Individual vs. Collaborative Methods of Crowdsourced Transcription

Samantha Blickhan, Coleman Krawczyk, Daniel Hanson, Andrea Simenstad, Amy Boyer, Victoria van Hyning

Pre-print version: 6/18/2019

Individual vs. Collaborative Methods of Crowdsourced Transcription

Samantha Blickhan[1*], Coleman Krawczyk[2], Daniel Hanson[3], Amy Boyer[1], Andrea Simenstad[3], and Victoria Van Hyning[4]

1 The Adler Planetarium, USA
2 University of Portsmouth, UK
3 University of Minnesota, USA
4 Library of Congress, USA

*Corresponding author: sblickhan@adlerplanetarium.org

Abstract

While online crowdsourced text transcription projects have proliferated in the last decade, there is a need within the broader field to understand differences in project outcomes as they relate to task design, as well as to experiment with different models of online crowdsourced transcription that have not yet been explored. The experiment discussed in this paper involves the evaluation of newly-built tools on the Zooniverse.org crowdsourcing platform, attempting to answer the research question: "Does the current Zooniverse methodology of multiple independent transcribers and aggregation of results render higher-quality outcomes than allowing volunteers to see previous transcriptions and/or markings by other users? How does each methodology impact the quality and depth of analysis and participation?" To answer these questions, the Zooniverse team ran an A/B experiment on the project *Anti-Slavery Manuscripts at the Boston Public Library*.

This paper will share results of this study, and also describe the process of designing the experiment and the metrics used to evaluate each transcription method. These include the comparison of aggregate transcription results with ground truth data; evaluation of annotation methods; the time it took for volunteers to complete transcribing each dataset; and the level of engagement with other project elements such as posting on the message board or reading supporting documentation. Particular focus will be given to the (at times) competing goals of data quality, efficiency, volunteer engagement, and user retention, all of which are of high importance for projects that focus on data from galleries, libraries, archives and museums. Ultimately, this paper aims to provide a model for impactful, intentional design and study of online crowdsourcing transcription methods, as well as shed light on the associations between project design, methodology and outcomes.

keywords

crowdsourcing; text transcription; experiment design

INTRODUCTION

In his seminal work *The Wisdom of Crowds* (2004), James Surowiecki argued that a heterogeneous group of individuals, including experts and non-experts, could more accurately and efficiently make decisions or offer solutions to complex problems than the average expert. His case studies ranged from guessing the weight of a prize ox at a county fair to determining where to look for the remains of a plane crash on the ocean floor. The term "crowdsourcing" was put forward by journalist Jeff Howe and his editor Mark Robinson in the journal Wired in

2006, to describe how companies and cultural heritage organizations could "tap the latent talent of the crowd" to source new ideas and innovative approaches to problems, and lower the cost of surfacing and utilizing those ideas (Howe 2006). Howe's definition has since been extended and applied to academic research, and scholarly definitions of crowdsourcing range from concise ("the process of leveraging public participation in or contributions to projects and activities," Hedges & Dunn 2017, 1) to all-encompassing ("a type of participative online activity in which an individual, an institution, a non-profit organization, or a company proposes to a group of individuals [the "crowd"] of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task," Estellés-Arolas & González-Ladrón-de-Guevara 2012, 9-10). The latter recalls Surowiecki's emphasis on the importance of heterogeneity and independence of decision makers, and most academic definitions include some acknowledgement of crowd diversity emphasizing the importance of non-specialist participation (see Van Hyning 2019, 4-5; for further discussion of the definition of crowdsourcing, see also Brabham 2013; Ridge 2014; and Terras 2016).

Though integral to Surowiecki's theory, variety among the crowd can add to the challenge of designing crowdsourcing projects as a whole, let alone specific tools that are intended for used across multiple projects with similar goals. As projects continue to proliferate, practitioners must ensure that the tools they are providing are not only appropriate for the crowd, but also support the production of high-quality data, and are useful for—and accessible by—the research communities for whom they are created.

This paper will specifically look at methods employed on the Zooniverse,[1] the world's largest platform for online crowdsourced research. The Zooniverse platform has its roots in scientific research, specifically astrophysics. The first project, *Galaxy Zoo* (Lintott et al. 2008), launched in 2007, and while humanities and collections-focused projects began to appear on the platform in 2011 (Williams 2014; Grayson 2016; Belknap 2018; BrodeFrank, Blickhan, & Rother 2019, Van Hyning 2019), the majority of the more than 200 projects that have launched on the platform have been from scientific disciplines. Text transcription as a task, however, is not unique to humanities research projects. *Old Weather*, one of the first Zooniverse text transcription projects, invites volunteers to transcribe weather data from ship logbooks, and the results are being used to carry out quantitative climatological research (Brohan 2012).

Zooniverse projects all follow the same general format: each item in a project, be it an image, audio or video file, is independently assessed by multiple individuals. The responses are then aggregated together for "consensus" (typically majority rule). For example, in the *Snapshot Serengeti* project,[2] which aims to identify animal species from camera trap images, 25 people tag each image with the name, number, and behaviors of the species present. If 24 out of the 25 volunteers tag the image as having one resting Zebra present, the resulting confidence on the consensus result is high (Swanson et al. 2016).

On the Zooniverse platform, the number of volunteers that must complete each task for a given image, audio, or video file before it is considered "complete" and ready for "retirement" varies from project to project, and is typically tied to the complexity of the task. A lack of specialist knowledge or instances of misclassification can admittedly lead to errors in volunteer-generated data (Freitag et al. 2016), but the consensus-based approach used by projects on the Zooniverse platform has a substantial track record of producing high-quality results that have been accepted and used by the scientific community (e.g. Lintott et al. 2008; Schwamb et al. 2013; Johnson et al. 2015; Hennon et al. 2015; Kuchner et al. 2017; Zevin et al. 2017). Over 150 peer-reviewed publications have been produced using Zooniverse data.[3]

[1] https://www.zooniverse.org.
[2] https://www.snapshotserengeti.org.
[3] https://www.zooniverse.org/about/publications.

When building Zooniverse projects in the humanities, researchers have thus far continued to follow these methods derived from scientific practice for non-textual data, but a recent grant-funded effort awarded our team the resources to test the efficacy and efficiency of such methods on academic crowdsourcing projects for text transcription. This paper will discuss the process of evaluating tools for text transcription in the humanities that were initially developed for non-text-based research projects on the Zooniverse platform, including the design, development and implementation of an experiment to test independent classification versus collaborative classification methods. In doing so, the authors sought to interrogate the role of independent decision-making—one of the fundamental principles of crowdsourcing.

I TRANSFORMING LIBRARIES AND ARCHIVES THROUGH CROWDSOURCING

In 2016, the Zooniverse team was awarded a National Leadership Grant from the Institute of Museum and Library Services (henceforth IMLS) for the project, "Transforming Libraries and Archives Through Crowdsourcing" (Van Hyning, Blickhan, Trouille, & Lintott 2017).[4] The award would support the creation of bespoke transcription projects, during which the Zooniverse team could develop methods to better support Galleries, Libraries, Archives and Museums (henceforth GLAM).

The research questions and project goals laid out in the proposal included the identification of best practices for GLAM-led crowdsourcing projects and, once identified, communication of these practices back to the GLAM community. As part of this exploration, the proposal included the evaluation of existing Zooniverse tools for online crowdsourced text transcription:

> Does the current Zooniverse methodology of multiple independent transcribers and aggregation render better results than allowing volunteers to see previous transcriptions by others or indeed collaborate to create a single transcription? How does each methodology impact the quality of data, as well as depth of analysis and participation?[5]

The proposal listed the creation of two bespoke text transcription projects as an outcome of the grant effort. These projects, hosted on the Zooniverse platform, would facilitate investigation of the research questions listed above. A call for text transcription project proposals that closed in February 2017 yielded more than 30 responses. After the selections were made by the Zooniverse team, *Anti-Slavery Manuscripts at the Boston Public Library* (henceforth *ASM*) was the first to undergo development.[6] The project aims to shed light on the Boston Public Library (henceforth BPL) Anti-Slavery Collection; "one of the largest and most important collections of abolitionist material in the United States," containing "roughly 40,000 pieces of correspondence, broadsides, newspapers, pamphlets, books, and memorabilia from the 1830s through the 1870s."[7] The dataset for *ASM* is made up of the correspondence from the Anti-Slavery Collection: 11,742 letters in total.

In order to address the research question evaluating individual versus collaborative methods of crowdsourced text transcription, the Zooniverse team chose to run an A/B experiment, in which 50% of visitors to a website are shown one version of the site, and 50%

[4] https://www.imls.gov/grants/awarded/lg-71-16-0028-16.
[5] https://www.imls.gov/sites/default/files/grants/lg-71-16-0028-16/proposals/lg-71-16-0028-16_proposal_documents.pdf.
[6] The project proposal was submitted by Tom Blake, Manager of Content Discovery at the Boston Public Library.
[7] https://www.antislaverymanuscripts.org/about.

are shown a different version. This practice allows researchers to observe reactions to a website before making a final decision about which version to ultimately use. For *ASM*, the A/B split involved creating two versions of the transcription interface, and evaluating the quality of transcription data produced through each version. Alongside the evaluation of data quality, the Zooniverse team tracked certain user behaviors using Google Analytics, and also gave the volunteer community the opportunity to provide feedback.

II EXPERIMENT DESIGN

The design of the A/B experiment was devised by Zooniverse team members in 2017. The A version of the interface was modeled after existing Zooniverse text transcription projects using individual transcription methods, and the B version consisted of a new method, featuring collaborative transcription. The A method has its roots in the development of the Zooniverse platform, which adheres to the principles of independence of assessment and heterogeneity of the crowd, as per Surowiecki's analysis of the power of distributed individuals to make better decisions in aggregate than the average solo expert (Surowiecki 2004). As noted in the Introduction, individual classification has been an essential part of the Zooniverse approach, to ensure that volunteer assessments of a subject are not influenced by the opinions of other volunteers. For short micro-tasks, it seems unproblematic to invite multiple volunteers to assess the same data, but for longer, more involved tasks such as text transcription, the necessity of independence (for the reduction of bias) is harder to justify, in part due to the sheer volume of data produced by the transcription task (on the level of page, line, word, character) and the range of potential variation between individual responses. Specifically, difficulties arise when considering 1) the time-consuming and often-subjective nature of transcription;[8] and 2) the difficulty of aggregating strings of text, particularly text with additional markup such as tags to indicate deletions, superscripts, or unusual page layout. [9]

Previous Zooniverse projects that inspired the A version include *Shakespeare's World*[10] and *AnnoTate*.[11] *Shakespeare's World*, launched in 2015, is a collaboration with the Folger Shakespeare Library and the *Oxford English Dictionary*, in which volunteers transcribe sixteenth- and seventeenth-century handwritten documents, such as letters and recipes. *AnnoTate*, also launched in 2015, is a partnership with Tate Britain and Tate Archive in which volunteers transcribe British and émigré artists' diaries, letters and sketchbooks. These projects were developed in tandem, and tested two similar—but slightly different—approaches to transcription. In *Shakespeare's World*, volunteers are shown a digital image of a manuscript and asked to use the cursor to place dots on either side of a word, phrase, or whole line of text they feel confident enough to transcribe.

[8] Within the field of palaeography, the process of interpreting handwritten text has historically been described by scholars like M.R. James and Bernard Bischoff as an "art" that requires time and experience to master. While this belief has been reconsidered in more recent scholarship, particularly in regard to pedagogical approaches to palaeography, it remains true that the transcription and interpretation of text can be less than straightforward, and even seasoned professionals can disagree at the best of times. For further reading, see especially Derolez 2003.
[9] Issues with text markup are not unique to the Zooniverse platform, nor is aggregation the only difficulty it presents. The team behind University College London's *Transcribe Bentham* project (http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham) has published on their experience using XML markup, particularly their finding that it has almost certainly been a barrier to participation (Causer & Terras 2014).
[10] https://www.shakespearesworld.org.
[11] https://anno.tate.org.uk.

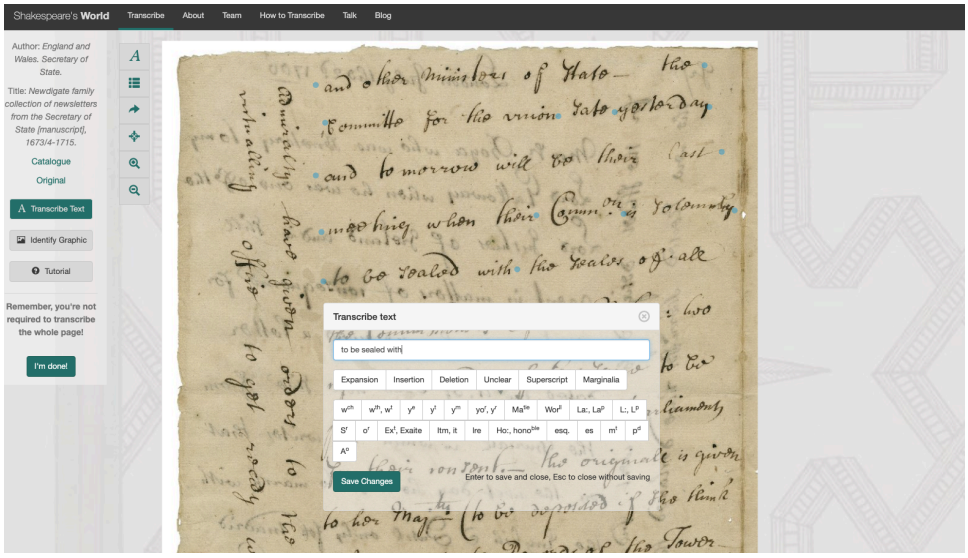

Figure 1: *Shakespeare's World* transcription interface.

AnnoTate volunteers place dots on either side of whole lines of text. The placement of the second dot triggers a pop-up transcription box, into which volunteers transcribe the text contained between the two dots. Volunteers can transcribe as much or as little as they like, ideally skipping words or lines they are not confident about.

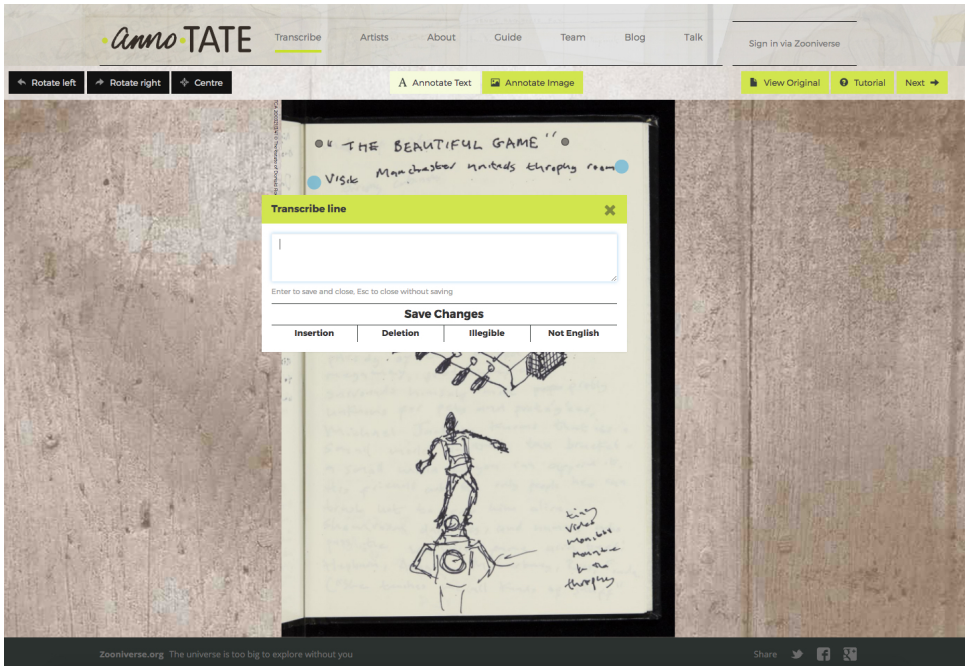

Figure 2: *AnnoTate* transcription interface.

Each line is transcribed by 3 or more people and only "retired" (considered complete and removed from active transcription) when the lines reach a given level of consensus. When they achieve consensus, lines are surrounded with grey dots to indicate to volunteers that those lines are complete (see the first line of the transcription shown in Figure 2). In a closed or independent system where volunteers are not able to see what other people have transcribed before, the grey dots steer volunteers to unfinished sections of a page and reduce wasted effort. Rather than show each page to 10 people, these experimental projects hoped to achieve complete transcriptions using fewer than five contributors per line. As in all Zooniverse projects launched before 2017, volunteers submit their classifications independently of one another, meaning the

raw transcription data must go through a process of aggregation to determine the majority assessment or transcription for each line on a page. Another intended use for the dots was to gather spatial data for each line, in the hopes that this information, along with the transcriptions, could be useful for future investigation into Handwritten Text Recognition (HTR) methods and machine learning.

The process described above is unusual for online crowdsourced transcription projects. Indeed, Zooniverse may be unique in this approach. The consideration of our method within a larger space of public humanities-focused crowdsourcing projects was one inspiration behind the research question about independent vs. collaborative transcription methods. We also wanted to determine whether the difficulty of aggregating multiple transcriptions from the individual transcription method is justified by the quality of the resulting data. The difficulty is not only related to the time-consuming task of transcribing handwritten text, but how to approach the annotation method (dot placement) in a way that provided useful data without adding too much additional effort on the part of the volunteer. As noted above, the *Shakespeare's World* and *AnnoTate* projects asked volunteers to place dots on either side of a word/phrase/whole line, or whole line, respectively. The *Shakespeare's World* method tested well in alpha and beta testing stages of the project, but proved difficult to aggregate after launch, when higher volumes of data were produced by a larger and more diverse volunteer base. For this reason, both versions of the *ASM* A/B experiment asked volunteers to consider physical lines of text as basic units for transcription, i.e. transcribe entire lines at a time, rather than split them into sections based on volunteer confidence in their ability to transcribe the contents. If a volunteer could not read a whole line, the hope was they would skip it altogether. An early version of the annotation design for the A interface asked volunteers to place multiple dots for each line, clicking between each word in the line of text they were annotating, as a way to assist with text alignment to facilitate aggregation. However, the response from beta testers was that this annotation method was far too cumbersome, and so both versions of the *ASM* interface featured an annotation tool that was similar to the *AnnoTate* annotation method with endpoint dots on either side of a whole line.

Another difference with previous Zooniverse projects in the annotation approach was that, for both *ASM* workflows, the two dots are connected by a line. This decision was made to help encourage volunteers to only transcribe individual lines, and reduce the likelihood of accidentally skipping lines or transcribing them twice (known as eye-skip), or attempting to transcribe entire paragraphs in a single pass. The latter behavior was seen in both *Shakespeare's World* and *AnnoTate*.

As noted above, the A version of the transcription interface is based on the *AnnoTate* project, and features individual transcription methods. A screenshot of the interface can be seen below, in Figure 3. Annotations for transcriptions completed by the current user are shown in blue. Annotations for transcriptions that are in progress are green.

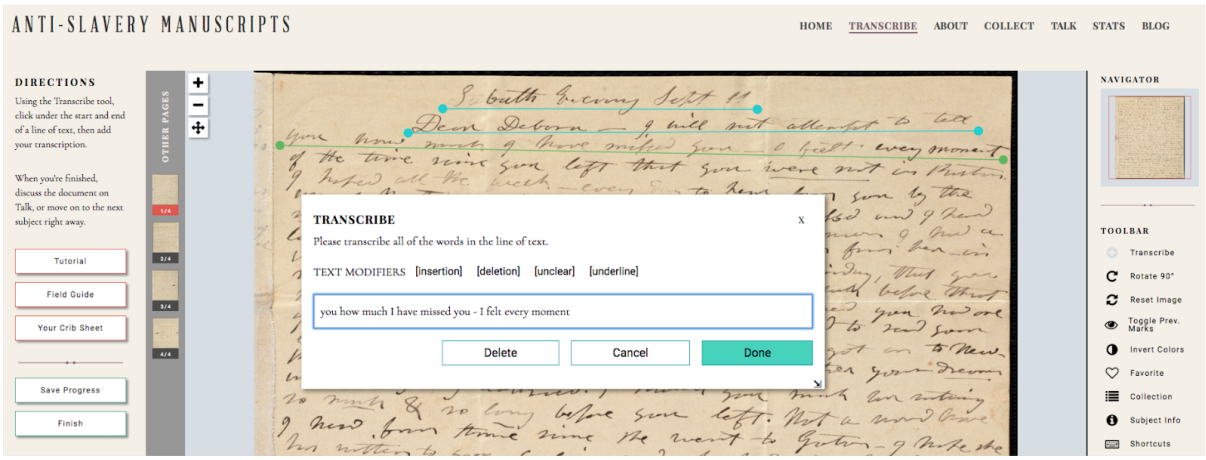

Figure 3: *ASM* individual transcription "A" interface.

The B version of the transcription interface was designed to be a mid-point between the individual method used by previous Zooniverse transcription projects, and what we will call the "open" interface used by many other transcription platforms, including *Transcribe Bentham*, the Smithsonian Transcription Center suite of projects, the Newberry Library's *Transcribing Modern Manuscripts*, and projects built using FromThePage and Scripto.[12] In those projects, a single transcription is created and edited by one or more volunteers who can see one another's work in real time, and then flagged for review when complete. The review process is often completed by volunteers as well, though some projects require volunteers to complete a certain number of transcriptions or achieve a certain status before they can review as well as transcribe. Once the transcriptions have passed peer review, some require editorial verification from experts such as cultural heritage staff or academic project team members.

The model described above removes the need for aggregation of results (which is useful for practitioners), but also removes the requirement of group consensus, as it is ultimately up to an individual such as the final transcriber or the reviewer to mark a transcription as complete. We felt it was important to carry out the individual transcription method to its logical conclusion: to see whether text could be treated the same as other types of data on the Zooniverse platform, and whether the method was viable and justifiable in terms of resources. We also wanted to explore a possibility that would allow volunteers to collaborate in the transcription process while retaining the algorithmic consensus model. Therefore, in lieu of replicating the open transcription method used by other platforms listed above, the Zooniverse team designed a third option (the B version of *ASM*), which combines the line-by-line functionality of the A method with the ability to view and interact with previous transcriptions.

Though the annotation tool in the B version is visually the same tool as the A version (endpoint dots connected by a thin line), the functionality is different. In the A version, a volunteer places the dots, transcribes into the popup box, and presses "Done" for each transcription they wish to enter. In the B version, a volunteer goes through that same process, but only if they are the first person to transcribe a letter. Subsequent transcribers also see the annotation dots and lines placed on the image by the first transcriber. If they disagree with the placement of the previous annotations, they can choose to ignore them and add new ones of their own (with the aid of the "Hide Previous Marks" option on the toolbar). If they agree with the placement of previous annotations, they can click on those annotations and see previous transcriptions via a dropdown menu. When previous transcriptions are selected by a transcriber, they automatically populate the text entry field in the transcription pop-up, where they can be

[12] http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham; https://transcription.si.edu/; https://www.newberry.org/transcribing-modern-manuscripts; https://fromthepage.com/; https://scripto.org.

submitted without alteration, or edited by the current transcriber. The interface, as seen by a volunteer encountering an already-transcribed subject, is shown in Figure 4. Annotations for previously-transcribed lines are red, annotations for transcriptions that are currently in progress are shown in green, and annotations for transcriptions completed by the current user are shown in blue.

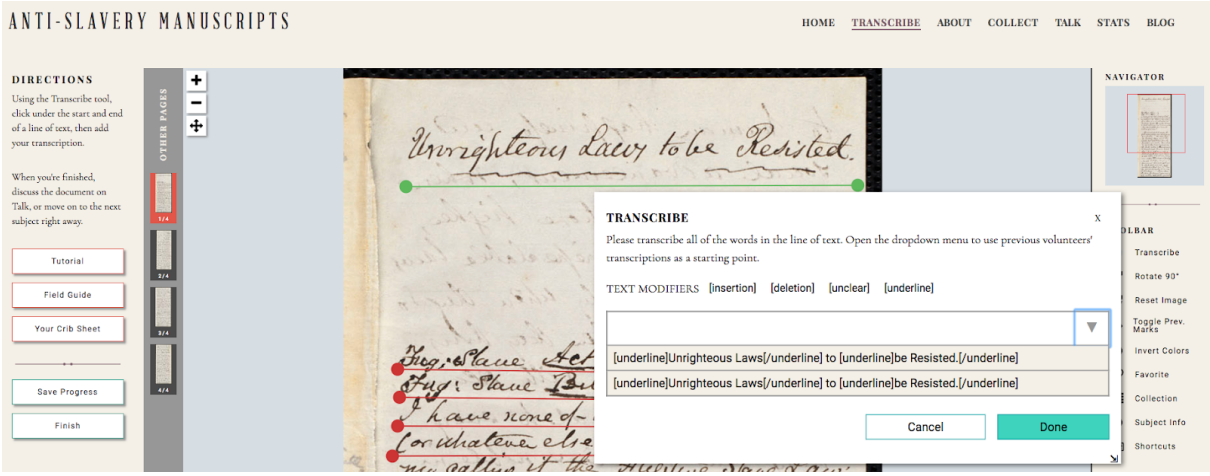

Figure 4: ASM collaborative transcription "B" interface.

2.1 Aggregation, Consensus & Grey Lines[13]

When transcription projects first launched on the Zooniverse platform, there was evidence that volunteers did not always complete an entire page of transcription. Thanks to the method of multiple independent transcribers, that was acceptable in principle. In practice, if all volunteers started at the top of a document, it meant that the beginning of a page was frequently over-transcribed, while the end of a page would receive fewer or no transcriptions. In an example from *Operation War Diary*,[14] which launched in 2014 and set out to transcribe 1.5 million pages of British World War I war diaries, the over-transcription of certain sections of text can be linked to the type of content being transcribed.

[13] Some contents of this section have been adapted from Krawczyk 2018.
[14] https://www.operationwardiary.org.

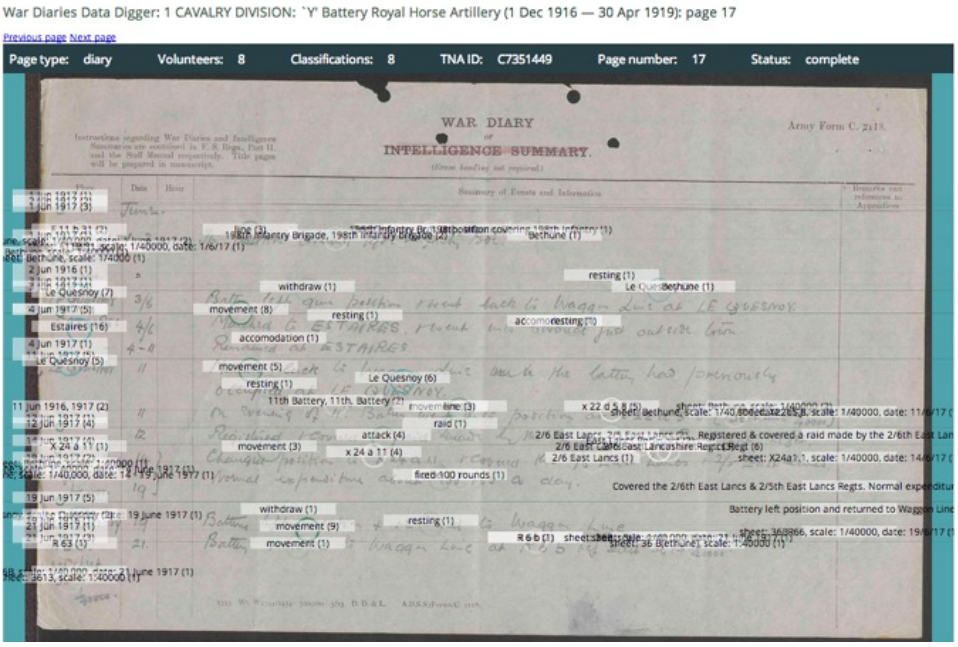

Figure 5: Transcription data from *Operation War Diary*, including number of transcriptions received per unit.

Figure 5 shows a completed page from *Operation War Diary*, with transcription data overlaid on the image. Each white label contains an aggregated transcription, including a number in parentheses which shows how many people transcribed that particular unit of text. In this example, the units that are most frequently transcribed are those that do not require specialist knowledge: dates, times, places, etc. Many units of text only received transcriptions from one or two users.

In an attempt to help ensure an even spread of transcription effort for *Shakespeare's World* and *AnnoTate*, the Zooniverse team developed a method in which a grey annotation would appear on a subject to show when a line of text had received enough transcriptions, so volunteers could focus their time on text that had not yet been completed (Van Hyning 2016). The greying-out method is used in both the individual and collaborative workflows on *ASM*, and consists of a two-step process: annotation aggregation, and text aggregation.

2.1.1 Annotation Aggregation

Annotation aggregation is achieved using data clustering, which works by identifying high-density areas in the data produced by the volunteers—the places where volunteers have marked lines on an image. The aggregation parameters are the same for both workflows, though the way that volunteers experience the aggregation outcomes within the project website will be different, depending which workflow they were sorted into.

When multiple volunteers underline the same line of text on an image, those lines are grouped together to make a single line. High-density areas are found using the DBSCAN algorithm from the python module Scikit-learn.[15] This algorithm has two parameters: the minimum density, and minimum number of points. For *ASM*, the minimum number of points is set to 1, to ensure all transcriptions in the collaborative workflow are shown to the next volunteer who sees a particular image. The density is based on clustering, and is the same across all pages and letters; the value was chosen by inspecting the data collected during the beta testing phase of the project, and identifying a value that worked well for the majority of cases.

[15] https://scikit-learn.org/stable/about.html#citing-scikit-learn.

The lines are first clustered by the slope of lines drawn by volunteers, which separates horizontal text from text written in the margins, cross-writing, or at unusual angles. Then the code scans for columns in the text—there is usually a single column of body text in the center of a page—and identifies column breaks, typically where no lines are drawn across the middle of the page. Next, the code finds lines of text for each of the identified slope clusters by clustering in the direction perpendicular to the lines of handwriting. Finally, clustering done parallel to lines of handwriting detects the start and end points for each line. For *ASM*, the slope clustering distance is set to 3 degrees, the "gutter tolerance" is set to 30 pixels (i.e. lines that make up neighboring columns can overlap by up to 30 pixels and still be considered a gutter), the line clustering distance to 20 pixels, and the endpoints to 30 pixels. Typically, these distances can be transformed into densities when combined with the minimum number of points. In this case, the number of points must be ≥1 (see below). In short, these densities can be viewed as how close the annotations of the transcribed lines need to be to one another in order to be grouped together in a single cluster.

In the collaborative workflow, we want to see all previous transcriptions for a page, so the minimum is set to ≥1 for all the clustering—once a single person has annotated and transcribed a line of text, the code will retain those data to show the next volunteer. For the individual workflow, the minimum is also set to 1, but the project transcription interface filters out what is shown to volunteers, to make sure annotations are not shown until they have enough annotations and transcriptions to be considered "complete," and need to be greyed out. Because the minimum for both workflows is set to ≥1, this means that the density parameter is the only way to control the clustering, and that parameter will need to be applied to every letter in the dataset being transcribed, meaning for both workflows there are cases in which the aggregation will fail. For examples of this type of failure, see Krawczyk 2018 and section 3.1 of this paper.

2.1.2 Consensus Score

To determine the consensus score for a line of text the aggregation code tokenizes each transcription on whitespace, then aligns the resulting tokens using the Collatex python package (an example of this is shown below in Table 1).[16]

&	tell	who	I	know-		[deletion]I	can	not	think[/deletion]
&	tell	what	I	know	-	[deletion]I	cannot		think[/deletion]
&	tell	what	I	know	-	[deletion]I	cannot		think[/deletion]
&	tell	what	I	know	-	[deletion]I	cannot		think[/deletion]
4	4	3	4	3	3	4	3	1	4

Table 1: Example of alignment by word.

The consensus line is created by taking the word from each column with the most votes. For Table 1, the consensus line would be: "& tell what I know - [deletion]I cannot not think[/deletion]". The bottom row of Table 1 also lists the number of "votes" received for the most common word from each column. The word "not" only received a single vote, but is still included in the consensus line. This is because while testing the code, we noticed that filtering out words with low consensus often resulted in losing words that should have been included in the final transcription. In cases like the one shown in Table 1, it is easier for a research team to

[16] http://interedition.github.io/collatex/pythonport.html; https://collatex.net/. The full python package (consensus_txt.py) used for *ASM* is available at https://github.com/ckrawczyk/ASM-consensus-text.

identify and remove an extraneous word from an aggregate transcription that it would be to identify and re-insert one that is missing.

By adding up the number of total votes and dividing by the number of words, we form the consensus score, which indicates the average number of transcribers who agreed on the transcription for this line. In the case of Table 1, the consensus score is 33 votes divided by 10 words, or 3.3. If everyone agreed on the line of text, the consensus score would be equal to the number of volunteers who transcribed the line, so this value can be viewed as the average number of volunteers who agreed on the text. In the figures below, these values are presented as "[x/y]" where x = consensus score and y = number of volunteers.

For both the individual and collaborative workflows, lines are considered complete when the consensus score is 3 or higher. Retirement of an entire letter requires three different volunteers to mark a subject as "complete" when they submit their transcription (this question is built into the transcription interface), or happens automatically when 15 people have worked on the letter without reaching consensus. This backup retirement method is intended to keep particularly difficult subjects from preventing the completion of a project. We learned from *Shakespeare's World* that there are some documents that are considerably more difficult than others to transcribe—typically due to poor handwriting, unusual or cramped layout, and the often-subjective nature of transcription—and which therefore do not reach consensus. The retirement metric is noted in the project data export, and individual lines which have not reached consensus (or have low consensus scores) can be identified by research teams for expert review.

The text aggregation described above can be affected by text markup tags, such as [insertion] and [deletion], which volunteers have the opportunity to include in their transcriptions as needed. To ensure that the aggregation is successful, we ask volunteers to be sure that there is whitespace separating the closing tag from the following word. If the whitespace is not inserted, for example, the following errors can occur:

to	let	[deletion]Sar[/deletion]Angelinea			give	[underline]all[/underline]the	
to	let		am[/deletion]	Angelina	give	all	the
to	let	[deletion]Sar[/deletion]Angelinea			give	[underline]all[/underline]the	
to	let	[deletion]Sar[/deletion]Angelinea			give	[underline]all[/underline]the	
to	let	[deletion]Sar[/deletion]Angelina			give	[underline]all[/underline]the	
5	5	3	1	1	5	4	1

Table 2: Example of misaligned text due to spacing errors.

The text in Table 2 has a consensus score of 3.125, meaning that by the *ASM* standards it would be considered complete, and greyed-out, but the line generated via consensus would make no sense:

"to let [deletion]Sar[/deletion]Angelinea am[/deletion] Angelina give [underline]all[/underline]the the".

The cases described above show the difficulty of developing a one-size-fits-all approach to text aggregation, not to mention the importance of reviewing results before accepting crowd-generated transcriptions as complete or correct. However, the consensus score, alongside the individual transcriptions, can be a good starting point for review, particularly when working with large datasets. For example, finding the average consensus score of a dataset and focusing effort on lines for which the consensus score is below average can expedite the review process.

2.1.3 Gold Standard Comparison

To evaluate the data quality from the individual and collaborative workflows we randomly selected letters from the first set of 2,713 letters to go into the *ASM* project, comprised of 19 pages of text overall (Blickhan 2018). These letters were transcribed by volunteers in each of the workflows being tested, as well as by an expert, who provided gold standard data.

Once the gold standard data was transcribed and exported from the project, we converted it to .txt format for the comparison process; testing aggregate versions of the data from the individual and collaborative workflows against the gold standard data. To generate .txt files for the aggregate results of each workflow, we exported the data from each workflow, then ran each set of data through the consensus_txt.py package. The resulting .txt files include the aggregate result arranged in "reading order" (top to bottom, left to right, and horizontal text before angled text).

To prepare the data for quality comparison, we took the following steps (for individual vs. gold standard and collaborative vs. gold standard):

1) Compare two files with the same name in two folders.
2) Remove metadata tags.
3) Get the Character error rate, as described below, for each line of text.

The output of this process was the following:

1) Matching lines of text between the Gold Standard and Experimental Cohorts (Individual and Collaborative).
2) Character error rate for each line of text with and without ignoring case.
3) Unweighted average character error rate for each page of text.

The character error rate measures the number of mistakes in a text string and divides it by the length of the string (in characters). Mistakes are classified as insertion, substitution, or deletion errors. A standard definition of character error rate is CER = (i + s + d) / n, where n = total number of characters, i = number of insertion errors, s = number of substitution errors, and d = deletion errors (Carrasco). The total number of mistakes (i + s + d) between the string of text being examined (in either the individual or collaborative transcription) and the complementary string of text in the gold standard transcription is also known as the Levenshtein distance, which refers to the number of single-character edits that would be required to transform one string into the other. Dividing the Levenshtein distance by the total number of characters (n) produces in the character error rate.

Metadata tags were ultimately removed for the purposes of this experiment because of their tendency to disproportionately raise the CER. For example, the presence of the metadata tag "[underline]" in one line of text but not another would result in a difference of 11 characters. To test the effect of metadata tags on transcription aggregation using this method of string comparison, we would recommend replacing the tags with unique single-character identifiers, though this would require research teams to make specific decisions about how to weigh the importance of metadata tag-related errors in transcription against other errors such as spelling.

III INDIVIDUAL VS. COLLABORATIVE DATA QUALITY RESULTS

The *Anti-Slavery Manuscripts* A/B experiment ran from January 23, 2018 to September 01, 2018. Volunteers who logged in using their Zooniverse accounts were sorted into either the individual or collaborative version of the transcription workflow. Non-logged-in volunteers were automatically sent to the individual workflow. Identical datasets of 2,173 letters were uploaded into each of the workflows. From this dataset, five letters (a total of 19 pages of text)

were selected at random to be used as a sample set. This number was chosen to balance the need for a large enough sample to perform a t-test, while recognizing the time limitations of generating "gold standard," or correct, transcriptions. The gold standard transcriptions of the test letters were provided by a content expert from the Massachusetts Historical Society. The five letters were completed by the individual and collaborative workflows, then compared with the gold standard data using the methods described below. When the letters were uploaded to the Zooniverse platform, they were each automatically given a unique identifier in the form of an 8-digit number. These numbers will be used throughout the section for clarity.

In the case of this experiment, the context of the transcription would remain intact even if case was ignored. Therefore, for this study we chose to work with the data for which case was ignored. There were also levels of granularity to choose from in terms of units of measure. Do we test this process across an entire letter? Page? Or do we go even closer and examine by line, word or character? Though many of the methods we used are functioning at the level of character or word (for example, the tokenization and alignment used to generate confidence scores, as well as the Levenshtein distance used to examine the quality of transcription results), we felt it was important to look at units that allowed the text to remain in some context. At first, we planned to run the quality comparison for units of entire letters:

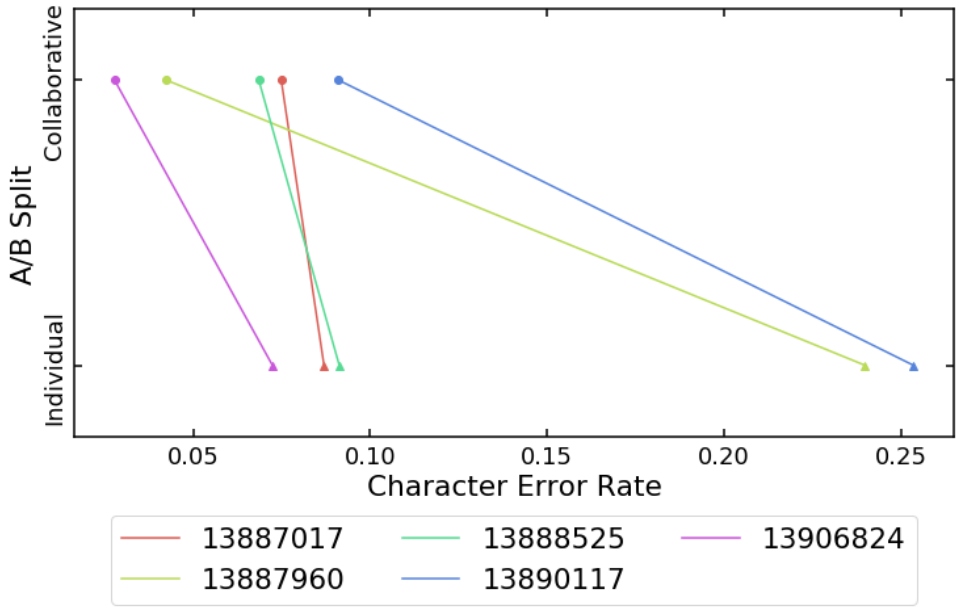

Figure 6: ASM data quality comparison results by letter.

Figure 6 shows the variation in character error rate (per letter) between the collaborative and individual results, as compared with the gold standard data. The 8-digit numbers in the figure legend are unique identifiers for individual letters While the results shown above were generally informative, we chose to run a statistical test with units of the 19 individual pages rather than entire letters. The results of the comparison by page are shown in Figure 7. The legend shows the Zooniverse ID for each letter, and the single-digit number following the ID indicates the page number.

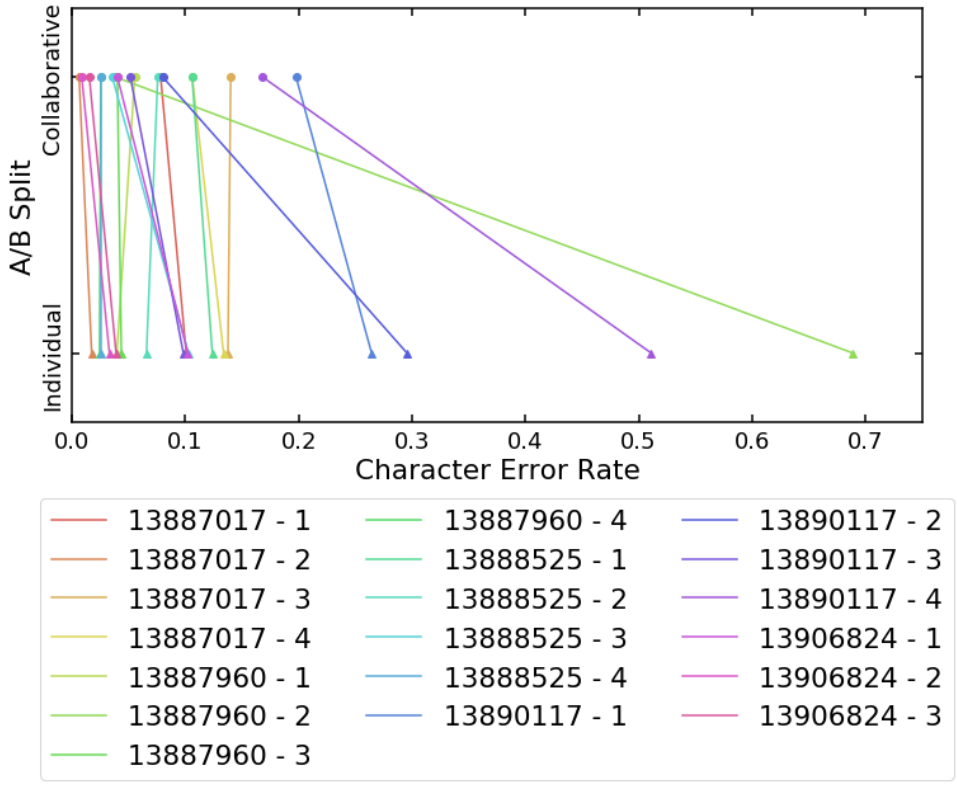

Figure 7: ASM data quality comparison results by page.

A matched pairs t-test (used to see whether there is significant difference between two sets of paired data) showed that the collaborative results had statistically significantly lower variation from the gold standard transcription than the individual results ($t = 2.184$, $df = 18$, $p\text{-value} = 0.02122$). For this test, the matched pairs were the matching pages of the letters from each of the collaborative and individual results. The collaborative workflow produced results closer to the gold standard data for all but five of the 19 test pages. Of those five pages, one produced equal results.

	13887017 p3	**13887960 p1**	**13887960 p4**	**13888535 p2**	**13888525 p4**
Individual	0.1367011	0.03904634	0. 02344643	0. 06121427	0.025
Collaborative	0.1381923	0. 05339208	0. 02448626	0. 07321403	0.025

Table 3: Pages for which the individual cohort produced data closer to gold standard.

Figure 8 shows a plot of the pages for which the individual cohort produced higher-quality data than the collaborative (the final row of Table *3* is not included, as the values are equal).

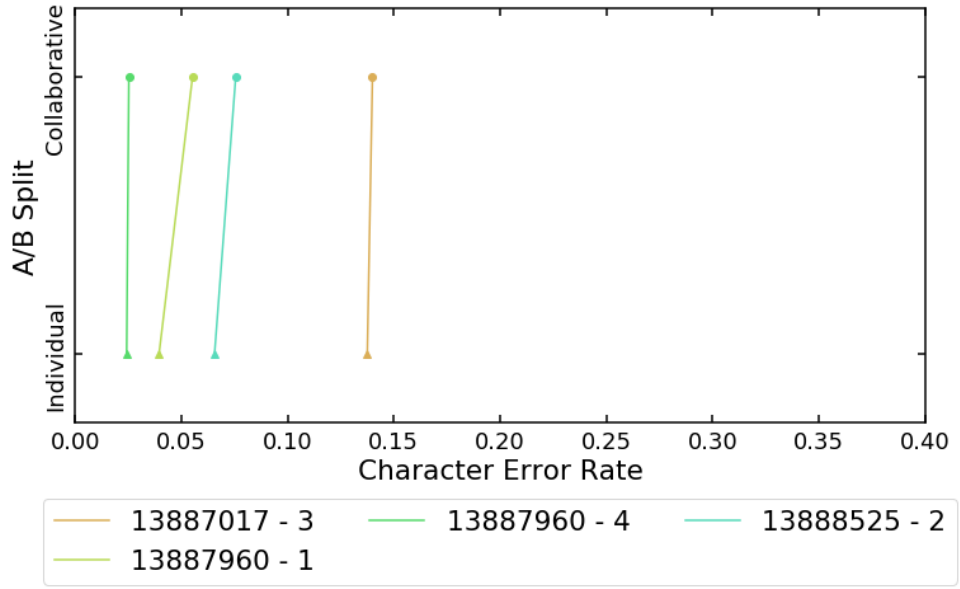

Figure 8: Pages for which the individual cohort produced data closer to gold standard.

Even when looking at the instances above in which the individual cohort produced higher-quality data, the difference is so slight that—when compared to other metrics like time and engagement—it is not necessarily higher enough to outweigh the benefits of those other elements.

3.1 Examining Instances of High Variation Between Individual and Collaborative Results

Returning to the data analysis by page (Figure 7), the majority of the results are contained within a relatively close range. There are only three outliers; pages for which the individual results were considerably further than the collaborative results from the gold standard transcription. The three examples are shown below.

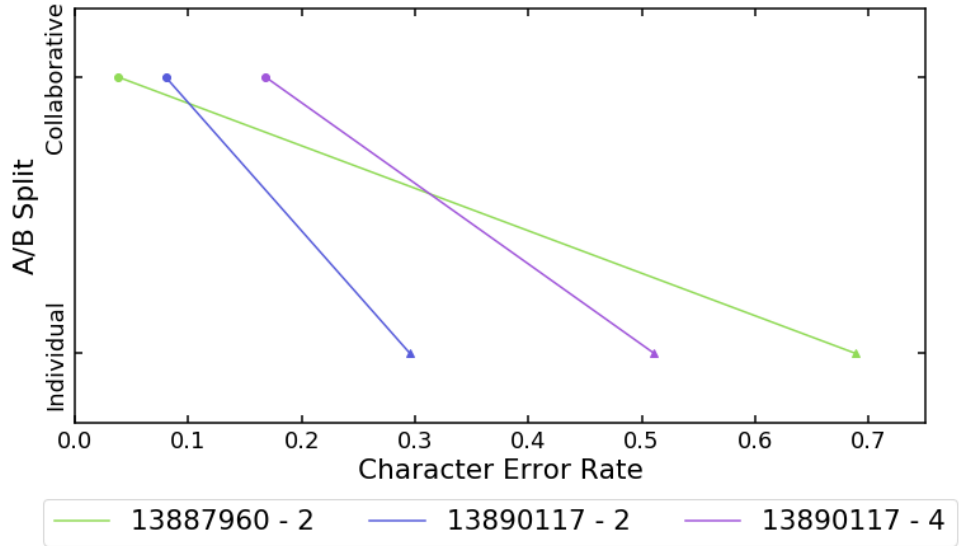

Figure 9: Outliers from original ASM data comparison by page.

Outlier 1: Subject 13890117, p.2

Subject 13890117 is a letter from Charles Fitch to Amos Augustus Phelps, likely written in 1833 (Fitch 1833). From the Boston Public Library's *Digital Commonwealth*: "Charles Fitch writes to Amos A. Phelps that he cannot sign Phelps's circular with his present views and feelings on the given subject. Phelps's circular, on the margins of which Fitch writes, declares slaveholding a sin and argues for Immediate Emancipation rather than Gradual Emancipation or Colonization. Fitch agrees slavery is a sin, but supports colonization."

Two of the pages in Figure 9 are from Subject 13890117. In both cases, the variation in output between the individual and collaborative results is due to the presence of printed text on the page (the "circular" referred to in the catalogue description). The printed text is much smaller than the handwritten text, meaning the annotations for those lines would need to be very close to one another—much closer than handwritten text would typically be—thereby having a negative effect on the clustering algorithms. Visualizations of the raw and aggregate line annotations on each page show that all three example pages feature instances in which data have been mistakenly grouped together by the aggregation code.

In the examples below, the clustered lines are represented by blue dots and lines on the original image (at left). The filled-in dots represent the first dot placed, and the empty circles represent the second. The visualization (at right) shows each aggregate line as well as the breakdown of words in the string for each transcription submitted. The gold standard visualizations only feature the breakdown for one transcription, as the data were provided by an expert transcriber. The individual and collaborative visualizations should feature multiple breakdowns for each line, as multiple people submitted transcriptions for each line of text.

In the case of Figure 10, the printed text was problematic because the small text size resulted in the lines being too close together for the aggregation method to identify them as separate from one another (i.e. the density was too high). The clustering code assumes the height of a line of text is about the same across all the letters in the dataset, which is clearly not the case on this page.[17] Because of this, though the transcriptions for each line of printed text were submitted correctly into the transcription interface, the attempt by the aggregation engine to align the text vertically has gone awry; it has attempted to generate a single line of text based on the entire printed section. The resulting line, "of proposed the names of others. I have taken entailed - just not will be attained is you will sign your large to recommend the lectures in slave cases a course of is not Immediate Emancipation fear the effectual means of his and its If inalienable rights - that such emancipation is safe for the within, ground &c it by", is made up of words that feature in the printed text, but is unintelligible, and has a very low consensus score: 1.78 out of 14 total transcriptions.

Toward the bottom of the original letter shown in Figure 7, it looks like the expert transcriber has missed annotating several lines of text, as the blue dots are missing for a few lines (fourth from bottom, beginning with "But I am…"; penultimate line, "to remove this great evil…"; final line "-ans now in operation…"). An examination of the raw data exports shows that this is not the case; transcriptions were submitted for all those lines, but the close proximity of the written text as the writer attempts to finish a thought while running out of room on the page has resulted in the clustering algorithm to combine the last four lines of text into a single line: "But in operation - I believe a decided do to the this is evil, than far other system of means." As in the previous case above, the consensus score for this line is low: 1.25 out of 4 total transcriptions.

[17] We are currently experimenting on new methods that can handle clustering of different densities within a letter (HDBSCAN or OPTICS instead of DBSCAN). See section V for more information.

The collaborative transcription visualization in Figure 11 shows similar problems as the gold standard version, but which are amplified due to the increased volume of data, as this page features transcriptions from multiple volunteers. The printed text has been similarly clustered together, resulting in a nonsensical transcription with a low consensus score (6.5 out of 60), and the bottom four lines of handwritten text have been clustered into a single, low-consensus line (2.65 out of 12).

The individual transcription visualization in Figure 12 shows a problem with the printed text (similar to the ones in the collaborative and gold standard examples above), but is also an example of how collaborative transcription methods can be useful in the case of handwriting that is particularly difficult to read. Looking at the raw transcription exports (the blue annotations on the left side of the image), the lack of annotations on the bottom third of the image indicate that many volunteers did not even attempt to transcribe this handwritten section. Therefore, the subject was retired after 15 different volunteers had submitted work, though many lines never reached consensus.

Outlier 2: Subject 13890117, p.4

Lack of completion on difficult pages (when transcribers are left in isolation) can also be seen on page 4 of Subject 13890117 (shown in Figures 10-12). As was the case with page 2 of this letter, the casual script, small text size, and close proximity of the lines (particularly toward the bottom of the page) proved to be off-putting to many transcribers in the individual cohort, who chose not to attempt a transcription of these lines.

Outlier 3: Subject 13887960, p.2

The third example from the plot shown in Figure 9 comes from Subject 13887960, a letter from Anne Warren Weston of Philadelphia to Deborah Weston, written in January of (likely) 1838 (Weston 1838). The first two pages of the letter are cross-written. Anne Weston begins her letter with, "It is after 9 Sunday evening, my dear Debora, but if I did not write now I know not when I should[.]" After writing four pages of text, she signs the letter. However, it seems that the next morning she had more to add, but rather than adding a new sheet of paper, she rotated the first page 90 degrees and began writing horizontally across the page, continuing from where she had left off the previous evening: "Monday morning. Mary, Mrs Philbrick & Miss Paul Rave gone in James Mott's carriage with Mr Thom to see the Water works."

The second page of the letter is used similarly; the second page of the original Sunday night missive cross-written with an update from "Monday noon" in which Anne tells Deborah that she has returned from the dedication of the Pennsylvania Hall and delivers news of that event, as well as several other additions to the original letter.

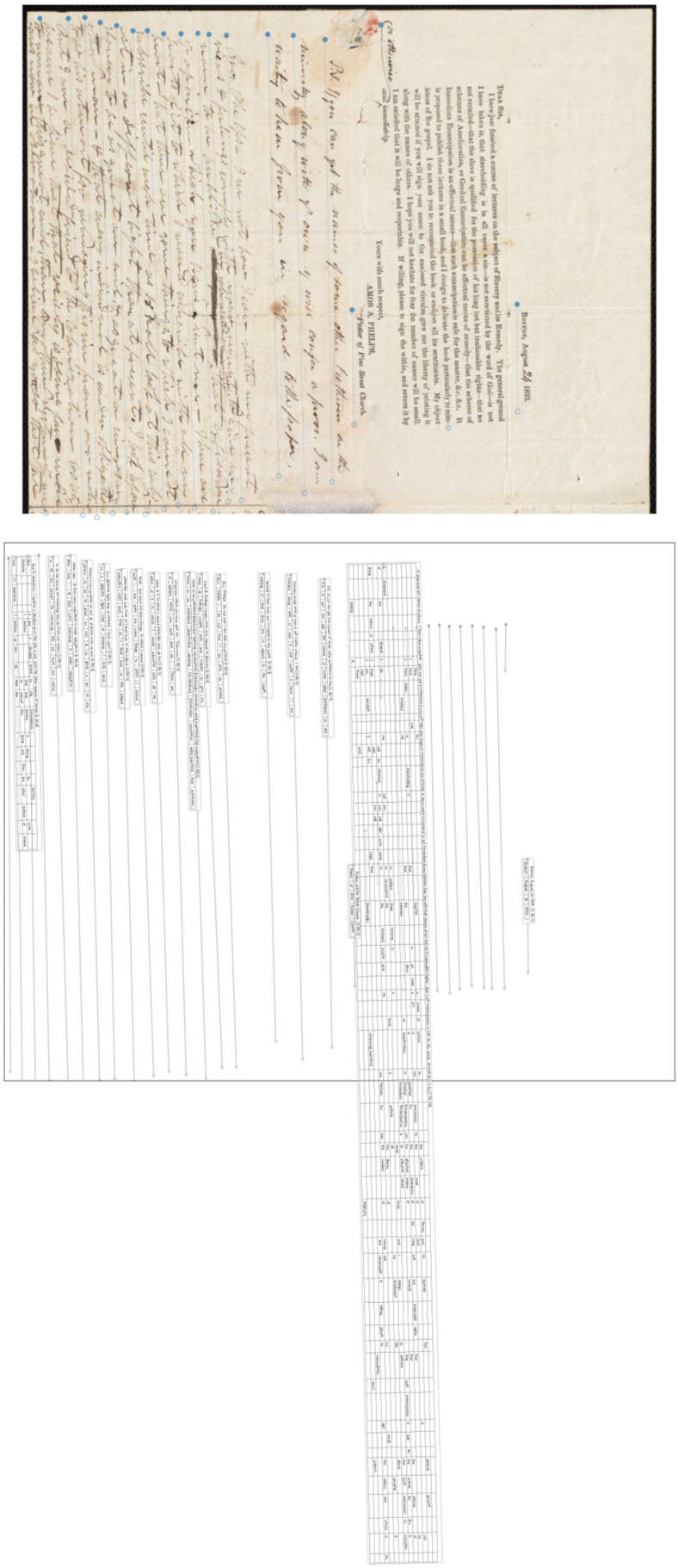

Figure 10: Subject 13890117, gold standard transcription, p.2.

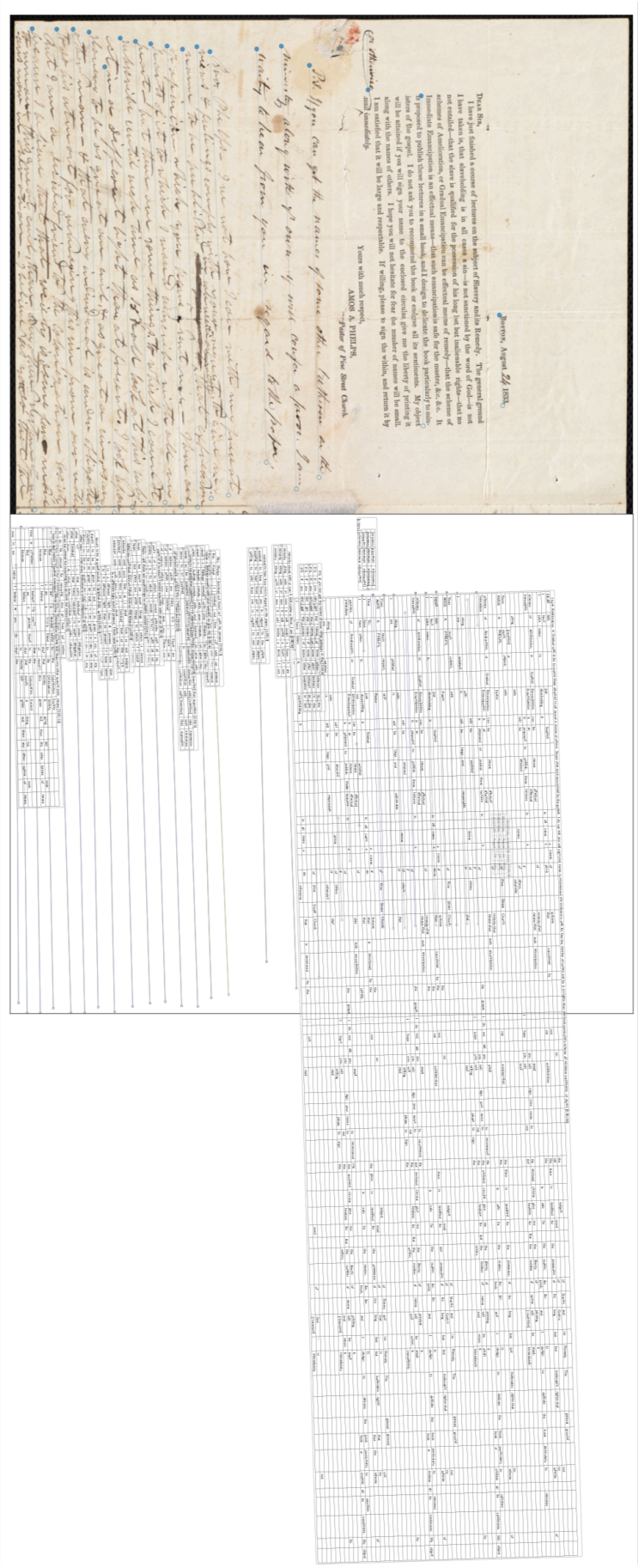

Figure 11: Subject 13890117, collaborative transcription, p.2.

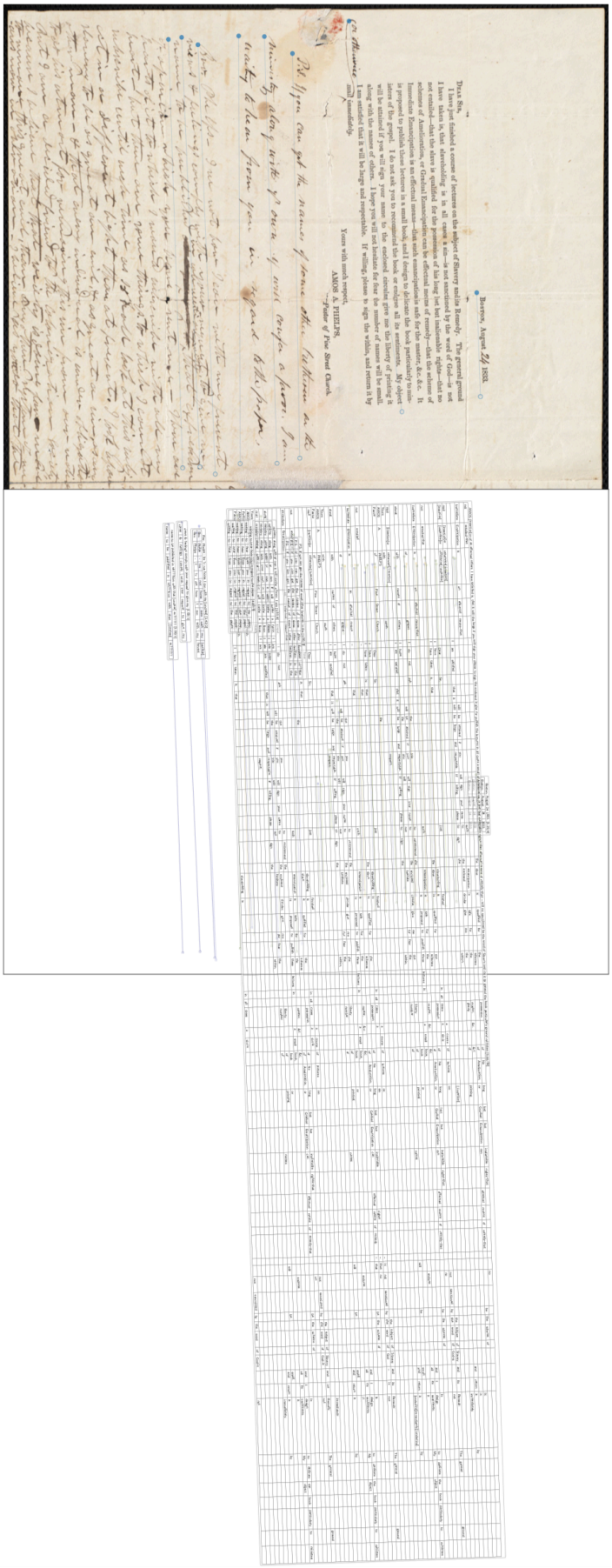

Figure 12: Subject 13890117, individual transcription, p.2.

Journal of Data Mining and Digital Humanities
ISSN 2416-5999, an open-access journal
http://jdmdh.episciences.org

Figure 13: Subject 13890117, gold standard transcription, p.4.

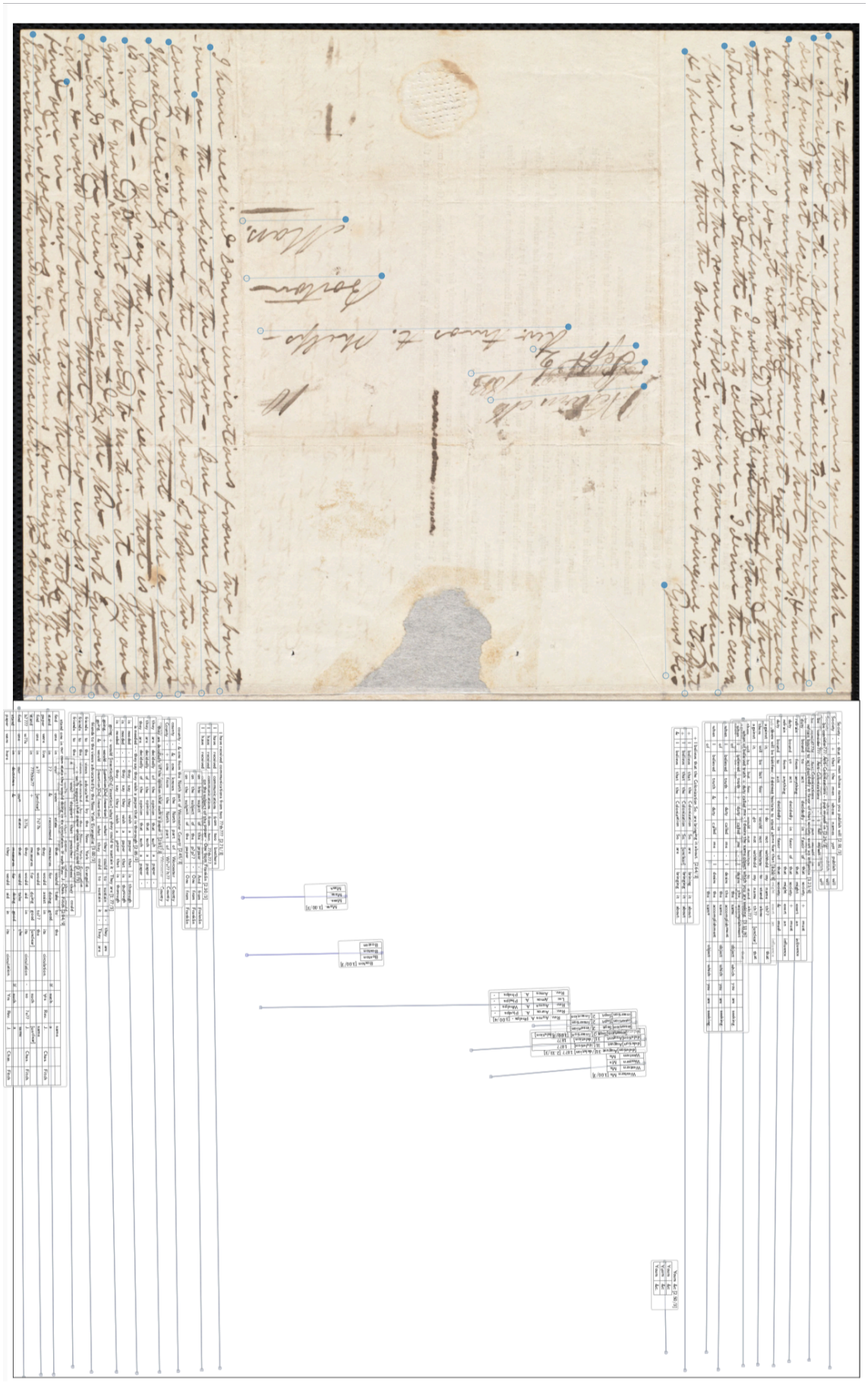

Figure 14: Subject 13890117, collaborative transcription, p.4.

Figure 15: Subject 13890117, individual transcription, p.4.

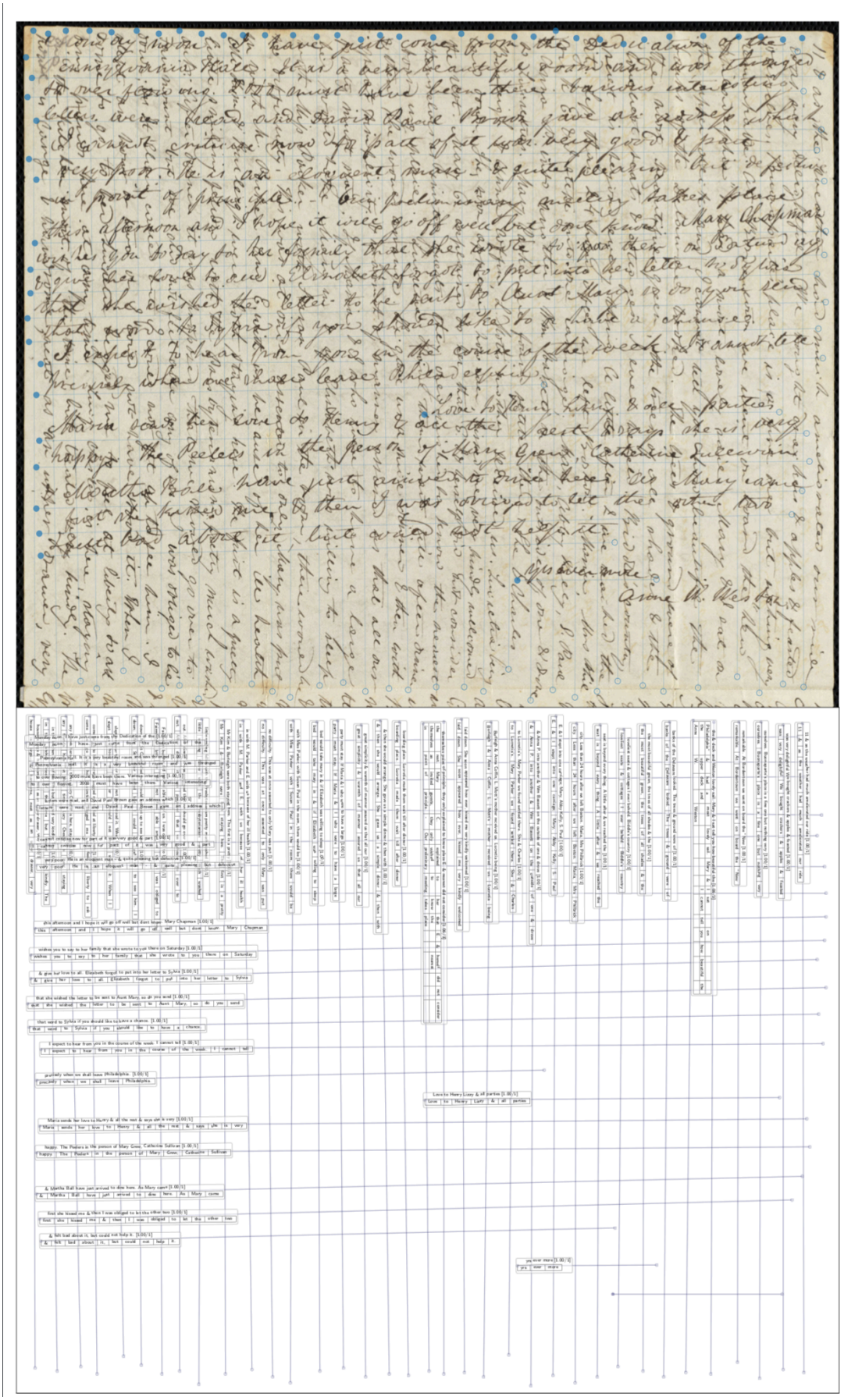

Figure 16: Subject 13887960, gold standard transcription, p.2.

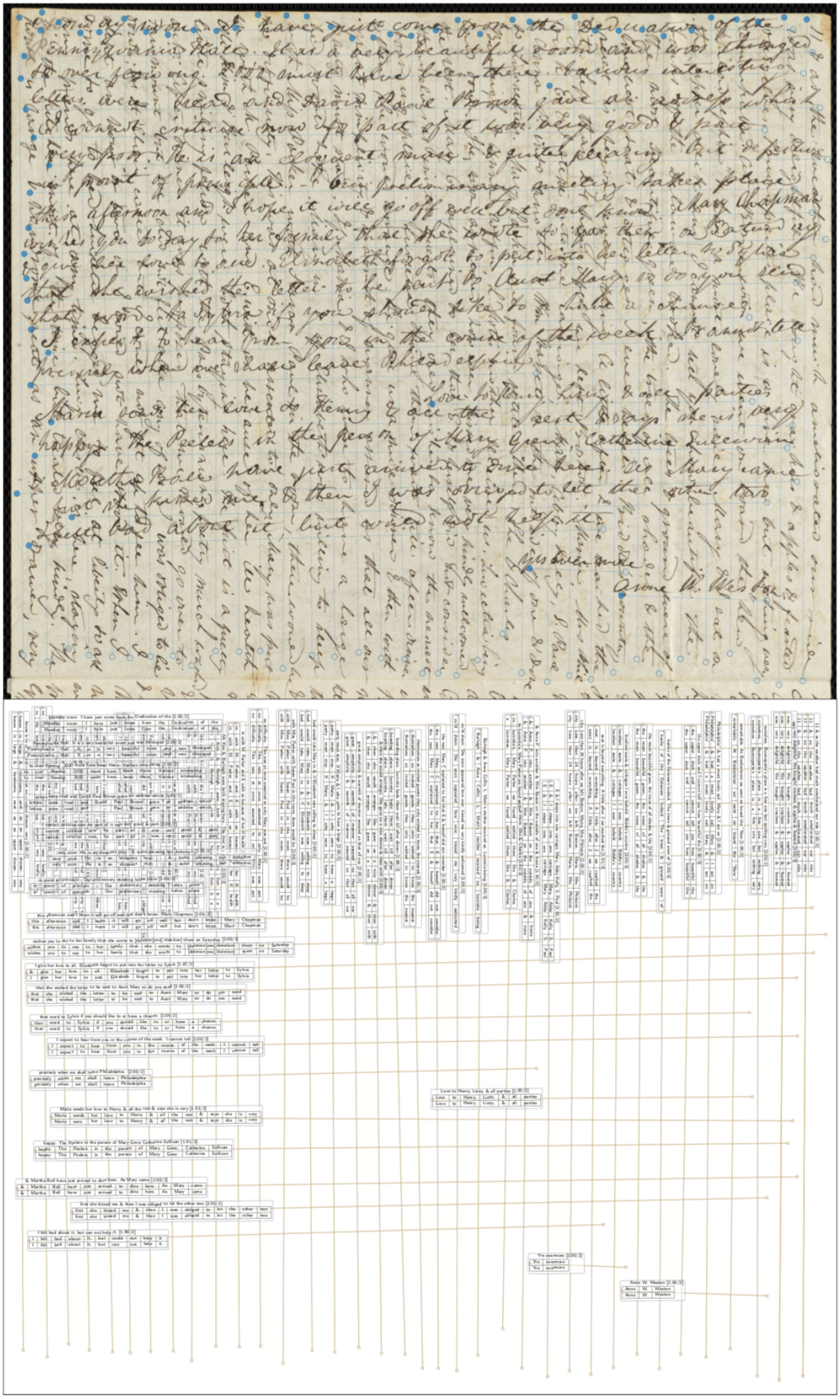

Figure 17: Subject 13887960, collaborative transcription, p.2.

Though the letter is cross-written, Anne Weston leaves plenty of space between her lines, and the gold standard and collaborative reduction visualizations show each line clearly. The individual visualization, on the other hand, is empty. Unfortunately, the system did not link the page in the individual workflow, so we do not have it for comparison.

3.2 Analysis with Outliers Removed

Removing the outliers discussed above from the sample confirmed the original results. Without the outliers, the collaborative results had even lower variation from the gold standard transcription than the initial results ($t = 3.102$, $df = 15$, $p\text{-}value = 0.003643$).

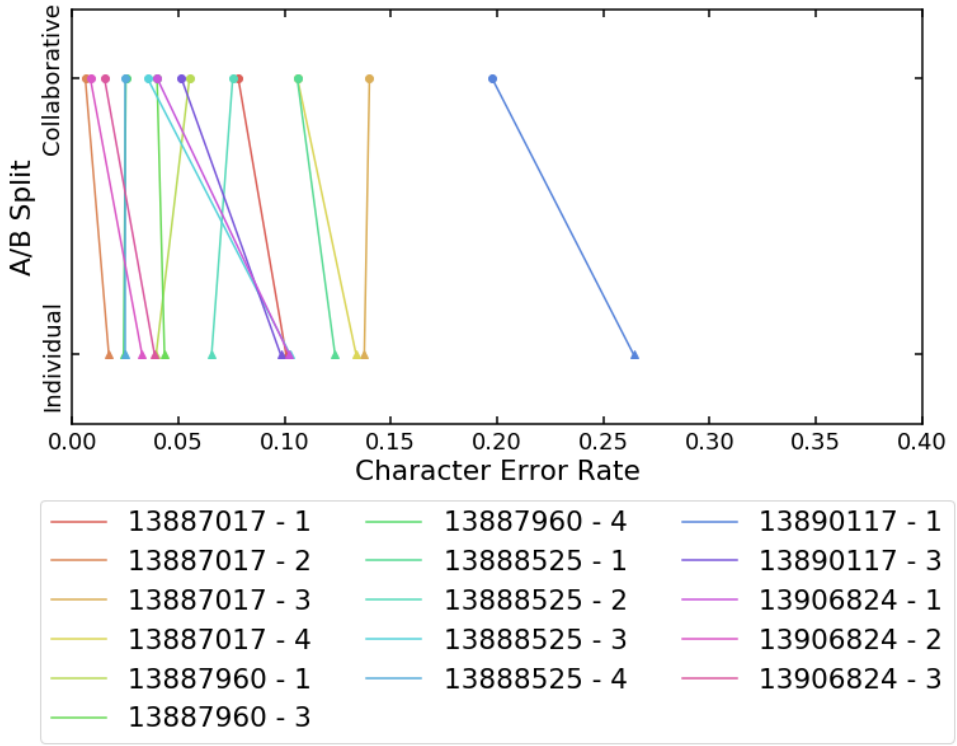

Figure 18: ASM data quality comparison results by page, with outliers removed.

IV EVALUATING BEHAVIOR & ENGAGEMENT

The second half of the IMLS research question being discussed in this paper asks: "How does each methodology impact the …depth of analysis and participation?" To answer this question, we used Google Analytics to track the behavior of volunteers on the *ASM* project, in both the individual and collaborative cohorts. The tracking was limited to logged-in users, and the use of Google Analytics on the Zooniverse platform is explicitly stated in the Privacy Policy (which users are required to read before creating an account).[18]

[18] "[W]e may use software such as Google Analytics that collects statistics from IP data. This software can determine what times of day people access our site, which country they access the websites from, how long they visit for, along with technical details of their computer (browser, screen type, processor)."

The purpose of this tracking was to record information that is not available from classification exports, such as agreement with other users, canceling a transcription, or average duration of classification sessions. We also wanted to see whether either method resulted in higher volunteer engagement with additional project elements, including transcription aids like the project Tutorial and the Field Guide—which offers examples of the type of material being transcribed, and difficult-to-approach text types like tabular or cross-written text—and supplementary information like the About page, which gives background about the project and archive, the message boards (called "Talk"), and the project blog.

The following sections will present the actions measured during the A/B experiment, which ran from January 23, 2018 - September 1, 2018.

4.1 Cohorts by Users and Sessions

	No. Users	No. Sessions	Avg. Session Length	Avg. Sessions per User
Individual	1,598	5,532	00:31:34	3.46
Collaborative	1,543	6,015	00:34:56	3.9

Table 4: Data on user numbers and session duration by cohort.

Users were split almost evenly into the individual and collaborative cohorts, with 50.9% in the individual cohort, and 49.1% in collaborative. The collaborative cohort produced 52.1% of the total transcription sessions, while the individual cohort produced 47.9%. The average session duration and number of sessions per user were slightly higher for collaborative than individual.

The numbers show that the collaborative cohort produced exponentially more transcription data than the individual cohort during the same time period. The collaborative cohort did this with the same number of participants than the individual cohort (in fact, approximately 50 fewer people), and without requiring a considerable increase in the average number of sessions per user, or in the average amount of duration for those sessions.

4.2 Cohorts by Classification Numbers

	Completed Classifications	Finished Annotations
Individual	7,608	186,789
Collaborative	9,976	245,038

Table 5: Number of classifications and annotations per cohort.

A comparison of classification and annotation events shows much more variation between the cohorts. A "completed classification" is when a volunteer finishes all the transcription they plan to do on a single letter, and submits this work. A "finished annotation" is when a volunteer annotates and transcribes a line, then presses "Done" in the transcription pane. A completed classification can be made up of multiple finished annotations.

Though the numbers of participants were similar, and the total number of sessions relatively close as well, the collaborative cohort produced 25% more completed classifications and finished annotations than the individual cohort (56.7% of the total completed classifications and total finished annotations, compared to 43.3% of the total of each for the individual cohort).

4.3 Collaborative Cohort Behavior

	Novel Transcription	Click Previous Line	Click Dropdown	Agreement
Collaborative	132,604	112,895	110,734	89,730

Table 6: Collaborative cohort behavior.

Within the collaborative cohort, the majority of engagements with previous volunteer transcriptions were agreements—selecting a previous transcription and submitting the line without editing. Approximately 80% of the transcriptions in which collaborative users clicked into the dropdown menu were selected and submitted without being edited. This information, combined with the data analysis from the previous section, suggests that volunteer transcribers are producing high-quality transcriptions on their own, and the ease of collaboration allows this process to move much more quickly than it would with individual transcription methods, without lowering the quality of transcription data produced.

4.4 Engagement by Cohort

	Done & Talk	Open Field Guide	Open Tutorial	Click About Page	Click Blog	Click Talk Header
Indiv.	1,062	1,855	1,295	209	54	845
Collab.	829	1,714	1,273	231	66	499

Table 7: Engagement by cohort.

Based on the information shown in Table 7, the individual cohort seemed to engage more with the Talk boards: 62.8% of visits to the Talk board from the header menu (which can be seen at top right of the transcription interface shown in Figures 16 and 17) were from the individual cohort, and 14% of all submitted transcriptions from the individual cohort were via the "Done & Talk" option shown in Figure 19, whereby a volunteer completes their transcription and is invited to the project discussion forum to discuss the letter they just worked on. Volunteers can submit their work and continue transcribing by selecting "Done," or submit their work and post a discussion topic by selecting "Done & Talk."

There could be a few reasons for this increased Talk board engagement by the individual cohort. The first is a desire for confirmation that they are doing things correctly; individual transcription can be isolating, and it would seem logical that transcribers who are not feeling as confident about their work would turn to a project community for advice or support, whereas the collaborative transcribers gain feedback in the form of others' transcriptions.

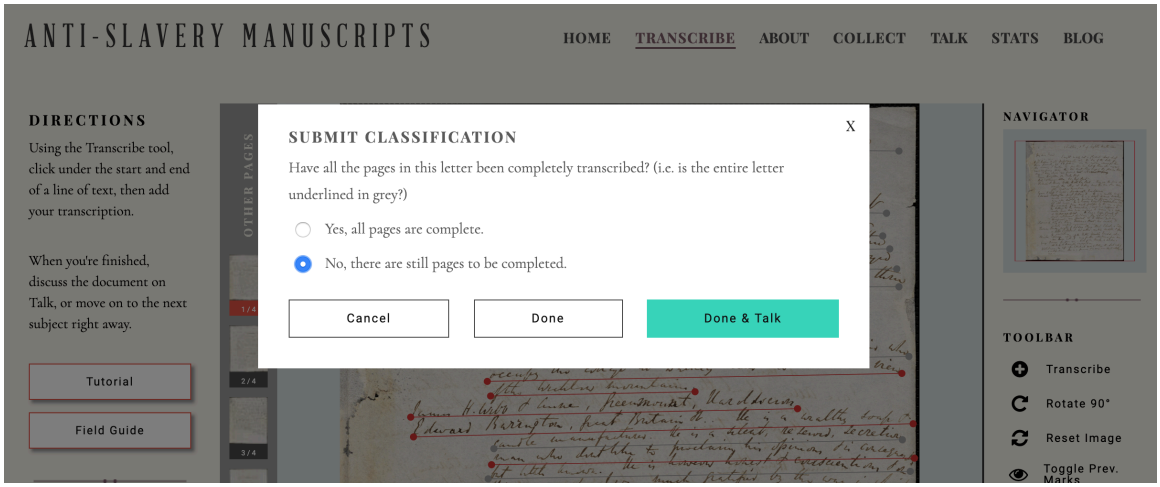

Figure 19: ASM submit classification prompt.

No significant differences were seen between cohorts in their engagement with the Field Guide, tutorial, About page, or blog. The Field Guide, as the name suggests, is a section that provides information about the content of the letters being transcribed, as a way to help transcribers when they encounter material that might be challenging or unfamiliar, such as

antiquated characters like the long s (ſ), which is often interpreted by modern-day transcribers as the letter f. The tutorial provides a step-by-step walkthrough of how to work through the transcription interface, and is automatically shown to volunteers the first time they visit the project. The About page provides the historical context of the project, the collection of letters being transcribed, and the team behind the project.

There is a distinct difference in the amount of engagement with tools and supplementary material that directly affect transcription (helper tools), and those that provide additional information about the project. Similarly, the Talk engagement was higher through the project interface than from volunteers clicking on the link in the header menu.

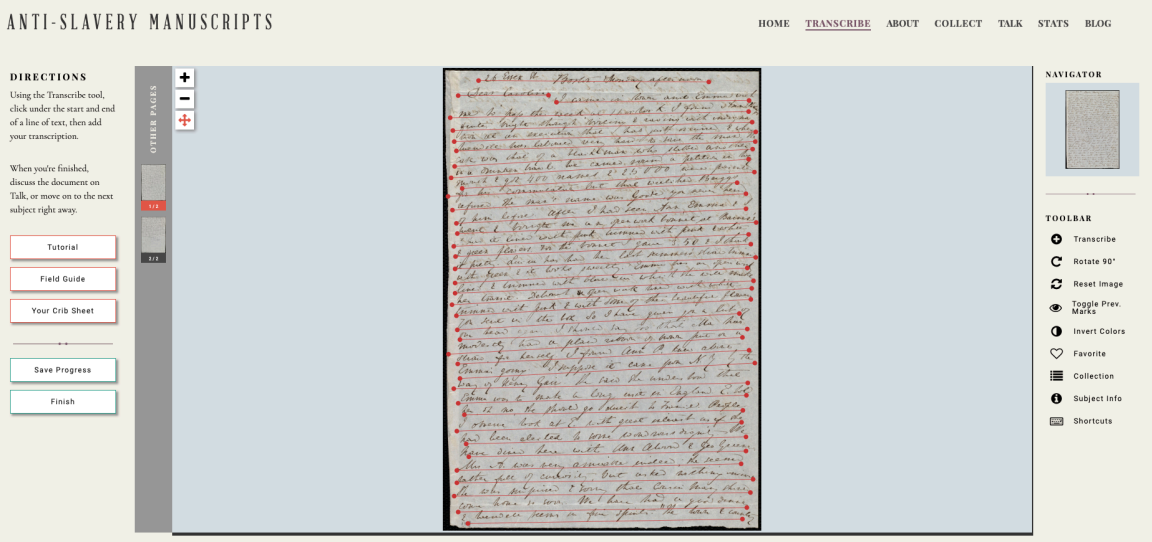

Figure 20: *ASM* transcription interface with toolbars and header options shown.

The method of access could possibly be a cause of the lower numbers for certain options shown in Table 7. The Done & Talk, Field Guide, and Tutorial buttons are all built into the transcription interface: the Done & Talk option is part of the transcription popup, and the Field Guide and Tutorial buttons are directly under the Directions and right above the frequently-used Save Progress and Finish buttons (see Figure 20). The links to the About page, Talk board, and project blog are in the header menu on the top right of the screen, and require the volunteer to navigate away from the transcription interface in order to interact with the content of those pages. Low click-rates from the header menu suggest that we may need to consider making the information from the About page more obvious within the transcription interface.

V CONCLUSION

In this A/B experiment, the cohort of volunteers sorted into the collaborative workflow produced transcription data that had significantly less variation from the gold standard transcription data than the transcription data provided by those sorted into the individual workflow. The collaborative cohort completed their transcriptions in a much shorter time span: the collaborative workflow was finished on October 22, 2018, approximately 10 months after the project launched. On that same date, the individual workflow was not yet 50% complete. Based on these results, after the first round of data analysis post-A/B experiment, we closed the individual workflow, meaning the entire *Anti-Slavery Manuscripts* project is now running on the collaborative mode of transcription. We will continue to monitor the results in collaboration with the research team at the Boston Public Library so that we can refine the aggregation

algorithms based on long-term examination of the resulting transcription data. For example, ongoing work indicates using HDBSCAN or OPTICS instead of DBSCAN allows for clustering of handwriting with different densities (e.g. handwriting that grows smaller as the writer reaches the bottom of the page, which in previous experiments was frequently incorrectly clustered together). At the time of writing, volunteers have finished transcribing 1,203 letters from the second set (out of five total sets of letters), bringing the total number of transcribed letters up to 3,376 since the project's launch. The current number of registered volunteers who have participated in the project is 4,775. Based on the results of the experiment, we will be building a generalized version of these tools for collaborative transcription into the Zooniverse Project Builder toolkit, so we can monitor the effectiveness of this transcription method on other datasets containing handwritten text.

After the A/B experiment ended, but before we closed the individual workflow, we opened up both workflows to the public so that volunteers sorted into the collaborative workflow could experience the individual method, and vice versa. We designed a short survey and asked volunteers to participate if they were interested in giving feedback on the different transcription methods. We received 58 responses, which will be analyzed and presented in a separate article.

5.1 Next Steps

The experiment described in this paper definitively shows that the collaborative method of text transcription devised for this experiment produces better quality data than the individual method used in older projects deployed on the Zooniverse platform, and requires less time to achieve these results. But work still needs to be done to determine where this collaborative version stands in comparison to the open collaborative methods of online crowdsourced text transcription discussed in section II of this paper.

First, we would invite the opportunity to engage with data about the metrics involved in other crowdsourcing projects. In general, online crowdsourcing projects tend to give regular updates on how many volunteers are participating and how many documents have been transcribed over the course of a project's lifecycle. In addition to these metrics, it would be useful to have access to additional information, such as how much time is spent on initial transcription compared to review. If review is required, is this step also completed by volunteers, or does it require content experts? If community-based review is used, is there a "final" review step before transcriptions are published? How long is the average timeline between uploading a document into a project and having the final transcription made available to the public via a content management system or some other means? Crowdsourcing projects have matured to the point where we now have enough data to systematically compare methods and results. Comparative, cross-project studies are needed in order to make the best use of volunteers' time and the research and institutional funding that supports crowdsourcing projects.

Second, there needs to be more information available about the quality of transcriptions that result from crowdsourcing projects, and the relationship between the results and the tools and methods used in their creation. Once made available to the public, how frequently are transcriptions flagged as containing errors that need to be corrected? How much maintenance do these databases require in the long term to guarantee that any necessary updates can be made in a timely manner? There is some existing information about quality and results which is extremely useful, though non-exhaustive and now slightly outdated. For example, in 2012, Ben Brumfield of *FromThePage* published a blog post on quality control in transcription projects, which includes descriptions of known crowdsourcing methods and examples of each type (Brumfield 2012). The *Transcribe Bentham* team have also published on the quality of their

project results in previous stages of the project, noting that (at the time of publication) the time spent by research associates on project moderation and editorial practice, if diverted to transcription alone, could have outpaced the volunteer output. These metrics were used to improve the transcription tools being used, but the team also noted that the need for editing considerably reduced the cost-effectiveness of crowdsourced transcription (Causer at al. 2012).

The question of accuracy or quality raised here is not intended to suggest that crowdsourcing cannot be an efficient, engaging way to produce digital versions of handwritten texts, nor is it meant to provoke mistrust in the results of crowdsourcing projects. It is instead proposed as a way to evaluate methods for collecting transcriptions, with a mind to reducing volunteer effort and ensuring usefulness of publicly-available results. It is entirely possible that a singular method will not be the best approach for all projects which aim for transcribed text as an outcome. If the results of differing methods are found to be comparable in terms of data quality, what can—or should—be used as the tie-breaker? Time spent transcribing? The facility of the back-end pipeline (i.e. the steps required for a completed transcription to be made available in a public repository)? Community preference? What, if any, are the differences between providing tools for other research teams to use (as is the case with the Zooniverse Project Builder), and creating a resource for use by a single institution for internal collections (as in the case of the Smithsonian Transcription Center)?

Finally, practitioners and researchers must continue to engage with methods that facilitate human and computer interaction as means of generating transcription data, to ensure that transcription tasks make the best use of volunteer time. Engagement with materials is an invaluable outcome of crowdsourcing projects, as well as arguably an ethical requirement; Trevor Owens has written about the importance of providing meaningful work within the context of crowdsourcing projects (2014, 278-279). Transcription provides a wonderful opportunity for in-depth engagement with source materials, but the massive volume of documents that need transcription is undeniable. The Zooniverse team has begun to experiment with incorporating machine learning techniques into text transcription projects (Hanson 2018; Hanson & Simenstad 2018; Verma 2019). This work builds off of the machine learning advances being made by research teams in scientific projects on the Zooniverse (Wright et al. 2017; Zevin et al. 2017; Bahaadini 2018; Willi et al. 2018; Trouille et al. 2019) as well as information provided in resources like Smith and Cordell's "A Research Agenda for Historical and Multilingual Optical Character Recognition" (Smith and Cordell 2018), and existing tools like Transkribus, developed as part of the READ project, which leverages crowdsourcing for HTR use.[19] New projects like *Living with Machines*, a multi-institution collaboration including the Alan Turing Institute and the British Library that brings together STEM and GLAM professionals in order to "devise new methods in data science and artificial intelligence that can be applied to historical resources, producing tools and software to analyse digitised collections at scale for the first time," are forging the way ahead for those who hope to apply human-computer optimization techniques for text-based data on a massive scale.[20]

It is our hope that by continuing to publicly re-evaluate the tools we offer for text transcription, we can acknowledge the needs of those using the resulting transcription data while also allowing our growing community of researchers and transcribers to participate in decisions being made about the tools we are creating and providing. Through these methods and others, we believe that online crowdsourced text transcription can remain a useful, accessible, and reliable method for creating digital versions of handwritten text for many years to come.

[19] https://transkribus.eu/Transkribus/; https://read.transkribus.eu.
[20] https://www.turing.ac.uk/research/research-projects/living-machines.

Acknowledgements

The study described in this publication would not be possible without the help of the thousands of volunteers who participated in this project, and who continue to do so to this day. Particular thanks must be given to the volunteer Moderators on the *Anti-Slavery Manuscripts* Talk boards: Gerry Gault, Holly Pence, and Joanna Treasure. Kathy Griffin of the Massachusetts Historical Society provided the gold standard transcription data for comparative analysis.

Anti-Slavery Manuscripts at the Boston Public Library was designed by Becky Rother, lead designer for Zooniverse, and built by Zooniverse web developers Will Granger and Shaun A. Noordin. We also acknowledge the contributions of the Zooniverse teams based at the Adler Planetarium, the University of Minnesota, and the University of Oxford. Funding for research and development was provided by the Institute of Museum and Library Services (grant LG-71-16-0028-16).

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and a grant from the Alfred P. Sloan Foundation.

References

Bahaadini, S. et al. (2018). Machine learning for Gravity Spy: Glitch classification and dataset. Information Sciences, 444, pp. 172-186. Available at https://doi.org/10.1016/j.ins.2018.02.068

Belknap, G. (2018). Illustrating natural history: images, periodicals, and the making of nineteenth-century scientific communities. *The British Journal for the History of Science*, 51(3), pp. 395-422. Available at https://doi.org/10.1017/S0007087418000511 (Accessed 19 May 2019).

Blickhan, S. (2018, April 4). Anti-Slavery Manuscripts: How We're Dividing the Data [Blog post]. Available at https://www.bpl.org/blogs/post/anti-slavery-manuscripts-how-were-dividing-the-data/ (Accessed 12 May 2019).

Brabham, D. C. *Crowdsourcing*. The MIT Press (Cambridge & Lindon), 2013.

BrodeFrank, J., Blickhan, S., and Rother, B. (2019). Crowdsourcing Knowledge: Interactive Learning with Mapping Historic Skies. *Museums & the Web 2019* [online]. Available at https://mw19.mwconf.org/paper/crowdsourcing-knowledge-interactive-learning-with-mapping-historic-skies/ (Accessed 1 June 2019).

Brohan, P. (2012, July 23). One million, six hundred thousand new observations [Blog post]. Available at https://blog.oldweather.org/2012/07/23/one-million-six-hundred-thousand-new-observations/ (Accessed 24 May 2019).

Brumfield, B.W. (2012, March 05). Quality Control for Crowdsourced Transcription [Blog post]. Available at http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html (Accessed 2 June 2019).

Carrasco, R.C. (n.d.) 2.3 Computing error rates [Online]. Available at https://sites.google.com/site/textdigitisation/home (Accessed 15 May 2019).

Causer, T., Tonra, J., and Wallace, V. (2012). Transcription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham*. *Literary and Linguistic Computing*, 27(2), pp. 119-137. Available at https://doi.org/10.1093/llc/fqs004 (Accessed 1 June 2019).

Causer, T., and Terras, M. (2014) 'Many hands make light work. Many hands together make merry work': Transcribe Bentham and crowdsourcing manuscript collections. In M. Ridge (Ed.). *Crowdsourcing our Cultural Heritage* (pp. 57-88). Farnham: Ashgate.

Derolez, A. (2003). *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. Cambridge: Cambridge University Press.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2), pp. 189-200. Available at https://doi.org/10.1177/0165551512437638 (Accessed 4 May 2019).

Freitag, A., Meyer, R., and Whiteman, L. (2016). Strategies Employed by Citizen Science Programs to Increase the Credibility of their Data. *Citizen Science: Theory and Practice* 1(1), p. 2. Available at http://doi.org/10.5334/cstp.6 (Accessed 3 June 2019.

Fitch, C. (1833). *Letter from Charles Fitch, to Amos Augustus Phelps* [Online]. Available at https://ark.digitalcommonwealth.org/ark:/50959/5h740422b (Accessed 1 June 2019).

Grayson, R. (2016). A life in the trenches? The use of operation war diary and crowdsourcing methods to provide an understanding of the British army's day-to-day life on the western front. *British Journal for Military History*, 2(2), pp. 160-185. Available at https://bjmh.org.uk/index.php/bjmh/article/view/96 (Accessed 1 June 2019).

Hanson, D. (2018). Combining Human and Machine Transcriptions on the Zooniverse Platform. Unpublished master's thesis. University of Minnesota, Minneapolis, US.

Hanson, D. and Simenstad, A. (2018). Combining Human and Machine Transcriptions on the Zooniverse Platform. *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 215-216). Available at https://www.aclweb.org/anthology/W18-6129 (Accessed 2 June 2019).

Hedges, M., and Dunn, S. (2017) *Academic Crowdsourcing in the Humanities: Crowds, Communities and Co-production.* London: Routledge.

Hennon, C., et al. (2015). Cyclone center: can citizen scientists improve tropical cyclone intensity records? *Bulletin of the American Meteorological Society*, 96(4), pp. 591–607.

Howe, J. (2006, June 1). The Rise of Crowdsourcing. *Wired*. Available at: https://www.wired.com/2006/06/crowds/ (Accessed 24 May 2019).

Johnson, L. et al. (2015). PHAT stellar cluster survey. II. Andromeda project cluster catalog. *The Astrophysical Journal*, 802(2), p. 127.

Krawczyk, C. (2018, November 26). Aggregating Annotations in the ASM Project [Blog post]. Available at https://www.bpl.org/blogs/post/aggregating-annotations-in-the-anti-slavery-manuscripts-project/ (Accessed 14 May 2019).

Kuchner, M.J., et al. (2017). The First Brown Dwarf Discovered by the Backyard Worlds: Planet 9 Citizen Science Project. *The Astrophysical Journal Letters* 841, L19.

Lintott, C., et al. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society,* 389(3), pp. 1179-1189. Available at https://doi.org/10.1111/j.1365-2966.2008.13689.x (Accessed 19 May 2019).

Owens, T. (2014). Making Crowdsourcing Compatible with the Missions and Values of Cultural Heritage Organizations. In M. Ridge (Ed.). *Crowdsourcing our Cultural Heritage* (pp. 269 - 280). Farnham: Ashgate.

Ridge, M. (Ed.). (2014). *Crowdsourcing our Cultural Heritage.* Farnham: Ashgate.

Schwamb, M. E., et al. (2013). Planet Hunters: A Transiting Circumbinary Planet in a Quadruple Star System. *The Astrophysical Journal* 768, p. 127.

Smith, D.A. and Cordell, R. (2018). A Research Agenda for Historical and Multilingual Optical Character Recognition [Online]. Available at http://hdl.handle.net/2047/D20298542 (Accessed 04 June 2019).

Swanson, A., Kosmala, M., Lintott, C., and Packer, C. (2016). A Generalized Approach for Producing, Quantifying, and Validating Citizen Science Data from Wildlife Images.

Conservation Biology, 30, pp. 520-531. Available at https://doi.org/10.1111/cobi.12695 (Accessed 10 June 2019).

Terras, M. (2016). Crowdsourcing in the Digital Humanities. In S. Schriebman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities, 2nd ed.* (pp. 420 – 439). New York: John Wiley & Sons Inc.

Trouille, L., Lintott, C.J., Fortson, L.F. (2019). Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human-machine systems. *PNAS*, 116(6). Available at https://www.pnas.org/content/pnas/116/6/1902.full.pdf (Accessed 10 June 2019).

Van Hyning, V. (2016, February 24). 'What's up with those grey dots?' you ask [Blog post]. Available at https://blog.shakespearesworld.org/2016/02/24/whats-up-with-those-grey-dots-you-ask/ (Accessed 26 May 2019).

Van Hyning, V. (2019). Harnessing Crowdsourcing for Scholarly and GLAM Purposes. *Literature Compass,* 16(3-4). Available at https://doi.org/10.1111/lic3.12507 (Accessed 26 May 2019).

Van Hyning, V., Blickhan, S., Trouille, L., and Lintott, C. (2017). Transforming Libraries and Archives through Crowdsourcing. *D-Lib Magazine,* 23(5/6). Available at http://www.dlib.org/dlib/may17/vanhyning/05vanhyning.html (Accessed 15 May 2019).

Verma, S. (2019). Digitization of Transcription on the Zooniverse Crowdsourcing Platform. Unpublished master's thesis. University of Minnesota; Minneapolis, MN.

Weston, A.W. [1838]. *Letter from Anne Warren Weston, Philadelphia, to Deborah Weston, 5 Mon.[?], January evening [1838?]* [Online]. Available at https://ark.digitalcommonwealth.org/ark:/50959/wm117z54p (Accessed 1 June 2019).

Willi, M. et al. (2018). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1). Available at https://doi.org/10.1111/2041-210X.13099 (Accessed 25 May 2019).

Williams, A.C. (2014). A Computational Pipeline for Crowdsourced Transcriptions of Ancient Greek Papyrus Fragments. *2014 IEEE International Conference on Big Data* (pp. 100-105). Available at https://doi.org/10.1109/BigData.2014.7004460 (Accessed 12 May 2019).

Wright, D. et al. (2017). A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society*, 472(2), pp. 1315-1323. Available at https://doi.org/10.1093/mnras/stx1812 (Accessed 25 May 2019).

Zevin, M., et al. (2017). Gravity Spy: integrating advanced ligo detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6), 064003.