Open Philology at the University of Leipzig

Frederik Baumgardt[1], Giuseppe Celano[1,2], Gregory Crane[1,2], Stella Dee[1], Maryam Foradi[1], Emily Franzini[1], Greta Franzini[1], Monica Lent[1], Maria Moritz[1], Simona Stoyanova[1]

[1] Department of Computer Science, University of Leipzig
Leipzig, Germany
[2] Perseus Project, Tufts University
Medford, MA, United States
E-mail: {baumgardt, celano, crane, dee, efranzini, franzini, moritz, simona.stoyanova}@informatik.uni-leipzig.de, {maryam.foradi, monica.lent}@uni-leipzig.de

Abstract

The Open Philology Project at the University of Leipzig aspires to re-assert the value of philology in its broadest sense. Philology signifies the widest possible use of the linguistic record to enable a deep understanding of the complete lived experience of humanity. Pragmatically, we focus on Greek and Latin because (1) substantial collections and services are already available within these languages, (2) substantial user communities exist (c. 35,000 unique users a month at the Perseus Digital Library), and (3) a European-based project is better positioned to process extensive cultural heritage materials in these languages rather than in Chinese or Sanskrit. The Open Philology Project has been designed with the hope that it can contribute to any historical language that survives within the human record. It includes three tasks: (1) the creation of an open, extensible, repurposable collection of machine-readable linguistic sources; (2) the development of dynamic textbooks that use annotated corpora to customize the vocabulary and grammar of texts that learners want to read, and at the same time engage students in collaboratively producing new annotated data; (3) the establishment of new workflows for, and forms of, publication, from individual annotations with argumentation to traditional publications with integrated machine-actionable data.

Keywords: crowd sourcing, linguistic annotations, CALL

1. Open Greek And Latin Project (Big and Linked Open Data)

The Open Greek and Latin Project (OGL) is one of the three complementary efforts constituting the Open Philology Project.

OGL is currently collecting and scanning editions of classical texts in an effort to build the largest and most comprehensive library of classical philology to date, concurrently contributing to the expansion of the Google Books classics collection. This open access and open source library is designed to feature searchable images of extant, copyright free editions of classical texts as well as translations in multiple languages, all encoded in accordance with the latest TEI EpiDoc (Text Encoding Initiative Epigraphic Documents) standards for optimal interchange of scholarly data across different projects and initiatives worldwide. The encoding runs in parallel with the ongoing conversion of the Perseus Digital Library files to EpiDoc. Having the Perseus Digital Library and OGL texts in EpiDoc will facilitate the linking of data with the epigraphic and papyrological databases which are already in EpiDoc, and the ones currently being converted (e.g. EAGLE).

In particular, OGL aims at providing at least one version for all Greek and Latin sources produced during antiquity (through c. 600 CE). Where existing corpora of Greek and Latin have generally included one edition of a work, the OGL corpus is designed to manage multiple versions of a work, including translations in multiple languages. For this purpose OGL has set up collaborations with academic teams in several countries, namely Bulgaria (University of Sofia), Georgia (State University of Tbilisi), Croatia (University of Zagreb), the US (University of Nebraska) and Italy (Università del Piemonte Orientale).

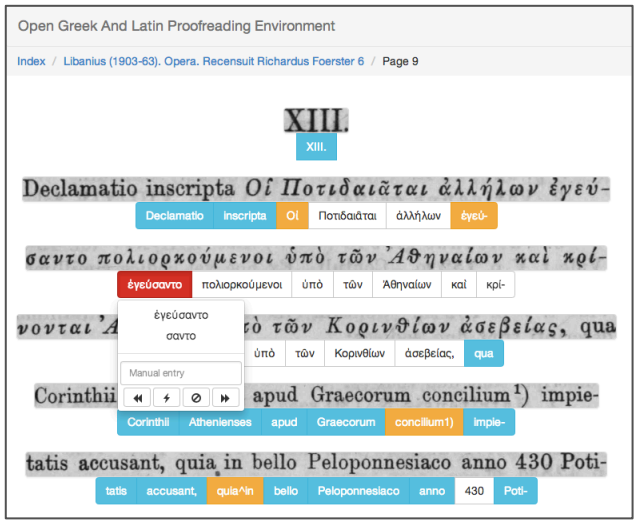

Figure 1: *Open Greek And Latin Proofreading Environment* for OCR corrections

This effort involves using OCR technology to create an open corpus that is reasonably comprehensive for the ca. 100 million words produced up to c. 600 CE. Due to the public nature of OGL, its OCR pipeline has been devised with a set of user interfaces that enable e.g. the students of the Digital Philology course to participate in the digitisation effort (see below). Built on top of the Oracle Grid Engine, the pipeline consists of three main

components: an interchangeable core of one of three OCR engines (Gamera, Tesseract, OCRopus), an optimization layer which runs the engine at varying configurations and selects and merges the best results (Robertson & Boschetti, 2013), and a module to align scan results with known editions and enable semi-manual corrections (Figure 1) (Boschetti, 2013). The computational track of the pipeline is complemented with a scheduling and reporting system that provides a page-by-page overview of OGL's progress and facilitates contributions from external sources (e.g. researchers, students, citizen scholars/scientists).

2. Historical Languages e-Learning Project (Historical Language Resources and Teaching)

The Historical Languages eLearning Project is building a web application to support users who want to learn a historical language through reading a primary source text written in that language. The system will give them appropriate vocabulary, grammar, and exercises for a text of their choice, by leveraging open access morphosyntactic data. We will supplement this with our own data, in order to reach users in more languages and to tailor the content to an internet-based audience. Users will also level-up to making original annotations to primary sources, such as morphosyntactic analyses, and composing and aligning translations, thereby making scholarly contributions to open access data throughout their learning experience (Perseids).

Our trial course will prepare students to read the Pentecontaetia section of Thucydides' Peloponnesian War in Ancient Greek (Figure 2). Since we know that identifying syntax and morphology of text indicates an understanding of the text itself (Harrington, 2012), we will use exercises like treebanking to evaluate users' grasp of the source material. One added benefit of such exercises is that they can be automatically assessed for accuracy, which allows us to provide individualized feedback to users. Therefore, one of our main goals is to make these kinds of pedagogical activities fun, rewarding, and engaging.

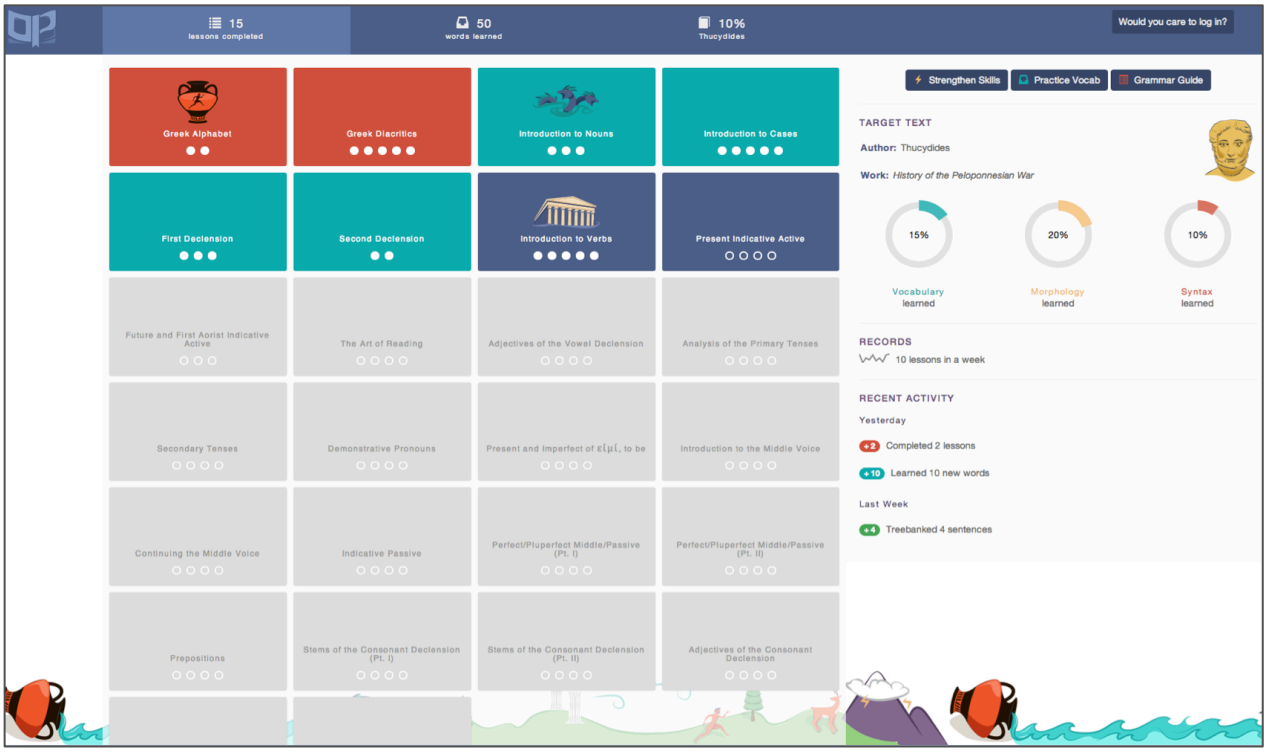

Figure 2: *Historical Languages eLearning Project* Homepage

In addition to exercises, the web interface itself must also be flexible and adaptive to individuals' needs. For example, we expect users to access the interface through a variety of devices and platforms. Through HTML5, CSS3, and responsive web design, we will serve an appropriate learning experience to users regardless of how they access the system (Figure 3). This approach, coupled with front and back-end templating systems, will allow us to create a system that is essentially language-independent, enabling access in any L1 at the structural level.

Figure 3: *Historical Languages eLearning Project* on a mobile device

Since the system relies on complex and structurally varied textual data, the challenge lies in storing this data in a scalable way, which stays true to the real world objects (e.g. sentences and documents). Our chosen graph model is capable of both supporting the user interactions in the application as well as representing this textual data in a way that the application can use it to generate dynamic exercises. Owing to its schema-free nature, graph databases are also easier to extend and maintain than traditional SQL databases.

Since the possibilities for research and learning depend on available repositories of richly annotated text, we will scale our resources to enable users to contribute to this data during their learning process. In this way, they can generate data that advances other fields, such as Natural Language Processing, which often rely on the accuracy provided by manually-curated data (e.g. disambiguating geographical place names) (Bamman and Crane, 2011; Nivre, 2008). Students will be able to collaboratively translate, annotate, and publish morphosyntactic analyses for additional primary sources, beginning with the data being generated by the Open Greek and Latin Project.

3. Publishing (Linked Open Data, Language Teaching and Evaluation)

One of the main tasks of the Open Philology Project is to establish a new model of scholarly publication in a born digital environment. Such a task is accomplished through Perseids, which is a collaborative platform of the Perseus Project that extends pre-existing open source tools for editing and annotating TEI XML documents in Classics. Perseids is a shared environment where users can edit, translate, and produce commentaries on ancient source documents, including inscriptions and manuscripts. Perseids' outcomes concern not only scholars but also students, who have the opportunity to work with original documents and contribute to the results of the scholarly community (Almas and Beaulieu, 2013). Such work is also being developed in the Digital Philology Course held at the University of Leipzig. The focus of this course is to analyze how the ancient goal of philology should be re-imagined given the challenges and opportunities of new digital media and the needs of an interlinked, global society. Students learn to deal with many different forms of publication familiar from print, such as editions, lexica, encyclopedias, commentaries, indices, grammars, etc. They also focus on particular challenging themes, as linguistic annotation of historical sources, representation of fragmentary authors, or tracking themes across language barriers.

One of the main publication models within Perseids is the Fragmentary Texts Editor (FTE), which is a shared environment for multi-level annotations of quotations and text reuses of ancient lost works (Almas and Berti, 2013a and 2013b). The goal of this editor is to produce open publications of fragmentary authors, which means annotating information pertaining to lost works embedded in surviving texts. FTE addresses many different requirements for creating a dynamic representation of quotations and text reuses of fragmentary authors, using various methods of inline and stand-off markup to produce stable ways for identifying and annotating text re-uses, including canonical citations, morphosyntactic analysis, translation and text reuse alignments. In particular, we are combining TEI, the Open Annotation Collaboration (OAC) core data model, and the CITE Architecture to represent text re-uses via RDF triples, in order to produce new dynamic, data-driven representations of the aggregated information.

3.1 Open Data Revenue models

With open data meaning by definition free access for all users, the Open Philology Project team is in the process of devising effective ways for it to be financially sustainable for years to come. The team is working to provide models to sustain and maintain distributed open source learning and discourse. The core principle is to move away from charging for monopoly access to data, to charging instead for services that allow users to identify, analyze and then contribute to increasingly complex open data.

The project aims at providing services for faculties, students and for the interested public with tools covering all areas from publishing, e-Learning and Online Assessment services, set at recognized and affordable price points. The team's objective is to widen the

audience, to radically reduce the unit costs, and to increase the available revenue for digital services, while still providing an open data platform.

4. References

Almas, B. and Beaulieu, M.C. (2013) 'Developing a New Integrated Editing Platform for Source Documents in Classics', *Literary & Linguistic Computing*. [Online] DOI: 10.1093/llc/fqt046.

Almas, B. and Berti, M. (2013a) 'Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors', *DH-Case 2013. Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*. Florence, September 10, 2013. ACM Proceedings.

Almas, B. and Berti, M. (2013b) 'The Linked Fragment: TEI and the Encoding of Text Re-uses of Lost Authors', *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*. Rome, Università la Sapienza, 2-5 October.

Almas, B., Berti, M., Choudhury, S., Dubin, D., Senseney, M., Wickett, K.M. (2013) *Representing Humanities Research Data Using Complementary Provenance Models* [Poster]. Exhibited at: *Building Global Partnerships – RDA Second Plenary Meeting* in Washington DC, 16-18 September 2013.

Bamman, D., Crane, G. (2011) 'The Ancient Greek and Latin Dependency Treebanks', in: *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, pp. 79-98. Berlin Heidelberg: Springer-Verlag. [Online] Available at: http://nlp.perseus.tufts.edu/docs/latech.pdf (Accessed: 27 August 2013).

Boschetti, F. (2013) *Components for Collaborative Philology*. [Online] Available at: https://github.com/CoPhi (Accessed: 26 August 2013).

Crane, G. et al. (2012) *Student Researchers, Citizen Scholars and the Trillion Word Library*. [Online] Available at: http://www.humanities.ufl.edu/pdf/Crane-%20Student%20Researchers,%20Citizen%20Scholars,%20and%20the%20Trillion%20Word%20Library.pdf (Accessed: 28 August 2013).

Harrington, M. (2012) *Treebanking as pedagogy: the role of syntactic control in language acquisition* [PowerPoint presentation], Tufts University, 7 August 2012.

Meurers, D. (2011) Enhancing Authentic Texts for Language Learners. Workshop on *Corpora in Teaching Language and Linguistics (CTLL)* at Humboldt-Universität zu Berlin, 6 January 2011.

Nivre, J. (2008) 'Treebanks', in Lüdeling, A. and Kytö, M. (eds.) *Corpus Linguistics: An International Handbook*. Berlin: De Gruyter, pp. 225–241. DOI: 10.1515/9783110211429.2.225 (Accessed: 23 October 2013).

Perseids: A Collaborative Editing Platform for Source Documents in Classics (n.d.) Available at: http://sites.tufts.edu/perseids/ (Accessed: 27 August 2013).

The Prague Dependency Treebank 2.0 (n.d.) Available at: http://ufal.mff.cuni.cz/pdt2.0/ (Accessed: 26 August 2013).

Robertson, B. and Boschetti, F. (2013) *Rigaudon: Polytonic Greek OCR*. [Online] Available at: http://heml.mta.ca/rigaudon (Accessed: 26 August 2013).

Rydberg-Cox, J. (n.d.) *A Hybrid Online System for Teaching Ancient Greek: A Digital Tutorial for Ancient Greek Based on John William White's First Greek Book*. [Online] Available at: http://daedalus.umkc.edu/FirstGreekBook/about/AHybridSystemforTeachingAncientGreekPreprint.pdf (Accessed: 24 June 2013).