Citation:

Brown, Travis, Jennifer Guiliano, and Trevor Muñoz. "Active OCR: Tightening the Loop in Human Computing for OCR Correction" National Endowment for the Humanities, Grant Submission, University of Maryland, College Park, MD, 2011.

Licensing:

Collaborating Sites:

University of Maryland
Maryland Institute for Technology in the Humanities

Team members:

Maryland Institute for Technology in the Humanities
Travis Brown
Paul Evans
Jennifer Guiliano
Trevor Muñoz
Kirsten Keister

Acknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the collaborating institutions or the National Endowment for the Humanities.

Active OCR: A Level II Start Up Grant

Enhancing the humanities through innovation: Over the past several years, many large archives (such as the National Library of Australia and the National Library of Finland) have attempted to improve the quality of their digitized text collections by inviting website visitors to assist with the correction of transcription errors. In the case of print collections, an optical character recognition (OCR) system is typically used to create an initial transcription of the text from scanned page images. While the accuracy of OCR engines such as Tesseract and ABBYY FineReader is constantly improving, these systems often perform poorly when confronted with historical typefaces and orthographic conventions. Traditional forms of manual correction are expensive even at a small scale.

Engaging web volunteers—a process often called *crowdsourcing*—is one way for archives to correct their texts at a lower cost and on a larger scale, while also developing a user community. The scope of crowdsourcing is still critically limited, however. For example, the NLA's Australian Newspaper Digitisation Project reports that 2,994 users corrected 104,000 articles during the first six months of the program. While 104,000 articles corrected is substantial, it is proportionally insignificant compared to the amount of text digitized by the HathiTrust Digital Library, for example, which currently holds over 3.3 billion pages, or even to the 182,000 texts that are included in Gale's Eighteenth-Century Collections Online (ECCO). Additionally, the success of crowdsourcing for OCR correction depends on the amount of public interest in the domain of the collection. A historical newspaper archive, for example, is likely to attract many more committed contributors than a collection of eighteenth-century theological tracts.

Active learning is a machine learning approach that attempts to maximize the effectiveness of a human annotator by making the annotation process iterative. Instead of learning a model from a static labeled data set, as in the standard form of supervised learning used in most OCR engines, an active learning system interactively queries the annotator about the instances it finds *most difficult* in some sense. The system prompts the annotator to select an appropriate label and/or provide input as to the example that most closely approximates the one the computer system is attempting to identify. This allows the system to learn more effectively from the human in the "human computing" loop by creating an opportunity for the human to intervene in the data assessment by improving the analytical function of the algorithm itself, not just correcting the individual data point. In traditional crowdsourced transcription applications, a user might spend ten minutes correcting an article, and the product of that labor would be a correct transcription of the article. In an application based on active learning or a similarly iterative approach, on the other hand, the user could identify dozens or hundreds of difficult characters that appear in the articles from that same time period, and the system would use this new knowledge to improve OCR across the entire corpus.

A portion of our funded efforts will focus on the need to incentivize engagement in tasks of this type, whether they are traditionally crowdsourced or engaged through a more active, iterative process like the one we propose. We intend, as we develop an OCR correction system designed to create opportunities for users to improve the system and not just the text itself, to examine how explorations of a users' preferences can improve their engagement with corpora of materials. While our hope is to decrease the number of errors as the algorithm improves, initial efforts (like most OCR engines) will rely on a large amount of annotator intervention, and potential volunteers will be reluctant to spend significant time correcting errors on documents they are uninterested in. We will explore how analyzing each user's preferences for certain types of documents and suggesting individually tailored queues of relevant materials to annotate may help to create a more engaged and dedicated user community.

We propose a proof-of-concept application that will experiment with the use of active learning and other iterative techniques for the correction of eighteenth-century texts (primarily but not exclusively written in English) provided by the HathiTrust Digital Library and the 2,231 ECCO text transcriptions released into the public domain by Gale and distributed by the Text Creation Partnership (TCP) and 18thConnect. Through 18thConnect, a scholarly organization focusing on the "long 18th century" in literary scholarship, we will have access to a number of users who would form a potential pool of annotators and evaluators of our proof-of-concept application. We do not see the focus on this period or language as being required for the technical success of the approach, but rather as a way to take advantage of existing resources and collaborations in order to tackle more directly the interesting problems involved in developing these iterative algorithms and the software to implement them. Once the system is developed it could be applied without substantial adaptation to other languages (at least languages supported by the underlying OCR engines) and historical periods. To demonstrate this capability we will test the system on a small multilingual set of early twentieth-century printed materials (including pages digitized by the Whitman Archive) in the final stages of the project.

Environmental scan: Our proposed **Active OCR** system would be an experiment in active learning for crowdsourced transcription correction. The effectiveness of such a system is described in Abdulkader and Casey (2009). They argue that although the mean word level error rates for currently functioning OCR rates is between 1 and 10 percent (a seemingly low number), the effect of these rates is prohibitive to information retrieval that has become fundamental to modern research. Using an error estimation algorithm, they run the same document through multiple OCR engines and estimate the error value. The higher the level of error, the greater the need for intervention by human annotators. After clustering the results with those of similar rates, human annotators are asked, from a set of images of individual words taken in context, to select the OCR engine whose algorithm functioned nearest to their own estimation of the word or input their own text. These results are then used to produce the corrected OCR text. We are not aware of any publicly available or demonstrated implementations of this approach—i.e., that uses the input of human annotators to improve the OCR system itself, not just the OCR-transcribed text. None of the existing transcription correction systems that we are aware of (including 18thConnect's TypeWright) currently use volunteer contributions as training material, and while OCR engines and toolkits such as Tesseract and Gamera offer OCR training interfaces, these are typically not interactive in the way we describe, in that they require the user to select the examples that are used for training, and they are not collaborative. Furthermore, they do not attempt to identify users' preferences in order to suggest additional annotator interactions.

History and duration of the project: MITH has a rich history of research and development on projects involving facsimile images and textual transcriptions, including archives such as the NEH-funded Shakespeare Quarto Archive and tools such as the Text-Image Linking Environment (TILE), a web-based application for creating and editing image-based electronic editions and digital archives of humanities texts (also funded by the NEH). The application we propose will take full advantage of this work and expertise. The project also builds on work done by Travis Brown with Matt Cohen at the University of Texas at Austin on projects for the Walt Whitman Archive. In this work, which Brown and Cohen outlined in a paper presented at the 2011 conference of the Society for Textual Scholarship, Brown used the training capabilities provided by Tesseract along with error estimation techniques to improve OCR output for two large corpora being digitized, transcribed, and encoded by the Whitman Archive. These projects demonstrated the effectiveness of an iterative training-transcription workflow, and the application proposed here will streamline this workflow and make the selection of materials for human consideration more focused and efficient. The project also builds directly on the open-source TypeWright system for collaborative OCR correction developed by Performant Software for our partners at 18thConnect, and we intend to re-use elements of the TypeWright design and code, in addition to drawing on the expertise of Laura Mandell, the director of 18thConnect. Finally, the focus on the collaborative curation of texts is closely aligned with the goals of Project Bamboo, an international partnership of ten universities, including the University of Maryland. Over the last year MITH has hosted three workshops on the design of corpora-focused functionality in the next phase of Project Bamboo, and the collections and tasks described in this proposal—as well as the TypeWright application specifically—have played an important role in this planning process. For example, the Woodchipper application for text visualization and exploration developed at the first of these workshops operates on HathiTrust and TCP-ECCO texts from the eighteenth and nineteenth centuries, many of which have almost unusable high rates of OCR errors. While that project was focused on exploration, not correction or other forms of curation, we discovered that Latent Dirichlet Allocation topic modeling provided a potentially useful means of characterizing and identifying classes of OCR errors. This kind of interplay between corpus curation and exploration is a core focus of Project Bamboo, and the proof-of-concept application proposed here (and more generally the approach we will explore) could potentially be extended and supported in the future as a part of that project.

Work plan: The first stage of development (Months 1-4) will involve the creation of a web-based user interface for editing character box data. A *character box* in this context is simply a pair of coordinates that identifies a rectangular subset of an image containing a single character of the transcription. Different OCR programs store this data in different formats, but these are generally equivalent and easy to map to each other. For example, if the first character on a facsimile page is "T", the character box file produced by Tesseract may contain a line like the following:

```
T 124 1298 135 1320 0
```

The Tesseract format uses a coordinate system with the origin at the bottom-left corner of the page, so this line indicates that the bounding box containing this character has (124, 1298) as its bottom-left corner and extends to (135, 1320) at its upper right.

There are many open source editors—including several web-based editors, such as a PHP tool provided by Tesseract—that allow character box data files to be edited in a graphical user interface, so that the user can drag box outlines to fit the character in the facsimile image, instead of being required to enter coordinates manually. These tools also generally support common operations such as splitting or merging character boxes (which is necessary, for example, when the OCR system incorrectly recognizes the characters “rn” as “m”). The TypeWright tool also supports word-level bounding box annotation in a collaborative setting.

The development in this first stage will involve selecting and adapting one or more of these existing open source tools that will provide the functionality we need and be compatible with the rest of our software platform. We anticipate that a full-time developer could accomplish this work in approximately two weeks. This work will be shared by Travis Brown (the lead developer) and the research assistant, with the precise division of duties depending on the expertise of the research assistant. All software will be documented by the lead developer and research assistant, and will be released (early and frequently) to the public through a GitHub repository under Version 2.0 of the Apache License.

The next stage of development (Months 5-7) will involve the creation of several key back-end components, including primarily uniform interfaces to several open source OCR engines that will allow us to gather additional information about their output, such as the value associated with the system’s confidence about each character. In the case of the Tesseract engine, Brown has already done much of this work in projects for the Whitman Archive. Another component will support error estimation, and may be based in part on work done for MITH’s Woodchipper project. These components will potentially operate in tandem, with the following high-level description representing one possible workflow:

1. The error estimator is used to identify the “worst” documents in the sub-corpus of interest to the current user.
2. Multiple OCR engines are used to re-transcribe these particularly erroneous documents, with points of disagreement or low confidence being highlighted as needing prioritization for annotator attention.

Related back-end error analysis services could also be developed during this stage. Because this work involves the development of several components that do not directly depend on each other, it will also be shared between the lead developer and the research assistant, and will be completed by the end of the semester during which the research assistant is employed.

The final stage of development (Months 8-10) will complete the human-computer loop, allowing the corrections of particularly challenging materials by human annotators to be used by the underlying OCR systems as training material. While this is a relatively simple extension to the interface in the case of Tesseract, for example, the challenge will be making this kind of iterative training computationally feasible in a multi-user collaborative environment. We imagine that the largest part of the lead developer’s investment of time will be committed to this stage of development.

The final stage of the project will involve testing and evaluation (Months 10-12). Volunteers will be invited primarily from the 18th Connect community, and all corrections and other data generated by users will be anonymized and made publicly available in a micro-format similar to that used by Tesseract. Evaluation will be managed by the lead developer and will focus on comparing the transcriptions generated by contributors and the system against traditionally corrected texts, using standard metrics for OCR evaluation.

Staff: The proposed project is fortunate to benefit from a variety of substantial relationships at the University of Maryland and MITH. Core project staff will include: Travis Brown, lead Research and Development Software Developer at MITH, who will lead the project and supervise all project activities and Mr. Trevor Muñoz, Assistant Dean of the University of Maryland Libraries and Associate Director of MITH, will be responsible for data management and curation throughout the lifecycle of the project.

Final product and dissemination: Our project will release all of its work as open-source code thereby encouraging other researchers to use our work, benefit from our investment of resources, and alter our code to extend the efficacy of Active OCR. We will release a reflective white paper at the end of the grant detailing the challenges and successes of experimenting with machine learning in the digital humanities focusing on OCR. As part of this white paper, we will also gather and archive volunteer contributions on our project website. These, along with archived social media streams likes blogs and twitter, will form a useful thread throughout our final reporting efforts as we attempt to extrapolate potential understanding of how active learning can challenge the current approach to OCR. These extrapolations will form a core set of recommended best practices for other developers wishing to join us in further development.