

# Quantifying the Impact of the ECCO Remediation Project

Chris Houghton

May 2023

## Introduction

In May 2023, Gale concluded more than a year's work when it released remediated Optical Character Recognition (OCR) text in its two original digital archives: ECCO, the *Eighteenth-Century Collections Online*; and the *Times Digital Archive*.

Both of these archives have had significant value to the global academic community over the past twenty years. Given their ubiquity and the increasing scrutiny on OCR with the growing popularity of digital humanities, Gale determined that they would be the first of its archives to have sections of their OCR remediated.

[This Q&A](#) with Gale senior product manager Megan Sullivan explains the rationale behind the decision, and how the documents to be remediated were chosen given the size of these archives and the prohibitive cost of re-OCRing the whole collection.

This report examines the impact of remediating ECCO OCR on accuracy, OCR confidence and the practical results of text mining this data.

## Why Remediate?

ECCO part I was released by Gale in 2003, with part II following in 2007. In 2003, ECCO was the most ambitious digitisation project ever attempted, and parts I and II totalled approximately 33 million pages and 180,000 volumes, comprising most significant books published in English in the 18<sup>th</sup> Century.

The technology used to digitise primary source archives has improved significantly since ECCO was first developed, particularly the software used to create text transliterations of the scanned images, or Optical Character Recognition (OCR). [This page](#) gives more detail on the technical processes behind a digital archive.

Improving OCR quality meets one of the most repeated pieces of scholarly feedback. Better OCR leads to more accurate search results and potentially better text mining of these collections, including in Gale's leading text and data mining platform, *Gale Digital Scholar Lab*. This report seeks to test this hypothesis.

## The ECCO OCR Remediation Project

"In both The Times and ECCO, we used a combination of user data and research trends to identify content sets for OCR improvement. In many instances, we relied on user search trends and document retrieval data to better understand what topics and time periods users search and interact with. We also relied directly on user feedback. Our product team logs every user request that comes through to us. We took these requests into consideration when identifying this content. Finally, we looked at broader trends in scholarship to analyze current research. For example, we identified works by women and BIPOC authors as top candidates for OCR improvement."

Megan Sullivan, Senior Product Manager Gale

The ECCO Remediation project comprised:

**Total Documents Remediated: 17,068**

**Total Pages Remediated: 2,841,331**

## OCR Accuracy vs OCR Confidence

It is difficult to measure the accuracy of all documents at the scale of a Gale archive. For example, in ECCO this would involve human transcribing all 33 million pages and comparing them to the machine-created OCR – prohibitive both in time and cost.

Calculating the accuracy of archives therefore often involves taking samples of documents and hand checking them or running them through multiple OCR engines and comparing the results.

OCR Confidence is a score that is generated by the OCR engine being used to transliterate the scanned image. Essentially, it is a measure of *how well the engine thinks it has done* – the more time the engine has to take deducing what the letters on the page are, the lower its Confidence Score becomes.

OCR Accuracy and Confidence can be affected by combinations of many factors, including:

- Quality of the original document
- Clarity of the digital facsimile
- Where and when the document was scanned, and by whom
- Age of the document
- Age of the OCR algorithm

**OCR Accuracy and OCR Confidence are not the same** – older OCR engines are often more confident about their work and overestimate the quality of their results. OCR Confidence can also be significantly affected by non-text elements in a document like an illustration or photograph – the OCR engine will try to read this as text and fail, bringing its Confidence score down even though the text in the document might be transliterated very well.

Analysing the new OCR Confidence Scores for the ~2.8M remediated ECCO pages and comparing them to their original Confidence Score shows:

**Average OCR Confidence Change, all pages -31.3%**

## Quantifying the Impact of the ECCO Remediation Project

This seems like a drastic fall, but it is explained by a general improvement in OCR accuracy and should be viewed as an unequivocally good thing for researchers.

These results are consistent with the development of OCR technology in the past 20 years. As OCR technology advances, the algorithms are likely to be more precise in their character recognition. This increased precision may result in lower confidence scores for certain characters or words, as the software is more cautious about its predictions.

Similarly, modern OCR software possesses more complex algorithms and models for character recognition. This increased complexity can result in lower confidence scores for specific documents, as the software weighs a more extensive range of possible interpretations.

**CONCLUSION: OCR Confidence is still the best indicator of OCR Accuracy, but should be considered critically**

## The Effects of ECCO Remediation on OCR Accuracy and Confidence

As part of the remediation project, an external vendor was commissioned to test the accuracy of the remediated OCR, compared with the original OCR. To ensure a manageable project, they randomly selected two pages from each of the first 1380 documents to be remediated to compare accuracy before and after for a total sample of **2760 pages**.

Comparing the new OCR Documents (c.2023) with the original OCR Documents (created between 2002 and 2007) in the sample showed:

**Change in OCR Confidence: -19.04%**

**Change in OCR Accuracy: +24.48%**

The fall in confidence score in the sample reflects that seen in the whole population of ECCO documents remediated.

**CONCLUSION: Remediating ECCO OCR has led to a ~25% Improvement in Accuracy**

## Newer OCR Algorithms are more accurate but more cautious

The change in nature of OCR engines with regard to confidence scores can be seen by calculating the difference between OCR Confidence and OCR Accuracy in the original and new OCR instances. A positive differential between Confidence and Accuracy indicates an algorithm that is overstating its effectiveness in accurately transliterating a facsimile document. A negative differential indicates an algorithm that is more cautious, even if it is significantly more accurate.

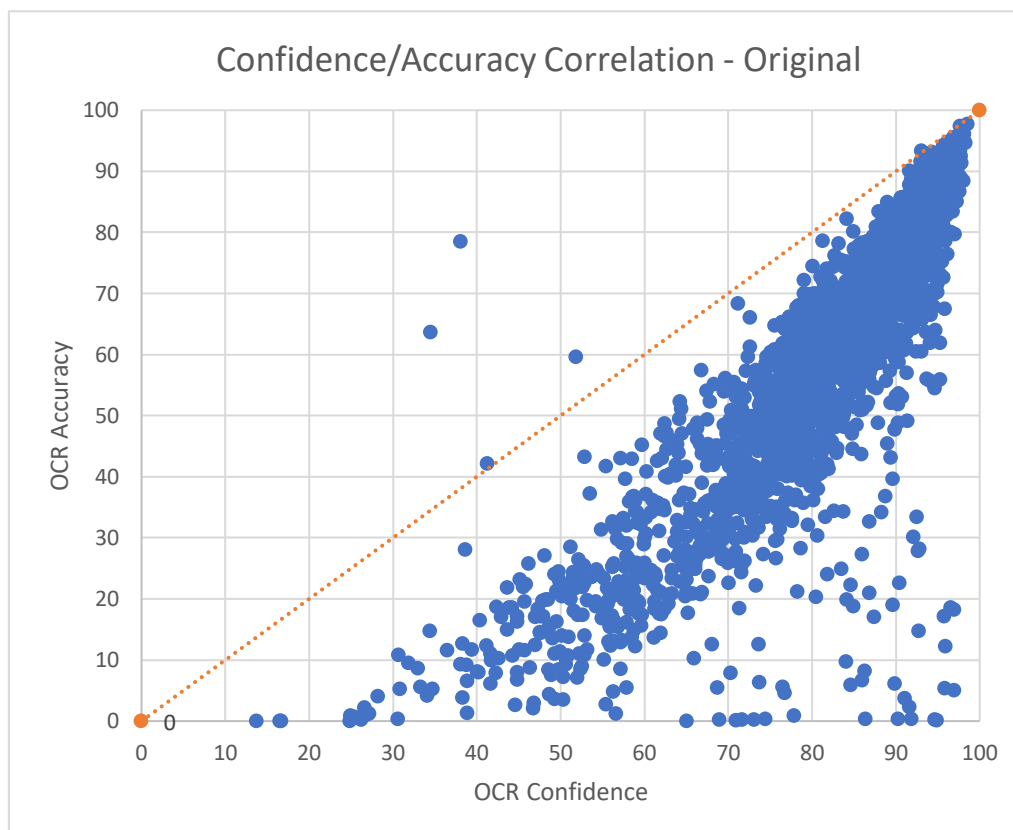
Comparing the differentials for the original OCR and the remediated OCR shows:

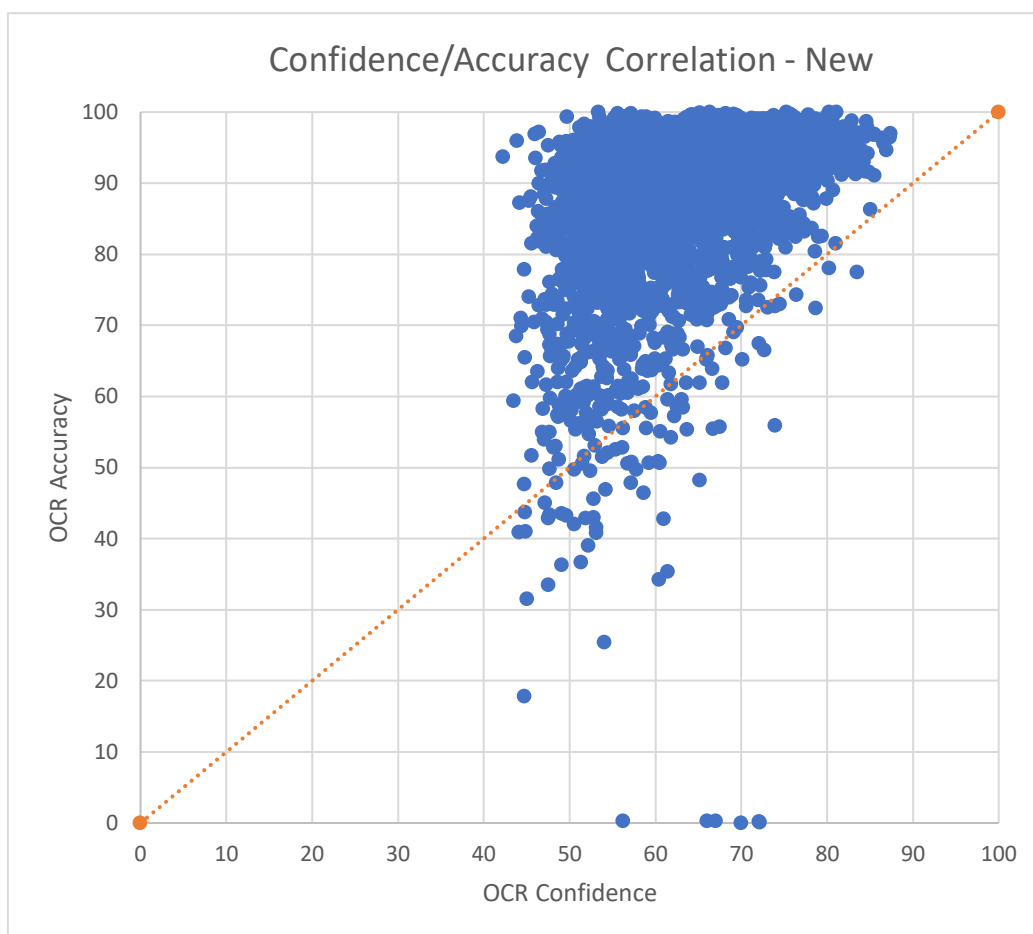
**Old Confidence/Accuracy Differential: +20.42%**

**New Confidence/Accuracy Differential: -23.09%**

## A Tighter Correlation Between OCR Accuracy and Confidence

Plotting OCR Confidence and OCR Accuracy scores for both the original version of the OCR and the new demonstrates the relative caution of the newer OCR algorithm:





A number of inferences can be drawn by comparing these graphs (the line indicates where OCR Confidence = OCR Accuracy):

1. This provides further evidence of the relative sophistication and caution of the newer OCR algorithm – the vast majority of documents are above the line in the newer OCR, meaning that most documents have OCR Accuracy scores higher than their OCR Confidence.
2. This also provides evidence that older OCR engines are more bullish about their capabilities - almost all of the original OCR documents have an OCR Confidence higher than Accuracy.
3. There are far fewer very poor (Accuracy <30%) OCR documents in the new OCR. The fact that there are still some pages with ~0 Accuracy suggests that the original physical document may be illegible.
4. The Confidence Scores in the new OCR set are a far better predictor of OCR Accuracy – many more documents in this set are close to the Confidence = Accuracy line.
5. Although they skew cautious (OCR Confidence < OCR Accuracy), the documents in the new OCR set are far more tightly grouped, suggesting that there is significantly less chance of the OCR making mistakes in accurately transliterating the text of the document.

**CONCLUSION: OCR Confidence, while more cautious, is a better predictor of Accuracy in the Remediated OCR**

## The Impact of more Accurate OCR on Text Mining

*Gale Digital Scholar Lab* is the leading integrated text and data mining platform, incorporating the OCR for hundreds of millions of Gale primary source documents with powerful text mining tools.

Running some of these tools on a sample of the new OCR documents for ECCO and comparing with the original OCR for the same documents illustrates the value of more accurate OCR to anyone undertaking a text mining project with this data.

To ensure a manageable experiment, a sample of 257 documents was taken. This sample was defined by being the first documents uploaded to a development instance of the *Lab*, which were the only documents available for mining when this experiment was conducted. By May 2023, the fully remediated set of 17,068 ECCO OCR documents had been added and were available for mining for anyone with access to the *Lab* and ECCO.

Two separate content sets of 257 documents were created – one containing the original OCR for the documents, and one containing the remediated OCR. Tools were then applied to both of these content sets with the same configurations, and the results examined. The cleaning tool in the *Lab* was not applied to any tool configuration so that all analyses were attempted on the raw data.

### Sentiment Analysis

This tool works by identifying tokens (words or characters) within a document set and then comparing them with the AFINN lexicon, a list of human labelled tokens, that can range in affective score from 5 to -5. Results can be visualised over time or on a plot of sentiment scores, and the analysis data can be downloaded.

Applying the Sentiment Analysis tool to the sample data gave:

	Total Tokens Identified	Tokens Scored for Sentiment	% of Total Tokens Scored
Original OCR	70,825,995	1,987,630	2.8%
New OCR	68,669,106	2,386,487	3.5%
Change Old>New	-2,156,889	398,857	

The fall in tokens (2.1M fewer) identified in the new OCR content set can be explained by the increased accuracy of the OCR identified earlier in the report. This suggests errors, including words that may have had a space inserted in the word in the original OCR (i.e. discontinued as discontinued), are now at least partially corrected in the new OCR. In the original OCR, discontinued would be two tokens, while in the new OCR, discontinued (no space) would be one.

The significant positive is the increase in tokens that are scored for sentiment, even as total tokens identified falls. This suggests that more accurate OCR enables the AFINN lexicon that powers the Sentiment Analysis tool to score more words, providing a better analysis of the texts. This can also be seen in the rise in tokens scored as a percentage of total tokens, from 2.8% in the original OCR, to 3.5% in the new OCR.

## CONCLUSION: More Accurate OCR enables more comprehensive Sentiment Analysis

### Ngram Analysis

The ngram tool in *Gale Digital Scholar Lab* can be configured to extract the most commonly occurring words or phrases from a content set.

Analysing the most occurring 1000 ngrams from both the original and new OCR content sets returned the following:

	Original OCR	New OCR	% Change
Total frequency Top 1000 ngrams	38,873,173	39,940,665	3%

Similar improvements to those in Sentiment Analysis can be observed – a small but significant increase in the ability of the tool to identify words or tokens in the OCR.

Further research will determine the impact of improving OCR on particular academic enquiries, but a sample of the ngram analysis demonstrates that this improvement in OCR accuracy has an impact on the order of the most commonly occurring words and phrases in the content set:

Position	Ngram (Original)	Count (Original)	Ngram (New)	Count (New)
1	the	3,012,816	the	3,164,959
2	of	2,204,346	of	2,201,623
3	to	1,563,126	a	1,561,738
4	a	1,558,437	to	1,554,239
5	and	1,326,546	and	1,384,955
6	in	976,614	in	1,010,831
7	or	693,982	or	701,359
8	is	575,977	is	577,109
9	of the	486,249	of the	514,962
10	that	427,785	that	452,905
11	by	379,682	be	397,771
12	be	376,211	by	389,506



## Quantifying the Impact of the ECCO Remediation Project

13	it	344,367	;	382,899
14	with	333,157	with	365,644
15	as	320,726	it	335,476
16	.	309,088	which	329,724
17	for	301,758	as	322,132
18	his	300,241	for	310,989
19	which	299,920	his	300,201
20	A	267,514	from	278,829

- Most discrepancies in the comparison of these two lists is relatively minor, i.e. the swap in order for 'be' and 'by'
- In the new OCR content set, ';' is the 13<sup>th</sup> most common ngram, but doesn't appear in the original OCR ngram list until position 40. This could be due to better accuracy in capturing the semi-colon – conversely, the high position in the original OCR of '.' could suggest that some of these full stops in the original OCR were actually mis-captured semi-colons, although this conclusion would need further close reading to confirm
- The longer Ngrams: 'which', 'of the', 'that' see the biggest increases in count in the new OCR content set

### Better Accuracy in identifying Ngrams

Analysing change in Ngram frequency between Original and New OCR (see Appendix A and B) provides the following:

**Change in top 50 Longest Ngram Frequency: +16%**

**Change in top 50 Shortest Ngram Frequency: -14%**

This suggests that more accurate OCR enables a quantitative tool like Ngram to better identify whole words, especially at longer lengths (where more errors are likely to occur due to the increased number of characters).

The inverse of this is that shorter Ngrams are less frequent in the remediated set of content as they often represent errors in the OCR ('t he' instead of 'the') – fewer OCR errors means fewer single letter Ngrams.




**CONCLUSION: More Accurate OCR improves Ngram Analysis somewhat by enabling better word identification**

## Named Entity Recognition

This tool extracts Named Entities from documents within a Content Set. The current implementation uses spaCy's Annotations Named Entity Recognition (NER) module. Entities are grouped into specific entity categories or "classes".

As an illustration, here are the top entities identified in the 'Geography' category within the Original and New Content Sets:

**Named Entity Recognition**  
Pre-remediation ECCO OCR

 Add Note
  Download
  Help

Entity ↕	Category ↕	Documents ↕	Count ▼
Earth	Geography	106	2285
Europe	Geography	173	1899
Sea	Geography	99	1229
Africa	Geography	122	887
Jupiter	Geography	76	811
South	Geography	91	639
Asia	Geography	115	614
East	Geography	117	534
River	Geography	52	529
Sun	Geography	64	486
Greeks	Geography	90	437
Latin	Geography	60	377
North	Geography	84	298
Mars	Geography	79	293
Heb	Geography	5	253
Venus	Geography	74	231

Named Entity Recognition Post-remediation ECCO OCR			
Entity	Category	Documents	Count
Earth	Geography	108	2451
Europe	Geography	169	1903
Sea	Geography	82	1255
Jupiter	Geography	88	1081
Africa	Geography	127	1012
South	Geography	93	812
East	Geography	134	766
Asia	Geography	130	706
Sun	Geography	63	578
River	Geography	51	562
Greeks	Geography	93	492
Heb	Geography	8	388
Island	Geography	54	373
North	Geography	84	357
Mars	Geography	81	313
Latin	Geography	58	309

Analysis of these results shows:

- Most entities are found in more documents in the New OCR set than in the Old, with the exception of 'Europe', 'Sea', 'Sun', and 'Latin'. Of these, 'Europe', 'Sea' and 'Sun' have been identified more in the content set
- As expected, the tool finds more of each entity within the post-remediation content set
- Some of the largest rises in documents identified and count are for the terms 'Africa' (122/887 > 127/1012) and 'Asia' (115/614 > 130/706). Given the current importance of studying colonial and post-colonial history, tools that more accurately identify these geographic terms are welcome.

## Conclusions

One of the biggest effects of remediating over 2.8 million pages of 20-year-old OCR data has been the OCR Confidence Score of the documents. Analysing every one of the pages in the 17,068 documents re-OCRd shows a drop of 31.3% in OCR Confidence. This should be viewed by researchers as a significant positive as it suggests modern OCR software that is more sophisticated and less over-confident when calculating its Confidence Score.

Conversely, the accuracy of the OCR has increased significantly. Close analysis of a sample of 2760 pages demonstrates an improvement in accuracy of nearly 25%, even as OCR Confidence falls by 19% in the same sample.

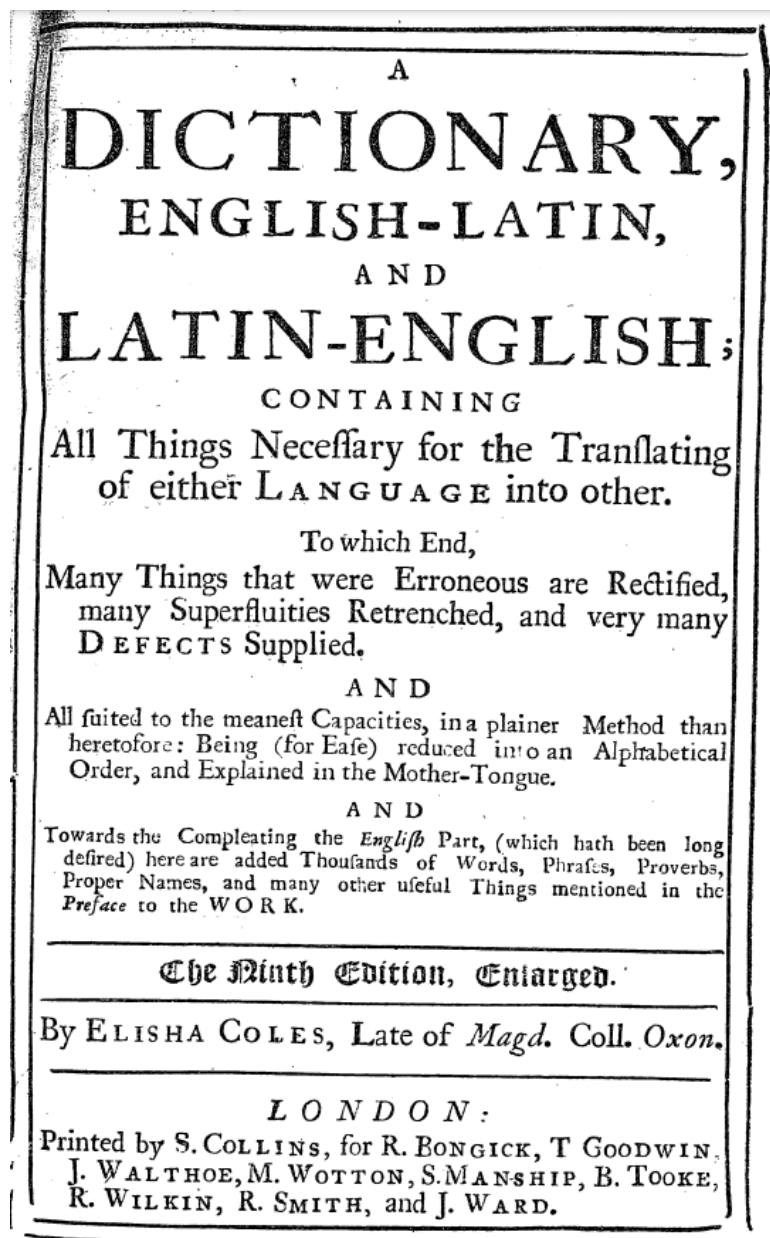
An important outcome of the remediation is the significantly closer correlation between OCR Confidence and Accuracy to the extent that OCR Confidence is a considerably improved indicator of OCR Accuracy in the newly created OCR corpus. The remediation project has generated much more accurate OCR with far fewer outliers in the data, either high confidence/low accuracy or low confidence/high accuracy documents.

While improving OCR Accuracy increases the effectiveness of retrieving documents through search, generating OCR Confidence Scores that are better indicators of OCR Accuracy ensures scholars can better trust it as an indicator of which documents to examine or discard in a text mining context. Having said that, whether the OCR was created in 2023 or 2003, it remains the case that, apropos a better mechanism, OCR Confidence remains the best indicator of OCR Accuracy but still should be considered with a critical eye.

Applying some of the tools in *Gale Digital Scholar Lab* to the original and newly remediated data gives the unmistakable conclusion that more accurate OCR leads to better outcomes in natural language processing tools. From enabling more tokens to be matched against a sentiment lexicon to improving Ngram analysis by correctly identifying more words, Gale's remediation project will improve outcomes for all researchers.

## Appendix

### Appendix A: An Comparative Example of OCR remediation



Coles, Elisha. A dictionary, English-Latin, and Latin-English, Containing all things necessary for the translating of either language into other. To which end, many things that were erroneous are rectified, many superfluities retrenched, and very many defects supplied. And all suited to the meanest capacities, in a plainer method than heretofore: being (for ease) reduced into an alphabetical order, and explained in the mother-tongue. And towards the compleating the English part, (which hath been long desired) here are added thousands of words, phrases, proverbs, proper names, and many other useful things mentioned in the preface to the work. The ninth edition, enlarged. By Elisha Coles, Late of Magd. Coll. Oxon. Printed by S. Collins, for R. Bongick, T. Goodwin, J. Walthoe, M. Wotton, S. Man-Ship, B. Tooke, R. Wilkin, R. Smith, and J. Ward, [1720?]. Eighteenth Century Collections Online, [link.gale.com/apps/doc/CB0126489824/ECCO?u=gale&sid=bookmark-ECCO&xid=156bdad2&pg=1](https://link.gale.com/apps/doc/CB0126489824/ECCO?u=gale&sid=bookmark-ECCO&xid=156bdad2&pg=1). Accessed 7 June 2023.

### Original OCR

DICTIONAR Y, EN GL IS H - LA TIN "I AND .LATI N-ENGLISH; C= ON T A IN  
INP~ G All Things Necessary for the Translating of eithe| LANG UAGE  
into other. Tob ·ihich En7d, Many Things that were Erroneous are  
]Rectified, many Superfluties Retrenched, and very many DE SECTs  
Supphied. All faited to the tneaneft Capacities, in a plainer Method  
than heretofore: Being (for Ease) reduced inro an Alphabetical Order,  
and Explained in the Mother-Tongue. Towards the· Compleating the  
Englijb Part, (which hath been long detred) here are added Thouflands  
of Words, Phira~ts, Proverbs, Proper Names, and many oth;er useful  
Things mentioned in the Preface to the WO~R K. ke~ifeFt~tit Eg~i-  
ftion, ~EmargeD. -By EL IS HA CO L E s, Late of Magd. Coll. Oxon.  
·Printed by S. COL L Ius, for R. BON GICK, T GoonwIN- J. WTaAL T HOE,  
M. WOT T ON, S.1%A N-8 H IP, B. TOOKE, R. WILKIit, R. SMITn, and J.  
WARD.

### Remediated OCR

DICTIONARY,,ENGLISH-LATIN,,AND,LATIN-ENGLISED,CONTAINING,All Things  
Necessary for the Translating of either Language into other.,To which  
End,,Many Things that were Erroneous are Rectified, many  
Superfluties Retrenched, and very many Defects Supplied.,AND,All  
suited to the meanest Capacities, in a plainer Method than  
heretofore: Being (for Ease) reduced into an Alphabetical Order, and  
Explained in the Mother-Tongue.,AND,Towards the Compleating the  
English Part, (which hath been long desired) here are added Thousands  
of Words, Phrases, Proverbs, Proper Names, and many other useful  
Things mentioned in the Preface to the WORK.,She Kmth Evitton,  
Enlarges.,By El is ha Coles, Late of Magd. Coll.  
Oxon\*,LONDON:,iPrinted by S. Collins, for R. Bongick, T Goodwin, J.  
Walthoe,M. Wotton, S.Man-ship, B. Tooke, R. Wilkin, R. Smith, and J.  
Ward.

- As expected, the newer OCR engine captures many words much better than the original version ('Englijb', 'Supphied', 'detred')
- The newer OCR engine is significantly better at capturing the capital letters at the start of the page, and has inserted far fewer spaces into words ('EN GL IS H' vs 'ENGLISH').
- The newer engine seems to cope much better with marks on the page. When capturing the word 'either' in the sentence "All Things Necessary for the Translating of either Language into other", the original OCR has read a mark above the 'r' as the dot in an 'i', so captures 'eithe|'. This error is not made in the remediated version.
- Faded or poorly scanned text is captured more effectively. The word 'into' in the sentence "...reduced into an Alphabetical Order..." is now captured correctly ('inro' in the original) despite the word being faded in the scan.
- The newer OCR engine still struggles with Gothic script in "The Ninth Edition, Enlarged", but makes a far better job of capturing it than in the original ('She Kmth Evitton, Enlarges.' Vs 'ke~ifeFt~tit Eg~i-ftion, ~EmargeD.').

## Quantifying the Impact of the ECCO Remediation Project

- The word 'LONDON' above the list of Printers isn't captured in the original, but captured perfectly (even capturing the colon following it) in the remediated version.
- The Printers names, while not completely perfect (Man-ship for Manship) are significantly better than in the original (S. COL L lus, for R. BON GICK, T GoonwIN- J. WTaAL T HOE, M. WOT T ON, S.1%A N-8 H IP, B. TOOKE, R. WILKlit, R. SMITn, and J. WARD.)

### Appendix B: Change in Ngram frequencies in 50 longest Ngrams

ngram	Length of Ngram	Count in old	Count in new	% change
according to the	16	6377	32%	-100%
according to	12	14365	15983	11%
belonging to	12	9817	12919	32%
particularly	12	6612	7288	10%
part of the	11	12795	14460	13%
between the	11	9733	11602	19%
in order to	11	7673	8356	9%
in the fame	11	7220	N/A	
through the	11	6864	8213	20%
of the fame	11	6515	N/A	
which they	10	12446	13847	11%
one of the	10	12063	13239	10%
particular	10	10921	12049	10%
before the	10	9996	11228	12%
afterwards	10	9948	11489	15%
out of the	10	9512	10624	12%
that which	10	8853	10426	18%
account of	10	7368	7411	1%
as well as	10	7249	7784	7%
themselves	10	7167	11456	60%
applied to	10	7073	7717	9%
called the	10	6517	7465	15%
frequently	10	6438	7391	15%
where they	10	6250	6896	10%
which the	9	21404	24232	13%
have been	9	17565	18856	7%
the other	9	17097	19386	13%
according	9	15929	17640	11%
which are	9	14620	14343	-2%
under the	9	14160	15944	13%
the whole	9	13787	15917	15%
therefore	9	13695	15300	12%
that they	9	13449	14690	9%

## Quantifying the Impact of the ECCO Remediation Project

different	9	13427	16977	26%
where the	9	12054	13889	15%
a kind of	9	11943	13263	11%
belonging	9	11565	15230	32%
the first	9	11342	22923	102%
any thing	9	11138	12868	16%
they were	9	11134	12236	10%
after the	9	11064	11884	7%
generally	9	10918	12007	10%
sometimes	9	10892	18506	70%
and other	9	10178	11050	9%
should be	9	10144	23373	130%
there are	9	10140	9232	-9%
about the	9	9693	10981	13%
among the	9	9384	10612	13%
they have	9	9323	10138	9%
following	9	9191	10249	12%
			Average	16%

## Appendix C: Change in Ngram frequencies in 50 shortest Ngrams

ngram	Length of Ngram	Count Old	Count New	Change %
a	1	1558437	1561738	0%
.	1	309088	93748	-70%
A	1	267514	1561738	484%
l	1	243339	219045	-10%
,	1	188612	151454	-20%
i	1	178031	219045	23%
;	1	168133	382899	128%
-	1	149874	87978	-41%
'	1	136148		-100%
t	1	106347	12072	-89%
f	1	94098	67104	-29%
1	1	77482	40843	-47%
o	1	73398	21228	-71%
l	1	72214	21341	-70%
r	1	63322	23898	-62%
e	1	62520	24099	-61%
"	1	60433	44487	-26%
:	1	56666	93652	65%
3	1	52582	39934	-24%
s	1	51725	19669	-62%



## Quantifying the Impact of the ECCO Remediation Project

3	1	51506	39934	-22%
2	1	49325	56577	15%
n	1	41504	9601	-77%
C	1	40677	12676	-69%
S	1	40581	19669	-52%
1	1	39742	40843	3%
c	1	39334	12676	-68%
*	1	39176		-100%
2	1	38832	56577	46%
E	1	35018	24099	-31%
(	1	34211	28284	-17%
4	1	32992	27567	-16%
4	1	32689	27567	-16%
y	1	32089	29611	-8%
d	1	30972	16574	-46%
L	1	29277	21341	-27%
)	1	29102	25408	-13%
6	1	27484	19408	-29%
v	1	27108	22986	-15%
5	1	27038	31862	18%
P	1	26226	11227	-57%
6	1	26040	19408	-25%
T	1	25431	12072	-53%
R	1	24423	23898	-2%
5	1	24296	31862	31%
&	1	23338	19193	-18%
7	1	22770	19075	-16%
h	1	20892	9880	-53%
J	1	20516	53603	161%
N	1	20484	9601	-53%
			Total	-14%