John's preaching. The Lace software has the emperor hearing that the roman empire will be ἐκPιζωθήσεσθαι. All one must do to fix this is highlight the word in question and change that P into a proper *rho* – suddenly the king misunderstands the correct word, ἐκριζωθήσεσθαι, and as a result tries, unsuccessfully, to have John killed.[6] Another common mistake comes in the form of accents. The Lace scan consistently read μετὰ as μετά. This may seem like a small issue but ask any classical Greek instructor and you will find yourself on the receiving end of speech on the importance of acute versus grave accents.

The final two issues one often finds in Lace texts concerns circumflexes and breathings. For example, in the Acts of John, every ποῦ is rendered as πού. It takes an even keener eye to notice the breathing marks. Where the Lace program reads a rough αὐτὸν, the OGL intern needs to smooth it over with an αὐτὸν. Again, while these may seem like relatively small issues, a stray circumflex or breathing mark can alter how readers understand the text.

The mistakes are so tedious that even the most zealous apocrypha lover would be forgiven for facing some fatigue after a few hours of editing. While Lace does a nice job of catching most mistakes, editors best go word by word to ensure accuracy. As we continue to correct more and more texts in OCR, the software learns. About a year passed between when I finished the Acts of John and started the Acts of Thomas and I can say that Thomas had significantly fewer mistakes than John. So, despite the sometime

[4] For a discussion of the Acts as a loosely associated genre, please see Judith Perkins, "Fictional Narratives and Social Critique." Pages 46-69 in *Late Ancient Christianity: A People's History of Christianity*, Vol.2. Edited by Virginia Burrus and Rebecca Lyman (Minneapolis: Fortress Press, 2005).

[5] Maximillanus Bonnet, *Acta Apostolorum apocrypha*, Part II, vol. 1. Lipsiae : apud H. Mendelssohn, 1891-1903.

[6] Acts of John in Rome: 5-12.

tedious nature of text editing, opening the Acts of John and correcting it over Sunday brunch is still an excellent morning activity.

Open Greek and Latin (OGL) Local: Bringing it Back Home to You

Amid building the First 1K Greek, the team realized that Open Access, CTS compliance, and the digital tools for pedagogy and alignment that we offered were greater assets than the creation of yet another corpus. To that end, the members of the OGL Group began to craft a workflow that was locally relevant using the University of Virginia as the experimental site.

For the past three years, the University of Virginia Library has funded five student interns a year to edit XML markup and work on the GitHub repository for the First 1K Greek. The interns, both graduate and undergraduate students, were trained at Harvard's Center for Hellenic Studies and later online with our colleagues at Leipzig and Tufts Universities. It is also important to mention the local OGL group at the University of Virginia which is comprised of two Metadata Librarians specializing in XML and TEI, our Data Science Institute Lab Manager who taught us how to be comfortable with GitHub, IT engineers who kept the software running, and the Classics and Archaeology Librarian. The students received paid internships funded by the UVA Library, but the librarians and engineers from the library gave their time voluntarily until their interaction with students in OGL was recognized as an asset to their jobs.

Student participation and even ownership has been the hallmark of the Open Greek and Latin project in its three years at the University of Virginia. The student interns have not only made real contributions to the core collection, but they have also been involved in the selection of texts based on their subject expertise, writing documentation of the OGL workflow with the librarians, and encouraging the expansion of the OGL at numerous venues including IFLA and the Society for Classical Studies Pre-conference Workshop held at Tufts University in 2018. They accomplished all of this while building the necessary digital skills for the current job market. Several of the graduate student interns have remained with the project for three years and are committed to mentoring new members.

The OGL Managing Group began to see the value of channeling student involvement in the process and proposed the creation of self-sufficient "pods" of interns working with librarians in many countries to produce a customized Open Access resource called Open Greek and Latin. The goal of OGL "Local" is to further a local version of the First 1K Greek whereby adding a digital text is managed by the individual institution with only limited intervention.

The OGL Local group chooses the edition of a text that has not been previously added to either the Perseus Digital Library or the First 1K Greek and adds it to the picklist. The selected text must comply with the host country's copyright laws in order to be digitized. Quality scanning and post-processing are done on site. Simultaneously, the digital text is given a CTS URN and descriptive metadata is verified. At the UVA Library, we found that we had several 18th century classical texts that we wanted as part of our digital library, so we chose to send those to our "medium-rare" digitization queue and on through the OGL process.

At present, the OCR work is done centrally at Mt. Allison University, but the UVA interns have been involved in providing ground truth for the OCR engine to "learn to read" 18th century Greek typeset. The text is then, edited by the interns in Lace, marked up in TEI EpiDoc and sent to the editors at Tufts and Leipzig via a GitHub pull request for inclusion in the repository.

In conclusion, we have presented our OGL workflow and mentioned some of both the challenges and the opportunities that we encountered. We have documented our path and found the workflow and tools we produced to be applicable to any language making our process available so many varied voices can be heard. We have created a relevant Open Access Library for Greek and Latin with equitable participation and have provided the model for local sustainability. The OGL can be, and indeed, already has been expanded to include other languages and cultural heritage. Finally, we have also

opened the door to an Open Educational Resource (OER) for use in the classroom and in pedagogy writ large. Perhaps, more importantly, Open Greek and Latin has encouraged an extensive collaboration between library, faculty, and students in their own scholarship and mastery of digital skills.

References

Acta Apostolorum apocrypha ,trans. by Maximillanus Bonnet (Lipsiae : apud H. Mendelssohn, 1891-1903) Part II, vol. 1.

Blackwell, C., Koentges, T., and Smith, N. "CITE Exchange Format (CEX): Simple, Plain-Text Interchange of Heterogenous Datasets" in *Digital Humanities 2018: Conference Abstracts*, (Mexico City: Universidad Nacional Autónoma de México, 2018) 541–543.

Blei, D. "Probabilistic Topic Models". *Communications of the ACM* 55, 4 (2012) 77–84.

Breuel, Thomas M. "The OCRopus Open Source OCR System." in *Electronic Imaging*, 2008, 68150F – 68150F – 15. International Society for Optics and Photonics.

Breuel, T. M., and U. Kaiserslautern. "The hOCR Microformat for OCR Workflow and Results." In *Document Analysis and Recognition*, 2007. ICDAR 2007. Ninth International Conference, 2:1063–67.

Koentges, T. (2016). "Topic Modelling of Historical Languages in R". Available at: http://www.dh.uni-leipzig.de/wo/topic-modelling-of-historical-languages-in-r/ .

Koentges, T. (2018a). ThomasK81/ToPan: "The Knights Who Say t-SNE" (Version 0.5). Zenodo. http://doi.org/10.5281/zenodo.1289084.

Koentges, T. (2018b). "Research Report: Computational Analysis of the *Corpus Platonicum*". *CHS Research Bulletin* 6(1). http://www.chs-fellows.org/2018/04/30/report-corpus-platonicum/ .

Koentges, T., C. Blackwell, J. Tauber, N. Smith, and G. Crane. "The CITE Architecture: Q&A regarding CTS and CITE". Accepted for publication in S. Bond, P. Dilley, and R. Horne, eds, *Linked Open Data for the Ancient World: A Cookbook* (New York: ISAW Papers, forthcoming).

Perkins, Judith. "Fictional Narratives and Social Critique." in *Late Ancient Christianity: A People's History of Christianity*, Edited by Virginia Burrus and Rebecca Lyman (Minneapolis: Fortress Press, 2005), vol.2, 46-69.

Robertson, Bruce. "Optical Character Recognition for Classical Philology." in *Digital Classical Philology* edited by Monica Berti, (Berlin: De Gruyter, 2019) 117–36.

Robertson, Bruce, and Federico Boschetti. "Large-Scale Optical Character Recognition of Ancient Greek." *Mouseion*, 14, no. 3 (2017): 341–59.

Smith, N. "Citation in Classical Studies". *Digital Humanities Quarterly* 3, no.1 (2019), http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html .