Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study

Hill, Mark John

Oxford University Press
2019

Hill, M J & Hengchen, S 2019, 'Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study', Digital Scholarship in the Humanities , vol. 34, no. 4, pp. 825–843. https://doi.org/10.1093/llc/fqz024

http://hdl.handle.net/10138/328229
10.1093/llc/fqz024

Quantifying the impact of dirty OCR on historical text analysis

Eighteenth Century Collections Online as a case study

Mark J. Hill
COMHIS, Department of Digital Humanities, University of Helsinki, Finland
Simon Hengchen
COMHIS, Department of Digital Humanities, University of Helsinki, Finland

Abstract

This article aims to quantify the impact OCR has on the quantitative analysis of historical documents. Using ECCO as a case study, we first explore and explain the differences between the OCR corpus and its keyed-in counterpart, created by the Text Creation Partnership. We then conduct a series of specific analyses common to the digital humanities: topic modelling, authorship attribution, collocation analysis, and vector space modelling. The paper concludes by offering some preliminary thoughts on how these conclusions can be applied to other datasets, by reflecting on the potential for predicting the quality of OCR where no ground-truth exists.

Introduction

This paper[1] compares the impact optical character recognition (OCR) has on the quantitative text analysis of historical documents. Specifically, we look at the 18th-century texts which have been made available in a keyed format as part of the Eighteenth Century Collections Online (ECCO) Text Creation Partnership (TCP) project, and the corresponding OCR documents held by Gale.[2] The analyses in this paper are divided into three parts. First, we quantitatively examine the distinctions between the two ECCO corpora. Second, we compare the texts through the outputs of four different textual analysis methods - topic modelling; collocation analysis; vector space modelling; and authorial attribution. Third, we offer some preliminary thoughts on the potential for predicting the quality of OCR.

There are at least three reasons for an analysis such as this. First is the oft-reported fact that researchers spend 80% of their time pre-processing data, and only 20% analysing

---

[1] This research is part of the Helsinki Computational History Group's (COMHIS: https://www.helsinki.fi/en/researchgroups/computational-history) larger project on ECCO and ESTC. We would like to thank Gale for providing our group with ECCO data.

[2] Although the OCR errors found in ECCO have been previously discussed, a systematic analysis of the data has not taken place. See: Spedding (2011), Bullard (2013), Prescott (2018). The closest to a systematic analysis we are aware of is a blog post by Hine for the Linguistic DNA project (https://www.linguisticdna.org/ecco-ocr/, retrieved October 10, 2018). For a broader overview of historical text analysis see: Piotrowski (2012).

it.[3] Estimations with regard to the required level of cleanliness necessary for textual analysis could be helpful in narrowing this ratio. Second, there are epistemological benefits to understanding the makeup up the data itself. Prescott, with reference to ECCO specifically, recently criticized scholars (such as de Bolla and Colley) who, he claims, pay no heed to the quality of their data.[4] Finally, there are numerous, inconclusive, claims with regard to the effect OCR errors have on humanistic analyses. Linguistic DNA, who did make a small quantitative study of ECCO for their own research aims, concluded that the error rates make the corpus unusable, while others have claimed that texts with as little as 20% OCR correctness provide enough signal to achieve better-than-random results, and that at 80% clean, texts are not noticeably worse than texts that are 100% accurate.[5] While these claims are both domain and method specific, the variation is itself evidence for the necessity of further study. Thus, by using a gold standard (TCP), this paper aims to measure how accurate different analyses are at different levels of OCR correctness. It is hoped that these analyses will aid researchers interested in quantitative text analysis (like the authors), both with regard to what quality of data may or may not be sufficient for specific interests, as well as support research which makes use of OCRed data.[6]

Data

The data being used in these analyses comes from two versions of ECCO. The first will be referred to as ECCO-OCR, a dataset which 'contains over 180,000 titles (200,000 volumes) and more than 32 million pages' according to Gale.[7] This equates to 537 GB of XML data, which correspond to 207,613 documents; 405,195,912 paragraphs; 771,738,286 sentences; 383,913,142 types appearing 10,548,490,456 times; and 58,429,423,917 total characters. Of those documents, 184,363 have corresponding English Short Title Catalogue (ESTC) IDs.[8] As can be inferred from the enormous number of types, OCR noise in this dataset is quite high.

The second dataset being used is ECCO-TCP. This is a keyed subset of ECCO, compiled with the support of 35 libraries and made up of 2231 documents.[9] Keying was carried out by external companies, and quality assurance was fulfilled by editorial teams in Oxford and Michigan: 5% (or 5 pages, whichever is the greater) of every work was proofread and documents that did not meet the QA standards were sent back. According to our analyses, all documents but two are in English: one work (33 pages) is in French, and

[3] Although definite evidence is difficult to find on this point, it is an often-repeated trope which has made its way into numerous papers and presentations. Having said that, from our own research we feel it may underestimate the ratio.

[4] Prescott proposes one solution: the comparison of results with manually crafted lists. However, he does not go beyond an assessment of search results (Prescott, 2018).

[5] Traub *et al* (2015), Strange *et al* (2014) Franzini *et al* (2018), or Eder (2013).

[6] For more on the principle of "fitness for use" see: Boydens (1999).

[7] https://www.gale.com/primary-sources/eighteenth-century-collections-online, retrieved April 27th, 2018.

[8] The authors wish to thank Prof. Eetu Mäkelä for these figures, as well as for drawing the ECCO OCR accuracy estimations to our attention.

[9] Although not included here, the editorial declaration available in TCP TEI headers states details of, and reasons for, inclusion of texts.

another (98 pages) is in Welsh.[10] Having said that, we are aware that there are significant chunks of text in other languages (in particular Latin) which are not tagged in the TEI TCP texts, but these do not impact our quantitative analysis of OCR, as this is done on the character level.[11]

While the overall quality of ECCO-TCP is very good, it is not perfect. Common errors and differences due to editorial decisions include: encoding problems (in the versions available on the OTA website); missing words (mostly catchwords); incorrect TEI page attributions; phrases such as *(page in other language)* included in the text; pages which simply state *(duplicate page)*; and missing pages.[12] Nonetheless, it is as good of a standard as is available for a corpus of this size, quality, and importance.

For our analyses, we constructed corresponding corpora at the corpus, document, and page level. Creating the per-page corpora was a particularly difficult task, and required an intensive mix of parsing, extracting, and correction. This was largely automated, but thousands of pages were also examined (and at times corrected) by hand so as to ensure the corpora matched. We estimate that we have very good matches for more than 99.5% of all pages in our testing corpus (n = 336,651).[13]

1. Analysing the Data

Corpus representativeness[14]

When looking at the distribution of characters and tokens in both corpora, the OCR corpus does not initially stand out - in fact, by some measurements (in particular characters) it is remarkably accurate. The figures are available in Table 1.

Table 1 Corpus differences between ECCO-TCP and ECCO-OCR

	Raw	**Punctuation and numbers removed**	**Punctuation, numbers, and stop words removed**
TCP Characters	343,993,778	328,401,977	231,502,888
OCR Characters	394,440,756	328,516,319	230,226,975

[10] This information was taken from parsing the TEI P5 files made available by the Oxford Text Archive, more precisely the `<text xml:lang="eng">` tag. The header of the TEI files mentions the existence of works in 'other languages, notably Latin and Welsh'. After parsing the whole collection for the above-mentioned tag, the only three languages found were English, French, and Welsh.

[11] Non-Latin sections of text are encoded in TEI.

[12] To offer three examples: K036193.000 contains encoding errors; K033495.000 has mislabelled pages; and K072003.000 has pages missing.

[13] Bad matches are most easily identified by extreme differences in token counts. This largely appears to be caused by mistakes in the TEI encoding of the TCP editions. The code for the automated page extraction process, as well as the rest of the code used in this paper, is available at https://github.com/COMHIS/ECCO-TCP_ECCO-OCR.

[14] As the aim at this stage was simply to measure and compare the TCP and OCR corpora, minimal processing was done. We accept that different processing techniques may improve (or worsen) these results. For more on the potential impact of processing text see Denny and Spirling (2018).

Difference	-50,446,978	-114,342	1,275,913
TCP Tokens	87,298,605	72,653,703	37,990,623
OCR Tokens	95,390,984	75,405,835	40,476,712
Difference	-8,092,379	-2,752,132	-2,486,089
TCP Types	765,275	749,866	749,693
OCR Types	2,703,969	2,675,725	2,675,551
Difference	-1,938,694	-1,925,859	-1,925,858

In terms of types, while the OCR tail is, unsurprisingly, very long, the overall curve of the distribution mimics both the TCP corpus, and what would be expected with regard to Zipf's Law. Thus, the introduction of OCR errors does not distort the expected distribution. However, most humanists are interested in the words these numbers represent. When looking at the top features in the two corpora, illustrated in Table 2, further differences emerge – in particular with the removal of stop words.

Table 2 Top 20 features. Left columns describe features with stop words, right columns without. Italics represent mismatches.

TCP	OCR	TCP	OCR		TCP	OCR	TCP	OCR
the	the	*it*	*his*		*time*	*c*	*give*	*p*
of	of	with	with		*mr*	*t*	*think*	*l*
and	and	he	he		*sir*	*mr*	*c*	*little*
to	to	as	as		*little*	*time*	*love*	*de*

a	a	for	for		*part*	*e*	*day*	*part*
in	in	*was*	*be*		*king*	*sir*	*people*	*n*
that	*i*	*be*	*was*		*lord*	*o*	*long*	*d*
i	*that*	by	by		*life*	*s*	*p*	*king*
his	*is*	which	which		*know*	*fame*	*found*	*lord*
is	*it*	*not*	*this*		*s*	*r*	*place*	*know*

When examining the top-500 types with stop words removed, one-fifth are not shared.[15] These errors are particularly important for two reasons: qualitatively, one can find tokens of historical interest being misrepresented. Quantitatively, not only does an incorrect token in the top feature list indicate additional noise (as a false positive, and the reason why there are five times as many types in the OCR corpus), but they also represent corresponding corruption (false negatives) elsewhere in the corpus. It is, therefore, necessary to measure these differences in a more robust way if one wants to understand corpora representativeness.

The first problem which needs to be addressed when comparing OCR documents with keyed partners is the design of an accuracy measurement when a corresponding index does not exist. That is to say, while a general index of every token and its location would be ideal, this is not possible when working with messy OCR, as corresponding tokens are often missing, malformed, or include additional white spaces and/or punctuation. Thus, it is necessary, if one wants to compare the two corpora, to rely on counted data. To this end, our comparisons are based on the number and range of features (character, types, and tokens) that exist in the TCP version, and how well these are represented in the OCR version. In this way we were able to estimate true positives (features found in both corpora), false positives (features found in the OCR corpus, but not the TCP) and false negatives (features in the TCP corpus but not found in the OCR).

[15] These are: 'ac', 'according', 'act', 'ad', 'afterwards', 'answer', 'authority', 'b', 'beauty', 'become', 'business', 'cafe', 'case', 'character', 'com', 'con', 'conduct', 'continued', 'dif', 'duty', 'ed', 'effect', 'en', 'english', 'ex', 'f', 'fall', 'fame', 'fate', 'fay', 'fays', 'fee', 'fide', 'fight', 'fit', 'fix', 'foul', 'g', 'generally', 'greatest', 'h', 'ha', 'hall', 'happiness', 'heaven', 'history', 'ihe', 'ihould', 'ill', 'immediately', 'ing', 'interest', 'iv', 'j', 'james', 'justice', 'k', 'kingdom', 'la', 'laid', 'lie', 'loft', 'lost', 'm', 'making', 'married', 'master', 'n', 'necessary', 'object', 'obliged', 'page', 'parliament', 'passion', 'per', 'pro', 'purpose', 'queen', 'r', 're', 'reft', 'regard', 'respect', 'rest', 'same', 'say', 'says', 'self', 'short', 'side', 'sight', 'something', 'soul', 'strong', 'subject', 'suppose', 't', 'ten', 'tion', 'u', 'un', 'used', 'vol', 'w', 'william', 'wish', 'women', 'x', 'y', and 'z'.

With these calculations we are able to offer some initial claims with regard to general representativeness: 6.8% of the OCR corpus (5,128,153 tokens) is made up of types which do not exist in the TCP corpus, while 0.6% of the TCP corpus (451,948 tokens) is missing from the OCR corpus. These are the easily identifiable errors, however. What these numbers do not account for are extra or missing tokens which are not unique types. Again, at the corpus level, we can see that 29.1% of the OCR data is made up of types which exist in both corpora, but for which there are 9.1% fewer tokens in the OCR corpus (additional false negatives), and 31.2% of the corpus is made up of types which match, but for which there are 7.1% extra tokens (additional false positives). Only 26.8% of the two corpora match both types and token counts. However, as these are corpora level estimates, a smoothing out of errors is almost certainly taking place, and the reality is likely to be even worse. Thus, we turned to the page level to continue our analysis.

Page level comparison

Due to the number of errors in the OCR data, a per-token or per-sentence investigation was impossible. Instead, the page is the smallest level of analysis available. Mean tokens per page in ECCO-TCP are 259, and in ECCO-OCR there are 283, while the mean number of characters per page is 1,487 and 1,274 respectively.

Making use of counted and comparative data - specifically true positives, false positives, and false negatives - we were able to estimate accuracy via precision, recall, and F1 Scores at the page level. At the token-level, the mean precision of ECCO-OCR pages is 0.744, recall is 0.814, and the F1 Score is 0.774. Precision appears to be negatively impacted by OCR noise, whereas recall demonstrates the general ability of the OCR software to correctly recreate tokens. Depending on the type of analysis one is conducting, either precision or recall may be a better confidence test, but for the sake of generalizing results, this paper focuses on F1. The total number of tokens per F1-scored page is available in Figure 1.

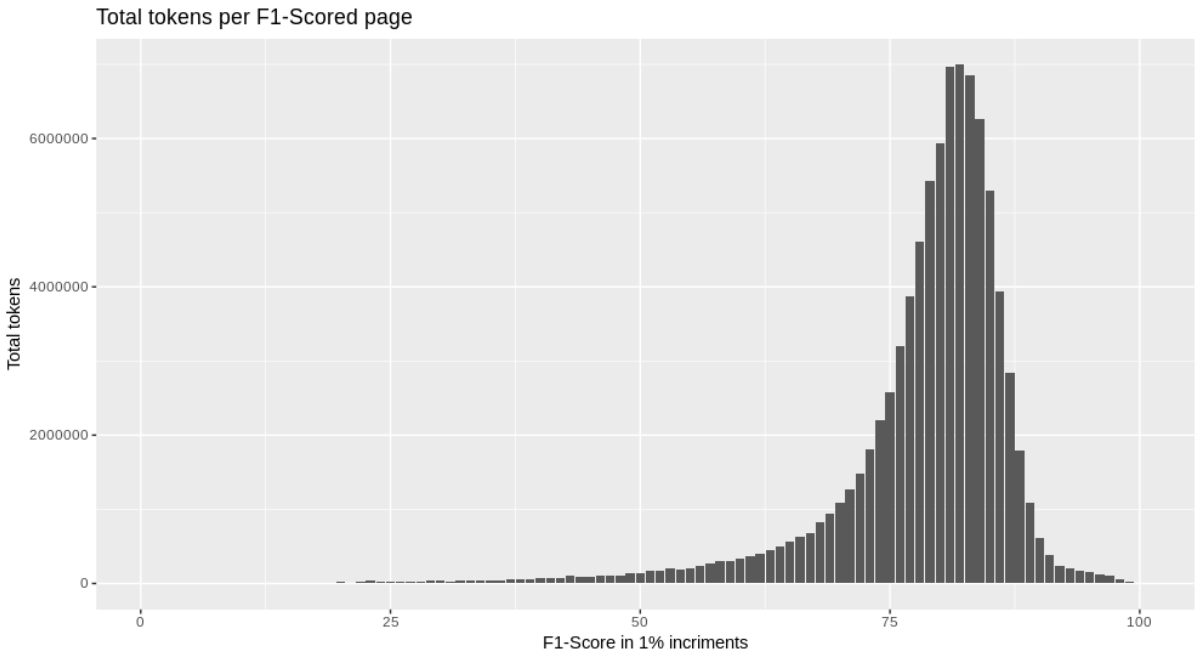

Fig. 1 Total tokens per F1-scored page

In the case of ECCO, we do have one additional accuracy metric available: the original OCR software's per-page confidence estimates. However, because it is difficult to know exactly what is being measured here, the scores were not used in our analysis. Nonetheless, when comparing them to the calculated F1 scores we did notice a promising statistical relationship ($p < 0.001$), although the OCR software overestimates its accuracy - both in terms of maximums achieved, variance, and shortness of tail. This is illustrated in Figure 2.

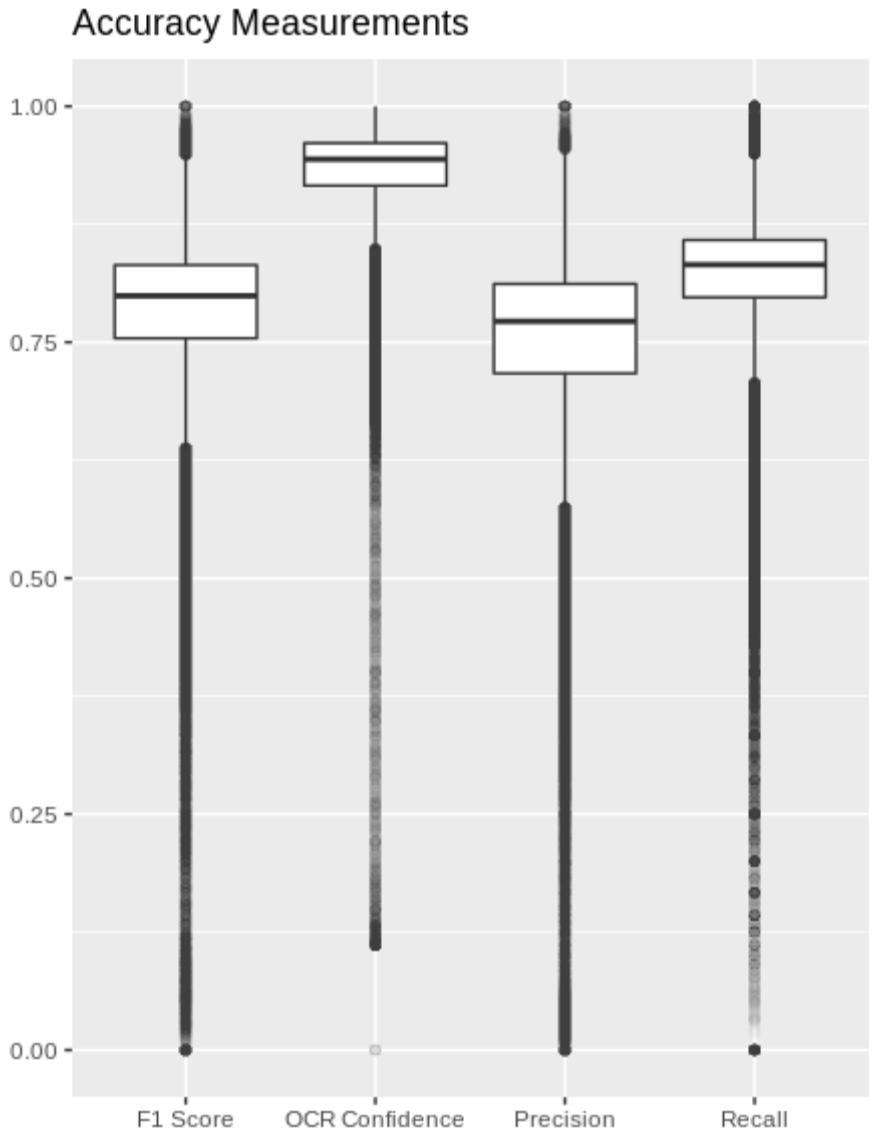

Fig. 2 Different measures of accuracy

By establishing F1 scores on a per-page basis, we were able to create F1-specific sub-corpora which could be used to measure distinctions between the corpora, as well as the outputs from specific tests. The sizes of the differing sub-corpora range between 1,334 pages (0.95 F1) and 121,749 pages (0.8-0.85 F1), as illustrated in Figure 3. However, because the tests conducted are between the TCP and OCR pairs, rather than between F1 sub-corpora, this is generally not a problem, and when the size of a corpus could impact the reliability of a test (for example, with authorship attribution), sub-corpora were merged. The total number of, and difference between, tokens is shown in Figure 4.

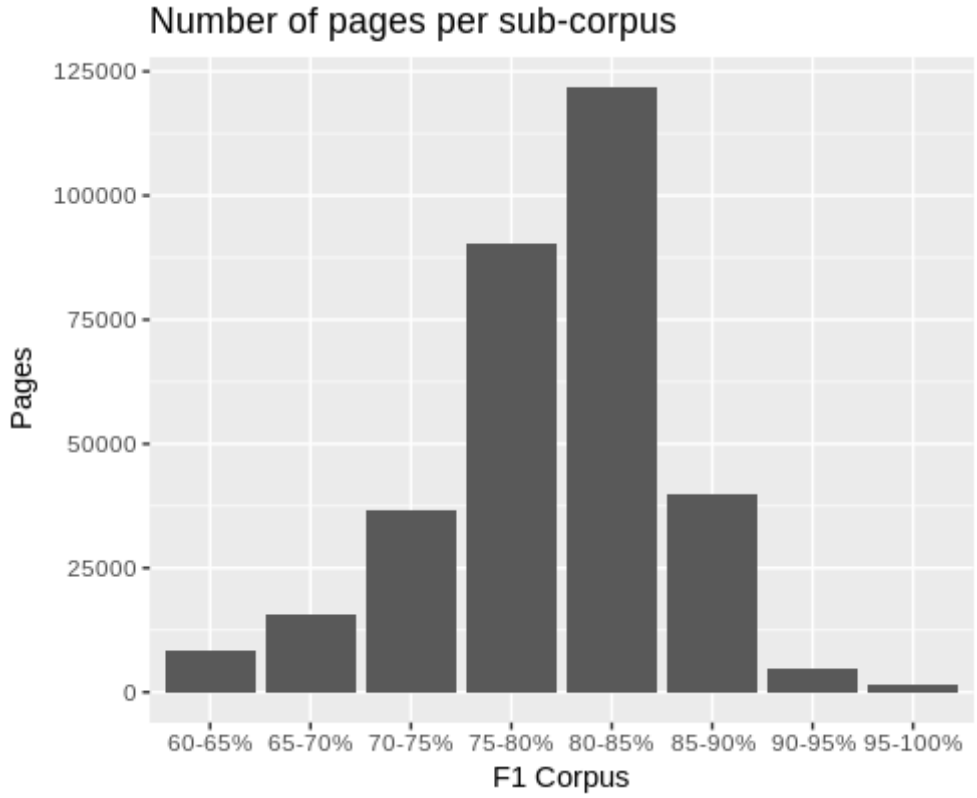

Fig. 3 Number of pages per F-score based sub-corpus

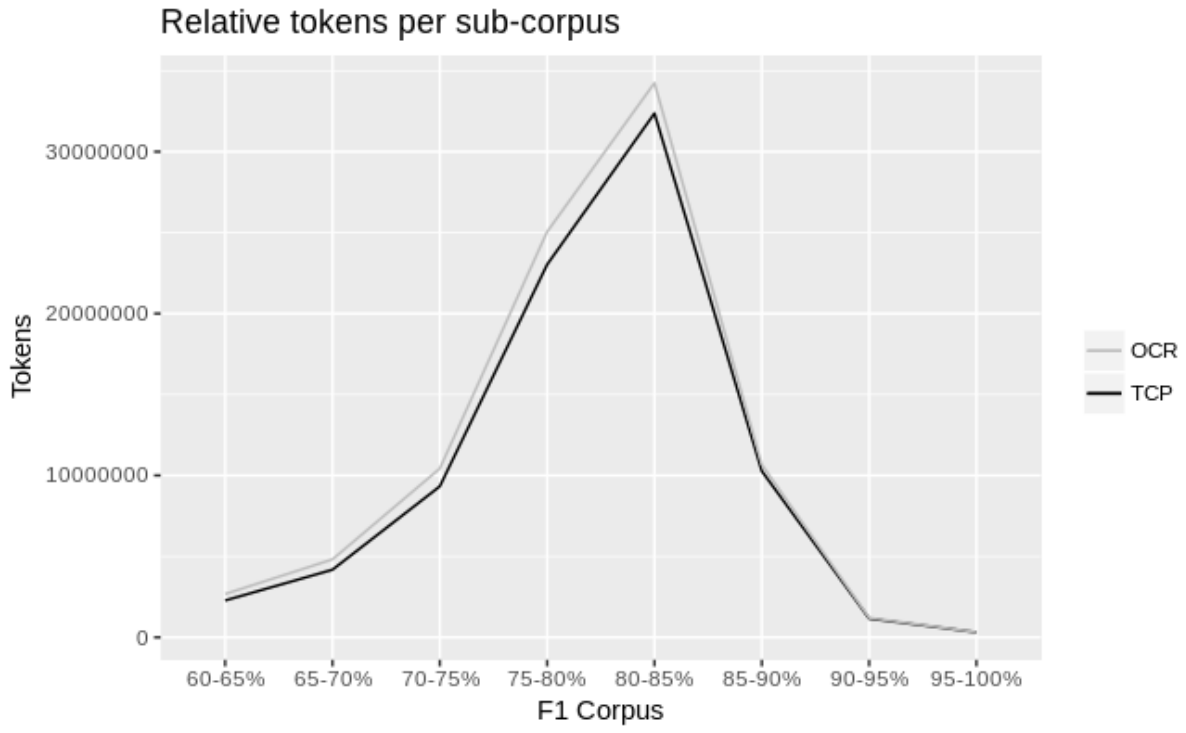

Fig. 4 Relative number of tokens per F-score based sub-corpus

Statistical analysis of OCR errors

Having estimated the overall accuracy of the OCR corpus, we wanted to see if it was possible to statistically identify specific causes of OCR errors beyond those resulting from the physical qualities of a text (i.e., smudges, damage, bleed/shine-through, annotations, fading, etc.).[16] As most eighteenth-century scholars familiar with OCRed editions of historic texts are aware, the long-s (*ſ*) and ligatures (multiple characters conjoined and compressed to fit single print blocks) are often misidentified. To investigate the impact these may be having on the overall OCR process, we created a list of measurable variables (individual letters, ligatures, time of publication, and length of words) to test against the frequency of accurately OCRed words.[17]

Again, because it is impossible to directly match TCP words with their OCR counterparts, we created a dataset made up of all words in the TCP corpus, counted the number of instances they were found in the OCR corpus, and extrapolated accuracy rates for individual letters and ligatures. This allowed us to create a count variable for every correct and incorrect instance of a particular letter and ligature being used in a word. The results of this analysis are displayed in Figures 5 and 6.

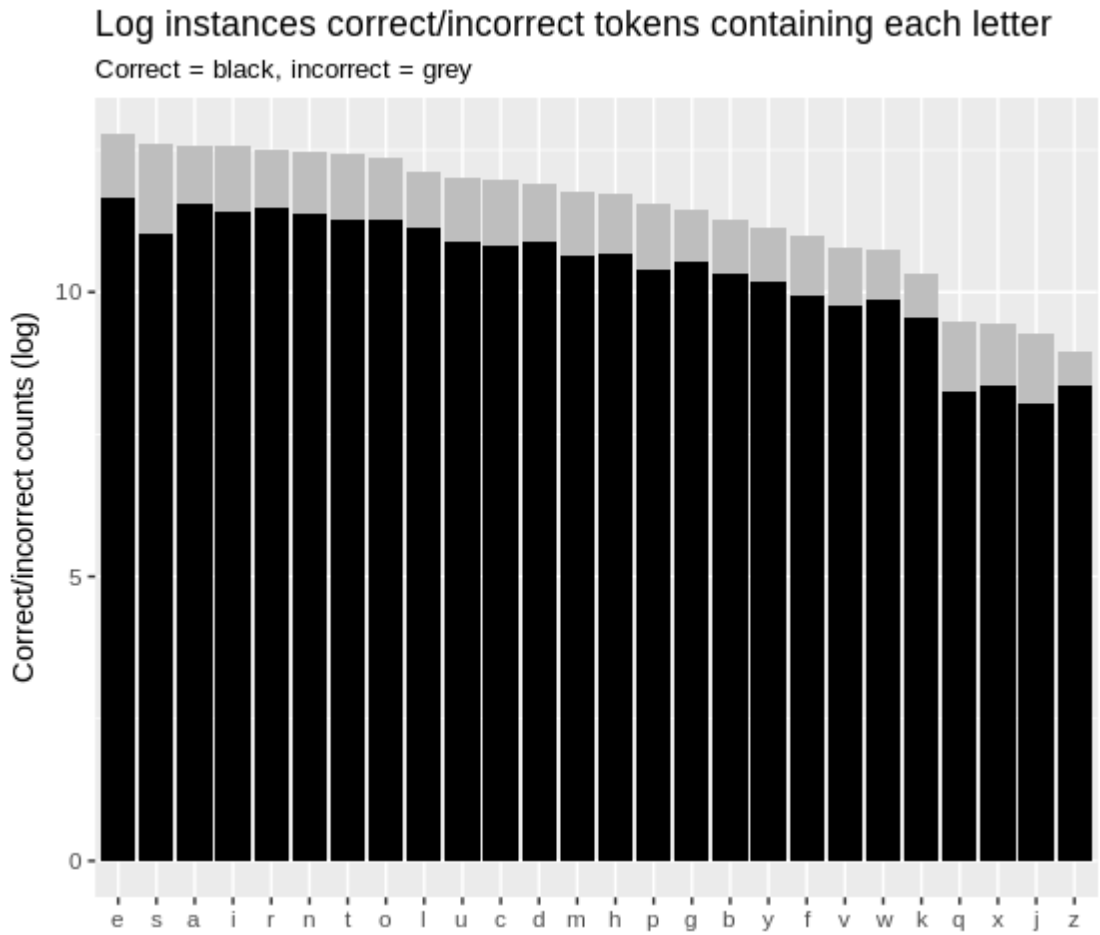

Fig. 5 Per letter log counts of correct and incorrect tokens

[16] The EU-funded IMPACT project (http://www.impact-project.eu/) has done work on the impacts of these aspects of historical texts when it comes to the OCR process.
[17] Time of publication did not have a statistical relationship to OCR accuracy.

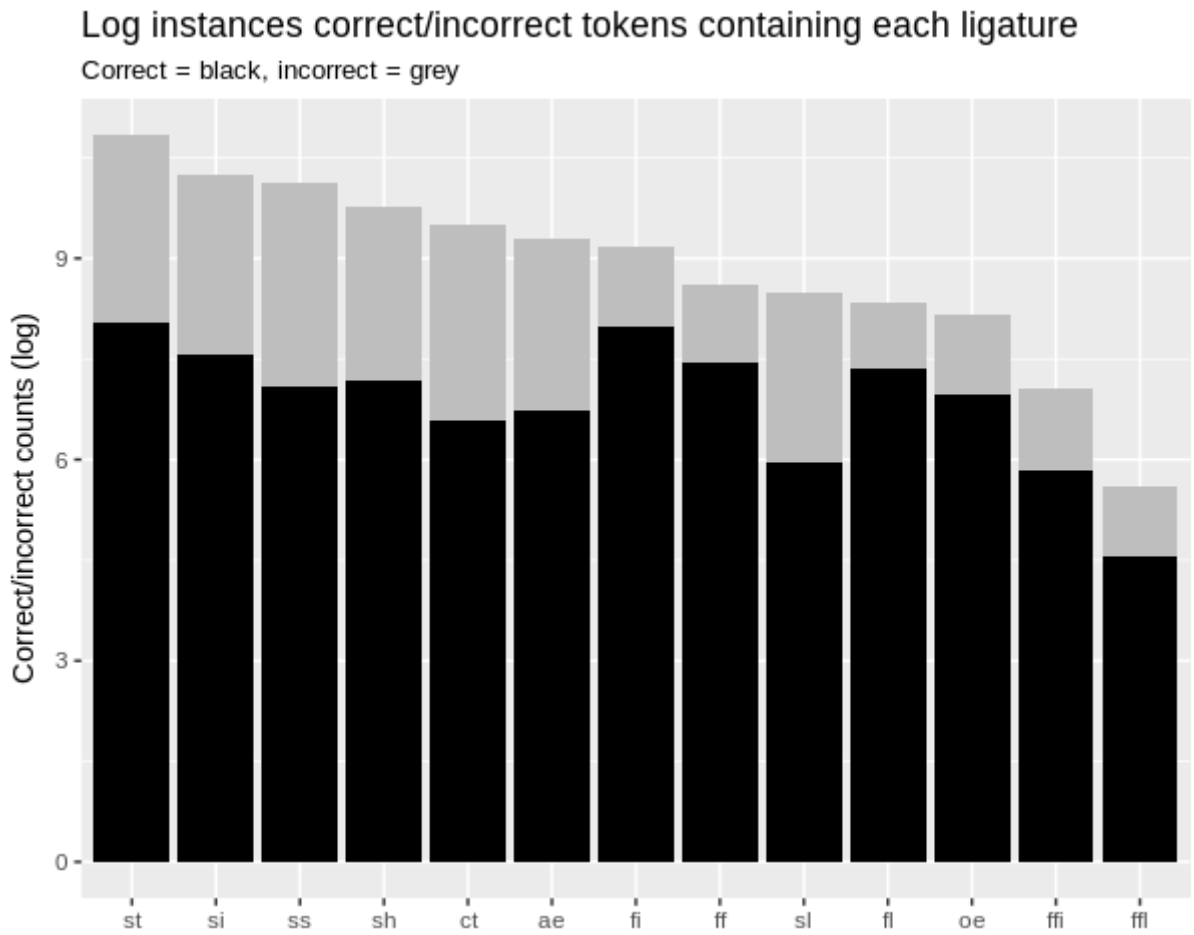

Fig. 6 Per ligature log counts of correct and incorrect tokens

At a first glance, the general dispersion of errors for letters does not appear to be particularly problematic in any one direction (although a variance amongst ligatures is much more obvious). At the statistical level, however, a clearer picture emerges: the only letter which had a statistically measurable negative relationship on words correctly being replicated through OCR was the letter 's' ($p < 0.001$).[18] In fact, comparatively, every other character was more likely to be found in a correctly OCRed word. Similar results were found with historically used ligatures. Interestingly, however, not all ligatures were equally problematic; those - unsurprisingly - containing a long-s, as well as 'ct', 'ff', and 'ffl', were statistically more likely ($p < 0.001$) to be part of words which had been incorrectly OCRed.

It is worth noting that the longer a token is, the more likely it was to be incorrect in the OCR corpus. However, there was no evidence that length itself was a factor. As shown previously, OCR software is remarkably good at identifying individual characters - discounting punctuation, the margin of error between the two corpora was 0.00348%. Instead, the problem with longer tokens is the potential for errors to be introduced – be it through physical defects, or the possibility of a long-s or problematic ligature being included.

While in some ways these results are not surprising. The OCR software used for ECCO clearly had not been trained to recognize the long-s or ligatures, and it should have a

[18] Due to the dataset being overdispersed, non-normally distributed, and count based, we used a negative binomial regression model to test the normalized error frequency per word against each letter in the alphabet and the length of the token. To test robustness, the test was duplicated using other models which, although not generally as well suited for the data, confirmed the results.

better-than-random chance of correctly recognising the characters it had been trained for. However, there are two important conclusions from this analysis: first, it allows us to further understand the ECCO corpus and its OCR errors. In this case, 51.7% of all types contain the letter 's' or the ligatures 'ct', 'ff', and 'ffl'. This means 26.9% of all tokens in the corpus were statistically likely to be corrupted during the OCR process. Secondly, if one were interested in the post-processing of historical OCRed documents - and it is clear that many are[19] - one should begin by correcting these statistically problematic errors. For example, there are 79,235 types with an "f" in the TCP corpus (11% of corpus), but 901,704 (33.7%) in the OCR corpus. One could easily replace instances 'f' in the OCR corpus in cases where that token has no TCP (or historical dictionary) match (*eg*, 'fame' and 'same' would not be converted, but 'rofe' would).[20] This is particularly promising as, rather than correcting the 'naturally' occurring errors which may be normally distributed and therefore less problematic, one can focus on statistically problematic cases which have a greater impact on analyses. If one is interested in achieving, and maintaining, a statistically balanced corpus, these distributions of errors must be taken into account.

2. Analysing with the Data

In this section of the paper we run a number of tests commonly associated with the digital humanities on the two corpora to compare outputs. Specifically, we make use of quantitative text analysis using topic models; collocation analysis; vector space models; and authorial attribution.

Topic Modelling

Topic models were trained on both the TCP and OCR corpora, and pre-processed in the same way: tokens were lowercased and stemmed, and stop words removed.[21] To ensure comparable results, the same parameters were used for both models - we trained topics by using a spectral initialisation, fixed a seed, and automatically estimated a K. ECCO-OCR and ECCO-TCP yielded 77 and 65 topics, respectively. This is not surprising, given the enormous difference in vocabulary size between the corpora, even after stemming. However, despite this divergence, topics present in ECCO-TCP are all present in ECCO-OCR. Furthermore, in both models topics appear in virtually identical coordinates on the

[19] See: Philosophical Integrator of Computational and Corpus Libraries (https://github.com/LanguageMachines/PICCL); eMOP post-processing OCR correction workflow (http://emop.tamu.edu/about); PoCoTo - An Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts (Vobl *et al*, 2014); Alex et al (2012); DataMunging (https://github.com/tedunderwood/DataMunging); Holley (2009).

[20] There are 5,411 words in the OCR corpus for which this type of transformation would result in losing a legitimate type, and thus should be excluded from this type of process. However, there are also outliers which may be worth correcting even if it does introduce new errors. For example, there are 336 instances of 'os' in the TCP corpus, but almost three million instances of 'of.' The instances of 'os', which is a word in French, do not come from the document in French.

[21] Although research shows that stemming produces no meaningful improvement, and, in fact, can degrade topic stability (Schofield and Mimno, 2016), it (and the more linguistically aware step of lemmatising) is usually a recommended step when carrying out topic modelling in digital humanities. Because of this, and the fact that the vocabulary size of our OCRed data made it computationally very heavy, this step was still undertaken.

intertopic distance map.[22] These topics share a highly similar probability distribution for the top words, although some divergence occurs for topics at the centre of the map. For these topics, the outlier words are often grammatical elements that were not considered as stop words and perhaps point to a style of writing more than a specific topic (usually auxiliaries such as 'will', 'shall', 'may', or conjunctions such as 'though'). Interestingly, these outliers are also more relevant within topics which have top words that, in a metaphorical sense, directly relate to poetry ('love', 'eye', 'sweet', 'heart', etc.).[23] This may indicate that poetry as a genre is more difficult to OCR, or that the topic modelling algorithm had trouble incorporating the genre of poetry, with its different style of writing and metaphorical, rather than literal, use of words.

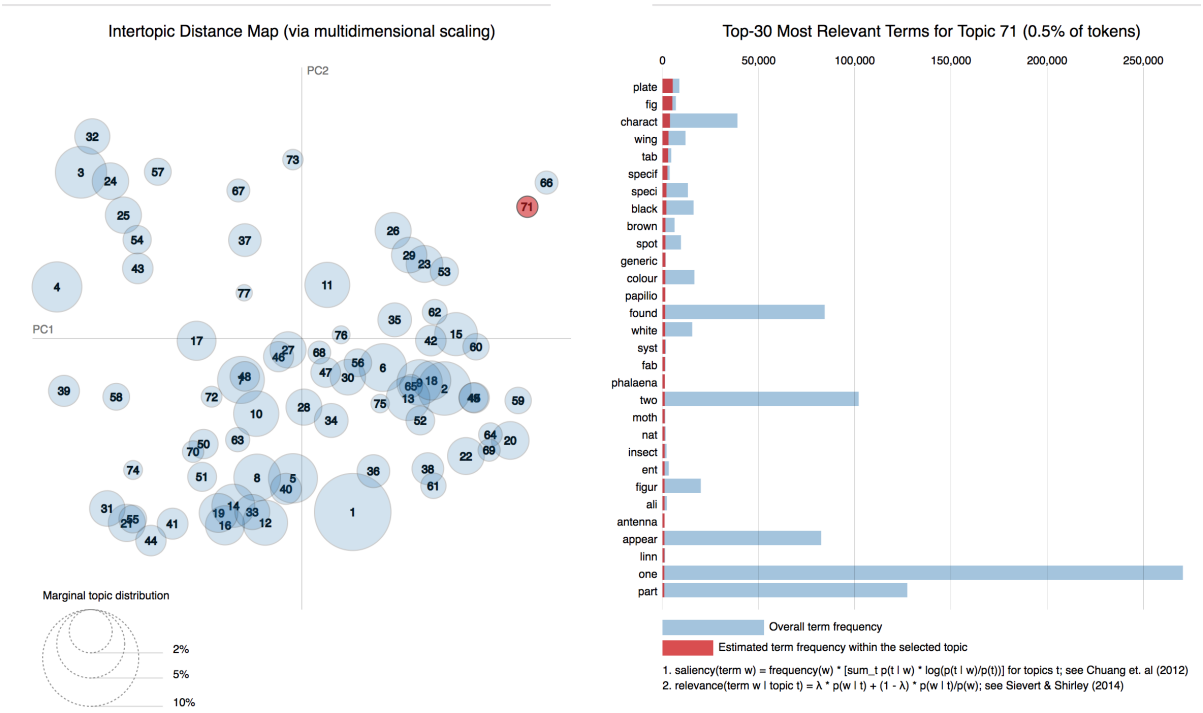

Fig. 7 LDAVis visualisation of topic 71 of ECCO-OCR

Figure 7 illustrates topic 71 in ECCO-OCR, which seems to refer to entomology. For this topic, the two models share 28 out of the 30 most relevant tokens. The two tokens that did not make it in to ECCO-OCR are 'genus' and 'shell' - both words contain the problematic letter 's' (although, interestingly in this case, the long-s was generally *not* used at the end of a word). The TCP words were replaced by 'ali' and 'appear', at positions 25 and 27 respectively. This indicates that whilst 'shell' and 'genus' are not deemed as important for the topic in ECCO-OCR as they are in ECCO-TCP (where they were ranked 14 and 26, respectively), their replacements do not take an important place either.

The pattern described above is replicated throughout the models. Topics continue to show the same pattern of tokens containing an 's' or a problematic ligature being in the model trained on ECCO-TCP, yet absent from ECCO-OCR. Most importantly, however, these additions and subtractions do not impact topic interpretability by humans, making the

[22] The intertopic distance map is generated through LDAvis. The distances are calculated through Jensen-Shannon divergence and scaled via principal component analysis. They are then projected onto two dimensions, as per Chuang *et al* (2012).
[23] The metric used for relevance is the one introduced by Sievert and Shirley (2014).

OCR topics both extremely good replications of the TCP topics, as well as qualitatively excellent.

Collocations

As collocations have become important topics of research in a number of fields, our next analysis turned to them.[24] We first looked at the impact OCR had on collocations found in the entire corpus. Using the lambda collocation scoring metric, a minimum count of 10, and no stop words, we identified 605,569 collocations in the complete OCR corpus and 490,623 in the TCP corpus.[25] Of these two lists, 319,440 did not overlap, and roughly 70% of these unique collocations were found in the OCR corpus. Thus, the OCR corpus both lost statistically significant collocations through OCR corruption, and gained non-existing collocations through the introduction of noise.

To understand the impact this may have on historical research, we looked at the collocations which contained the words 'public' and 'publick', of which there were 765 in the TCP corpus and 750 in OCR. Of these, 305 did not intersect, and many of the missing collocations are those which would be of interest to eighteenth century historians.[26]

Moving away from the corpus as a whole, we turned to the F1-scored sub-corpora, to examine the impact different levels of OCR quality had on collocation analysis. A summary of the figures is displayed in Table 3, and a visualisation is available in Figure 8.

Table 3 Comparison of collocations between ECCO-OCR and ECCO-TCP

Corp/F1	60-65%	65-70%	70-75%	75-80%	80-85%	85-90%	90-95%	95+%
TCP	16,600	31,590	70,690	163,241	217,092	76,356	9,267	2,132
OCR	11,484	24,442	60,554	151,411	208,355	73,772	8,944	2,084
Distinct	9,024	14,686	28,424	57,728	62,499	14,446	933	76
%	54.3%	46.5%	40.2%	35.4%	28.8%	18.9%	10.0%	3.6%

[24] See: Gablasova *et* al (2017), Evert (2008), Uhrig *et al* (2018), Brezina *et al* (2015).

[25] Lambda based on Blaheta and Johnson (2001).

[26] These included: 'publick diversions', 'publick assemblies', 'public spectacles', 'public notoriety', 'public institutions', 'public estimation', 'public nuisance', 'public thanksgiving', 'publicly sold', 'public lectures', 'public acts', 'public spectacle', 'public resolutions', 'public affections', 'publick revenues', 'public edict', and 'publick entertainments'.

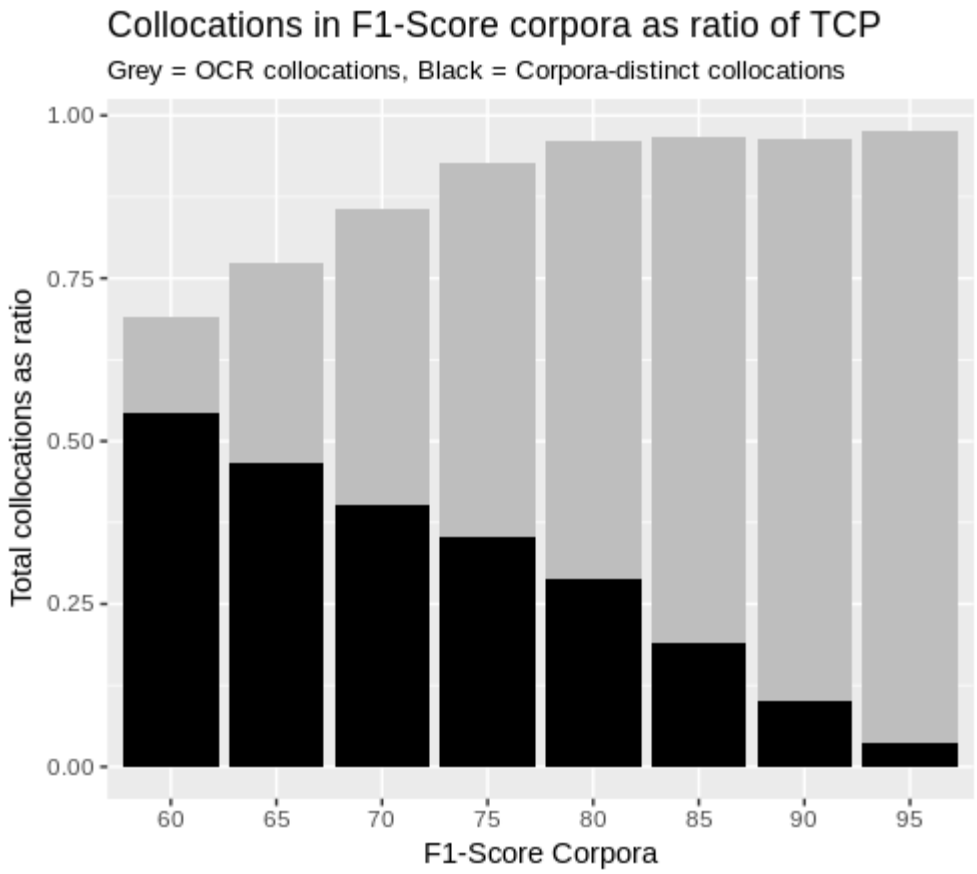

Fig. 8 Total collocations as ratio, per F-score sub corpus

Interestingly, unlike results found for entire OCR corpus, the subsets resulted in fewer statistically significant collocations than their TCP pairs. Thus, the OCR collocations were likely coming from the very bad OCR. Nonetheless, noise remains a concern – in fact, a majority of collocations in the OCR corpora are noise until around the 65% F1 range, and although the total number of collocations begin to approximate each other from around the 75% mark, it is not until around the 90% mark that error rates in actual collocation matches are at the 10% level.

The conclusions to be drawn from these results need to be tailored to each researcher's specific ends. If one is interested in diversity and frequencies, then a higher level of accuracy will certainly be desirable. However, if the aim is to find popular collocations, then perhaps a lower threshold is sufficient (albeit, with the problem of the long-s and ligatures taken into consideration).

Vector Space Analysis

The third analysis undertaken made use of similarity and distance measurements in vector space.[27] To do this we made use of the quanteda package for R, and analysed pages, documents, and features at different F1 levels.[28]

[27] This analysis was also conducted using word2vec (Mikolov *et al* (2013)). However, as Recchia (2016) have noted, word embeddings result in "opaque mathematical representations, creating difficulties for researchers attempting to use them to draw conclusions about the use of particular words" whereas associations in count-based vector space can be "more clearly and rigorously

At the page level, similarity and distance measurements are essentially meaningless. While there is a strong relationship ($p < 0.001$) between the two (the greater the F1 score, the greater the similarity measurement), the variance is extremely high (Figure 9).

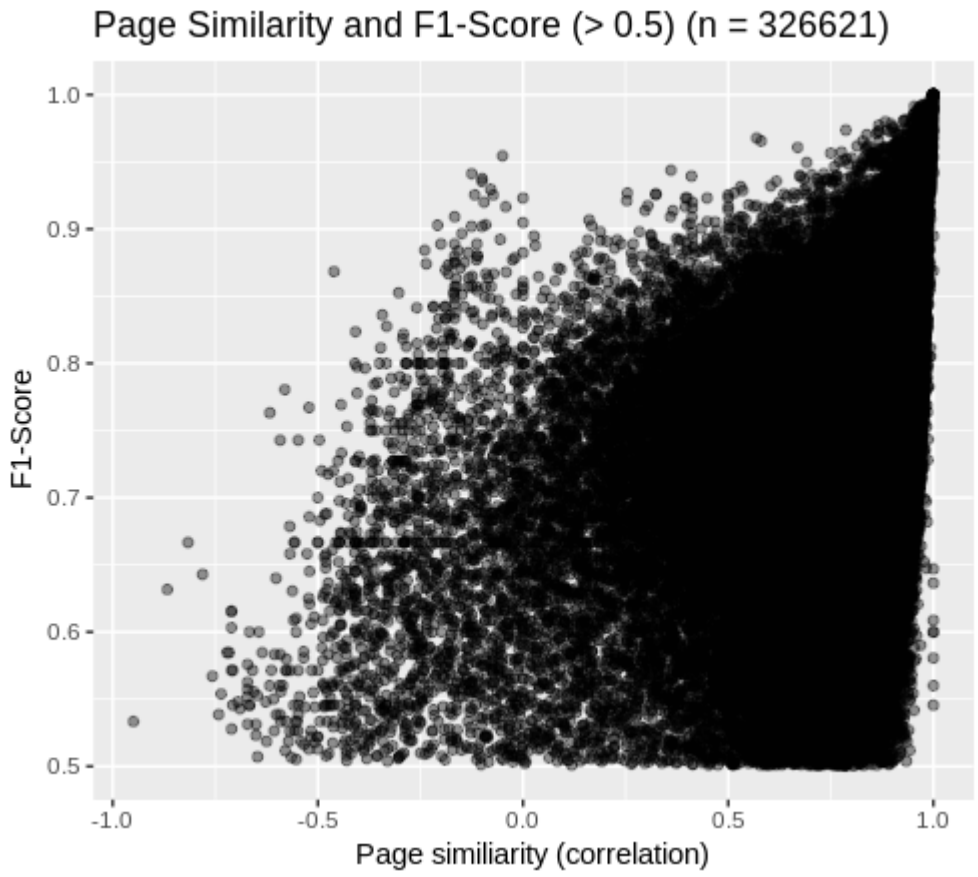

Fig. 9 Variance between F1 score and document similarity at the page level

These results become clearer when we move to the document level, at which point document length appears to be a much stronger - and understandable - predictor of document similarity than F1 scores (black points in Figure 11 represent the 100 lowest F1 scored documents).

investigated." In addition, Antoniak and Mimno (2018) point out the limitations of (neural) word embeddings for smaller corpora.

[28] Corpora were processed to remove punctuation, numbers, and quanteda's 175 default English stop words. Additionally, the long-s in TCP documents was converted into the short-s to aid automated matching tests. Similarity and distance measurements used for documents and pages are the default parameters used by the quanteda and proxy packages for R: correlation and Euclidean.

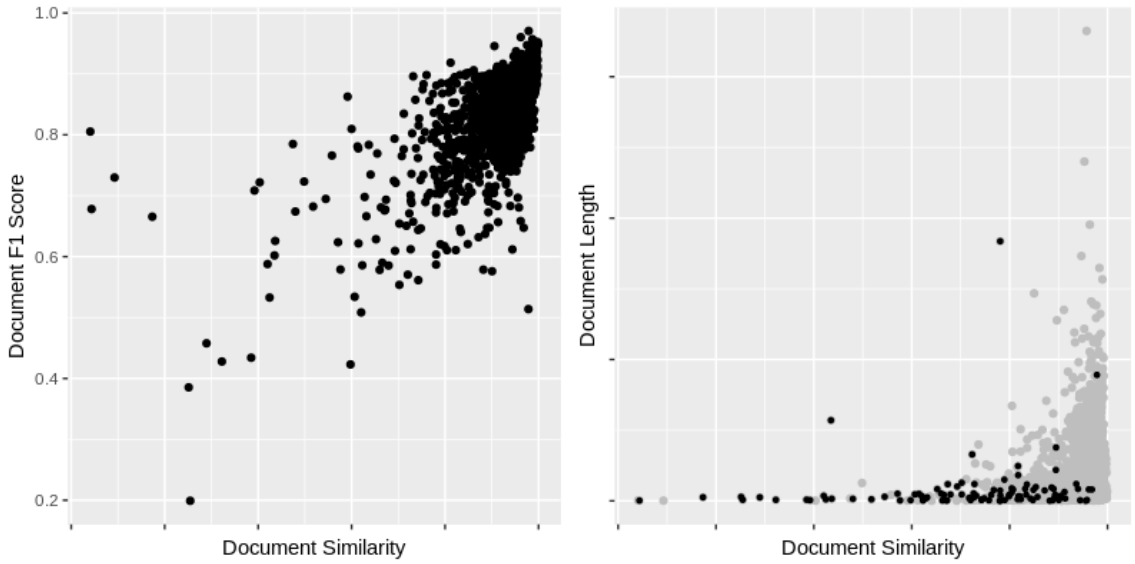

Fig. 10 F1-Scored documents and their similarity score (left)
Fig. 11 Document length vs document similarity, black points represent worse 100 F1 scores (right)

Thus, document length can be considered a larger concern in some cases than OCR quality. However, at the token level, it becomes much easier to recognize the impact OCR errors are having on quantitative text analysis.

To calculate the impact of bad OCR at the token level we took a two-part approach: first, we randomly selected 1000 features in 8 F1-ranked corpora, and returned the top 25 most similar words. With these we counted the number of matches between the OCR and TCP sub-corpora.

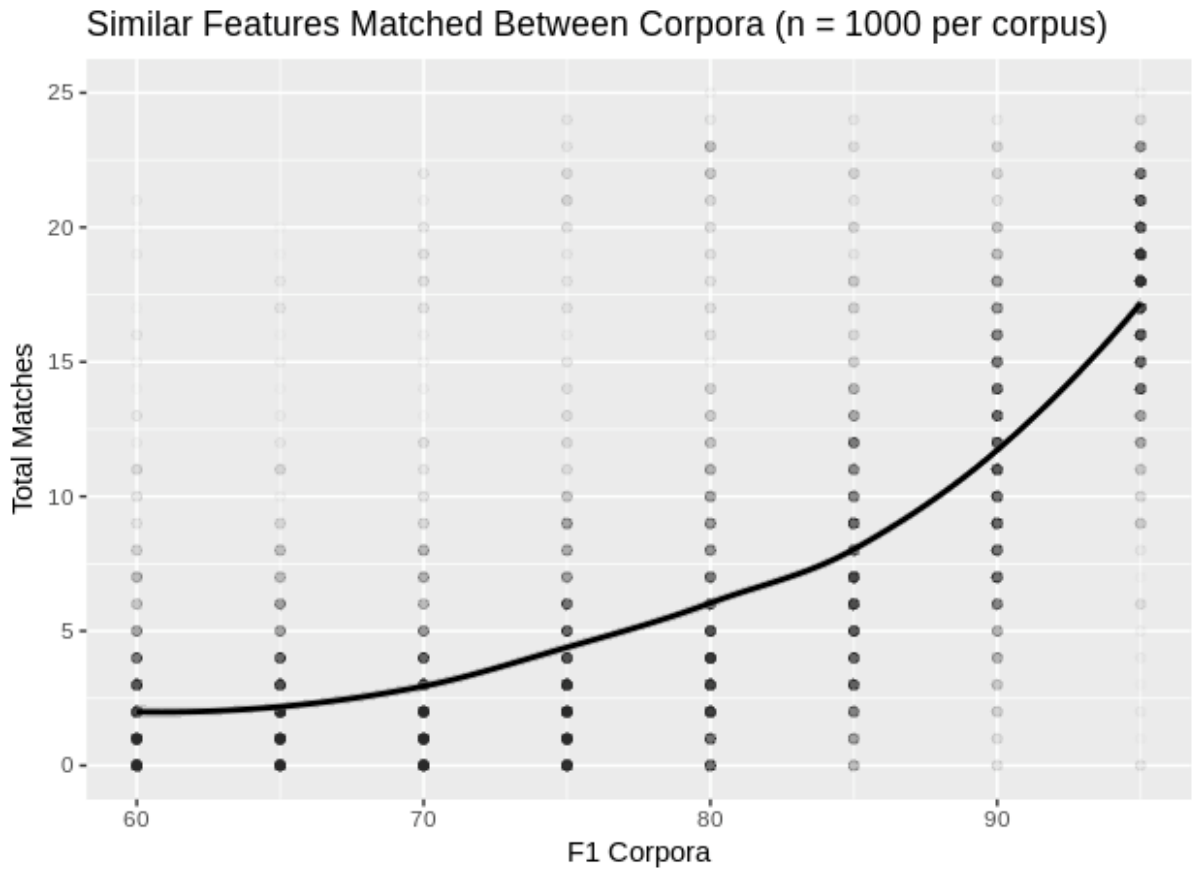

Fig. 12 Distribution of correct token similarity matches per F1 score sub-corpus

The results are clear: while there is a clear relationship between F1 scores and matched similar terms, it only becomes statistically noteworthy at the 75-80% F1-Score range ($p < 0.001$).

To investigate these results further – and in particular, to see whether matches were in fact meaningful – we chose a subset of tokens to qualitatively examine.[29] First, we examined the cardinal directions. In our experience, the ability of these words to return their siblings is a basic indicator of the content and quality of a dataset, and this case was no different. In the 60% F1 corpus 'south' was being returned as a match for 'north' and 'ea4l', and at 65% 'south', 'weft' (which was almost always returned instead of 'west'), and 'east' were the three most similar terms. For 'east', 65% was also the threshold for returning siblings, although it took until the 75% range for 'south' to be included.

The early appearance of 'weft' is worth noting. Again, due to the problems associated with the long-s in 'west', whenever it was the token from which similar words were being searched, results were very poor. In fact, it was not until the 85% F1 corpus that another cardinal direction was found in the top-25 matches – albeit one which was an OCR error: 'thesouth.' The majority of all results for 'west' were unrelated tokens and OCR errors. This indicates the very inclusion of 'west' in the OCR corpus is perhaps largely a result of OCR false positives, rather than true positives.

Additionally, we examined the words 'passion' and 'princess,' as both are qualitatively interesting, but technically problematic. The results were as one would expect: it was not until the 70% range that 'passion' returned anything which could be considered

[29] These were: "north", "south", "east", "west", "passion", "princess", "king", "public", and "religion."

contextually relevant - or words for that matter - when ‘passions’ and ‘poetry’ were matched (albeit behind ‘enthuiiafms’, ‘caufe'is, ‘derivd’, ‘entlhui’, ‘futceptible’, ‘jpane’, and ‘entbufiaftick’). From 75% it became easier to interpret some of the results (‘painting’, ‘emotion’, ‘object’, ‘mind’), although many of the tokens returned were derivations or OCR errors for passion itself (‘passions’, ‘paflion’, ‘pafion’, ‘and’paffions’). Results were even poorer for ‘princess’, for which mismatches and OCR errors dominated until the 85% range. In contrast, ‘king’ returned very good contextually relevant results from the 60% F1 range (‘scotland’, ‘earl’, ‘charter’, ‘lands’, ‘son’, ‘robert’, ‘daughter’, ‘married’, ‘england’, ‘lord’, ‘succeeded’, ‘william’, ‘alexander’, ‘great’, ‘david’, and ‘heir’).

Finally, we looked at two further words of interest to eighteenth century historians, but less problematic for OCR software: ‘public’ and ‘religion.’ In the former case, OCR errors dominated until the 70% F1 range, at which point the terms ‘private’, ‘people’, and ‘government’ were returned. At 75% ‘nation’, ‘country’, ‘revenue’, and ‘will’ emerged; and at 80% ‘state’, ‘general’, ‘national’, and ‘interest’ were added. In the case of ‘religion’ we find ‘christian’ as the first result at the 60% range (amongst 24 other results which were OCR errors). This appears to have been an outlier, however, as at 65% the results were entirely made up of OCR errors. At 70% the results become clearly relevant with ‘religious’, ‘church’, ‘protestant’, ‘christian’, ‘christianity’, ‘popery’, ‘god’, ‘true’, and ‘civil’ all being returned before any OCR noise, and from 75% the results are contextually, and OCR-wise, excellent.

Ultimately, this brief qualitative review confirms the quantitative tests above.[30] Although relevant results can be found in some cases immediately (“king”), these are generally outliers. Truly meaningful results were generally found after the 80% mark - with the very important exception of words containing the long-s and ligatures. In these cases, results remain very poor throughout. While this obviously impacts searches in which the keyword contains these features, it must also be remembered that these words are also likely to be underrepresented (if represented at all) as results. For example, while ‘king’ often returned ‘queen’ and/or ‘prince’, ‘princess’ was never returned (although this match was made in the TCP corpus).

Authorial Attribution

The final analysis was authorial attribution using the Stylo package for R. As our aim was to test the impact of OCR, rather than corpus sizes, we chose to use the 25 most prolific authors in ECCO-TCP in terms of documents included, with the aim of getting more varied texts. We used the next 25 authors as non-author training material. For the test-authors we created sub-corpora based on F1 scores.

[30] It should be noted that the makeup of the sub-corpus plays a role in the qualitative step, as matches judged relevant are dictated by the contents of the documents which make up the corpus (in addition to OCR). Thus, it is important that these results mimic the quantitative tests.

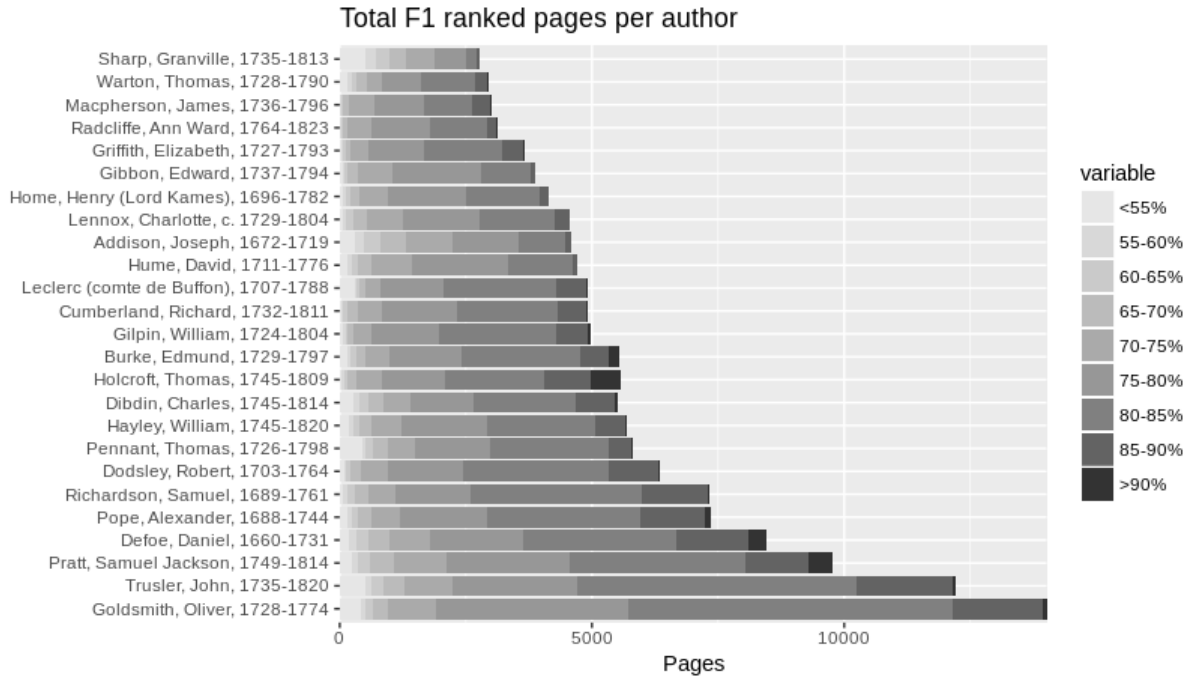

Fig. 13 Tested authors, total pages, and their F1 score distribution

To test authors, we extracted ranked F1 test sets (9), as well as matching TCP sets to test validity. Each of these sets were made up of the number of pages which best estimated token counts of 1,000, 2,500, 5,000, and 10,000.[31] In some cases the number of pages available were fewer than necessary for the desired token length. In these cases (73) that specific F1-based test was skipped. The rest of an author's work, minus subset pages, were used as training data. On this data we made use of three statistical tests (delta, k-nearest neighbour, and nearest centroid classifier), three different feature selections (single tokens, token bigrams, and character three-grams), and tested the most frequent 100, 200, 300, 400, and 500 features. In each test the OCR and TCP subset were measured to predict authorship, and instances in which the test failed to successfully identify the author of the TCP subset (7,158; 17.7% of all tests) the results for the OCR test were discarded.[32] The results below are for the remaining 30,864 tests (Figure 14).

[31] The size of a reliable corpus is debated (Eder 2013), but ranges generally between 1,000 (Biber 1990, 1993) and 10,000 words (Burrows 2007).

[32] Although not under the remit of this paper, this result is worth reflecting on: over 80% of all tests correctly identified the author in question with no human intervention beyond choosing the variety of tests.

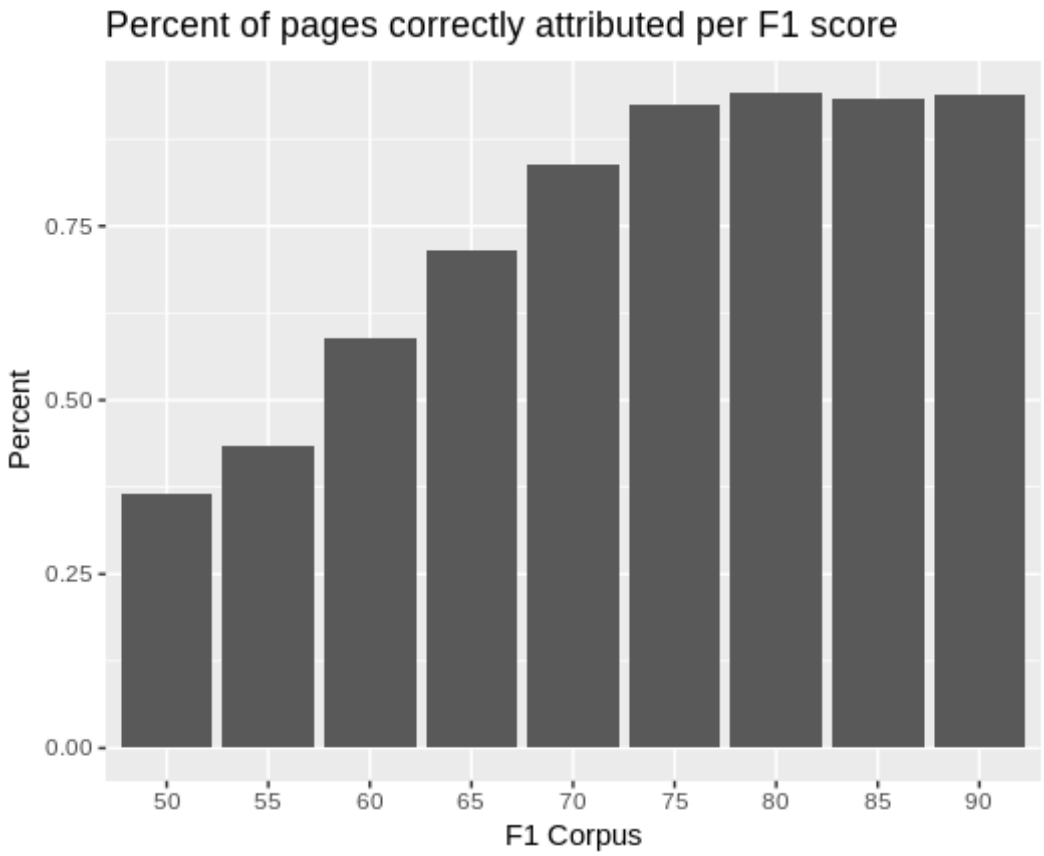

Fig. 14 – Percent of tests which correctly attributed authorship per F1 score

While the relationship between OCR quality and authorial attribution success is perhaps unsurprising, the threshold at which OCR quality impacts results may be. In our tests, there were negligible improvements to results due to cleaner data from the 75% range (success rates of 92.3%, 94.0%, 93.3%, 93.9%). Although the goal of this analysis was not to investigate the impact corpus size had on authorial attribution, but instead OCR errors, the results when examined by corpus size are also worth looking at.

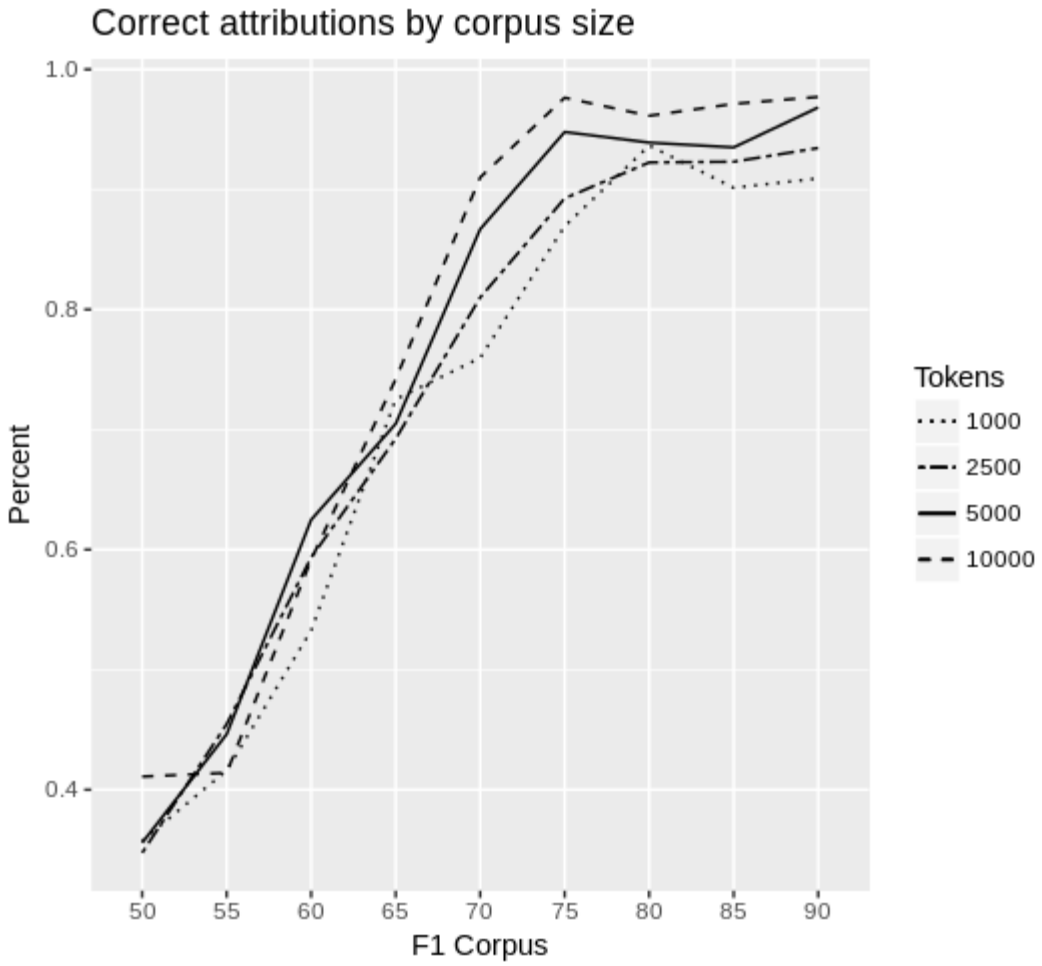

Fig. 15 Authorial attribution test accuracy per F1 corpus size

Table 4 Correct author attribution per corpus size and F1 score

F1 %	65-70%	70-75%	75-80%	80-85%	85-90%	>90%
1000t	0.7251351	0.7597368	0.8693027	0.9365512	0.9013661	0.9091731
2500t	0.6923742	0.8099839	0.8925926	0.9224638	0.9230964	0.9344928
5000t	0.7048077	0.8669003	0.9478691	0.9389922	0.9350242	0.9682927
10000t	0.7419811	0.9099843	0.9764079	0.9613636	0.9712871	0.9770492

It is not surprising is that, once OCR errors are no longer the key variable impacting authorship attribution (~75-80% and above), the size of the corpus emerges as a key indicator for test accuracy. However, what may be surprising is that, between accuracy ranges, corpus size has an impact on results which may indicate that, in some cases, a larger corpus may be as, or more, valuable than cleaner text.

3. How to Analyse Data

The results up until now demonstrate that OCR below the 70-75% accuracy level has, unsurprisingly, a strong negative impact on a number of analytical methods. However, what may be surprising is that OCR data with, what may appear subjectively to be, substantial issues – data in which as much as two or three words out of ten are misidentified – is still potentially very useful. However, there is an elephant in the room: how does one know the quality of their OCRed text? To this we provide three tentative responses.

First, many OCR outputs include confidence levels, and although these confidence scores are themselves a topic of much research and debate, in the case of ECCO, we have shown that there is a clear relationship between these estimates and actual F1 scores (albeit, the software overestimated accuracy). Thus, if one has these confidence figures, there is evidence that one can be confident in them.

However, if one does not have these figures (or if one wants an additional measurement of accuracy), there is the possibility of randomly sampling pages and counting errors to estimate precision. Our research shows that for ECCO-TCP, with regard to precision, the mean amongst all documents was 0.734 (ranging from 0.11 to 0.927). However, it should also be noted that the range of which precision could vary within documents averaged at 0.54.

The third response to this problem is one which points to future research: it may be possible to develop models which estimate OCR accuracy. Although proposing a gold standard for this is beyond the remit of this research paper, we did examine three preliminary possibilities. First, we began with the hunch that lexical diversity may work as an indicator, as OCR noise would throw off counts. To this end, we tried MTLD, HDD, and Maas measurements and compared the two corpora.[33] Although it is well argued that lexical diversity cannot be measured in a way which is independent of text length, the hope was the extreme tail in OCRed data would make these texts obvious. However, all three measurements failed to offer any indication of potential OCR quality, which means lexical diversity of bad OCR seems to resemble clean texts in this case. However, Tweedie and Baayen may offer a hint for moving forward: “partial randomisations, where text is permuted in sections to allow for confidence intervals to be constructed around the empirical values of the measures.”[34]

We also measured the distribution of the letter ‘s’ and the letters ‘ct’ in OCR documents (Figures 16-19).

[33] Methods chosen following McCarthy and Jarvis (2010)

[34] Tweedie and Baayen (1998), p. 324.

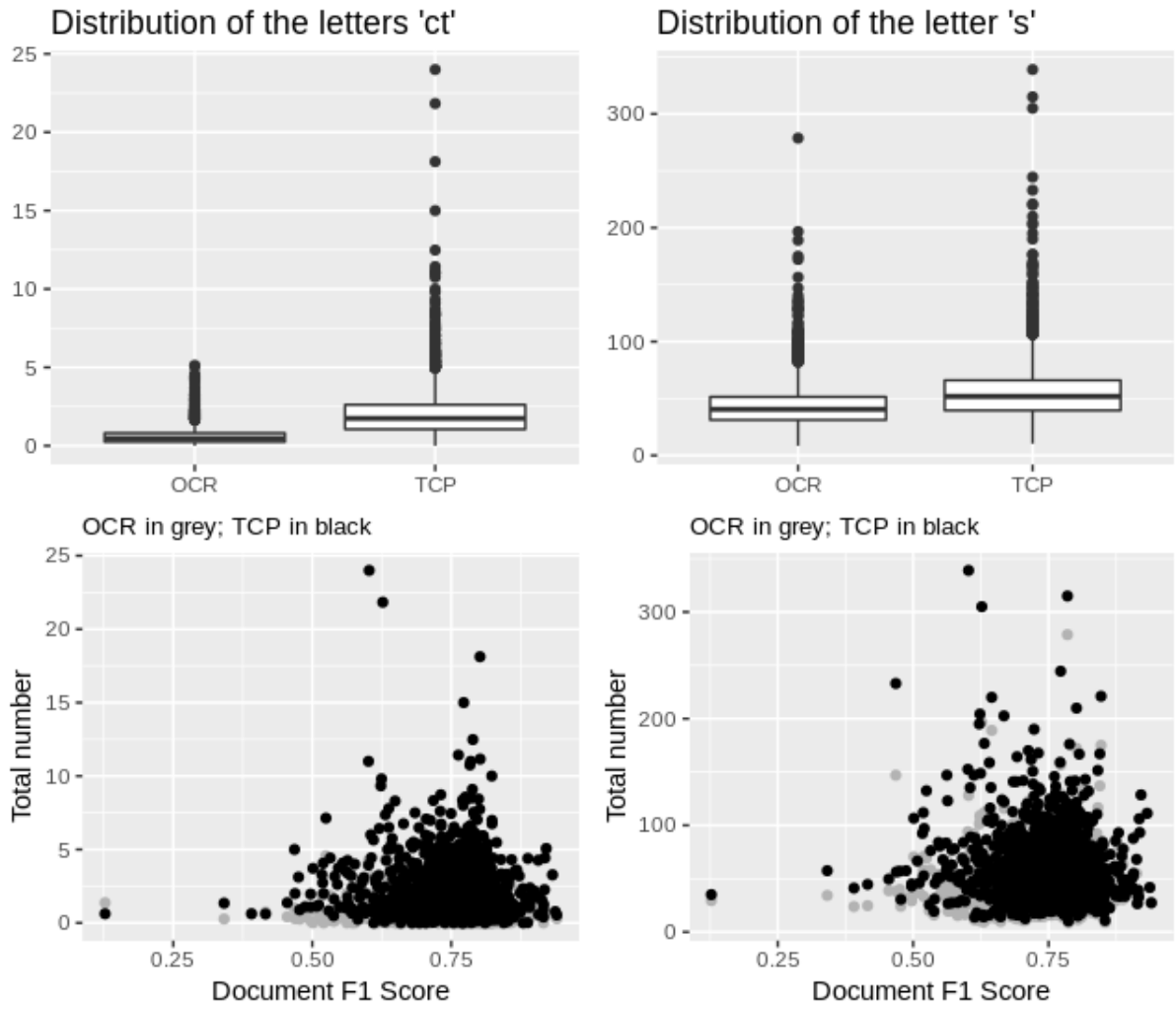

Fig. 16-19 Distribution of ‘s’ and ‘ct’ in OCR and TCP corpora

In both cases, there appears to be a noticeable tail at the low-end of the F1 table (< 0.6). While this may appear promising at first, it is worth noting that we may simply be witnessing the cause of bad OCR, rather than an indication of it. That is, the documents with lower F1 scores and fewer instances of the letter ‘s’ are there exactly because there are more instances of ‘s’ in the clean document. Thus, extrapolating from this point may have problematic side effects (for example, ranking documents which naturally have fewer instance of the letter ‘s’ as poor OCR). Additionally, the specificity of the test limits its value; results are tied to documents printed when the long-s was in use, in a language which made use of it, and were digitized with software that did not recognize it.

Finally, in an attempt to estimate OCR quality, we modelled the relationship between specific OCR artefacts and F1 scores. Due to the huge number of variables necessary to do this, we used the Glmnet R package to fit a Least Absolute Shrinkage Selector Operator (lasso) generalized linear model, a method which is purposefully conservative in terms of the variables it tests, and penalizes coefficients so as to keep their values minimal, making it very useful for feature selection in large datasets. To this end, we created a dataset which included every character-feature separated by whitespace in the OCR corpus, and counted their instances, per F1-scored page, adjusted for page frequency. This resulted in a matrix containing 2.7 million types with 101 separate occurrence counts. We also created a dataset in which all types found in the TCP corpus were purged (leaving 2.3 million features).

The results are promising, and a number of features are identified as statistically significant to F1 scores. However, many of these are clearly corpus specific (and those which do not appear to be so initially may still be).[35] Thus, these are preliminary results which suggest the possibility of creating a model which estimates OCR errors which is built upon the statistical likelihood of noise at various levels of F1 accuracy. However, to ensure this model is not corpus specific, it must be built with corpora additional to ECCO.

Conclusion

This paper has offered a number of insights into the impact OCR quality can have on quantitative historical-text analysis. First, we have provided an overview of the qualities of the ECCO TCP corpus which may be indicative of the corpus as a whole, as well as offer insights into large digitized corpora in general. These details are, as Prescott has pointed out, essential knowledge for researchers using ECCO. Second, this analysis allowed us to demonstrate that OCR errors are not neutral when it comes to eighteenth-century texts, and instead, the long-s and ligatures are statistically more likely to result in erroneously recognized words. This offers important insights to the many researchers interested in post-processing OCR data. Third, we have demonstrated that the impact of OCR, with regard to topic modelling, is perhaps less problematic than one may initially guess, and that OCRed data can be suitable for this type of analysis. In particular, if one aims for an exploratory view of data, a corpus as clean as ECCO is satisfactory. Fourth, the impact on collocations is problematic, but with the correct domain expertise (in terms of corpus makeup and research aims) the analysis remains viable with OCRed corpora - in particular for data above the 80% F1 point. Fifth, OCR errors are less of an issue when it comes to vector space models than the length of the document. Additionally, at the token level, useful data begins to emerge as early as the 70% range, but meaningful results (qualitatively and quantitatively) are generally found after the 80% mark, with the exception of tokens containing the long-s and ligatures. Sixth, when it comes to authorial attribution, the impact of OCR errors is greater below the 75% level, although this must be weighed against corpus size. Finally, we have attempted to provide some initial research into estimating OCR errors in corpora. For eighteenth-century texts, ligatures and the long-s appear to be a fruitful avenue for further research. More generally, lexical diversity measurements are not indicative of OCR quality out-of-the-box, but research into random sampling may change this. Finally, our initial tests on the ECCO-OCR corpus show GLM lasso regularization is able to identify a number of tokens which are predictors of OCR quality. Thus, a more general attempt to identify features is an area in which further research should be conducted.

[35] Some examples of statistically relevant features for predicting OCR noise which are clearly corpus specific include: ‘slatutes’, ‘felf-defence’, and ‘majefy's’. Those which may not be domain specific include: ‘~’, ‘¼’, ‘½’, ‘¾’, ‘ioi’, ‘i_’, ‘_i’, ‘/’, ‘]’, ‘·’, ‘$’, and ‘_’.

References

Alex, B., Grover, C., Klein, E. and Tobin, R. (2012). Digitised Historical Text: Does it have to be mediOCRe? In Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna, September 21, 2012 (pp. 401-409).

Antoniak, M., Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. Transactions of the Association of Computational Linguistics, 6, pp. 107-119.

Boydens, I. (1999). Informatique, normes et temps. Bruxelles: Bruylant.

Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In Proceedings of the international working conference on advanced visual interfaces (pp. 74-77). ACM.

Benoit, K. (2018). quanteda: Quantitative Analysis of Textual Data. R package version 1.3.0. http://quanteda.io.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. Literary and linguistic computing, 5(4), pp. 257-269.

Biber, D. (1993). Representativeness in corpus design. Literary and linguistic computing, 8(4), pp. 243-257.

Blaheta, D. and Johnson, M. (2001). Unsupervised learning of multi-word verbs. In Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations (pp. 54-60)

Bullard, P. (2013). Digital Humanities and Electronic Resources in the Long Eighteenth Century. Literature Compass 10(10), pp. 748–760

Burrows, J. (2007). All the way through: Testing for authorship in different frequency data. Literary and Linguistic Computing, 22(1), pp. 27–47.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), pp. 139-173

De Bolla, P. (2013). The architecture of concepts: The historical formation of human rights. Oxford University Press.

Denny, M. J. and Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis, 26(2), pp. 168–189.

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. Literary and Linguistic Computing, 28(4), pp. 603-614.

Eder, M. (2013). Does size matter? Authorship attribution, small samples, big problem. Digital Scholarship in the Humanities, *30*(2), pp. 167-182

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal 8(1), pp. 107-121. https://sites.google.com/site/computationalstylistics/stylo.

Evert, S. (2008). Corpora and collocations. In Lüdeling, A. and Kytö, M. (eds.), Corpus Linguistics. An International Handbook, article 58, pp. 1212-1248. Mouton de Gruyter, Berlin.

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J.K., Franzini, E., Byszuk, J. and Rybicki, J. (2018). Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm. Frontiers in Digital Humanities, 5, p. 4.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), pp. 1-22. http://www.jstatsoft.org/v33/i01/.

Gablasova, D., Brezina, V. and McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. Language Learning, 67(S1), pp. 155-179.

Hill, M.J. (2016). Invisible interpretations: reflections on the digital humanities and intellectual history. Global Intellectual History, 1(2), pp. 130-150.

Holley, R. (2009) How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine, 15(3/4).

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition (IJDAR), 12(3), pp. 141-151.

Luyckx, K. (2010). Scalability Issues in Authorship Attribution. Ph.D. thesis, University of Antwerp.

McCarthy, P.M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods, 42(2), pp. 381-392.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Piotrowski, M. (2012). Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies, 5(2), pp. 1-157.

Prescott, A. (2018). Searching for Dr. Johnson: The Digitisation of the Burney Newspaper Collection. In Brandtzæg S. G., Goring P., Watson C. (eds). Chronicles News and Newspapers from the Early Modern Period to the Eighteenth Century. Brill.

Recchia, G. (2016). The Utility of Count-based Models for the Digital Humanities, paper presented to Digital Humanities Congress, Sheffield, 8-10 September 2016. https://www.dhi.ac.uk/dhc/2016/paper/98.

Roberts, M.E., Stewart, B.M. and Tingley, D. (2018). stm: R Package for Structural Topic Models. http://www.structuraltopicmodel.com.

Rodriquez, K.J., Bryant, M., Blanke, T., and Luszczynska, M. (2012). Comparison of Named Entity Recognition tools for raw OCR text. Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna, September 21, 2012, pp. 410-414.

Schofield, A. and Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. Transactions of the Association for Computational Linguistics, 4, pp. 287-300.

Sievert, C., and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

Spedding, P. (2011). 'The New Machine': Discovering the Limits of ECCO. Eighteenth-Century Studies 44(4), pp. 437–53.

Strange, C., McNamara, D., Wodak, J. and Wood, I. (2014). Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers. DHQ: Digital Humanities Quarterly, 8(1).

Traub, M.C., Van Ossenbruggen, J. and Hardman, L. (2015). Impact analysis of OCR quality on research tasks in digital archives. In International Conference on Theory and Practice of Digital Libraries (pp. 252-263). Springer, Cham.

Tweedie, F. J. and Baayen, R. H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. Computers and the Humanities, 32(5), pp. 323-352.

Uhrig, P., Evert, S., Proisl, T. (2018). Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes**.** In Cantos-Gómez, P. and Almela-Sánchez, M. (eds.), Lexical Collocation Analysis: Advances and Applications, pages 111–140. Springer International Publishing, Cham