Human Computation (2014) 1:2:177-184

ISSN: 2330-8001, DOI: 10.15346/hc.v1i2.8

Participatory Philology: Computational Linguistics and the Future of Historical Language Education

GREGORY CRANE, University of Leipzig

STELLA DEE, University of Leipzig

ANNA KROHN, Tufts University

ABSTRACT

As part of the proceedings of the Citizen Cyberscience Summit 2014, this brief summarizes the presentation given by the Open Philology Project on its goals, work, and collaborators. With a focus on the interface between 21st-century philology and citizen science, this paper reviews the data we collect, why we gather that data, and the cohort that we engage for assistance with data production. The paper presents the work of the Historical Languages eLearning Project and the incorporation of pedagogy into resources for participatory philology and reviews a case study of a class at Tufts University that supports the viability of our approach. Above all, we seek to demonstrate the deep similarities of technical infrastructure and research processes between participatory philology and citizen science, despite fundamental differences in a humanistic versus a scientific approach to the subject matter. In so doing, we hope to help lay the foundation for increasing contribution by the humanities to the fields of citizen science and human computation.

1. INTRODUCTION

The Open Philology Project plans to rejuvenate the role of historical texts in the intellectual life of a global world, and to revolutionize the teaching of historical languages in support of a dialogue among civilizations. This plan is shaped by how we define philology; to paraphrase the definition given by Augustus Boeck in 1822, “philology is the analysis of the ancient world in its entirety, including everything in the physical and intellectual world through the use of written

sources." Philology includes not only how we understand the past, but how we arrive at that understanding, in addition to how we frame that perception to our contemporaries and future generations. The discipline encompasses an enormously broad range of sources, including inscriptions, engraved objects, and papryi, as well as books and manuscripts. This range of material requires an equally broad methodological approach; again in the words of Boeck, "no methodology is out of scope if it allows us to draw meaning from the words of the past--whether that methodology involves archaeological digs, irregular verbs, or probability theory."

The scholarly analysis of this amount of material is beyond the capacity of experts alone. Representing this material so that it might be accessible to a wider audience poses an even greater challenge. As a consequence, the Historical Languages eLearning Project proposes a participatory approach to philology. We directly engage student researchers and citizen scholars in the production and analysis of digital textual data (Crane et. al, 2012); our plan envisions a globally accessible laboratory environment designed to support students and interested citizens making immediate and attributable contributions to current philological research.

2. RELATED WORK

The Open Philology Project is split into two symbiotic sub-projects: Open Greek and Latin (OGL) and the Historical Languages eLearning Project. The former focuses on the digitization of textual data, beginning with at least one edition of every Ancient Greek and Latin text, while the later focuses on the creation of resources that allow students to edit and annotate that textual data while optimizing their own language learning. Each project draws on related work taking place across national boundaries.

Open Greek and Latin builds upon the work of the Perseus Digital Library at Tufts University, as well as the optical character recognition (OCR) software developed by Bruce Robertson and Frederico Boschetti. Ongoing collaboration concentrates on revising the Canonical Text Services (CTS) Protocol developed by the Homer Multitext Project as an addressing schema for discrete textual units (Smith, 2009), as well as the EpiDoc Team to develop a suitable subset of the TEI Guidelines for the OPP corpus (http://sourceforge.net/p/epidoc/wiki/Home/). We currently collaborate closely with the Billion Words Project at the University of Leipzig (http://billion-words.e-humanities.net/), as well as the work of the Pelagios and Pleiades Projects with respect to Named Entity Identification (http://pleiades.stoa.org/).

The Historical Languages eLearning Project proceeds in close conversation with the Alpheios Project (http://alpheios.net/), as well as Dr. Brian MacWhinney, a psychologist and specialist in second language acquisition at Carnegie Mellon University. Work on dynamic syllabi, ePortfolios, and supporting the reincorporation of student and scholarly annotation into the corpus proceeds in close collaboration with the collaborative editing platform being developed by the Perseids Project (http://sites.tufts.edu/perseids/). We remain in constant communication with an international group of historical language teachers at the secondary and post-secondary level,

including Dr. Neven Jovanović at the University of Zagreb, Dr. Irine Darchia at Tbilisi State University, and the Classics Department at Tufts University.

3. PARTICIPATORY PHILOLOGY

The philological research pursued by the Open Philology Project and its collaborators focuses on four categories of data. The Historical Language eLearning Project plans to enable students and citizens to contribute to the repositories of each data category in the process of learning the language in which the texts are written, while customizing language resources to the learning styles of users and studying the process by which users learn historical languages. The first section of this brief reviews the types of data and their promise for future linguistic research, while the next section presents a case study of classroom contribution to this data by students without prior study of Ancient Greek, to demonstrate the feasibility of lowering the barrier to entry for contribution. Finally, the brief highlights some of the technical and computational infrastructure that makes this work possible.

3.1 Annotation for Research

3.1.1 Translation Alignment

Translation alignment describes the process of annotating the correspondences between the words of a modern language translation with a historical language text. The kinds of translation alignment we hope to support fall into roughly three categories. First, the correction of computationally aligned texts; second, the alignment of scanned and OCRed existing translations; and third, the creation of new translations into modern languages, produced either individually or collaboratively.

This kind of data is particularly useful for computer scientists conducting Natural Language Processing (NLP) tasks over both time and language families, especially in refining the tools used for word sense disambiguation (Bamman and Crane, 2011). In what is essentially a special case of word sense disambiguation, aligned translations can also be used for the named entity identification described in section 3.1.4; if one student is able to distinguish whether it's Cairo, Egypt from Cairo, Illinois in the text of a translation, and another student is able to align the words of an English translation containing the word 'Cairo' with the original Arabic, the alignment will allow the reader or computer scientist to disambiguate Cairo in the original text.

3.1.2 Treebanking

Morphosyntactic annotation, or treebanking, is a type of grammatical analysis designed for languages with a word order possessing greater freedom, including Ancient Greek and Latin. Although directly concerned with morphology and syntax, the process of annotating a sentence also requires deep semantic understanding (Mambrini, 2013), perhaps most obviously in such

cases as ellipses, which are missing words implied by the context. For this reason, we plan to support multiple morphosyntactic hypotheses for all sentences, each of which has been determined valid through a process of distributed review by a global cohort of expert annotators.

From a pedagogical perspective, classroom experience at Tufts suggests that the ability to produce defensible treebanks is correlated with an ability to translate text. However, treebanks have the advantage of being able to be assessed computationally; moreover, the data itself are independent of the first language of the student, allowing students to work together on morphosyntactic analysis of the same historical text even if they do not share a common modern language. Once available, morphosyntactic data support research in a number of fields, including the determination of agency in narrative texts (Mambrini, 2013).

3.1.3 OCR Correction

The Optical Character Recognition (OCR) system for both Greek and Latin currently produces results that are useful, but not perfect. However, all stages of computational textual analysis rely on faithfully transcribed source text. Moreover, participating in proofreading of OCR can require minimal knowledge of the language, but rather only an ability to read the script. Therefore, we plan to open up the proofreading of OCR to citizen philologists, creating CAPTCHA games that work with original data from OGL texts.

3.1.4 Named Entity Identification

Named entity identification includes the process of categorizing textual references to named entities, such as people and places, in accordance with authority lists that distinguish between those entities sharing the same name, as in the Cairo example above (Smith and Crane, 2001). We plan to provide citizens and scholars a gamified interface to disambiguate references in primary sources against existing authority lists available in the Perseus Digital Library, as well as those produced by third parties such as the Pleiades Project, refining those authority lists as necessary in the face of new primary source data.

Having data with accurate and disambiguated named entities supports research such as that conducted by Maxim Romanov on the centers of scholarship in the Arabic-literate world over a period of centuries (Romanov, 2013); Romanov relied upon the toponymic and professional qualifiers often found in Arabic person names. Through our participatory philology efforts, we hope to generate data that can lead to similar research across a range of historical languages.

3.2 Annotation for Learning and Learning for Research

Initial research by the Open Philology Project places the worldwide number of high school students studying Ancient Greek and Latin at well over three million (Franzini, 2014). Given this reliable cohort of possible participants, high school and college classrooms are a critical first use-case for the tools and resources of the Historical Languages eLearning Project.

3.2.1 Translation Alignment and Treebanking for Novices

We have already begun to explore the utility of translation alignments and treebanking for novice students. In the spring of 2013 and 2014, Tufts University offered a Greek Literature in Translation course. The goal was to take students who we assumed had little to no exposure to Ancient Greek and have them work directly with the authors in the original language. We developed projects to introduce the students to the Perseus Digital Library and the Alpheios Project tools. The projects taught them how to make alignments between English and Greek and how to reference Greek treebanks to improve those alignments. Overall the analytical quality and depth of understanding exhibited by the student work was encouraging. Many students expressed confidence in their ability to gain insight and meaning from the original texts that would have otherwise been lost in a translation. In the future we are aiming to have all alignments and treebanks created and accessed through the Perseids platform, thereby creating ePortfolios for the students and archiving their work for further analysis. We now plan to abstract the projects and methods for use in other courses, including other introductory Classics and Political Science courses that make use of primary source materials in a variety of languages.

3.2.2 Dynamic Syllabi

We hope to enable the computational generation of dynamic syllabi or curricula that are customized to the needs of users and the content of texts. The pilot syllabus built by the Perseids Project and used by Marie-Claire Beaulieu for a Tufts mythology class serves as an example of the former; the syllabus was automatically populated by particular excerpts of texts from the Perseus Digital Library through the Canonical Text Services' Application Programming Interface (CTS API[1]). Meanwhile, the introduction to Ancient Greek through Thucydides being developed by the OPP E-Learning team exemplifies the latter; students will learn vocabulary and grammar in order of its frequency within the text of Thucydides. Ultimately, we hope to integrate the contribution of all types of annotation listed in Section 3.1 into OPP dynamic syllabi, and build the APIs that will allow that data to support the automatic creation of new curricula.

3.2.3 E-Portfolios

The material that each student has learned and contributed, including word lemmas, word forms, and original annotations, will be recorded in an ePortfolio according to its stable URI. This ePortfolio will not only serve as a tool for the student to communicate their ability to teachers and employers, but will itself support research by the Open Philology Project into second language acquisition of historical languages in a digital environment. Current progress on ePortfolios relies heavily on work undertaken by the Perseids Project at Tufts University (Almas and Beaulieu, 2013).

[1] For a brief introduction to this course, see the presentation given at a 2013 Digital Classicist seminar available at: http://www.digitalclassicist.org/wip/wip2013-08mb.pdf

3.3 Technical Infrastructure

The code-base of the Open Philology Project is openly available on Github (https://github.com/OpenPhilology) and draws upon the work of many of the collaborators mentioned above. The Historical eLearning Project more specifically draws upon the textual data produced by the Open Greek and Latin Project that in turn relies upon Bruce Robertson's Optical Character Recognition (OCR) for Ancient Greek, as well as the OCR Proofreader developed by Frederico Boschetti. The Project is compliant with the Canonical Text Services (CTS) Protocol developed by the Homer Multitext Project to address scholarly text and images (http://folio.furman.edu/projects/citedocs/ctsurn/specification.html), and, together with the Perseus Digital Library (https://github.com/PerseusDL), is working towards providing stable URIs for all items in the Perseus and Open Philology corpora, including texts, authors, lexical units, and named entities[2].

Work towards generating computational hypotheses for the types of annotation listed above relies on algorithms developed by Saeed Majidi at Tufts (Majidi and Crane, 2013); while development by David Bamman at Carnegie Mellon University informs work towards the automatic markup on documents through aligned translations (Bamman et. al, 2010).

4. FUTURE WORK

In addition to the ongoing work outlined herein, future work will concentrate on transforming accurate and complete annotations to games allowing users to contribute new annotations while deepening their own linguistic understanding. In addition, we hope to follow up on close collaboration with those in the field of experimental computer-assisted language learning (eCALL), those who actively test the effects of different virtual environments on language learning, to create resources that are responsive to individual differences in the preferences, abilities, and learning styles of users. Furthermore, we hope to expand and deepen our collaboration with an international cohort of classroom teachers, creating resources in response to their needs and that are able to be edited by teachers themselves.

5. CONCLUSION

In conclusion, the mandate of the Open Philology Project, namely the rejuvenation of philology in a digital age, faces many similar opportunities and challenges to those of laboratories that rely upon data collected or analyzed through citizen science tools. There simply are too much textual data in historical languages for scholars to analyze alone (Crane et. al., 2012), yet supporting outside contribution that is mutually beneficial to both participants and academic researchers requires a great deal of time and resources. In the case of the Open Philology Project, the

[2] For more on the current state of these URIS, see: http://sites.tufts.edu/perseusupdates/beta-features/perseus-stable-uris/

technical and instructional infrastructure is largely in place; the immediate future will see the consolidation and integration of resources rather than their development. In the longer term, the Project is optimistic that participatory philology will be able to support new research in both the sciences and humanities.

6. ACKNOWLEDGEMENT

We would like to acknowledge the Alexander von Humboldt Foundation, the Europäischer Sozialfonds (ESF), the Sächsische AufbauBank (SAB), and Tufts University for making this work possible.

7. REFERENCES

Almas, B., Beaulieu M. (2013). "Developing a New Integrated Editing Platform for Source Documents in Classics." In Literary and Linguistic Computing; doi: 10.1093/llc/fqt046.

Bamman, D., Babeu, A., and Crane, G. (2010). Transferring structural markup across translations using multilingual alignment and projection. In JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries, pages 11-20, New York, NY, USA. ACM.

Bamman, D. and Crane, G. (2011). Measuring historical word sense variation. In JCDL '11: Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries, pages 1-10, New York, NY, USA. ACM.

Bamman, D. and Crane, G. (2011). The Ancient Greek and Latin Dependency Treebanks . In Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series, pages 79-98, Berlin Heidelberg. Available at : http://hdl.handle.net/10427/75560

Crane, G., B. Almas, A. Babeu, L. Cerrato, M. Harrington, D. Bamman, H. Diakoff (2012). Student Researchers, Citizen Scholars and the Trillion Word Library. In Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital libraries, pages 213-222, Washington, D.C. : ACM Digital Library. Available at: http://hdl.handle.net/10427/75559

Franzini,E. (2014). CALL FOR COLLABORATION: Could you help us quantify the total number of students studying Latin and Ancient Greek in the World? Digital Humanities, Universität Leipzig Blog. Available at: http://www.dh.uni-leipzig.de/wo/call-for-collaboration-could-you-help-us-quantify-the-total-number-of-students-studying-latin-and-ancient-greek-in-the-world/

Majidi, S. and G. Crane (2013). "Committee-Based Active Learning for Dependency Parsing." in Research and Advanced Technology for Digital Libraries: Lecture Notes in Computer Science Volume 8092: pp. 442-445.

Mambrini, Francesco. (2013) "Thucydides 1.89-118: A Multi-layer Treebank." CHS Research Bulletin 1, no. 2 . http://nrs.harvard.edu/urn-3:hlnc.essay:MambriniF.Thucydides_1.89-118_Multi-layer_Treebank.2013

Romanov, M. (2013). Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunni World (661-13000 CE)(Doctoral dissertation). Department of Near Eastern Studies, University of Michigan.

Smith, N. (2009) Citation in Classical Studies. Digital Humanities Quarterly: Changing the Center of Gravity: Transforming Classical Studies through Cyberinfrastructure 3(1). Available at: http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html

Smith, D.A. and Crane, G. (2001). Disambiguating geographic names in a historical digital library. ECDL 2001: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, pages 127-136, Darmstadt.