Unraveling Confidence: Examining Confidence Scores as Proxy for OCR Quality

Mirjam Cuper(✉), Corine van Dongen, and Tineke Koster

KB, National Library of the Netherlands, Prins Willem-Alexanderhof 5, 2595 BE The Hague, The Netherlands
mirjam.cuper@kb.nl

Abstract. While performing Optical Character Recognition (OCR), most engines provide confidence scores. These scores give an indication on how certain an engine is that a word or character is correctly determined. The practical application of this score is not yet clear and various studies have discussed the (un)usability of these confidence score as an estimation of OCR quality. Using a dataset of 2000 historical Dutch newspapers we investigated different aspects of the confidence score as provided by ABBYY Finereader, while also looking for a way to use the confidence score as an indication of quality. Such an indication could be used by institutions to determine which part of their collection would benefit from re-OCRing or post-processing. We found that the reliability of the confidence score as a measure of quality is largely dependent on the way the engine has been configured. In addition we show that when there is a high enough correlation between the word confidence and the Word Character Error (order independent) the word confidence can be used to calculate a proxy measure for categorizing digitized texts. However, such a measure must be recalculated for individual OCR engine set ups and producers. For our dataset this proxy measure performs well for the separation of digitized texts into categories of those with a very good and those with a very bad quality with total accuracy of 83%.

Keywords: ocr quality · confidence score · quality indication · digitisation workflow

1 Introduction

Mass digitization is often used by heritage institutions to digitize their textual collections. As there is a continuous growth in methods and techniques to analyze these digitized texts automatically, the need for high quality digitized texts is increasing [11,12,14,19]. However, automatic quality measurement of these texts is still an unsolved issue, and the question of how to automatically determine what parts of collections are candidate for re-OCRing or post-processing is still unanswered [3,11,16].

Some of the frequently used Optical Character Recognition (OCR) software engines that produce these digitized texts provide a measure, the confidence

G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14191, pp. 104–120, 2023.
https://doi.org/10.1007/978-3-031-41734-4_7

score, that indicates how certain the engine is that the suggested word is the correct word [7]. Numerous studies mention the use of these confidence scores as an indication of OCR quality, however there is no consensus on the applicability. Some studies report a relation between the word confidence and quality [4,5,12,17], but other studies suggest caution when using them [11,15,19]. Overall, there is no agreement about whether these confidence scores are a reliable indication of quality. To make matters more complex, there is little to no transparency on how the confidence scores are calculated, and this calculation may vary between producers or software versions [5,6,16,19]. This makes it unlikely to find a single solution for all OCR engines.

In order to improve our understanding of the usefulness of the confidence score we conducted several exploratory analyses. From our institutions' perspective the quality of the whole page is very important, as this can be used to determine which pages need to be re-OCRed of post-processed. We therefore try to determine a way to measure if the confidence score can be used as an accurate proxy measure for quality on the page level.

Furthermore, in order to get a better insight into the specific performance of the OCR engine, we analyzed the results of the OCR on word level, comparing OCR results with a Dutch lexicon and to Ground Truth.

2 Related Work

Finding information about the confidence scores of OCR engines is extremely hard, and there appears to be no standard. The documentation on the confidence score and what a confidence score exactly represents differs per company. ABBYY Finereader describes their word confidence as an estimated probability that the chosen word variant is correct. However, they point out that this metric is only useful for in-page word comparisons [2]. The documentation of Kofax stated little information about their word confidence. They mention that 0% means low confidence and 100% means a high confidence. Furthermore, they mention that the confidence measure is not comparable between different engines [10]. For other engines we could not find information about whether they use a confidence score or how the confidence score is calculated.

We were unable to find any documentation on how the different OCR engines actually calculate the confidence scores. Some engines mention that bonuses and/or penalties are also incorporated, but a precise method is not given. It appears that the user is also able to influence the confidence calculation by changing certain settings [2]. Due to the lack of standardization and transparency for the confidence calculation a comparison between engines is not useful in our setting. In fact, as confidence scores are apparently defined by the recognition engine, even similar confidence scores may not translate well between engines and represent a different confidence for each engine.

Various studies have mentioned the use of confidence score as a indication of OCR quality. Some studies report positively about the usability of confidence scores as quality indication. [17] showed that the mean confidence score correlates

well with the character error rate (CER), as well as with lexicality. Similar research was done by [5]. While they did not use the confidence scores as it was not clear what exactly was being measured, they did see a promising correlation. [12] performed an extensive survey of post-OCR processing approaches where they saw an important role for the confidence score in detecting where post-OCR improvements could be made. Word confidence has also been used to finetune the selection of bounding boxes by [4] where a multimodel relation was shown between the word confidence and the noise within bounding boxes.

On the other hand, experiments from [15] show that the word confidence score as provided by the OCR engine deviates from the true word recognition. Consequently, they conclude that the word confidence score are a limited estimator. Supporting this, [11] compared the OCR confidence score for each page with the CER and found that the confidence score had at best a slight correlation with the CER and was not useful as a parameter for quality. [19] also found no clear use for the confidence score provided with OCRed texts, not only from a theoretical point, but also its practical implementation as many user interfaces do not support filtering or automatic extraction of the confidence. They also noticed that the method of calculation and documentation for the calculation of the confidence score is quite nontransparent, making it hard to trust in the measure.

[6] mentions that confidence scores are often used as a substitute for accuracy by lack of other, more accurate, metrics. They also present a method to transform the confidence score into a proxy measure. They suggest taking a subset from which the quality is known, and from which the confidence scores are available. Using this subset an algorithm needs to be written to correlate the two scores and calculate a proxy accuracy measure. This proxy measure can then be used on the collection as a quality indication. Regular checks on the algorithm are necessary, to see if it still fits the collection.

3 Methodology

In this study, we analyse the confidence scores using various different viewpoints to get a better insight in the usability of the confidence scores as quality indicators.

First, we examine to what extent the average confidence score on page level can be used by institutions to support decision making regarding which parts of the collections are most suited to be re-OCRed or post-processed. We therefore follow the suggested approach of [6]: to correlate the actual accuracy with the extracted confidence scores and use these scores to create an algorithm that provides a proxy measure.

Furthermore, to get a more detailed view on the performance of the OCR engine and to better understand the word confidence, we zoom in on the word level, using a lexicon lookup and the Ground Truth to find discrepancies between word confidence and the probable ‘correctness’ of a word.

3.1 Dataset

We used a set of 2000 newspaper pages ranging from 1631 to 1995 [20]. These pages were originally digitized by two different companies, using three different versions of ABBYY Finereader [1]. The full 2000 pages were re-OCRed by a third producer with a newer version of ABBYY Finereader leading to a dataset of 4000 pages digitized by three different producers and four different ABBYY Finereader versions. In addition to the OCRed documents, there is also manually created Ground Truth available for each document.

The data contains newspapers from a broad range of years and contains both scans from microfilms and scans from paper. The years are divided in three time ranges, based on spelling changes in Dutch.

Table 1 shows an overview of the pages divided by producer and ABBYY version. The OCR is stored in Alto XML files. These files provide the confidence score of each detected word. These word confidence scores are used to calculate the average word confidence of per page.

Table 1. Distribution of pages among producer and ABBYY Finereader version.

Producer	ABBYY version	Number of pages
A	8.1	1325
A	9	31
A	10	92
B	10	552
C	12	2000

3.2 Analysis on Page Level

For our analysis on page level, we compared the average word confidence per page with the WER_oi per page. The average word confidence was based on the individual word confidence of all words on a page. The WER_oi was determined for all OCRed document with the use of the Ground Truth and the ocrevalUAtion tool [8] with the default settings. We choose to use the WER_oi for the comparison instead of the WER, as for this analysis, we are more interested in the correct prediction of each individual word than in the correct order of words.

To determine if the average word confidence is a reliable indication of the quality of a page, we calculate the correlation between the average word confidence and the WER_oi. If these measures correlate it could imply that the word confidence can be used in a similar way as the WER_oi. In order to exclude that certain characteristics have a strong influence on the correlation, we also determined the correlation between the word confidence and WER_oi for each producer, ABBYY Finereader version, year group and whether it was a microfilm scan.

Selection of the appropriate correlation type is done by plotting the average word confidence and the WER_oi in a scatterplot. These plots are then manually inspected to determine which correlation coefficient to use. If a linear relation is detected Pearsons r will be used. If a non-linear monotonic relation is detected, Spearmans rho will be used. If no relation is detected, the page level analysis will be aborted.

Based on the outcome of the correlation, we will either continue with exploring a proxy measure (see Sect. 3.3) or discard the exploration and continue to an analysis on word level (see Sect. 3.4). We consider a correlation with a coefficient of 0.8 or higher as a strong enough correlation.

3.3 Exploring a Proxy Measure

Based on the outcomes of the previous section, we choose a set corresponding to an ABBYY Version with an high enough correlation (>0.8) and a sufficient amount of pages. We choose an engine based selection as this is the most practical selection method to use as institution. If other stratification levels, such as year group, point to a significant disturbance of the correlation, these are removed from the dataset before continuing further analysis.

For this subset, we will explore if a proxy WER_oi measure can be calculated with use of the average word confidence, to classify the quality of a page.

To do this, we will compare three approaches:

1. Naive conversion to convert the average word confidence (wc) into a proxy WER_oi (proxy) measurement with the formula:

$$proxy = (1 - wc) * 100 \tag{1}$$

2. The use of simple linear regression to convert the average word confidence (wc) to a proxy WER_oi measurement with the formula:

$$proxy = a * wc + b \tag{2}$$

3. The use of polynomial regression to convert the average word confidence (wc) to a proxy WER_oi measurement with the formula:

$$proxy = a * wc^2 + b * wc + c \tag{3}$$

Before calculating the proxy WER_oi, we divided our dataset randomly in a train (70%) and test (30%) set. The train set was used to determine the formula for calculating the proxy WER_oi measure, the test set was kept apart for testing the formula.

As the word confidence represents the confidence that a word is correct, and the WER_oi is the percentage of word errors, OCR confidence score were inverted into a measure of 'unconfidence' or confidence that the word is wrong. This is then multiplied by 100 to achieve a WER_oi proxy measure on the same scale as the WER_oi.

For both linear regression and polynomial regression the formula obtained from the regression is used to predict the WER_oi proxy measure based on the average word confidence.

The performance of the three methods was determined by calculating the Mean Square Error and R^2 based on the test set. The method with the best results was chosen for further analysis.

These comparisons utilize a continuous scale. However, for most institutions it is more relevant to have a broader classification. For example, knowing if a page is good enough to be presented online, or if it is desirable to re-digitize it. Therefore, to get an indication of the practical usefulness of our best performing method, a step was added in which the classification performance of the selected method was tested.

We started by categorizing the pages based on the WER_oi. We based our cut-offs on the recommendations of [18]. They recommended an OCR quality of at least above 80%, but preferably over 90% for downstream tasks and analysis. To translate this to WER_oi, the inverse of these quality cut-offs is taken as the WER_oi corresponds to the percentage of faults and their cut-offs are based on the percentage quality. This results in cut-off values of WER_oi equal than or lower than 10 for the desired ('Good') category, between 10 and 20 for the minimal required ('Average') category, or bigger than 20 as low quality category.

We then calculated the WER proxy measure using the correlation formula of the best performing method and categorized all pages into on of the above categories. The categorization performance of the method was determined using a confusion matrix from which the accuracy, precision and recall were calculated.

3.4 Analysis on Word Level

For our analysis on word level, we chose one or multiple potential interesting subsets based on the results of Sect. 3.2. All words of the subset were extracted with their corresponding word confidence score. Then, we pre-processed the words to be able to perform a decent lexicon comparison. We executed the following steps:

- If a word is more than one character long, and it starts or end with a punctuation mark, this punctuation mark is removed;
- If there are any uppercase characters in the word, they are transformed to lower case.

Then, we performed a lexicon comparison for all words. The lexicon we used was a combination of a modern lexicon from OpenTaal [13] and a historical lexicon from the Dutch Language Institute [9]. We labeled every word as either 'found in lexicon' or 'not found in lexicon'. When a word was purely numeric, or contained only numbers and periods or commas, it was labeled as 'found in lexicon'.

A boxplot and a histogram are used to determine how well the two categories correspond to the word confidence.

Any anomalies that are found in the data, such as words with a high confidence that are not found in the lexicon, will be further investigated. Such words are compared to the Ground Truth and investigated to determine if these words have specific characteristics.

4 Results

4.1 Exploratory Analysis on Page Level

To determine if the word confidence score can be used as a proxy for quality, we started by investigating the correlation between the WER_oi and the word confidence. The analyses were stratified by various variables (manufacturer, OCR software version, year group and he presence of a microfilm) in order to investigate strong effects on the confidence score.

As can be observed in Fig. 1, there was a clear, slightly non-linear, monotonic relation between the average word confidence and the WER_oi. Therefore we used Spearmann's rho to determine the strength of the correlation of the total set and the various subsets. Table 2 provides an overview.

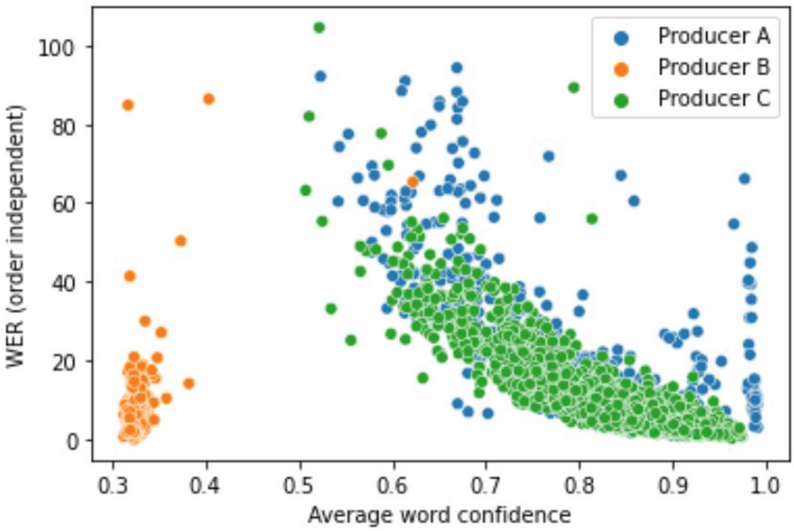

Fig. 1. Correlation between average word confidence and WER_oi per producer

As can be seen in the plot, the confidence score is strongly dependent on the producer and producer B appears to be a discrete subset (Fig. 1). Despite a high correlation of the larger subsets, the correlation of the total set is quite low with a confidence interval of −0.440 to −0.489. This is caused by the extreme low and inverted correlation of the word confidence from producer B to the

WER_oi (0.367). As the correlation of producer B is very different from the other producers it could have a distorting effect when determining differences in version, year or presence of microfilm. Therefore we excluded this producer for the remaining subsets.

Table 2. Distribution of pages per time period. For the sets marked with an *, producer B was excluded

	# pages	Spearmans' rho	p-value	confidence interval
Total	4000	−0.465	≤0.001	[−0.440, −0.489]
Stratified by producer				
Producer A	1448	−0.780	≤0.001	[−0.759, −0.799]
Producer B	552	0.367	≤0.001	[0.437, 0.292]
Producer C	2000	−0.854	≤0.001	[−0.841, −0.865]
Stratified by ABBYY Finereader version				
ABBYY 8	1325	−0.774	≤0.001	[−0.752, −0.795]
ABBYY 9	31	−0.934	≤0.001	[−0.866, −0.968]
ABBYY 10*	92	−0.876	≤0.001	[−0.818, −0.917]
ABBYY 12	2000	−0.854	≤0.001	[−0.841, −0.865]
Stratified by year group				
1631–1882*	517	−0.919	≤0.001	[−0.904, −0.931]
1883–1947*	1885	−0.786	≤0.001	[−0.768, −0.802]
1948–1995*	1046	−0.760	≤0.001	[−0.733, −0.785]
Stratified by microfilm				
microfilm*	841	−0.765	≤0.001	[−0.735, −0.791]
no microfilm*	2607	−0.834	≤0.001	[−0.821, −0.845]

There is some slight variation between the various ABBYY Versions. ABBYY 9 has the highest correlation (−0.934), but also the lowest number of pages. For the year groups, the oldest newspapers (1631–1882) had the strongest correlation (−0.919), while the most modern newspapers had the lowest correlation (−0.760). When looking at the presence of microfilm, the strongest correlation is on the group that does not have a microfilm (−0.834).

4.2 Exploring a Proxy Measure

For the exploration of a proxy measure, we used a subset of the total dataset including only ABBYY version 12.

This set was split into a train (1400 pages) and a test set (600 pages) and the three approaches (3.3) were performed on the train set. Scatterplots with trendlines for the three approaches are shown in Fig. 2. The mean square error and R^2 were calculated to evaluate the various models and are presented in Table 3.

Of the three methods, the polynomial formula has the best fit to the data and is best suited to determine the quality based on the word confidence score,

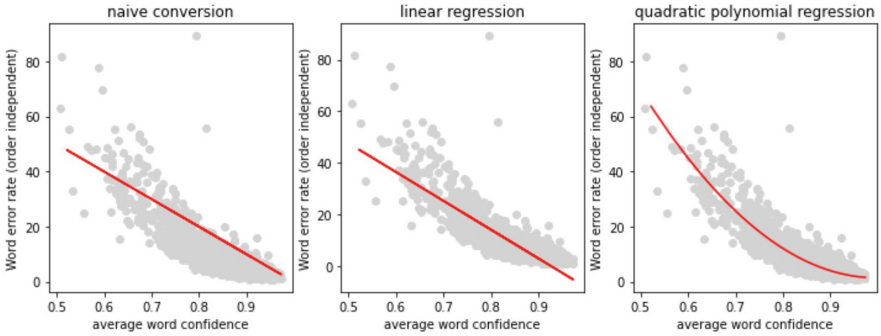

Fig. 2. Comparing proxy measures

Table 3. Comparing proxy measures

	Mean Square Error	R^2
Naive conversion	68.49	0.203
Linear regression	25.5	0.703
Quadratic polynomial regression	**17.75**	**0.793**

as it had the lowest error and the highest percentage of explained variance (17.75 and 0.793 respectively).

We therefore continued with this method, with the corresponding formula:

$$276.234x^2 - 550.55x + 275.748 \tag{4}$$

For the practical application using quality categories (Sect. 3.3) we determined the accuracy of the proxy measure test set using the polynomial regression. The results are presented in Table 4 and as a confusion matrix in Fig. 3.

We found that for ABBYY 12 the category with the lowest WER_oi, the 'Good' category, could be predicted with the highest precision and recall, followed by the 'Low' and the 'Average' categories. The total accuracy for the model is 0.83.

From the confusion matrix it is noticeable that there is almost no contamination between the two outer categories. The 'Good' quality category contains only one page (0,3%) with a true classification of 'Low' that is falsely classified as 'Good'. Similarly, the 'Low' category contains only one page (2%) with a true classification of 'Good' that is falsely classified as 'Low'. There is significantly more overlap between the 'Low' and 'Average' categories, and the 'Good' and 'Average' categories.

In most workflows the 'Low' and the 'Average' categories will likely undergo further checks as these are targets for re-OCRing or post-processing. A 'Good' that is falsely predicted as an 'Average' is therefore not as much of a problem as an 'Average' that is falsely predicted as a 'Good'. The confusion matrix shows that 28 pages with an actual 'Average' quality were falsely predicted as 'Good'

quality. Whether this can be a problem depends largely on how far the actual WER_oi is from the predicted WER_oi. To establish if these 28 false positives are closer to 'Good' or to 'Low' we determined the mean of the actual WER_oi. The mean actual WER_oi of this group was 12.86, meaning that these false positives were closer to 'Good' than to 'Low'.

Table 4. Performance of proxy measure for each category

	Precision	Recall	f1-score	# pages
Good	0.93	0.88	0.90	401
Average	0.59	0.73	0.65	126
Low	0.86	0.75	0.80	73

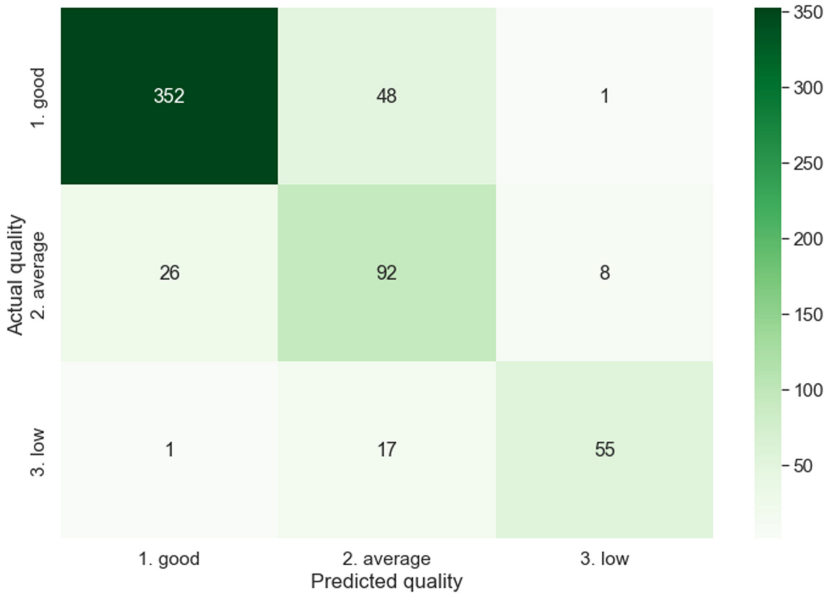

Fig. 3. Confusion matrix of proxy measure performance

4.3 Analysis on Word Level

Experiment 3a: Exploring ABBYY 12 Finereader for Producer C. From the 2000 pages of the ABBYY 12 version from producer C, we extracted all words with their corresponding word confidence. This resulted in a list of 7,227,304 words. For each word it was determined if it could be found in the lexicon. For both groups, found or not found in the lexicon, the word confidence was plotted (Fig. 4). As can be observed from the figure, it appears that the

higher the word confidence, the higher the chance that the words can be found in a lexicon. For words that have a word confidence of 0.9 or higher, 94.8 % can be found in the lexicon indicating a strong relation between word level confidence and 'correctness' of a word.

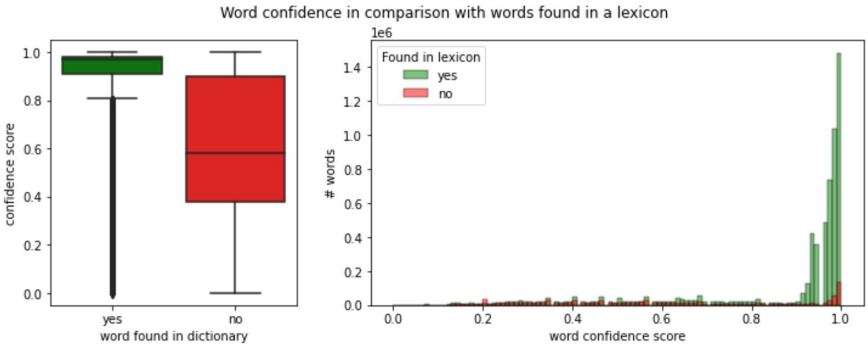

Fig. 4. Word confidence frequencies for words found or not found in a lexicon

As the histogram for the word confidence scores shows a left-skewed distribution, we decided to explore the difference between the samples in the tail and those in the body of the distribution. We therefore split the data into two specific cases:

1. a word has a high word confidence ($\geq$0.7) but was not found in the lexicon;
2. a word has a lower word confidence ($\leq$0.7) but was present in the lexicon.

For the cases that had a high confidence score, but were not present in the lexicon, 50.9% of the words with a confidence score of 1 (78,680 words) existed of only one character.

To determine if these may have been mistakes in the OCR, we counted the occurrence of each single-character word with a confidence score equal to or greater than 0.7 and compared these to the occurrence of the same character in the Ground Truth. From this comparison, we saw that some single character words, despite their high confidence, are likely OCR errors, while others are likely correct words. Some examples of single character words with their corresponding counts are shown in Table 5.

A negative Delta indicates that more occurrences of a character are found in the Ground Truth than in the OCR. When more single character words are found in the OCR than in the Ground Truth, the chance that these are OCR errors is high, while when there more single character words are found in the Ground Truth, these are likely correct. As we look at only a small percentage

of the OCR, the difference can be quite big but it can still be used as a rough estimation that likely they where correctly detected.

Table 5. High confidence, not found in lexicon

word	words with confidence ≥ 0.7	words in GT	$\Delta\%$	likely correct
,	9,385	987	0.895	no
’	2,298	31	0.987	no
:	1,544	527	0.659	no
;	1,415	210	0.852	no
f	3,263	10,328	−2.165	yes
f	1,935	10,205	−4.274	yes
–	3,186	6,227	−0.954	yes
a	1,771	2,100	−0.186	yes

Another noteworthy observation was that there were several small words, like articles and prepositions, with a wide variety of confidence scores. For some of these, more than half of the occurrences had a confidence score lower than 0.7. The total occurrence of these words in the OCR was close to the total of occurrences in the Ground Truth, indicating that they were likely correctly recognized, despite the variation of confidence scores. Some examples are shown in Table 6.

Table 6. Low confidence, found in lexicon

word	% with confidence ≤ 0.7	total words OCR	Total words in GT	$\Delta\%$
in	70.3%	123,822	129,013	−0.0419
en	68.6%	145,130	153,604	−0.0583
op	58.4%	64,041	63,721	0.005
de	55.7%	359,191	381,197	−0.061

Experiment 3b: Comparing ABBYY Version 10 for Producer A and B. In Sect. 4.1 we determined that the correlation between the average word confidence and the WER_oi was highly dependent on the producer. Where producer A and C have a very similar correlation, producer B had a very divergent result, with much weaker correlation between the average confidence score and the WER_oi.

To explore the influence of this producer variance, we selected all words that from both producer A and B that were digitized with the same ABBYY version (ABBYY version 10). The word confidence was analyzed for words occurring and not occurring in the lexicon. The word confidences per group are presented

in a boxplot and histogram, as shown in Fig. 5 for producer A and Fig. 6 for producer B.

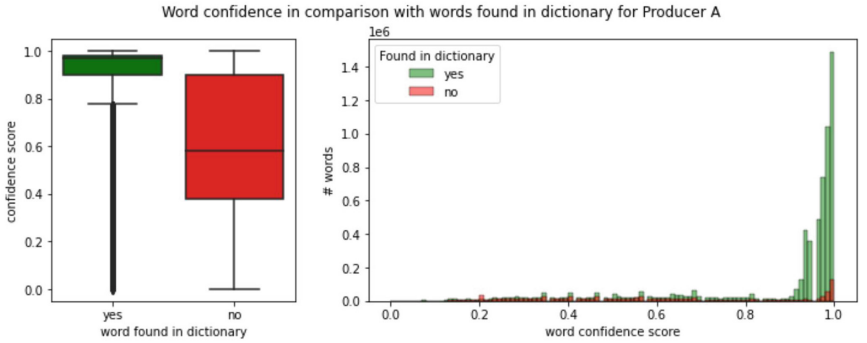

Fig. 5. Word confidence frequencies for words found or not found in a lexicon for producer A

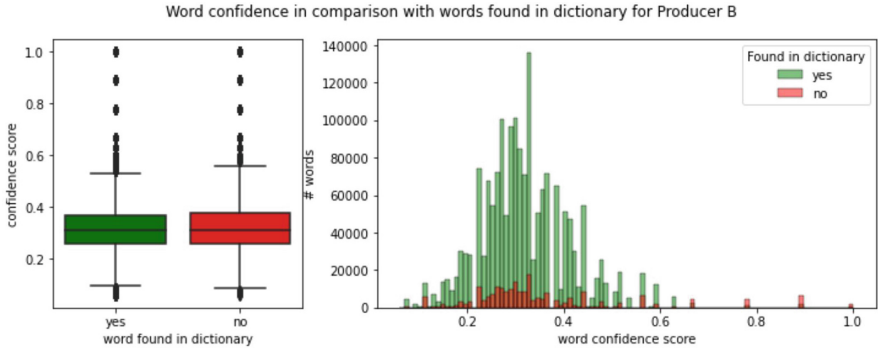

Fig. 6. Word confidence frequencies for words found or not found in a lexicon for producer B

There is a large difference in distribution between both producers. Producer A (733,4917 words) has a clear left-skewed distribution of the word confidence score, whereas producer B (1,934,017) has a more or less normally distributed word confidence score. For producer A most words that are found in the lexicon have a word confidence higher than 0.7 (80.8%), and most words that were not found in the lexicon had a word confidence lower than 0.7 (63.5%). Whereas for producer B the words 'found in the lexicon' and words 'not found in the lexicon' had similar word confidences, as can be observed from the overlapping

distributions in Fig. 6. For producer B, 99.8 % of the words found in a lexicon have a word confidence lower than or equal to 0.7 against 84% of the words that were not found in a lexicon.

5 Discussion

Our results suggest that under certain conditions the confidence score can be used as a proxy measure for quality. In practice this would mean that when a small sample shows a correlation between word confidence and WER_oi, (polynomial) regression can be used to create an indication of the quality of the digitized texts. In our study polynomial regression as a proxy measure was most accurate in detecting digitized texts with the lowest WER_oi, which corresponds to the texts with the highest quality.

The results as presented in the confusion matrix show that there is almost no contamination between the 'Good' and 'Low' categories. This implies that when a page is classified as either of 'Good' or 'Low' quality, the chance of belonging to the other of these two categories is very low. As cross-contamination between these two categories undesirable, this result is very promising.

However, these formulas do not seem to generalize very well between various versions of OCR software or various producers. Meaning that with every new producer or version, the formula must be recalculated. Nonetheless, organizations can use one small, standard set of Ground Truth to re-calculate the formula when a new engine or producer is introduced, making this a quickly obtainable method that needs little in the way of resources. When the scans corresponding to this Ground Truth set are re-OCRed for every new producer or engine, it can be used as a quality control to compare the new OCR results and performance.

An interesting and unexpected result was that the year group with the highest correlation between the proxy WER measure and the WER_oi was the 1631–1882 group. In this period the Dutch language was not yet standardized and newspapers from that time had large spelling variations and generally differ more from modern Dutch than the other year groups in our study. After some inquiries, it appears that the producer added a specific historical lexicon to the OCR engine which we suspect has contributed to a higher word confidence.

A more worrying result is that the confidence scores from producer B have such a different correlation compared to the other producers. After inquiry it became apparent that the ABBYY version of producer B was a non-optimized, 'from scratch', version. The other producers trained and optimized their ABBYY versions before use. Also, we strongly suspect that provider B did not use historical lexicons, whereas the other producer included these in their workflow.

All the above combined make it very important to know what engine, producer and personalisation of the OCR was applied before drawing conclusions. The usefulness of our method is therefore strongly influenced and limited by these variables.

A remarkable result was that a large part of the high confidence words (>0.7) were in fact single character words. Initially such words would be considered

incorrect, as these words were not found in the lexicon and there are only a few single letter words in Dutch. However, upon closer inspection with the Ground Truth, it appeared that a part of these single character words were identified correctly and consisted, among others, of currency marks (f and *ƒ*, which stands for florin, the Dutch currency from the 15th century until 2002) and parts of enumerations (the dash). In these cases the engine identified the correct 'word', even though the words did not exist as such. However, other single character words were present in much larger numbers in the OCR than in the Ground Truth, indicating that they were incorrectly identified, despite their high confidence. In addition, a large section of two character words had a low word confidence (<0.7) but could be found in the lexicon. These consisted of, among others, articles and prepositions. This may be explained by the fact that for short words it is more difficult to find the correct word when the original material is damaged or unclear. Missing one letter would mean that a large part of the word is uncertain, drastically lowering the confidence score. This shows that it is important to know what the contents of the texts are and that specific rules and conditions must be kept in mind for language, time period and type of text.

It is important to note that we used the WER_oi measure of the OCREvaluation tool [8]. As different toolings can have different results [11], it could be that our results are biased due to the choice of software for the calculation of the WER_oi. Furthermore, when looking at the results of the word level analysis, it is important to keep in mind that if a word is found in a lexicon it does not necessarily mean that the word is the correct word. Small substitutions can change a word into a different, 'correct', word that is not the word in the original text.

In this study we focused on the page level as this was most relevant for our institution. However, researchers generally prefer to know the quality at the article level as this helps them select what they need for their research. In the future we will expand the current research to include article level data. Also, We would like to replicate our findings using texts that were digitized with other engines to see how well the method generalizes.

As the success of OCR is largely dependent of the quality of the scan and the conditional of the original material, it would be interesting to determine if the confidence score does not only correlate to the quality of OCR, but also to the quality of the scan or original material. This could help institutions pinpoint where to repeat or improve scans, or where to redo the OCR.

6 Conclusion

In conclusion, when there is a correlation between word confidence and WER_oi, it is feasible to build a proxy WER_oi measure to classify texts into quality groups and use this groups to determine which pages are a good candidate for re-OCRing or post-processing or already good as it is. This can be done with a minimum of Ground Truth, making it very interesting method to quickly sort digitized texts into rough quality categories.

References

1. ABBYY: ABBYY FineReader Server (2022). www.abbyy.com/finereader-server/
2. ABBYY: FineReader Engine 12 for Windows Developer's Guide (2022). http://help.abbyy.com/en-us/finereaderengine/12/user_guide/introduction_startpage/
3. Anderson, N., Muhlberger, G., Antonacopoulos, A.: Optical character recognition: IMPACT best practice guide. www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf. Accessed 05 Oct 2022
4. Gupta, A., et al.: Automatic assessment of OCR quality in historical documents. Proc. AAAI Conf. Artif. Intell. 29(1) (2015). https://doi.org/10.1609/aaai.v29i1.9487
5. Hill, M., Hengchen, S.: Quantifying the impact of dirty OCR on historical text analysis: eighteenth century collections online as a case study. Digit. Scholarsh. Hum. **34**, 825–843 (2019). https://doi.org/10.1093/llc/fqz024
6. Holley, R.: How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Mag. Mag. Digit. Libr. Forum **15** (2009)
7. Impact Centre of Competence: Confidence Level (OCR) (2018). www.digitisation.eu/glossary/confidence-level-ocr/
8. IMPACT Centre of Competence: ocrevalUAtion (2019). http://github.com/impactcentre/ocrevalUAtion
9. Instituut voor de Nederlandse taal: INT Historische Woordenlijst (2012). http://taalmaterialen.ivdnt.org/download/tstc-int-historische-woordenlijst/
10. Kofax: Kofax documentation. http://docshield.kofax.com/KTA/en_US/740-uc0n6j0c5s/help/SD/ScriptDocumentation/c_Welcome.html
11. Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., Pletschacher, S.: A survey of OCR evaluation tools and metrics. In: The 6th International Workshop on Historical Document Imaging and Processing, HIP 2021, pp. 13–18. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3476887.3476888
12. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-OCR processing approaches. ACM Comput. Surv. **54**(6) (2021). https://doi.org/10.1145/3453476
13. OpenTaal: Nederlandse woordenlijst (2020). http://github.com/OpenTaal/opentaal-wordlist
14. Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., Varner, S.: Final report – always already computational: collections as data, May 2019. https://doi.org/10.5281/zenodo.3152935
15. Salah, A.B., Moreux, J.P., Ragot, N., Paquet, T.: OCR performance prediction using cross-OCR alignment. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 556–560 (2015). https://doi.org/10.1109/ICDAR.2015.7333823
16. Smith, D., Cordell, R.: A research agenda for historical and multilingual optical character recognition (2019). http://hdl.handle.net/2047/D20298542
17. Springmann, U., Fink, F., Schulz, K.: Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents (2016)
18. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: Assessing the impact of OCR quality on downstream NLP tasks. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH, pp. 484–496. INSTICC, SciTePress (2020). https://doi.org/10.5220/0009169004840496

19. Traub, M.C., van Ossenbruggen, J., Hardman, L.: Impact analysis of OCR quality on research tasks in digital archives. In: Kapidakis, S., Mazurek, C., Werla, M. (eds.) TPDL 2015. LNCS, vol. 9316, pp. 252–263. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24592-8_19
20. Wilms, L., Koster, T.: Historical newspaper OCR ground-truth data set (2020)