NEH Digital Implementation Grant

Reading the First Books: Multilingual, Early-Modern Optical Character Recognition for Primeros Libros

3. Abstract

Digital facsimile collections of early modern printed books (books printed on hand presses in the 15th-17th century) greatly improve access to these cultural heritage materials for scholars, students, and the general public. The utility and accessibility of these digital collections, however, has been limited by the challenges of transcribing early modern printed books: linguistic complexity, unstable orthography (spelling and punctuation), and uneven typesetting and inking make these books difficult to read for humans and machines alike. The goal of this project is to develop and implement groundbreaking methods in the automatic transcription of early modern printed books. This will increase access to books that are not just a vital record of this exciting period in European, colonial, and indigenous American history, but also reflect the development of a new, transformative technology – the printing press.

To address this challenge, we have developed an Optical Character Recognition (OCR) prototype that extends Ocular, an innovative system developed by Taylor Berg-Kirkpatrick et al in 2013. Ocular expands on pre-existing OCR systems by modeling the behavior of a hand press, including uneven printing and inking. Our prototype, which we will call Ocular+, extends Ocular by incorporating multilingual capabilities and tools for handling variable orthographies, like those of the newly developed indigenous orthographies of Mexico and Peru. Developed in the Fall of 2014, Ocular+ uses a trilingual model for Spanish, Latin, and Nahuatl (the dominant indigenous language of central Mexico). When tested on sample documents from a corpus of 16th century Mexican printed books, we achieved a significant improvement over the current state-of-the-art (see table 2.1 in the appendices).

With funds from the NEH Implementation Grant, we will implement this new digital tool for use in humanities research. In doing so, we will achieve three products of significant use for humanities scholars: an open-access digital tool for producing transcriptions, a statement of best practices for early modern automatic transcription, and new digital corpora. To produce the open-access digital tool, we will collaborate with the Initiative for Digital Humanities, Media, and Culture at Texas A&M University to incorporate Ocular+ into their Early Modern OCR Project (eMOP) workflow. eMOP (see appendix 3) is a Mellon Foundation grant-funded project to create an open source, automatic transcription workflow for early modern printed documents. eMOP, which has been used to transcribe 45 million pages, leverages and produces cutting-edge tools for analyzing facsimiles of texts printed in the 15th-18th century. Our collaboration will significantly increase the effectiveness of eMOP's transcription system by incorporating a new OCR engine into the workflow and by expanding its post-processing systems to include multi-lingual texts. This tool will open the way for new collections, particularly those that extend digital scholarship beyond monolingual corpora.

We will produce new digital corpora by implementing the new eMOP workflow for the *Primeros Libros* project, an effort to digitize all surviving exemplars of all books printed in the Americas before 1601, founded in 2009 as a collaborative international enterprise involving both The University of Texas at Austin and Texas A&M University. The collection serves as a unique reflection of the range of textual production in early colonial America by both European and indigenous intellectuals writing in languages from Spanish and Latin to Nahuatl, Mixtec, and Aymara. This includes, for example, the only Nahuatl

grammar from the period written by a native speaker of the language. Through the transcription process, we will produce a description of 'best practices for automatic transcription' based on conversations with scholars and other potential users. The end result will be newly transcribed corpora which will create new possibilities for scholarship and increase discoverability for users such as the approximately 1.5 million Nahuatl speakers living in Mexico and the United States today.

Statement of Innovation

Recent developments in OCR have produced tools that can transcribe documents printed in the early modern period. However, these tools are unable to handle documents that switch between languages, that are written in understudied languages, or that contain the obsolete spellings common in texts from this era. This project introduces new technical advancements drawn from our machine learning and natural language processing research that significantly improve transcription accuracy on these texts.

Statement of Humanities Significance

The transcription of documents printed in the early modern period unlocks these records for digital analysis, from simple searches to computational analysis. Our tools will expand the global reach of digital scholarship by significantly improving the automatic transcription of multilingual historical books. Our transcription of the *Primeros Libros* collection of American printed books will enable new scholarship on the first contact between indigenous and Spanish communities.