Tracking Reader Annotations in Printed Books by Collating and Transcribing Multiple Exemplars

Abstract

Most past digitization projects have focused on transcribing documents individually. With the availability of library-scale digital collections, we propose a Digital Humanities Advancement Grant (Level II) to develop computational image and language models to discover multiple copies and editions of similar texts and to correct each text using these comparable witnesses. We provide evidence that this collational transcription system can significantly improve optical character recognition on historical books. We also propose to use these collated editions to discover annotated passages in large digitized book collections. This approach will therefore not only mitigate the errors that reader annotations introduce into the OCR process but will also produce the first automatically generated database of handwritten annotations, Ichneumon. Methods and software developed by this project will thus benefit future research on automatic collation, book history, and historical reading practices.

1 Contents

2 List of Participants

- David Smith, PI; Northeastern University

3 Narrative

We propose to develop and evaluate tools and corpora for "OCR collation"—detecting and taking advantage of multiple copies and editions of books to improve transcription accuracy. We further propose to use these collations and improved transcriptions to train models to detect the presence and location of handwritten annotations in printed books. This project will therefore be complementary to existing work in bibliography, book history, historically-focused OCR, and advanced imaging techniques that seek to characterize and transcribe historical printed materials and to reconstruct historical reading practices.

We start by describing how manuscript annotations interact with current OCR systems, how redundancy in digitized collections can help improve transcription accuracy in texts with and without annotations, and how we can use this redundancy to train systems to detect manual annotations on page images (§3.1). We then connect our proposed work to existing research and systems for cataloguing historical annotations and for historical OCR (§3.2). We describe our work so far on detecting and aligning multiple editions (§3.3) and lay out the work plan for this eighteen-month project (§3.4). We close with a description of the software and datasets we propose to publish (§3.5).

3.1 Enhancing the humanities

The Boston Public Library holds nearly 3000 volumes collected by John Adams, the second president of the United States. These books, the cataloguers note, contain "significant annotations in Adams's hand". His copy of Rousseau's *Nouvelle Héloïse*, photographed and OCR'd by the Internet Archive, shows several annotations on v. 2, p. 14 (figure 1a). When Rousseau rails against those who seek to rule society by their "birth" or "riches", Adams inserts a caret and adds, in French in the margin, "or by beauty of face or figure". When, further, Rousseau says that aristocrats and plutocrats ought to be called out or punished (*décrier ou punir*), Adams notes (in English) that it is society itself that institutes aristocracy or plutocracy: "Peoples, Nations, not Individuals, are guilty of this. Riches and fame are Chimeras too."

The cataloguing and transcription of these sorts of annotations have provided a wealth of evidence to scholars of historical reading practices and other intellectual work, as we discuss below in the Environmental Scan (§3.2). These annotations are especially valuable when we have external documentary and internal paleographic evidence to link annotations to particular readers, as here, or particular time periods.

First, however, we focus on how annotations affect the automatic transcription of the printed text they accompany. Figure 1b, for example, shows the output of the ABBYY FineReader OCR system used by the Internet Archive on Adams's copy. The marginal annotations cause errors in detecting printed text in page images and, once included, add the noise of OCR'd handwriting to the output. This OCR output also shows a common case of mismatch between OCR models and data: the mistranscription of long s (ſ) as *f*.

In contrast, the same OCR system performs much better on a cleaner copy of the same 1764 Duchesne edition from the Thomas Fisher Rare Book Library at the University of Toronto (figure 1c). Both copies show the broken *d* in *décrier* and the spot over the second *a* in *naissance*. The OCR output does not, of course, show the noisy text from attempting to transcribe handwriting, but the overall base text is also cleaner (figure 1d). The advantage is not all on one side: the OCR of the Boston copy correctly transcribes the *c* in *justice*, where the Toronto copy reads *e*.

These examples illustrate several problems this project seeks to address. Specifically, we propose to:

- **detect and align** multiple OCR'd copies of the same edition (*manifestation* in FRBR terminology) and copies of different editions (*expressions*) of books to allow comparison of both OCR transcriptions and of manuscript modifications after a physical copy has been printed;

- **infer a consensus transcription** for each edition of each work;
- **align images from multiple copies of the same edition** to detect manuscript annotations and produce training data for annotation-detection models; and
- **produce a database of annotated passages** by applying annotation-detection models to both editions with multiple copies and with single copies.

As concrete deliverables, this project will produce both data—alignments of editions, improved OCR transcriptions, a database of manuscript annotations in books—and open-source software—for collating OCR output and for detecting manuscript annotations.

The most immediate benefit of this project for the humanities will be improved OCR transcripts of the books we align and improved models for OCR transcription and post-correction. Higher accuracy at OCR transcripts will improve downstream applications such as information retrieval and extraction, stylometric and sentiment classification, and text-reuse and linguistic analysis. We expect that these improvements will have greater impact on older books, where current font models perform less well, and on books with significant manuscript annotations, where comparison with different exemplars or modeling the location of handwritten marks, will facilitate more accurate OCR of the underlying printed text.

Beyond simply improving OCR accuracy, however, this project ultimately aims to build models for detecting, and databases for documenting, the prevalence and locations of manuscript annotations in print archives. As discussed in §3.2, many projects have set out to catalogue handwritten annotations, either in the libraries of a small group of readers (Jardine and Grafton 1990, Havens 2016, Chenoweth et al. 2018) or for a large number of copies of particular editions (Gingerich 2002, Palmer 2014). We expect this project to complement these efforts by compiling an extensive database of handwritten annotations across many editions and time periods.

3.1.1 Technical approach

Aligning comparable text. The first step in the proposed approach outlined above is to align pages from multiple copies of the same edition. How prevalent is such duplication in the digitized archives available to us? To answer this question, and to conduct preliminary experiments on the effectiveness of our proposed OCR collation techniques, we aligned 934 manually transcribed books from 1500–1800 from the Text Creation Partnership (University of Michigan Library 2018) to the OCR of three million public-domain books from the Internet Archive (Internet Archive 2018). To perform these alignments between noisy OCR transcripts efficiently, we used methods from our earlier work on text-reuse analysis (Smith et al. 2014, Wilkerson et al. 2015, Smith et al. 2015). An inverted index of hashes of word 5-grams was produced, and then all pairs from different pages in the same posting list were extracted. Pairs of pages with more than five shared hashed 5-grams were aligned with the Smith-Waterman algorithm with equal costs for insertion, deletion, and substitution, which returns a maximally aligned subsequence in each pair of pages (Smith and Waterman 1981). Aligned passages that were at least five lines long in the target TCP text were output.

For each target OCR line, there are thus, in addition to the ground-truth manual transcript, zero or more **witnesses** from similar texts, to use the term from textual criticism. Figure 2 shows an example line of text with the manual TCP and automatic OCR transcripts from the same (2b) and different (2c) editions. After this alignment process, we have 8.6 million pairs of lines from matched pages in the TCP ground-truth and Internet Archive books. Of these OCR'd lines in digitized editions, 5.5 million lines, or 64%, also align to an OCR'd line from *another* digitized edition in the Internet Archive. This analysis suggests that, at least for pre-1800 English books, we are likely to find a significant number of editions with multiple digitized copies.

Inferring consensus transcriptions. The next step in our proposed approach is to collate the OCR transcripts of different copies of books. We propose to use the sets of witnesses produced by the alignment approach above both to do **unsupervised training** of OCR correction models and to **collate witnesses** to infer a consensus transcription for each given copy of a text. As we saw above, in the early modern books that aligned to TCP texts, 64% of the lines had multiple witnesses. Our approach is therefore designed to correct OCR both in the presence of other witnesses and when none are available. We split the OCR output from the Internet Archive that was aligned to TCP books into training and test sets. We train character-level bidirectional recurrent neural network (RNN) sequence-to-sequence models with attention (Bahdanau et al. 2015) in both supervised and unsupervised conditions. For the supervised case, we used the manual TCP transcript as the correct output during training. For the unsupervised case, we took clean text and randomly applied insertion, deletion, and substitution edits to 10% of the characters, matching the average error rate.

Both the supervised and unsupervised sequence-to-sequence models can be applied unchanged to correct single OCR outputs. For lines with other witnesses, we extend the sequence-to-sequence model with attention by averaging the input representation at each decoding time step, as we proposed in earlier work (Dong and Smith 2018). This allows collation to scale linearly with the number of witnesses, instead of quadratically (as with comparable models for multi-document summarization) or exponentially (as with exact multiple sequence alignment approaches to collation). In experiments with the test set (described above) of aligned TCP books, the character error rate (CER) of the Internet Archive's ABBYY FineReader OCR was 10.7%, with a word error rate (WER) of 31.7%. On a test set of lines with both single witness and multiple witnesses (where available), the supervised model achieved 4.7% CER and 11.2% WER; the unsupervised model achieved 6.8% CER and 20.3% WER. These are the top outputs from our decoder, but when we use beam search to find the 1000-best hypotheses, the oracle error rates are about half of the top numbers. Additionally, we find that many of the errors of the unsupervised correction system result from the synthetic training data not modeling the distribution of short and long *s* (ſ) in the ground-truth data. We propose to close the gap between these current results and the lower bound of the oracle scores first by modeling these variant characters and by rescoring the 1000-best hypotheses by discriminative language models.

Aligning comparable images. For pairs of copies from the same edition, we will not only align their OCR transcripts, as described above, but also align the images of their pages. Forced alignment of page images has been used in OCR to incorporate different image preprocessing pipelines (Lund et al. 2013) or to correct for bleedthrough (Wang and Tan 2001). In this project, we can take advantage of the fact that each page has already been analyzed and transcribed. We can thus use the alignment of the transcripts, described above, to constrain the alignment of the underlying images and make it more efficient. We can then subtract from a given binarized image the pixels that are also present in the corresponding positions of other images. After correcting for small-scale noise, due to edge effects, dust on the page, etc., we can run edge-detection algorithms on the residual image in order to label regions as likely handwritten text or other annotations.

Detecting manual annotations. Regions marked as annotations by the image alignment above will then form a training and test set for detecting annotations in non-aligned images. Recent work on page layout analysis has taken advantage of general progress in image annotation and segmentation, such as fully convolutional networks (Long et al. 2015) and other neural network approaches. In preliminary experiments with page segmentation, we have found that sparser architectures such as SegNet (Badrinarayanan et al. 2017) are both more effective and more efficient. Although we are not, for this project, proposing to transcribe handwritten text, annotations in printed books have the advantage of being placed in juxtaposition to printed text that we can transcribe with OCR. We can thus take advantage of the single- and multi-input collation and correction techniques described above to produce more accuracy OCR of the printed "anchor text" of handwritten annotations. For both multi-copy editions, by image alignment, and for single-copy editions,

by image classification, we can then compile a database of the locations and printed text associated with annotations in public-domain Internet Archive books. This database, Ichneumon, is designed to complement in-depth studies of annotation performed by scholars in book history and related fields. We have also allotted time to perform exploratory data analysis of the books and passages most likely to receive annotation.

3.2 Environmental scan

The past few decades have seen many scholars from book history, literary studies, the history of science, and other fields focus on the interaction of readers (and writers) with printed texts (Snook 2013). To touch on only a few themes, some studies make historical arguments about reading and writing (Sherman 2008) or education and information culture (Blair 2010). Others have focused on the practices of individual readers, from, e.g., Gabriel Harvey and John Dee in the sixteenth century (Jardine and Grafton 1990, Havens 2016) to Jacques Derrida in the twentieth (Chenoweth et al. 2018). Still others have made bibliographic studies of the annotations in a large proportion of the copies of particular books, such as Copernicus's *De revolutionibus* (Gingerich 2002) or Lucretius's *De rerum natura* (Palmer 2014). While Bourne (2017), among others, points out the limitations of using, e.g., readers' marks in Shakespeare's First Folio in isolation, marks of known or unknown readers' interactions with texts form the starting point for many literary and historical studies.

The digitization of libraries of historical printed books has led to complementary efforts to catalogue and transcribe the annotations of particular readers—in projects such as the Archaeology of Reading (Havens 2016; figure 3 below) or Derrida's Margins (Chenoweth et al. 2018)—or annotations in particular genres, such as early modern drama (Munson 2016; figure 4 below). Some library users have even crowdsourced (via Book Traces) the indexing of handwritten annotations (figure 5). Our proposed project seeks to support these and similar efforts by developing tools and datasets for detecting and analyzing annotations. This project could also take advantage of these databases of annotated page images, where available, to evaluate and improve page-analysis models.

Mass digitization projects by government (e.g., the Library of Congress, Bibliothèque nationale de France), commercial (e.g., Google Books, GALE), and nonprofit (e.g., the Internet Archive, JSTOR) entities have made available large amounts of historical print materials, but these projects often use methods optimized for modern print. Several research projects such as IMPACT and OCR-D in Europe and the Early Modern OCR Project (eMOP) in the U.S. (Heil and Samuelson 2013, Gupta et al. 2015) have worked on improving the accuracy of OCR for early modern printed materials. In addition to supervised and semi-supervised training procedures for early modern typefaces, these projects have employed alignment of outputs from several input OCR systems. These previous approaches to alignment for OCR correction have used small numbers of witnesses that could be aligned exactly and voting methods that only worked when the correct output existed in at least one of the input noisy OCR transcripts (Lund and Ringger 2009, Wemhoener et al. 2013, Xu and Smith 2017). Most other common approaches to OCR post-correction used hand-labeled training data and only work to correct single inputs, not collated sets of witnesses (Kolak et al. 2003, Boschetti et al. 2009, Lund et al. 2011, Al Azawi et al. 2015). For detecting manual annotations, we propose to use convolutional networks. Unlike most OCR page-layout modules, our classification task is to find annotations both outside and inside of existing text regions. Alternative approaches that use recurrent networks to model dependence among page regions may thus be less effective (Breuel 2017).

3.3 History of the project

Our proposed work on aligning OCR transcripts of the same and comparable editions of books grows out of our earlier work on the NEH- and Mellon-funded Viral Texts and Proteus projects, which built open-source

tools for text-reuse detection and analysis. The work on text reuse in books for the Proteus project also allowed us to download three million public-domain books from the Internet Archive, which will ensure that this project will not need to start with a lengthy data-collection phase. In addition to publications on our work on text-reuse (Smith et al. 2014, Wilkerson et al. 2015, Smith et al. 2015), these projects also led to publications on inferring information propagation networks (Xu and Smith 2018) and collating aligned texts (Xu and Smith 2017). Our recently completed Mellon-funded project to conduct a survey and write a report on a research agenda for historical and multilingual OCR led to further improvements in learning and decoding algorithms for OCR collation (Dong and Smith 2018).

After the completion of this proposed eighteen-month project, we plan to use the collation of multiple copies of each addition to make direct improvements in OCR models. By exploiting the alignment of both manual and automatically produced consensus transcripts with page images, we hope to greatly expand the amount of training data available for many typefaces and scripts. This approach has become more realistic due to state-of-the-art OCR systems training on whole lines instead of isolated words or characters. In addition, we hope to use information about the location of handwritten annotations to improve general page-layout analysis models. Both of these tasks are of general interest to the OCR community (Breuel 2017; e.g.) and could receive NSF support. Starting from our database of manual annotations in printed books, we hope to forge collaborations with researchers in book history, literary studies, and the history of science who could help build models of information propagation and information seeking behaviors in books and tools for scholars of these phenomena. These collaborations, we hope, could receive further support from the NEH, the Mellon foundation, and others.

3.4 Work plan

We propose an eighteen-month project to develop and evaluate tools and corpora for OCR collation and annotation detection.

3.4.1 Tasks

Months 1–4 Align public-domain books from the Internet Archive on the basis of their one-best OCR transcripts using text-reuse analysis software developed during the Viral Texts and Proteus projects (Smith et al. 2014, Wilkerson et al. 2015, Smith et al. 2015). Pairs of books with matching page and line breaks are inferred to be copies of the same edition. Pairs of books with different pagination/lineation are inferred to be different editions. Where available, manual transcriptions of books (e.g., from the Text Creation Partnership) are also aligned for evaluating OCR accuracy. (Smith and RA)

Months 3–8 Collect sets of witnesses by aligning OCR transcripts from identical and similar editions. These witness sets are used to train unsupervised post-correction models. The approach developed by Dong and Smith (2018) is used as a baseline to collect candidates for possible corrections. We experiment with different constraints on the decoder and with estimating different language models to rerank the candidates. (Smith and RA)

Months 5–12 Align pairs of page images from matching editions at the pixel level to tag image regions belonging to the shared printed text or subsequently-added annotations. Character locations in first-pass OCR will be used to prune the search space. This will create a large training set for classifying other page images in the corpus. (Smith and RA)

Months 9–16 Train classifiers to detect handwritten annotations on page images, including manuscript text, underlining, and other marginal and interlinear marks, using the aligned page images as training and

evaluation data. (Smith and RA)

Months 13–17 Exploratory data analysis of annotation anchor text, building predictive models of which passages, in aggregate, are annotated. (Smith and RA)

Months 13–18 Compile and release Ichneumon database of annotated passages in public-domain books. (Smith and RA)

3.4.2 Risks

One class of risks relates to our ability to collect sufficient training data for annotation detection. If necessary, we could improve perfomance by manually marking page images for annotated regions. We could either produce these annotations from scratch or correct the output of initial models for detecting annotations. Another class of risks relates to the computational cost of aligning OCR transcripts and images from large numbers of books and of running annotation detection on a corpus of page images. To maximize the benefit of this project, we will prioritize books from before 1800, where the accuracy of baseline OCR systems is already very low and where manuscript annotations would be of greatest historical interest. As time permits, we would also include later books.

3.4.3 Evaluation

An important component of this project is using the existence of multiple scanned copies of the same editions of books to produce ground-truth for OCR correction and annotation detection. By aligning manual transcriptions of particular editions, such as those created by the Text Creation Partnership, to OCR output, we will create significant amounts of training and test data for OCR. By aligning images from different copies of the same edition, we will create ground-truth data for classifying image regions as underlying print or post-print markings. In addition, we hope to take advantage of existing annotation cataloguing databases referenced above (§3.2) to further evaluate our approach.

3.5 Final product and dissemination

All software source code written for this project will be released under an open-source license. In addition to keeping source code in a Northeastern-hosted git repository during development and after the completion of the project, the code will be stored in a public repository on GitHub for backup and dissemination. This software includes text-reuse analysis and anlignment software, whose development started with the Viral Texts and Proteus Projects, as well as code for collating and rescoring multiple OCR transcripts, for aligning multiple page images, and for training and deploying annotation detection models.

By the end of this project, we will release three datasets: a corpus of the identifiers of aligned page images from Internet Archive books, along with coordinate information for handwritten annotations discovered by the alignment process; the output of our collational OCR correction model for aligned text passages, to provide more accurate transcripts of the text; and a database, Ichneumon, of handwritten annotations detected in printed books, both from the aligned pages described above and from applying the annotation-detection model to unaligned pages. Since all of these data are derived from public-domain page images and OCR transcripts from the Internet Archive, we will release all of these datasets under an open license. The datasets will be hosted on Northeastern servers, in the University library's institutional repository. We will also discuss contributing improved OCR transcripts and other data back to the Internet Archive. Likely venues for many results arising from this and later phases of this project are hosted by the Association for Computational Linguistics, which makes all publications freely available.

4 Biographies

David Smith is an assistant professor in Northeastern University's College of Computer and Information Science. He is a founding member of the NULab for Texts, Maps, and Networks, Northeastern's center for research in the digital humanities and computational social sciences. He holds a Ph.D. in computer science from Johns Hopkins and an A.B. in Classics (Greek) from Harvard. In between, he worked as a research programmer for the Perseus Digital Library Project.

5 Project budget

6 Appendices

6.1 Bibliography

Mayce Al Azawi, Marcus Liwicki, and Thomas M. Breuel. Combination of multiple aligned recognition outputs using WFST and LSTM. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 31–35, 2015.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Ann Blair. *Too Much to Know: Managing Scholarly Information before the Modern Age*. Yale University Press, 2010.

Federico Boschetti, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane. Improving OCR accuracy for classical critical editions. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 156–167, 2009.

Claire M. L. Bourne. Marking Shakespeare. *Shakespeare*, 13(4):367–386, 2017.

Thomas M. Breuel. Robust, simple page segmentation using hybrid convolutional MDLSTM networks. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 733–740, 2017.

Katie Chenoweth, Alexander Baron-Raiffe, and Rebecca Sutton Koeser. Derrida's margins, version 1.0.0. `derridas-margins.princeton.edu`, 2018. Accessed 1 June 2018.

Rui Dong and David A. Smith. Multi-input attention for unsupervised OCR correction. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2018.

Owen Gingerich. *An Annotated Census of Copernicus'* De revolutionibus *(Nuremberg, 1543 and Basel, 1566)*. Brill, 2002.

Anshul Gupta, Ricardo Gutierrez-Osuna, Matthew Christy, Boris Capitanu, Loretta Auvil, Liz Grumbach, Richard Furuta, and Laura Mandell. Automatic assessment of OCR quality in historical documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1735–1741, 2015.

Earle Havens. The archaeology of reading. `archaeologyofreading.org`, 2016. Accessed 1 June 2018.

Jacob Heil and Todd Samuelson. Book history in the Early Modern OCR Project, or, bringing balance to the force. *Journal for Early Modern Cultural Studies*, 13(4):90–103, 2013.

Internet Archive. eBooks and Texts. `archive.org/details/texts`, 2018. Accessed 1 June 2018.

Lisa Jardine and Anthony Grafton. "Studied for action": How Gabriel Harvey read his Livy. *Past and Present*, 129:30–78, 1990.

Okan Kolak, William Byrne, and Philip Resnik. A generative probabilistic OCR model for NLP applications. In *Proceedings of the Conference on Human Language Technology of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 55–62, 2003.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

William B. Lund and Eric K. Ringger. Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 231–240, 2009.

William B. Lund, Daniel D. Walker, and Eric K. Ringger. Progressive alignment and discriminative error correction for multiple OCR engines. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 764–768, 2011.

William B. Lund, Douglas J. Kennard, and Eric K. Ringger. Combining multiple thresholding binarization values to improve OCR output. In *Proc. Document Recognition and Retrieval (DRR)*, 2013.

Rebecca Munson. Common readers. `www.commonreaders.info`, 2016. Accessed 1 June 2018.

Ada Palmer. *Reading Lucretius in the Renaissance*. Harvard University Press, 2014.

William H. Sherman. *Used Books: Marking Readers in Renaissance England*. University of Pennsylvania Press, 2008.

David A. Smith, Ryan Cordell, Elizabeth Maddock Dillon, Nick Stramp, and John Wilkerson. Detecting and modeling local text reuse. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2014.

David A. Smith, Ryan Cordell, and Abigail Mullen. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3), 2015.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

Edith Snook. Recent studies in early modern reading. *English Literary Renaissance*, 43(2):343–378, 2013.

University of Michigan Library. Text Creation Partnership. `www.textcreationpartnership.org`, 2018. Accessed 1 June 2018.

Qian Wang and Chew Lim Tan. Matching of double-sided document images to remove interference. In *CVPR*, 2001.

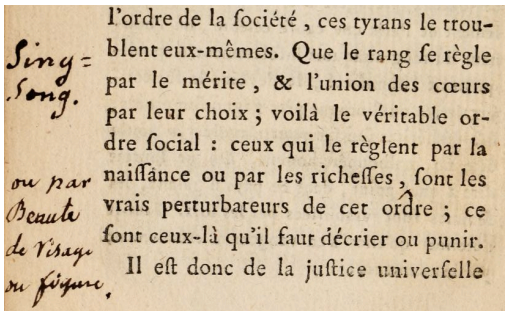
David Wemhoener, Ismet Zeki Yalniz, and R. Manmatha. Creating an improved version using noisy OCR from multiple editions. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 160–164, 2013.

John Wilkerson, David A. Smith, and Nick Stramp. Tracing the flow of policy ideas on legislatures: A text reuse approach. *American Journal of Political Science*, 2015.

Shaobin Xu and David A. Smith. Retrieving and combining repeated passages to improve OCR. In *Proceedings of the ACM+IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017.

Shaobin Xu and David A. Smith. Contrastive training for models of information cascades. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

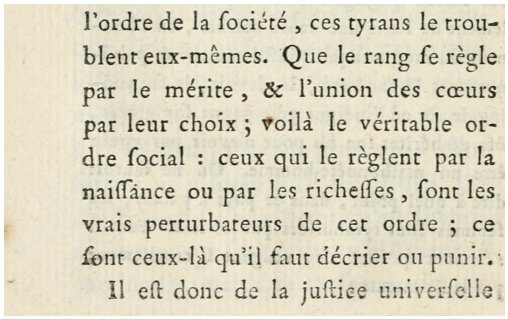
6.2 Example annotations and OCR



(a) John Adams' copy, Boston Pub. Lib.

```
Tordre de la fociété , ces tyrans le trou-
Si^ct^ ^^ei^f ^"x-mèmes. Que le rang fe règle
r^_^ par le mérite , & l'union des cœurs
J par leur choix ; voilà le véritable or-
dre focial : ceux qui le règlent par la '
iTM/ /7a.r naiflTânce ou par les richeffes , font \q&
tuutu ^^^^^ perturbateurs de cet or3re j ce
J Y^ ^onr ceux-là qu'il faut décrier ou punir.
/, II eft donc de la iuftice univerfeile
```

(b) ABBYY FineReader OCR of 1a

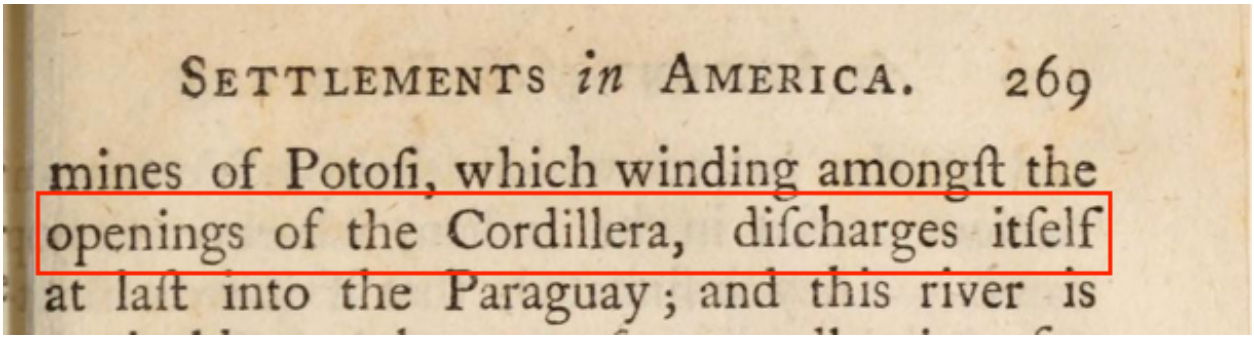


(c) Fisher Rare Book Library, Toronto

```
l'ordre de la fociété , ces tyrans le trou-
blent eux-mêmes. Que le rang fe règle
par le mérite , & l'union des cœurs
par leur choix \ voilà le véritable or-
dre focial : ceux qui le règlent par la
naiffance ou par les richelfes , font les
vrais perturbateurs de cet ordre ; ce
font ceux-là qu'il faut décrier ou punir.-
il eft donc de la juftiee univerfelle.
```

(d) ABBYY FineReader OCR of 1c

Figure 1: Images and OCR of two copies of a 1764 Duchesne edition of Rousseau's *La nouvelle Héloise*



(a) Page image, with bounding box of second line

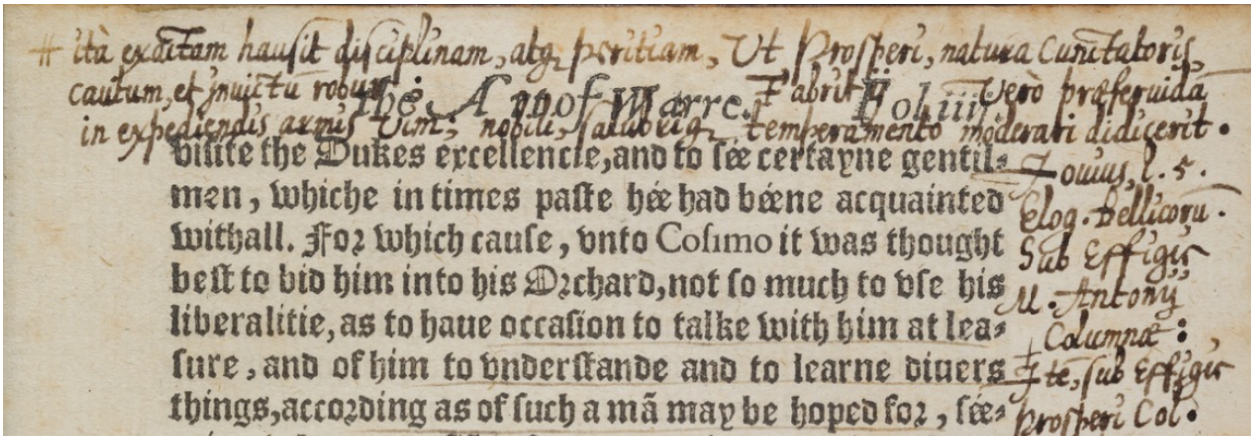
Source	OCR?	Text
TCP K023456.001	N	openings of the Cordillera, diſcharges itſelf
This image	Y	• openings of the Cordillera, difcharges itfelf
cihm_32607	Y	openings of the Cordillera, difcharges itfelf

(b) Three transcripts of this edition: the first is by hand by the Text Creation Partnership; the second is the OCR transcript of the image in 2a; the third is the OCR transcript of a different copy of this edition.

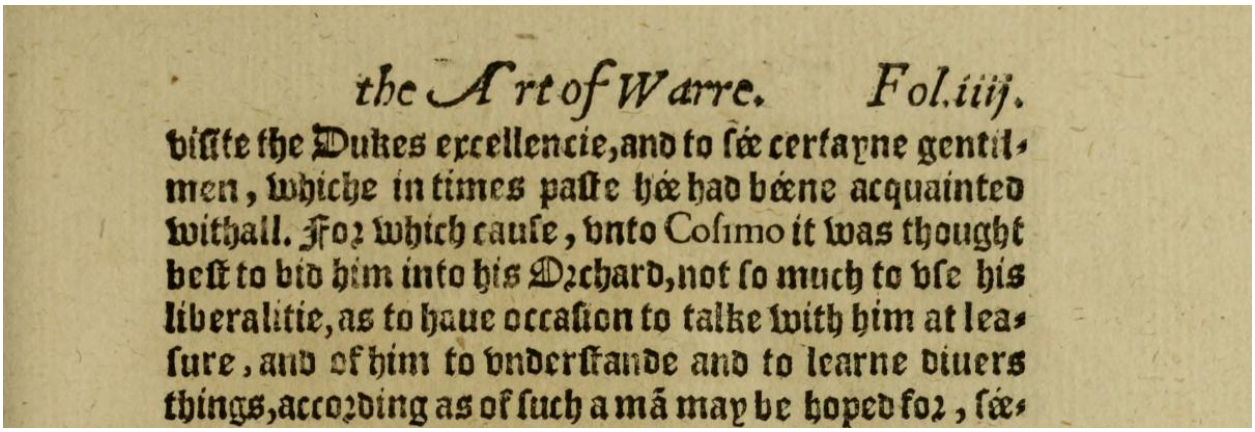
Book ID	Text
accountofeuropea01burk	openings of the\nCordillera, difcharges itfelf
anaccounteuropea01burkiala	openings of the\nCordillera, difcharges itfelf
cihm_32428	openings of the\nCordillera, difcharges itfelf
accountofeuropea001burk	openings of the\nCordillera, difcharges itfelf
cihm_49269	openings of the\n: Cordillera, difcharges itfelf
anaccountofeurop01burkiala	openings of the Cordillera, dif-\ncharges itfelf
cihm_48531	openings of the Cordillera, dif-\nmarges itfelf
cihm_28244	openings of the Cordillera, dif-\n€Jbaige|s itfelf
cihm_54658	openings of the Cordillera, dif-\ncharges itfelf

(c) Nine transcripts of several copies of different editions aligned to the text line in 2a. Note the varying location of the line breaks.

Figure 2: An image from one copy of Edmund and Willam Burke, *An account of the European settlements in America*. London: Printed for R. and J. Dodsley in Pall-Mall, 1757. In addition to the image, the alignment approach described above found one manual (TCP) and two OCR transcripts of the same edition, as well as nine OCR transcripts of different editions. Note that none of the OCR versions correctly transcribes the long s (ſ). Correcting this sort of mismatch in the character inventory is one goal of our proposed approach.

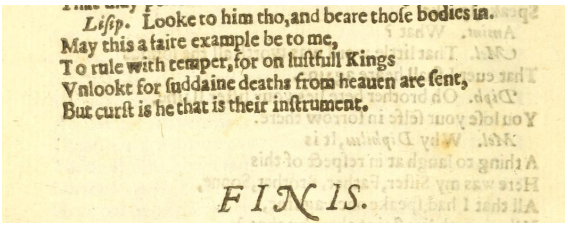


(a) Princeton copy, annotated by Gabriel Harvey

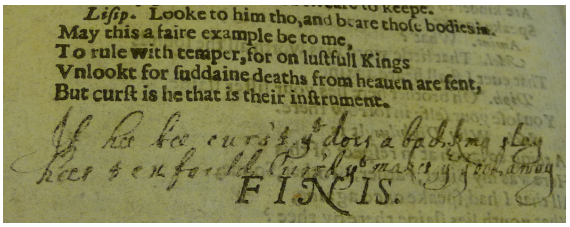


(b) Boston copy

Figure 3: Images from two copies, at Princeton University and the Boston Public Library, of Niccolò Machiavelli, *The arte of warre*. London: W. Williamson, 1573. The Archaeology of Reading project (Havens 2016) transcribed the manuscript annotations by Gabriel Harvey.

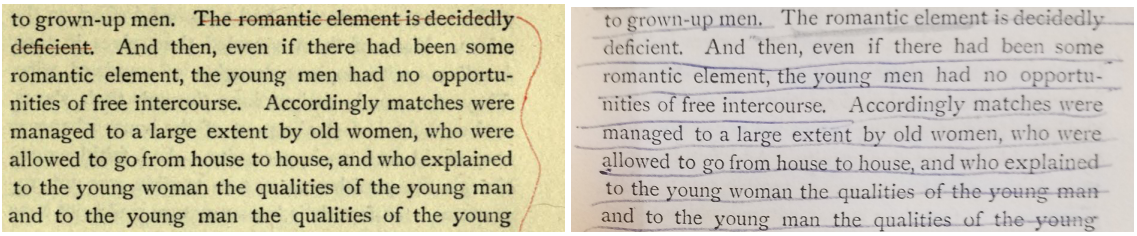


(a) Boston Public Library



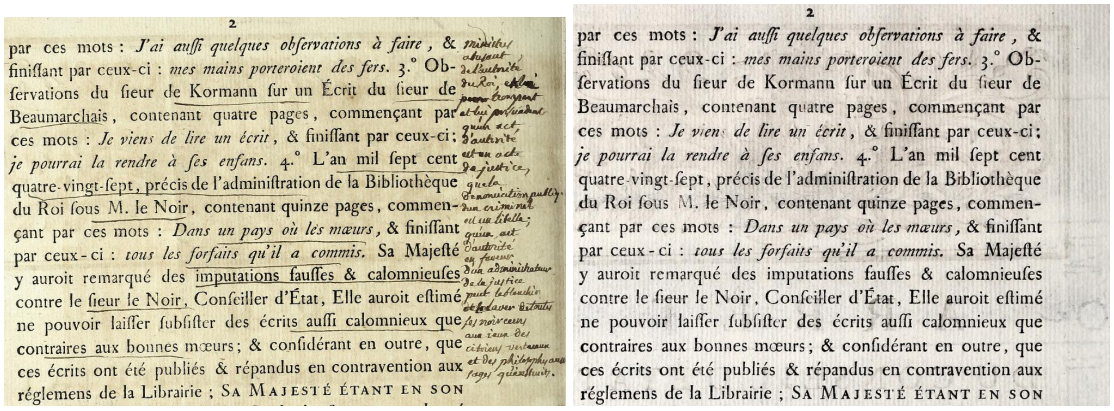
(b) Houghton Library, Harvard University

Figure 4: Two images of the last page of Beaumont and Fletcher's *Maid's Tragedy*. London: A. M. for Richard Hawkins, 1630. The Houghton image is from via Munson (2016).



(a) University of California Libraries (b) Image uploaded to `booktraces.org`

Figure 5: Two images of p. 53 of James Donaldson, *Woman ; her position and influence in ancient Greece and Rome, and among the early Christians*. London: Longmans, Green & Co., 1907.



(a) Annotated copy (b) Unannotated copy

Figure 6: Images from the second page of two Newberry Library copies of *Arrêt du Conseil d'État du roi, pourtant suppression de plusieurs libelles*. Paris: L'Imprimerie royale, 1787.

7 Letters of committment and support

8 Data management plan

All software source code written for this project will be released under an open-source license. In addition to keeping source code in a Northeastern-hosted git repository during development and after the end of the project, the code will be stored in a public repository on GitHub for backup and dissemination. This software includes text-reuse analysis and anlignment software, whose development started with the Viral Texts and Proteus Projects, as well as code for collating and rescoring multiple OCR transcripts, for aligning multiple page images, and for training and deploying annotation detection models.

By the end of this project, we will release three datasets:

1. a corpus of the identifiers of aligned page images from Internet Archive books, along with coordinate information for handwritten annotations discovered by the alignment process;

2. the output of our collational OCR correction model for aligned text passages, to provide more accurate transcripts of the text; and

3. a database, Ichneumon, of handwritten annotations detected in printed books, both from the aligned pages described above and from applying the annotation-detection model to unaligned pages.

Since all of these data are derived from public-domain page images and OCR transcripts from the Internet Archive, we will release all of these datasets under an open license. The datasets will be hosted on Northeastern servers, in the University library's institutional repository. We will also discuss contributing improved OCR transcripts and other data back to the Internet Archive.