DHQ: Digital Humanities Quarterly
2014
Volume 8 Number 1

Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers

Carolyn Strange <carolyn_dot_strange_at_anu_dot_edu>, Australian National University
Daniel McNamara
Josh Wodak <josh_dot_wodak_at_anu_dot_edu_dot_au>, Australian National University
Ian Wood <ian_dot_wood_at_anu_dot_edu_dot_au >, Research School of Computer Science, Australian National University

Abstract

Digital humanities research that requires the digitization of medium-scale, project-specific texts confronts a significant methodological and practical question: is labour-intensive cleaning of the Optical Character Recognition (OCR) output necessary to produce robust results through text mining analysis? This paper traces the steps taken in a collaborative research project that aimed to analyze newspaper coverage of a high-profile murder trial, which occurred in New York City in 1873. A corpus of approximately one-half million words was produced by converting original print sources and image files into digital texts, which produced a substantial rate of OCR-generated errors. We then corrected the scans and added document-level genre metadata. This allowed us to evaluate the impact of our quality upgrade procedures when we tested for possible differences in word usage across two key phases in the trial's coverage using log likelihood ratio [Dunning 1993]. The same tests were run on each dataset – the original OCR scans, a subset of OCR scans selected through the addition of genre metadata, and the metadata-enhanced scans corrected to 98% accuracy. Our results revealed that error correction is desirable but not essential. However, metadata to distinguish between different genres of trial coverage, obtained during the correction process, had a substantial impact. This was true both when investigating all words and when testing for a subset of “judgment words” we created to explore the murder’s emotive elements and its moral implications. Deeper analysis of this case, and others like it, will require more sophisticated text mining techniques to disambiguate word sense and context, which may be more sensitive to OCR-induced errors.

Introduction

Digitized historical newspapers have enriched scholars’ capacity to pose research questions about the past, but the quality of these texts is rarely ideal, due especially to OCR errors. Although scholarly concern about this problem in relation to the use of large, web-based historical resources is growing [Hitchcock 2013], research that involves the digitization of smaller corpora of original print documents confronts the same challenge. The perennial question in digital humanities research – how accurate must digitized sources be to produce robust results – arose in the course of our attempt to interrogate the meanings of a highly publicized murder case prosecuted in New York in 1873. This paper traces how we addressed this challenge: firstly by searching for the techniques best suited to digitize and to interpret news coverage of the trial; and secondly, by applying a statistical tool to test for possible shifts in popular appraisals of the case. 1

Our project transformed over several years, from an individual, archive-based inquiry into a text mining collaboration that required the conversion of original news accounts into searchable texts through OCR. Our initial scans had an error rate of 20%, which raised the prospect that text mining might not substantially enhance our capacity to analyze word usage in the case’s coverage. Consequently, we proceeded to reduce noise through manual correction of the scans and through the addition of genre metadata, thereby creating three datasets – the original OCR scans, a subset of the OCR scans selected through the addition of genre metadata, and the metadata-enhanced scans corrected to 98% accuracy. 2

By analysing each dataset with the log likelihood ratio statistical tool, we determined that the labour-intensive work of cleaning the data modestly improved the reliability of our test – to establish whether or not popular judgment of the case altered after controversial evidence was introduced in the course of the murder trial. However, the addition of genre-related metadata proved to be considerably more significant. Most importantly, the digitization of the previously unsearchable primary sources did make it possible to pose a research question that could not have been answered persuasively using unsearchable texts.

3 As Tim Hitchcock argues, the development of digital history into a discipline requires that we expose and evaluate the research processes that allow us to compose "subtle maps of meaning" from piles of primary sources [Hitchcock 2013, 20]. Accordingly, we begin with an account of the murder case in Section 1, and discuss how the digitization of its news coverage opened up new ways to mine its meanings. In Section 2, we discuss the nature of our corpus and the ways in which we produced machine-readable text using Adobe Photoshop Lightroom 3 and ABBYY FineReader 11. In Section 3, we outline the nature of the errors produced in the scanning processes, while Section 4 details the steps we took to correct them and to add genre metadata post OCR. Section 5 explains how we used the log likelihood ratio tool to analyze word frequency and to test the use of judgment words in trial coverage, using our three different datasets. The results of our tests appear in Section 6, which is followed by a discussion of the possible future directions of text mining research based on small- to medium-sized corpora. We conclude that text mining can enrich historians' capacity to analyze large bodies of text, even in the presence of OCR-induced errors. Supplementing digitized text with genre metadata permits a finer-grained and more reliable analysis of historical newspapers. Investing the time to produce clean data and metadata improves performance, and our study suggests that this is essential for more sophisticated analysis, such as language parsing. Finally, our project underlines the need for interdisciplinary teams to ensure the integrity of the digital tools used, as well as the reliability of their outputs' interpretation.

From Historical Analysis to Digital Historical Research

4 The Walworth murder project began in 2003 as a humanistic enterprise conducted by an historian of gender and criminal justice who read over four hundred newspaper accounts of the case on microfilm, which were reproduced by printing out hard copies of images. This body of unsearchable records was augmented as the number of digitized newspapers available through open-source and proprietary online databases grew exponentially over the 2000s, although many of those texts were unsearchable image files. [1] A research grant made it possible in 2012 to digitize the entire body of primary sources (the paper-based prints and PDF images of newspapers) through OCR scans. [2]. This funding meant that hypotheses developed in the course of the historian's earlier close reading of the case could be tested in a collaboration that included two hired computer scientists and a digital humanities scholar.

5 The Walworth case's extensive and sensational newspaper coverage indicated that the murder of Mansfield Walworth, a second-rate novelist and third-rate family man, stirred deep feelings, particularly because the killer was his son, Frank. The murder provoked troubling questions: Was it legally or morally excusable for a son to kill his father, no matter how despicable? And why, in a family filled with lawyers and judges (including the murdered man's father, Judge Reuben Hyde Walworth) had the law not provided a remedy [O'Brien 2010]? The event occurred on 3 June 1873, when Frank Walworth, a youth of nineteen, travelled to Manhattan to confront his father, who was recently divorced from his mother. Mansfield Walworth had sent a raft of letters to his ex-wife, full of murderous threats and mad ravings. After intercepting these alarming letters, Frank Walworth shot his father dead, then informed the police that he had done so to save himself, his mother and his siblings. Was the shooter an honourable son? A maudlin youth? Insane? Speculation swirled but the initial response to the murder was one of shock: a refined young man from a highly respectable white family had committed cold-blooded murder [O'Brien 2010]. Scores of headlines announced that this was no ordinary murder but a "PARRICIDAL TRAGEDY."

6 Our working hypothesis was that popular readings of the Walworth murder changed as the trial progressed – from initial horror over the crime of parricide, to an appreciation of domestic cruelty and the menacing nature of the victim. The trial resembled a real-life domestic melodrama [Powell 2004], and its turning point occurred when the victim's vile letters to his ex-wife were read into evidence, as the defence attempted to verify the threat Mansfield Walworth had presented to his son and ex-wife. At this point in the trial, the dead man's profane abuse was recorded by trial reporters for the nation

to read:

> You have blasted my heart and think now as you always thought that you could rob me of the sweet faces of my children and then gradually after a year or two rob me of my little inheritance. You will see, you God damned bitch of hell, I have always intended to murder you as a breaker of my heart. God damn you, you will die and my poor broken heart will lie dead across your God damned body. Hiss, hiss, I'm after you… I will kill you on sight.

What was the impact of this obscene evidence on the public's judgment of the case? Did it change over the course of the trial, and if so, how? These were research questions best addressed through the mining of the newspaper coverage, a substantial corpus beyond the capacity of human assessment. 7

Although the volume of the Walworth case's news coverage was modest compared to large-scale, institutionally funded text mining projects, the standards set in several benchmark historical newspaper text-mining research projects informed our approach. Many, such as Mining the *Dispatch*,[3] use manually double-keyed documentation, followed by comparison-based correction, to produce datasets with 98+% accuracy. Mining the *Dispatch* used a large corpus of nineteenth-century U.S. newspapers to "explore – and encourage exploration of – the dramatic and often traumatic changes as well as the sometimes surprising continuities in the social and political life of Civil War Richmond." 8

That project combined distant and close reading of every issue of one newspaper (112,000 texts totalling almost 24 million words) "to uncover categories and discover patterns in and among texts." [Nelson 2010] As its director explained, high accuracy levels were necessary to combine text mining with historical interpretation most productively: "the challenge is to toggle between distant and close readings; not to rely solely on topic modelling and visualizations." [Nelson 2010] Studies that pursue similar objectives must first determine the best methods to produce machine-readable text. The next section details the scanning process preliminary to our analysis. 9

The Production of Machine-searchable Texts

Using a variety of sources, including photocopied prints of microfilmed newspapers and PDF image files of stories sourced from several databases, we gleaned 600 pages, comprising approximately 500,000 words of digitized text. Figure 1 shows an example article ready for scanning. 10

Since optimal machine-readable text was integral to our analysis of the murder trial's meanings, we reviewed the quality control methods used by ten large public and private institutions, from scanning hardware and software through to a diverse array of OCR software.[4] We intended to use non-proprietary software,[5] but we ultimately selected ABBYY FineReader 11, since it allows for the customisation of features.[6] The first source of text was photocopied pages printed in 2003 from microfilmed newspaper images. Since online repositories of newspapers have subsequently become more numerous, we were able to replace the poor quality text in microfilm scans with PDF image files of the photocopies. However, this strategy proved to be too time-consuming, since each of these page-scan PDFs contained upwards of 60 paragraphs of text (an average of 5,000 words in small print), spread across upwards of 8 columns, which required laborious searches for references to the Walworth murder trial. Consequently, this replacement strategy was used only for the worst 10% of the photocopies. 11

All the scanned files viable to use from the existing scanned microfilm were imported into Adobe Photoshop Lightroom 3. This process involved batch scanning pages in black and white into 300DPI TIFFs according to newspaper, and then placing them in physical folders by newspaper, in the same order in which they had been scanned, to facilitate cross-referencing of particular pages with their corresponding files. Within Lightroom each file was then manually cropped one-by-one to select only those columns related to the Walworth murder.[7] Although Lightroom is designed for working with large catalogues of photographs rather than "photographs of text", it allowed us to batch process select images iteratively through non-destructive image editing, so that tests could be made to determine which combination of image processing was likely to achieve the highest OCR accuracy. 12

THE WALWORTH MURDER.

The Story of a Member of the Bereaved Family.

Scenes Beside Chancellor Walworth's Coffin.

A Letter Written in Blood and Sanded with Powder.

A Wife's Silent Agony—Another Life To Be Saved at the Expense of the Memory of the Dead—Sad and Stern Facts which Must Come Out.

The Walworth tragedy, which a few days ago startled the community to its depths, has not yet lost any of its morbid interest for the public. Every little detail and surrounding of the horrible murder, the antecedents of the victim and his slayer, even everything that remotely concerns the bereaved family, is discussed with that unhealthy gout which only a satiety of crime can produce. The demeanor of the wretched prisoner, the characteristics of the father and the sufferings of the unfortunate mother are still subjects of all-absorbing interest in gossiping circles.

THE Wal-**worth** MURDER.

The Story of a Member of the Bereaved Family.

Scenes Beside Chancellor Walworth's Coffin.

A **hotter** Written in Blood and Sanded with Powder.

A Wife'**i Bite** Agony—Another Life To Be **fisved** at the Expense of the Memory of the Dead-Sad and **Stem** Facts which **Mwt** Come Out.

The walworth tragedy, which a few days **agox** •**turned** the community to its depths, **hag** not yet **toat** any of its morbid **intercut** for the public. **Kverjr UlUe**

Figure 1. An example of one of the newspaper articles on the Walworth case (from the *New York Herald*, 10 June 1873), showing average level image degradation. The OCR output appears to the right, with errors highlighted in bold.

Training software to recognize patterns of font and content is one of the most challenging aspects of OCR, as it draws on Artificial Intelligence to "recognize" multitudes of shapes as belonging to corresponding letters. Our project revealed that ABBYY FineReader's training capacity is limited. After we exported files produced through Lightroom, newspaper by newspaper,[8] we trained ABBYY to recognize each newspaper's fonts, and each file was further "cleaned up" by straightening text lines and by correcting for perspective distortion. Although ABBYY appeared to "recognize" frequently occurring words, like the surname Walworth, the OCRd results produced variations, such as "Wolwarth" and "Warworth". It became evident that ABBYY cannot recognize that all such variations of "Walworth" should have been converted automatically to "Walworth", given the high statistical likelihood that they were in fact "Walworth". 13

Our process exposed the sorts of image degradation common in the digitization of historical newspapers, including: smudged, faded and warped text; ripped or crumpled originals; image bleed from the reverse side of the paper; crooked and curved text lines; and overexposed and underexposed microfilm scans. As a result, ABBYY's deficiencies required that customized automated corrections be applied in the post-OCR phase of our project. 14

Nevertheless, the production of machine-readable text resulted in a uniform dataset of newspaper articles that offers considerable granularity, including the capacity to analyze a corpus of articles on the Walworth murder trial according to date, a critical factor in our study, considering the admission of Mansfield Walworth's extraordinary letters into evidence. The coverage of the corpus, disaggregated by newspaper, is shown in Figure 2. 15

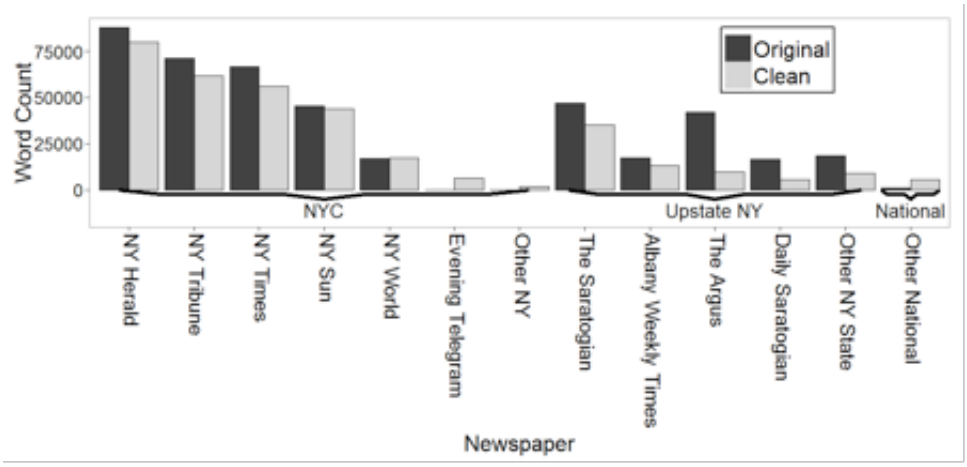

Figure 2. Word counts per newspaper, original text from OCR scans (left) and cleaned text (right). The correction process is discussed in Section 3.

OCR Errors and Quality Control for Text Mining

The variable quality of digitized historical newspapers has long been a challenge for digital scholarship [Arlitsch2004] Much of that variability is associated with the historical and contemporary resources of publishing houses, meaning that major metropolitan papers typically sit at one end of the legibility spectrum and smaller, regional papers sit at the other. In our study, articles from the *New York Times* yielded accuracy levels of 94.5%, and they are obtainable through the paper's own search engine. Furthermore, articles in the *Times* repository are cropped and cleaned, which means they are ready for OCRing with minimal image manipulation. In contrast, OCR scans from an important upstate New York newspaper based in the state capital, the *Albany Argus,* yielded results of only 65% accuracy. Unfortunately, due to the substantial additional labour required to raise this and other smaller papers' level of accuracy, these scans were mostly too poor to incorporate. Thus the variation in file quality from different newspapers presented a limitation on the project's initial ambitions. More broadly, this problem flags the significant impact that OCR quality can make in the range of sources used for text mining. OCR errors are part of a wider problem of dealing with "noise" in text mining [Knoblock 2007], which may also stem from other sources such as historical spelling variations or language specific to different media texts. 16

The impact of OCR errors varies depending on the task performed, however [Eder 2013]. The tasks of sentence boundary detection, tokenization, and part-of-speech tagging on text are all compromised by OCR errors [Lopresti 2008]. As Lopresti concludes: "While most such errors are localised, in the worst case some have an amplifying effect that extends well beyond the site of the original error, thereby degrading the performance of the end-to-end system." Another study performed document clustering and topic modelling on text obtained from OCR [Walker et al. 2010]. These authors found that for the clustering task the errors had little impact on performance, although the errors had a greater impact on performance for the topic modelling task. A study involving the task of stylistic text classification found that OCR errors had little impact on performance [Stein et al. 2006]. In contrast, Eder advises that "tidily prepared corpora are integral to tests of authorship attribution" [Eder 2013, 10]. Thus, the relevance of scanning errors remains a matter of debate. 17

Some studies of the effect of OCR errors [Lopresti 2008] [Walker et al. 2010] [Stein et al. 2006] have conducted comparisons by analysing two corpora, identical except for corrections of individual words. Our study was distinct in two respects. First, it analyzed the effect of OCR corrections on corpora at the word level, and it removed duplicate, irrelevant and very poorly scanned text. We then added genre metadata and verified newspaper and date metadata. From the point of view of a scientific experiment about the effect of OCR errors, these extra steps may be considered "confounding variables". In contrast, projects such as ours make these corpus preparation steps necessary, since questions of content as well as subtleties of word use are both critical. Second, rather than conducting "canonical" tests, such as document classification tasks through supervised machine learning, we selected key word analysis with log 18

likelihood ratio significance testing. These decisions situated our text analysis in a real-world digital humanities workflow.

19 The accuracy of character recognition at the word level is especially significant in projects that involve the interpretation of sentiment [Wiebe 2005]. Words that appear rarely, as opposed to ones that appear most frequently, tend to convey deep meaning, particularly words associated with intense emotions, such as anger or disgust [Strapparava and Mihalcea 2008]. Because we attempted to determine the Walworth case's meanings for contemporaries, including their moral judgments of the principals, we considered our initial scanning error rate of 20% to be unacceptable. This assessment led us to invest the time required to clean the text manually after the OCR process by correcting errors at the character level as well as removing duplicate and irrelevant text. Additionally, because we expected that opinion pieces such as editorials and letters to the editor would provide the clearest indication of public perception of the Walworth case, we added genre metadata to the corpus as a supplement to the cleaning process. We then conducted the log likelihood ratio comparison of word frequency across two phases of the case's reportage, both to analyze the impact of the cleaning process and the addition of genre metadata, and to test our historical hypothesis through text mining.

Correcting OCR-induced Errors and Adding Genre Metadata

20 This section discusses the strategies we undertook to reach a level of accuracy comparable to that achieved in benchmark historical newspaper text mining projects. It also explains how and why we added genre-based metadata before we performed analysis using log likelihood ratio [Dunning 1993].

21 Measuring the accuracy of OCR scans can be conducted at both the character and word level, which is determined by dividing the number of units that are correct by the total number of units [Rice et al. 1993]. Calculating such accuracy involves hand-labelling all characters and words with their correct values and is very time consuming, however. To avoid this evaluation step, a word accuracy approximation can be measured as a proportion of words appearing in a standard dictionary.[9] This approach does not consider two opposing factors: those words which are correct but not in the dictionary, and those that are incorrect but in the dictionary. Despite this limitation, a reliable indication of the digitized text's accuracy is possible.

22 Because the coverage of the Walworth case included proper names and archaic terminology, it was unrealistic to anticipate 100 per cent accuracy. Words that were split or joined through OCR errors were another confounding factor in this estimation of accuracy. For example, in one instance the word "prosecution" was split into two words ("prosec" and "ution") by the OCR scan, while in another case the words "was severely" were merged into one garbled word, "was" "Aevertyy". To assess this effect, we calculated that the average word length for the uncorrected (5.84 letters) and corrected (5.68 letters) texts had approximately a 3% error rate, which we deemed small enough to ignore for the purposes of our study.

23 Table 1 shows the approximate word accuracy calculated according to this method, both before correction and after the corrections, which we describe in the remainder of this section. The pre-correction accuracy was comparable to the 78% achieved for the British Library's 19th Century Online Newspaper Archive [Tanner et al. 2009]. The post-correction accuracy is near the target of 98% used by the National Library of Australia Newspaper Digitization Program [Holley 2009].

	Words	Words in dictionary	Words not in Dictionary	Approximate Word Accuracy
Original	478762	391384	87378	81.7%
Clean	345181	336779	8402	97.6%

Table 1. Effect of post-OCR correction on accuracy.

24 A process of manual correction was undertaken to remove the errors generated through OCR, because we sought a clearer signal in the analysis of the texts. Working with "noise", whether induced by OCR or from other sources such as

spelling variations or language variants used on social media, is common in the fields of text mining and corpus linguistics [Knoblock 2007] [10] However, historical interpretation relies on data sufficiently clean to boost the credibility of the analysis at this scale. Automated techniques were used in a limited way, but to achieve results at the high standard desired, we determined that manual correction was essential.

Manual correction offered the benefit of removing duplicate and irrelevant sections of text; in addition, it allowed us to add document-level metadata tags, which is a critical step in complex text analysis. For practical reasons a single corrector was used, but to achieve even greater accuracy, multiple correctors could be used and their results compared. 25

The post-OCR correction process entailed five steps: 26

1. Simple automatic corrections were made. These included: the removal of hyphens at line breaks, which are mostly a product of words appearing across lines; correcting some simple errors (such as “thb” → “the”); and the correction of principal names in the text, such as “Walworth” or “Mansfield”. Full stops not marking the end of sentences were also removed to permit the documents to be broken into semantically meaningful chunks using the full stop delimiter.
2. Articles with an approximate word accuracy below a threshold, set to 80%, were in general discarded to speed the correction process. However, those falling below the threshold, but hand-selected for their rich content, were retained.
3. The text was corrected by hand, comparing the original image file and the post-OCR text version of the same articles.
4. Duplicate and irrelevant text was removed.
5. Metadata tags for article genre were added, broken into four categories: “editorial”; “incidental reportage”; “trial proceedings”; and “letter to the editor”. Previously added tags for the name of the newspaper and date were also verified and corrected where required.

Automated correction using search and replace with regular expressions was necessarily limited to avoid introducing new errors, since we considered a garbled word preferable to a “correction” leading to a wrong word. We anticipated that the clear patterns in the observed errors would lend themselves to more sophisticated correction processes using supervised machine learning techniques. However this application proved beyond the scope of this project. 27

Given the modest size of the corpus and the research funding available it was feasible to hand-tag genre, delivering accuracy benefits over automated approaches. Although we considered automatic inclusion of metadata (for example, within the TEI standard) as well as automatic part-of-speech tagging (valuable for tasks such as document classification), we determined that plain text plus article-level genre/date/newspaper metadata was sufficient for keyword analysis in our project. 28

Manual correction is inescapably a time-consuming process, although it does offer collaborative benefits, since it involves all team members in the close examination of texts. Some projects opt to offshore OCR correction, but ethical considerations concerning the exploitation of foreign labour as well as quality control concerns ruled out this option in our study. The efficiency of inputting corrections was improved by using spelling and grammar error highlighting in Microsoft Word. This phase took approximately one hundred hours, at an average rate of 57.5 words per minute, which is comparable to that of an efficient typist. Although this procedure was efficient for moderately corrupted text, and easier to sustain over long work sessions, highly inaccurate scans rendered typing from scratch necessary, as it was quicker than correcting the garbled OCR output. When added to the lengthy OCR scanning process, the labour required to correct scanned text does raise the question of whether OCR is the most efficient way to digitize a medium-size corpus of historical newspapers to a high degree of accuracy. 29

As well as typing from original scans, we transcribed texts using a voice recognition program (Dragon Naturally Speaking 12), another option for the correction and input process. Typing was predominantly used, since it tends to be quicker than transcriptions of dictation for corrections. For inputting longer sections from scratch, dictation was slightly faster and more convenient to use. However, it tends to fail “silently”, in that it substitutes unrecognized words with other words, which a spell-checker cannot detect. Typographical errors, on the other hand, are more likely to form non-words 30

that spell checkers can identify. Dictation is also more likely to fail on names and uncommon words and proper nouns, precisely those words which the study is most interested in identifying.

In summary, while OCR achieves relatively accurate results (around 80%) on historical newspaper collections such as the one used in this study, manual correction is required to achieve high accuracy (around 98%). Depending on the corpus size and the resources at hand, this two-step process may be no more efficient than directly inputting the original texts from scratch. 31

Measuring the Effect of Post-OCR Correction Using a Sample Task

Determining the tenor of the Walworth case's newspaper coverage and testing for possible shifts over the course of trial was the object of our text mining analysis, but the methods we selected to do so are relevant to wider debates over the utility of OCR and post-OCR correction processes. In order to evaluate changes in the popular assessment of the case we created two subsets of the digitized corpus: Phase I (news accounts before the introduction of Mansfield Walworth's shocking letters), and Phase II (trial coverage subsequent to the letters' introduction, including Frank Walworth's conviction and sentence of life in prison).[11] We investigated which words varied at statistically significant rates from Phase I to Phase II, particularly those indicative of the sentiments stirred by the crime and the characters involved. To undertake this analysis we used a list of "judgment words". Through a close reading of the texts and knowledge of common words used in criminal trial reportage in this period the historian produced a preliminary list of words of moral judgment and character assessment, which we supplemented through the addition of similar words selected with the aid of topic modelling of the corpus. Finally, we further augmented our list by adding other forms of the selected words that appeared in the 2011 edition of the American English Spell Checker Oriented Word Lists.[12] We chose the statistical tool log likelihood ratio, since it is designed to measure variation in the word frequency between two sections of a corpus [Dunning 1993]. Most importantly, log likelihood ratio discerns statistically significant word frequency variations which are highly likely to appear as a result of true properties of the corpora, rather than by chance. By calculating log likelihood ratio across the two phases of newspaper reportage, we tested for changes in the popular judgment of the Walworth case; this test also allowed us to analyze the effectiveness of post-OCR cleaning by comparing the results of the task performed on the text before and after correction.[13] 32

Log likelihood ratio, which identifies meaningful variation in word frequency in one corpus relative to another [Dunning 1993], produces a p-value on the corresponding test statistic, which can be interpreted as the probability of the observed word frequencies, given the null hypothesis that there is no difference between the two corpora. For words with a p-value below some significance level (for example $p \leq 0.05$) the null hypothesis may be rejected; in other words, the difference in word frequency between the two corpora may be considered statistically significant when the variation is highly unlikely to be a result of chance. However, it is worth noting that while this holds for any given word, if we use a given significance level to select a set of words, it may still be likely that the result for at least one of the selected words may appear to be significant by chance alone. Multiple hypothesis testing provides a rubric for managing this phenomenon, for example by reducing the p-value used for individual words. We chose not to pursue this approach; instead we qualitatively analyzed words, identified by log likelihood ratio, which helped to detect the minority of words incorrectly identified as significant. Because we worked as an interdisciplinary team, the historian contributed to this critical examination of the output of a statistical technique. 33

Dunning introduced the log likelihood ratio as a tool for word frequency analysis that would be more robust than the previously prevalent chi-squared test for small samples of text. It has been used in previous studies comparing corpora, for example looking at the proceedings of a 19th century British murder trial [Archer forthcoming]; the distinctive lexicon used in a professional environment [Rayson and Garside 2000] [14]; historical spelling variations [Baron 2009]; and the lines of a particular character in a play [McIntyre 2010]. We did consider other tests, which have been proposed as alternatives to the log likelihood ratio. For instance, Fisher's exact test [Moore 2004] calculates exactly what the log likelihood ratio approximates but requires greater computational resources, while the Mann-Whitney Ranks test [Kilgarriff 2001] considers the distribution of word frequency within a corpus, as does the t-test [Paquot and Bestgen 2009]. After reviewing these options we decided that log likelihood ratio was the best option, due to its well-established 34

use in the comparison of corpora.

As described in Section 4, the correction process was enhanced through the inclusion of metadata about article genre. The two phases differed in genre mix, since the reportage from Phase II was dominated by coverage of the Walworth case trial proceedings. While we were interested in detecting changing public opinion, differences in genre could possibly have obscured the shift we anticipated. Where genre metadata was available, we restricted our analysis to opinion articles about the case, consisting of editorials and letters to the editor, since this genre is most likely to capture words of interest. Furthermore, technical judgment words appearing in trial proceedings – particularly those used by lawyers and the judge in court – indicate legal constructions that may not have reflected public opinion. This caveat was another reason for the restriction of the corpus to opinion articles using genre metadata. As Figure 3 shows, we compared the original text without metadata; the original text restricted using genre metadata; and the cleaned text also restricted using genre metadata. 35

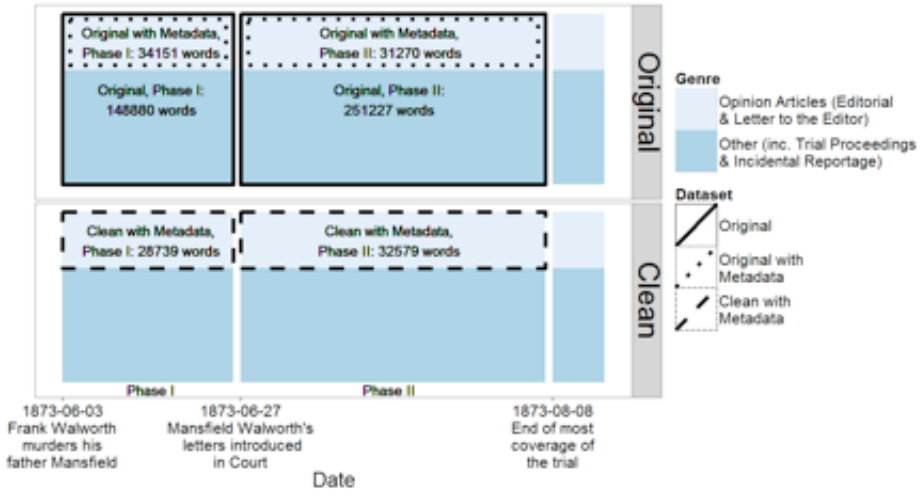

Figure 3. Schematic of the datasets used in the experiments presented in this paper. The word counts for each dataset for Phase I and Phase II are shown.

Pre-processing of the texts was performed to improve the quality of results returned. The following four steps were taken: 36

1. All text converted to lower case.
2. Punctuation removed and the possessive form "'s".[15]
3. Stopwords removed, such as "the" and "of", from a standard stopwords list[16]
4. Words identified from the custom-built list of 357 judgment words, consisting primarily of adjectives, adverbs and abstract nouns.

These steps, as well as the log likelihood ratio calculations, were performed using several open source tools.[17] The results of the experiments are detailed in the following section. 37

Results

The tests we conducted involved comparing newspaper coverage of the Walworth case over our two periods: Phase I (the crime, the arrest, the coroner's inquest and the trial's opening); and Phase II (subsequent to the letters' introduction up to the verdict and sentencing). By using data with and without post-OCR correction we were able to address our historical question and to evaluate the effect of this correction. 38

Figure 4 shows the ten Phase I words that appeared at most significant frequency, measured by log likelihood ratio, for the three datasets presented in Figure 3. The defendant is the focus of early reportage, with terms such as "young", and "son" appearing, as well as the negative word "murderer" (considering that his conviction had not yet occurred). The 39

opinion article genre focuses on the defendant's family background, including the word "chancellor" (Judge Reuben Hyde Walworth, Mansfield's father), "albany" (the state capital, where the defendant's uncle lived) and "literary", the last of which referred to Mansfield Walworth's career as a gothic novelist, rather than his negative character traits. Even without OCR correction, most of these words of interest were identified. Despite the presence of non-words caused by OCR error, they do not appear in these top few words. Without metadata, the words tend to focus on the minutiae of the murder scene, such as "stairs", "body" and "door", rather than on more substantive issues of character. This points to the need for historical researchers to consider adding genre metadata prior to calculating log likelihood ratios.

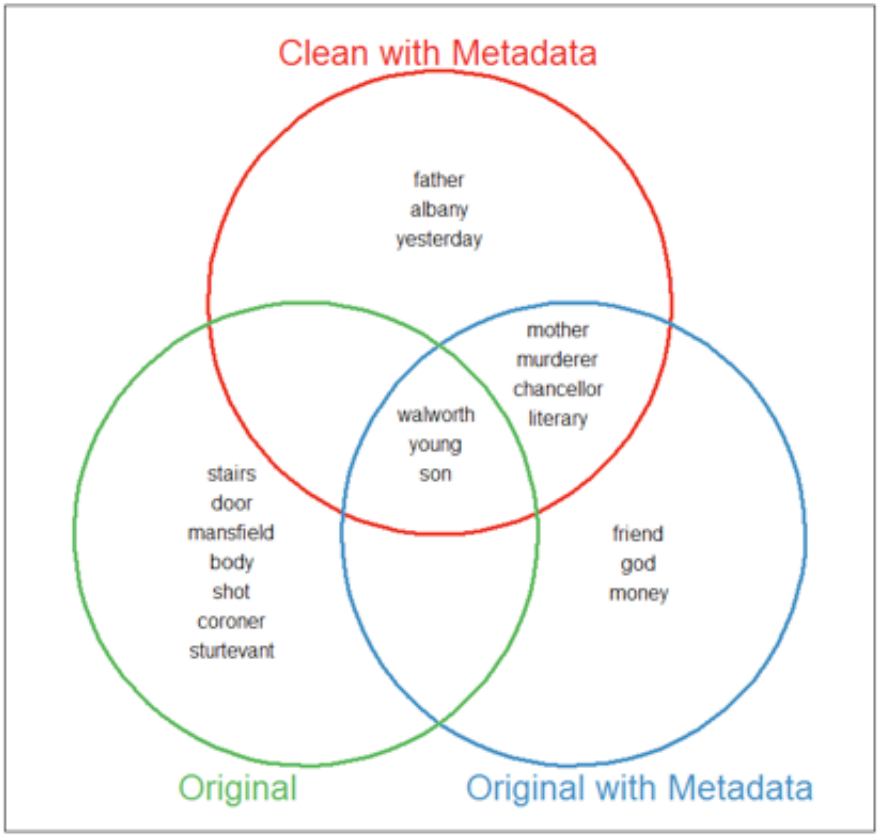

Figure 4. Top 10 Phase I words for each dataset described in Table 2, ranked by log-likelihood ratio. All words shown were significant at p≤0.0001. No judgment words appear.

The same approach for Phase II yielded primarily legal terms, which disclosed little about changing opinion. Therefore, we show in Figure 5 those words from our judgment word list that occurred more frequently in this second trial period at a statistically significant rate (using a significance level of p≤0.05). The words "insanity" and "insane" may refer to both Frank and Mansfield, since the defence suggested that the son may have suffered from a form of madness inherited from his disturbed father. The terms "threats" and "madman" reflect a new focus on the condemnation of Mansfield, although there is some possibility "madman" could also refer to Frank. The words "deliberation" and "deliberate" may negatively describe Frank's actions, but they may equally be procedural legal terms relating to the jury. The differences between the datasets are less pronounced in this experiment, aside from the fact that the original dataset contained more statistically significant words, since it includes substantially more words overall (see Table 1for details). Some of these words suggest a condemnation of Mansfield ("demon") and potential approval of Frank ("honor"), since he claimed he had killed his father to protect his mother. This pattern suggests that using judgment words may be an alternative to adding genre metadata, since these words implicitly refer to genre.

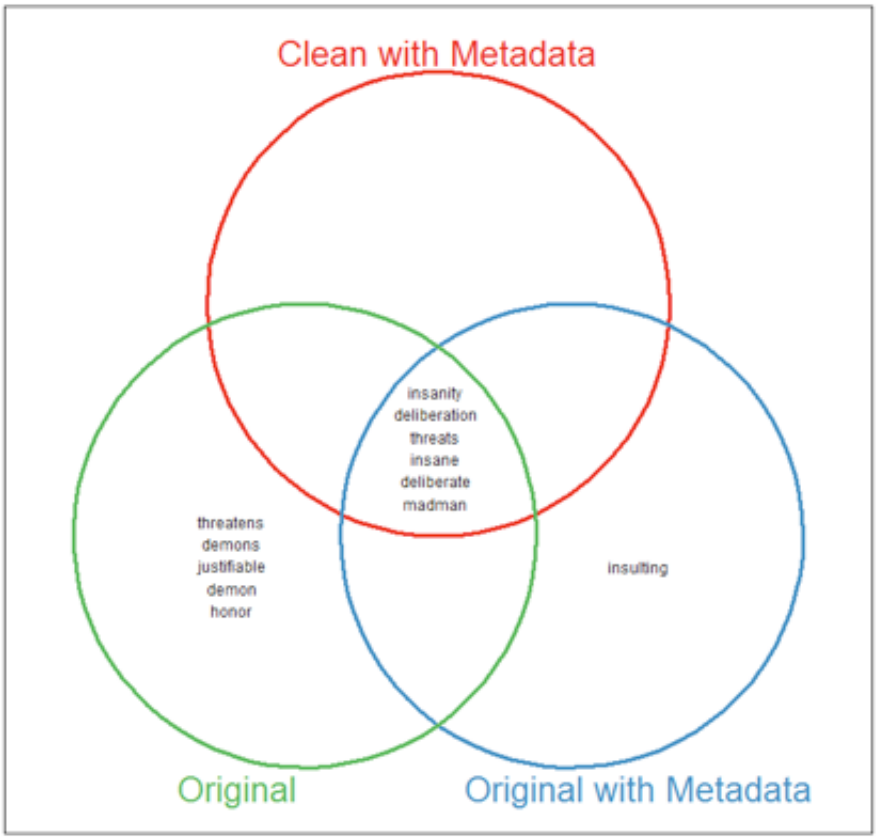

Figure 5. Frequent Phase II words for each dataset described in Table 2, using judgment words which are statistically significant at p≤0.05.

Fortunately, ambiguities such as whether “insanity” refers to Frank or Mansfield, or whether “deliberate” refers to Frank’s shooting, the judge or the jury, may be resolved using more sophisticated techniques. For example, words may be matched to characters in the case through sentence blocks, word proximity, or full-scale parsing for semantic structure. Such techniques typically require very high quality text to be effective. For words with relatively low frequencies, manually investigating the contexts of occurrences can also be used.

Significant Words	Precision	Recall	Significant Judgment Words	Judgment Words Precision	Judgment Words Recall	
Original with Metadata	751	0.48	0.66	23	0.65	0.93
Original	2852	0.12	0.61	38	0.34	0.81

Table 2. Performance of Original with Metadata and Original Datasets compared to Clean with Metadata dataset. The Clean with Metadata dataset returned 545 significant words of which 16 were judgment words. A significance level of p≤0.05 was used.

Table 2 shows the precision and recall of the results of words with a significance level of p≤0.05 for the original with genre metadata and original datasets compared to the “gold standard”, that is, the clean with metadata dataset. Recall refers to the proportion of words significant in the clean with metadata dataset, which are also significant in the original (with or without metadata) dataset. Precision is the proportion of words significant in the original (with or without metadata) dataset, which are also significant in the clean with metadata dataset. This evaluation methodology allowed us to drill down deeper than we could by using the small shortlist of words shown in Figure 4 and Figure 5, and it

revealed a strong discrepancy between the datasets.

The precision scores of 0.48 and 0.12 indicate that many words were incorrectly identified as significant, while the recall scores of 0.66 and 0.61 suggest that a substantial portion of significant words was missed. The original dataset approached the original with metadata dataset on recall, but it had much lower precision, indicating that it returned many results with limited usefulness. Some non-words induced by OCR error appear at a statistically significant level in the original and original with metadata lists, such as “ol” (instead of “of”) and “ihe” (instead of “the”). Using the judgment word list, the precision scores of 0.65 and 0.34 suggest, again, that many “false positive” words were identified, with the problem magnified without adding in genre metadata. The recall results of 0.93 and 0.81 were stronger for the judgment word list, however, which is of interest given the emotive nature of the case and its coverage. Overall, there was a substantial discrepancy in the words identified in the original datasets, with and without metadata, compared to the clean with metadata dataset. This confirms that OCR errors can, indeed, influence later analysis of this nature. 42

It is worth examining in detail one example in which an OCR error produced a judgment word found to be significant using the original with metadata dataset, but not by using the clean with metadata dataset. In the original with metadata dataset, the word “maudlin” was identified as occurring significantly more in Phase I, with a frequency of 3 compared to 0 in Phase II. However, there was one instance of “maudlin” occurring in Phase II in the clean with metadata dataset which was missed due to an OCR error – swapping “maudlin” for “inaudlin”. In the clean with metadata dataset the frequency counts of 3 for Phase I versus 1 for Phase II were not significantly different. While these frequencies may seem low, relative scarcity does not indicate low significance. In fact, our test indicated the opposite to be the case. 43

The contexts of “maudlin” appear in Table 3, which indicates that the uses of the term in Phase I occurred in the context of disapproval of Frank’s parricidal motive and cool demeanor. The use of the term in Phase II was different, we discovered, because it referred to the state of mind of another murderer in an earlier trial, in which a plea of insanity had been successful. This shows that further work is required to identify the implications of word use based on their contexts. Indeed, it seems that there was a suggestive change in the frequency of “maudlin” between Phase I and Phase II. 44

Newspaper	Date	Phase	Correct in Original?	Context
NY Tribune	1873-06-04	I	Yes	“We protest in advance against such resort to maudlin sentimentality”
NY Tribune	1873-06-05	I	Yes	“There’s a [sic] something indefinable about this maudlin sentimentalism that throws a glow of heroism round the murder”
The Saratogian (quoting *NY Tribune*)	1873-06-12	I	Yes	“We protest in advance against such resort to maudlin sentimentality”
NY Tribune	1873-07-04	II	No	“the maudlin sorrow of a drunkard” (referring to another case where insanity was successfully pled, an outcome the author critiques)

Table 3. Contexts of the judgment word “maudlin”.

Overall, the cleaning of the data was not essential to achieving results of interest on the two-phase comparison task, since many significant words could still be identified. Still, there were substantial differences between the results of the clean and original datasets, as significant words were missed and “false positives” were generated. The adding of genre metadata permitted the filtering of more significant words through the use of opinion articles, something that was not possible with the original dataset. Our list of judgment words likely performed a similar filtering function to the genre metadata, though it lacks the flexibility to detect unexpected words. 45

Future Work

Moving beyond the analysis presented in this paper, it would be desirable to identify which words refer to which characters in the case. With the clean corpus we now have at our disposal, this identification could be achieved through the automated tagging of syntactic metadata. This analysis would allow us to track public opinion at the level of the individual with greater precision. Words which may refer to multiple individuals may be disambiguated using techniques such as sentence blocks, word proximity and semantic parsing. It is expected that this more complex task would show greater differences between the raw OCR output and the corrected text, since it depends on the presence of grammatically well-formed sentences rather than word counts alone. 46

An alternative approach that may be useful in similar projects would involve identifying which terms are distinctive "hallmarks" of particular subcorpora (for instance, selected on the basis of date or news source). A feature selection metric such as mutual information could be used to identify which terms are most predictive in classifying documents as belonging to particular subcorpora. Turning from supervised to unsupervised learning, our team anticipates producing results based on topic modelling, a common strategy in the digital humanities which has been applied to US historical newspapers [Newman and Block 2006], [Yang et al. 2011], and [Nelson 2012]. The goal of such projects is to find topics in large volumes of newspaper reportage, and to track changes as indices of shifts in public discourse. The effect of OCR errors on such topic models is also an active subject of research, and our results suggest that scholars consider this issue thoroughly before undertaking large-scale projects [Walker et al. 2010]. 47

The broader ambition of this research project is to situate the Walworth case in its wider historical context. Can it be shown through text mining that prevailing understandings of masculine honour, morality and family values were challenged by this dramatic incident? In future work we will compare our digitized collection with larger newspaper corpora. Google n-grams [18] is a common and accessible choice for researchers, but its contents are different from our corpus in both format (the full text of its sources is unavailable) and in genre (it covers non-fiction, arcane technical writings and literary works). A more promising collection is Gale's Nineteenth Century Newspaper collection.[19] If it becomes fully searchable it will provide a vast dataset from which subsets of texts (such as editorials on domestic homicide) can be selected to evaluate the distinct and shared features of the Walworth case's coverage. Nevertheless, there are reasons for caution. In large corpora such as these, OCR induced errors will remain an issue, since hand correction of texts on a vast scale is infeasible. 48

Conclusion

Digital humanities scholars have been drawn to text mining as a technique well suited to the analysis of historical newspapers, since it allows for meaning to be drawn from volumes of text that would be unmanageable for an individual researcher to absorb and analyze. It provides a tool that can test hypotheses generated through traditional historical analysis, and ideally, generate new possibilities for study that could not have been generated through close reading alone. However, the digitization of historical texts is a complex and time-consuming process which is worthy of consideration in itself. Through the example of the Walworth murder case's newspaper coverage, this paper has outlined the two-step digitization process our team undertook: first, performing OCR scans from original newspapers and image files; and second, cleaning and post-processing to ensure that all text included is accurate, relevant, and labelled with genre metadata. We have provided an original, detailed methodology for conducting digitization of a medium-size corpus. 49

OCR, we determined, is effective in digitising historical newspapers to roughly 80% accuracy. However, to achieve high levels of accuracy (around 98%), the labour-intensive cleaning required to remove OCR errors means the two-step process may be no more efficient than manually inputting texts from scratch, a procedure that suits small- to medium-scale projects. While our research involved both scanning and cleaning texts, historical researchers more commonly perform keyword searches on existing databases of historical documents to conduct text mining analysis [Hitchcock 2013]. Importantly, our study shows that the result set may contain OCR errors, irrelevant and duplicate content; similarly, insufficient metadata can generate spurious results that are difficult to detect. Our method proposed for the cleaning process, as well as our appraisal of the value of this step, signals the way forward to overcome this problem. 50

The value of correcting OCR output from around 80% accuracy to near 100% is an important consideration for 51

researchers, in view of the labour-intensive process required. We demonstrated this empirically by performing a sample task of interest on both the clean and original versions of the corpora. This task involved finding words, including those from a list of pertinent judgment words, which changed in frequency across two phases of the case's reportage. Log likelihood ratio was used as a test for statistical significance. With the uncorrected OCR output it was possible to identify words appearing significantly more frequently in one time period relative to another, but a substantial proportion were missed and "false positives" were introduced. The cleaning was thus desirable but not essential. The addition of genre metadata led to results of greater interest, since it allowed a focus on articles more clearly relevant to the research question. This paper is unique in situating OCR error correction in a digitization workflow also involving content selection, document-level metadata enhancement and practical time and cost constraints, as it evaluates this text cleaning phase holistically.

Like many digital humanities projects, this study underlines the value of input from researchers across the disciplines of history and computer science to design the project, select the methodology, implement the tasks and interpret the results [Ayers 2013]; [Nelson 2012]. Without this combination of skills and expertise, as well as facilitative research funding, such studies are unfeasible. Our team's scientific expertise allowed us to customize software for text mining analysis rather than using off-the-shelf solutions, which gave us full control over the integrity of the tools used, while the historian posed the research question and critically examined test results against the initial close reading of the case. This collaborative, interdisciplinary model will continue to be critical to foster robust research in the field of digital humanities.

Notes

[1] The online sources used for this project were the Library of Congress's Chronicling America, Historic American Newspapers (http://chroniclingamerica.loc.gov/); America's Historical Newspapers (http://www.newsbank.com/readex/?content=96); Gale 19thC US Newspapers (http://gdc.gale.com/products/19th-century-u.s.-newspapers/); and New York State Historical Newspapers – Old Fulton NY Postcards (http://fultonhistory.com/Fulton.html). In addition, photocopies of microfilm were made in 2003 at the NY State Newspaper Project hosted by the New York State Public Library (http://www.nysl.nysed.gov/nysnp/).

[2] Research funding for this project was provided by the Australian Research Council

[3] Mining the *Dispatch* (http://dsl.richmond.edu/dispatch/Topics). Historian Robert K. Nelson directs the University of Richmond's Digital Scholarship Lab, which developed this project.

[4] The 10 public and private institutions are Improving Access to Text; Australian Newspapers Digitisation Program; The Text Creation Partnership; British Newspapers 1800-1900; Early English Books Online; American National Digital Newspaper Program; Project Gutenberg; Universal Digital Library Million Book Collection; and Gale Eighteenth Century Collections Online.

[5] Other commercial software that were evaluated, but were not found to be as suitable as ABBYY include: ExperVision OCR, Vividata, VelOCRaptor, Presto! OCR, OmniPage, Olive, and Prizmo. Other open source software that was evaluated for its suitability was OCRopus, hocr-tools, isri-ocr-evaluation-tools, Tesseract, and GOCR. For a comprehensive list of OCR software see http://en.wikipedia.org/wiki/List_of_optical_character_recognition_software.

[6] We are also grateful for advice provided by the Digitisation Facility at the National Centre of Biography, Australian National University (http://ncb.anu.edu.au/scanner).

[7] Since Lightroom cannot import PDFs all files were sorted and processed in the one application. Only the high quality PDFs of the *New York Times* were "clean" enough to be OCRed directly from the downloaded PDF, so did not require processing through Lightroom.

[8] One unfortunate downside to this workflow is that Lightroom cannot export greyscale images, so it only exported each 20MB TIFF as a 110MB TIFF.

[9] The dictionary was compiled from Kevin Atkinson's SCOWL wordlists (Spell Checker Oriented Wordlists) available at http://wordlist.sourceforge.net.

[10] The Corpus Analysis with Noise in the Signal 2013 conference (http://ucrel.lancs.ac.uk/cans2013/) is a good example.

[11] Frank Walworth was pardoned four years after his conviction, but this twist to the story attracted little attention from the press [Strange 2010].

[12] For information on the Spell Checker Oriented Word List see http://wordlist.sourceforge.net/. The list, or "dictionary", is a concatenation of word lists compiled for use in spell checkers. We are grateful for Loretta Auville's advice on this aspect of our study.

[13] A first principles approach to this question is also possible, but due to the mathematical complexity of incorporating OCR errors into calculations finding significant words with log likelihood ratio, we used an empirical approach in this paper.

[14] Apparently by accident, this study uses an incorrect variant of the log likelihood ratio. In the second equation the authors present on page 3, the sum should run over all four cells of the contingency table (rather than just those in the top row), and the observed and expected values for each of these should be calculated. With a large corpus size relative to word frequency, the ratio of observed to expected values for the bottom row cells will be approximately 1 and hence the contribution of these cells will be negligible. However, with a small corpus size relative to word frequency these cells make a substantial contribution and should not be ignored. Several open-source tools including Meandre (http://seasr.org/meandre), which we used in this study, repeat this error. We modified the source code in order to use the log likelihood ratio as it originally appeared in [Dunning 1993]. This confirms the need for humanities scholars to work with experts in computer science and digital humanities, to ensure a deep understanding of statistical techniques, rather than rely on off-the-shelf tools which may occasionally have inaccuracies in their implementation.

[15] A more thorough stemming or lemmatisation approach was not performed but may be useful in future.

[16] An in-house stopword list from NICTA (National ICT Australia) was used.

[17] Meandre (http://seasr.org/meandre), Monk (http://monkproject.org) and OpenNLP (http://opennlp.apache.org). These tools were sufficiently powerful for our study, though there are a range of other similar tools available, such as Wmatrix (http://ucrel.lancs.ac.uk/wmatrix/) and WordSmith (http://www.lexically.net/wordsmith/).

[18] http://books.google.com/ngrams

[19] http://gdc.gale.com/products/19th-century-u.s.-newspapers/

Works Cited

Archer forthcoming . Archer, Dawn. "Tracing the crime narratives within the Palmer Trial (1856): From the lawyer's opening speeches to the judge's summing up."

Arlitsch2004 Arlitsch, Kenning, and John Herbert. "Microfilm, paper, and OCR: issues in newspaper digitization."". *Microform and Imaging Review* 33: 2 (2004), pp. 58-67.

Ayers 2013 Ayers, Edward L. "Does Digital Scholarship have a Future?". *EDUCASEreview* 48: 4 (2013), pp. 24-34.

Baron 2009 Baron, Alistair, Paul Rayson and Dawn Archer. "Word frequency and key word statistics in corpus linguistics". *Anglistik* 20: 1 (2009), pp. 41-67.

Dunning 1993 Dunning, Ted. "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19: 1 (1993), pp. 61-74.

Eder 2013 Eder, Maciej. "Mind your Corpus: Systematic Errors in Authorship Attribution". *Literary and Linguistic Computing* 10: 1093 (2013).

Hitchcock 2013 Hitchcock, Tim. "Confronting the Digital, or How Academic History Writing Lost the Plot". *Cultural and Social History* 10: 1 (2013).

Holley 2009 Holley, Rose. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitization Programs". *D-Lib Magazine* 15: 3-4 (2009).

Kagan 2009 Kagan, Jerome. *The Three Cultures: Natural Sciences, Social Sciences, and the Humanities in the 21st Century*. Cambridge: Cambridge University Press, 2009.

Kilgarriff 2001 Kilgarriff, A. "Comparing Corpora". *International Journal of Corpus Linguistics* 6 (2001), pp. 97-133.

Knoblock 2007 Knoblock, Craig, Daniel Lopresti, Shourya Roy and Venkata Subramaniam, eds. "Special Issue on Noisy Text Analytics". *International Journal on Document Analysis and Recognition* 10: 3-4 (2007).

Lopresti 2008 Lopresti, Daniel. “Optical Character Recognition Errors and their Effects on Natural Language Processing”. Presented at *The Second Workshop on Analytics for Noisy Unstructured Text Data*, sponsored by ACM (2008).

McIntyre 2010 McIntyre, Dan, and Dawn Archer. “A corpus-based approach to mind style”. *Journal of Literary Semantics* 39: 2 (2010), pp. 167-182.

Moore 2004 Moore, Robert C. “On log-likelihood-ratios and the significance of rare events”. Presented at *The 2004 Conference on Empirical Methods in Natural Language Processing* (2004).

Nelson 2010 Nelson, Robert K. *Mining the Dispatch – Introduction*. *Mining the Dispatch*. 2010. http://dsl.richmond.edu/dispatch/pages/home.

Nelson 2012 Nelson, Robert K. *A Conversation with Digital Historians*. *Southern Spaces*. www.southernspaces.org/2012/conversation-digital-historians.

Newman and Block 2006 Newman, David J., and Sharon Block. “Probabilistic topic decomposition of an eighteenth-century American newspaper”. *Journal of the American Society for Information Science and Technology* 57: 6 (2006), pp. 753-767.

O'Brien 2010 O'Brien, Geoffrey. *The Fall of the House of Walworth: A Tale of Madness and Murder in Gilded Age America*. New York: Henry Holt and Company, 2010.

Paquot and Bestgen 2009 Paquot, Magali, and Yves Bestgen. “Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction”. *Language and Computers* 68: 1 (2009), pp. 247-269.

Powell 2004 Powell, Kerry. *The Cambridge Companion to Victorian and Edwardian Theatre*. Cambridge: Cambridge University Press, 2004.

Rayson and Garside 2000 Rayson, Paul, and Roger Garside. “Comparing corpora using frequency profiling”. Presented at *Workshop on Comparing Corpora*, sponsored by Association for Computational Linguistics (2000).

Rice et al. 1993 Rice, Stephen V., Junichi Kanai and Thomas A. Nartker. *An Evaluation of OCR Accuracy*. Information Science Research Institute, 1993.

Stein et al. 2006 Stein, Sterling Stuart, Shlomo Argamon and Ophir Frieder. “The effect of OCR errors on stylistic text classification”. Presented at *The 29th annual international ACM SIGIR conference on Research and development in information retrieval*, sponsored by ACM (2006).

Strange 2010 Strange, Carolyn. “The Unwritten Law of Executive Justice: Pardoning Patricide in Reconstruction-Era New York”. *Law and History Review* 28: 4 (2010), pp. 891-930.

Strapparava and Mihalcea 2008 Strapparava, Carlo, and Rada Mihalcea. “Learning to Identify Emotions in Texts”. Presented at *The 2008 ACM symposium on Applied computing*, sponsored by ACM (2008).

Svensson 2010 Svensson, Patrik. “The Landscape of Digital Humanities”. *Digital Humanities Quarterly* 4: 1 (2010). http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Tanner et al. 2009 Tanner, Simon, Trevor Muñoz and Pich Hemy Ros. “Measuring Mass Text Digitization Quality and Usefulness”. *D-Lib Magazine* 15: 7-8 (2009).

Walker et al. 2010 Walker, Daniel D., William B. Lund and Eric K. Ringger. “Evaluating Models of Latent Document Semantics in the Presence of OCR Errors”. Presented at *The 2010 Conference on Empirical Methods in Natural Language Processing*, sponsored by Association for Computational Linguistics (2010).

Wiebe 2005 Wiebe, Janyce, Theresa Wilson and Claire Cardie. “Annotating Expressions of Opinions and Emotions in Language”. *Language Resources and Evaluation* 39: 2-3 (2005), pp. 165-210.

Williams 2011 Williams, Jeffrey J. “The Statistical Turn in Literary Studies”. *The Chronicle Review* 57: 18 (2011), pp. B14-B15.

Yang et al. 2011 Yang, Tze-I, Andrew J. Torget and Rada Mihalcea. “Topic modelling on historical newspapers”. Presented at *The 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, sponsored by Association for Computational Linguistics (2011).