ReadOCR: A Novel Dataset and Readability Assessment of OCRed Texts

Hai Thi Tuyet Nguyen[1], Adam Jatowt[2], Mickaël Coustaty[3], and Antoine Doucet[3]

[1] Posts and Telecommunications Institute of Technology, Vietnam
tuyethai@ptithcm.edu.vn
[2] Department of Computer Science, University of Innsbruck, Austria
adam.jatowt@uibk.ac.at
[3] L3i, La Rochelle University, France
{mickael.coustaty, antoine.doucet}@univ-lr.fr

Abstract. Results of digitisation projects sometimes suffer from the limitations of optical character recognition software which is mainly designed for modern texts. Prior work has examined the impact of OCR errors on information retrieval (IR) and downstream natural language processing (NLP) tasks. However, questions remain open regarding the actual readability of the OCRed text to the end users, especially, considering that traditional OCR quality metrics consider only syntactic or surface features and are quite limited. This paper proposes a novel dataset and conducts a pilot study to investigate these questions.

Keywords: Readability assessment · OCR errors · Hierarchical attention network · BERT.

1 Introduction

Considerable efforts have been devoted to transforming historical documents into electronic form for better preservation and easier access. Applying modern OCR technologies on old documents often leads to noisy outputs which negatively affect reading, retrieving, and other processes in digitized collections [18,2]. Whereas there have been many studies conducted regarding the impact of OCR errors on IR and NLP tasks [14,4,17], the influence on the reading ease remains still an open question. Usually, common metrics such as word error rate (WER) and character error rate (CER) are used to validate the quality of digitised text, implicitly including its readability.

Text readability (aka. reading ease) is however affected by many factors, such as lexical sophistication, syntactic complexity, discourse cohesion, and background knowledge [6]. Several readability formulas and machine learning techniques have been suggested in the past to assess an input text's readability. However, all of these approaches are designed to work on clean text, i.e., one without any OCR errors. In other words, studies on readability of digitized texts are largely missing.

In this paper, we propose a novel dataset and conduct several experiments on this data in order to answer question how OCR errors impact readability of texts. The contributions of our work are as follows:

1. We introduce a novel dataset for readability assessment of OCRed texts. Studies on this dataset can help to understand the impact of OCR errors on reading. Additionally, based on the dataset, one can have more clues to decide whether the quality of target OCRed texts is acceptable for reading or not. Future systems can train and test their readability assessment models based on our proposed corpus. The corpus is publicly available and freely accessible[4].
2. We investigate the following: (i) the relation between readability reduction (i.e., the reduction of an original text's readability score due to OCR process) and standard error rates like WER, CER; (ii) the relation between the readability reduction and the readability of an original text; (iii) the impact of corrupted lexical, grammatical words, and two typical OCR error types on the readability.
3. Finally, we apply state-of-the-art methods for the readability assessment of noisy texts in our dataset.

2 Proposed dataset

2.1 Document collection

Several corpora have been proposed for studying text readability including Weebiz [20], OneStopEnglish [19], Newsela [21], and CommonLit[5]. Whereas Weebit [20] and Newsela [21] classify texts into specific classes according to the age group for which the text is designed, OneStopEnglish [19] includes texts of three reading levels: beginner, intermediate, and advanced. Instead of assigning age-dependent classes or broad reading levels like the three above datasets, CommonLit provides actual readability scores for each text.

Our objective is to observe the effects of OCR errors at different reading ease scores, hence, we investigate two relations: (1) one between the readability of OCRed texts and their error rate metrics, and (2) one between the readability of OCRed texts and the readability of their original (non-OCRed) versions. In order to do that, we need a corpus composed of both original texts labelled with their readability scores and of their OCRed versions which are corrupted to different degrees and are also labelled with readability scores. Since no such dataset exists, we have decided to create and share one.

According to our observation, a small error rate may not affect a coarse-grained reading level of a corrupted text (i.e., OCRed text). For example, changing WER from 10% to 15% of the same document may not affect the reading label of that document (e.g., the document may be still judged as being of the intermediate difficulty level). By using specific fine-grained readability scores rather

[4] https://tinyurl.com/ReadOCR
[5] https://www.kaggle.com/c/commonlitreadabilityprize/data

than coarse-grained broad labels, we can more reliably compute correlation levels for our analysis. Based on this requirement, we have used and adapted an existing readability dataset that provides the actual reading ease scores instead of the broad readability classes which the texts belong to. Among the popular readability datasets, only the CommonLit dataset meets our requirements, hence we utilized it in our experiments. This corpus contains literary passages from different time periods and their fine-grained reading ease scores. It includes 2,834 texts collected from several sources, such as Wikipedia[6], Africanstorybook[7], Commonlit[8], etc.

2.2 Proposed text corpus

Since we wanted to manually assess readability of OCRed texts with different error degrees, we have sampled a subset from the CommonLit dataset. We randomly chose 161 files and added noise to the data at the word and document levels to mimic varying quality of OCR under different word error rates. The original (non-corrupted) files are also included in the proposed corpus along with their 483 corrupted versions. In total, there are 644 files whose detailed information is indicated in Table 2. This corpus is split into two parts (84% for training and 16% for testing) for conducting experiments discussed in Section 3.

The way we corrupted the original texts at word-level and document-level is described below:

- Word-level: we first picked all words belonging to a common English dictionary, and we randomly replaced some of their characters with plausible characters mimicking OCR errors. The plausible characters have been deduced from a confusion matrix which was created based on the alignment data corpus which is publicly provided through the ICDAR 2017 competition for post-OCR correction [5]. This corpus has been compiled from nine sources, mainly from the National Library of France and the British Library. It also represents a relatively wide time range. Its documents have different degradation levels under independent preservation practices. All of these characteristics make this corpus representative for typical OCR errors. Examples of plausible characters are illustrated in Table 1.
- Document-level: According to [1], average WERs range from 1% to 10%. Yet, some datasets have higher WERs [12]. In order to study the readability on a wider range of WERs, we corrupted original data with six WER levels ranging from 7% to 32%, with a step of 5%. In particular, we randomly chose words from each document based on the required WER. Each chosen word was then replaced by its randomly selected corrupted version that was generated at the word level.

[6] https://en.wikipedia.org
[7] https://www.africanstorybook.org
[8] https://www.commonlit.org

Table 1: Examples of plausible characters (i.e., OCRed chars) of the original characters (i.e., GT chars) [13].

GT chars	OCRed chars
a	u, n, e, i,
c	e, o,
e	o, c,
f	t, l, i,
h	b, i, n,
o	e, a, c,

Table 2: The numbers of files and tokens of the constructed corpus and its split parts.

Stats	Parts	Original	Corrupted	Total
Files	All	161	483	644
	Train	135	405	540
	Test	26	78	104
Tokens	All	27,809	83,670	111,479
	Train	23,320	70,170	93,490
	Test	4,489	13,500	17,989

After creating the noisy versions of the original texts, we asked three volunteers[9] to read all the texts and assign a score for each noisy text to indicate how understandable it is in comparison with its original text. Similar to classical WER values we wanted to obtain fine-grained scores on the scale ranging from 0 to 100, hence, the annotators were asked to assign their scores within this range.

In order to assess the agreement on continuous data (i.e. detailed scores given by the annotators) we applied intra-class correlation coefficient - ICC [16,15], which ranges from 0 to 1. Among 6 cases of ICC, the best suited one for our case is ICC2k. ICC2k is the reliability estimated for the means of k raters with each text target rated by a random sample of k raters. We have obtained an ICC2k value of 0.865 which indicates a good reliability in our corpus according to a guideline of selecting an intraclass correlation [10].

In addition to the assigned scores, the annotators made also notes to justify their scores. By closely examining the obtained comments, we found that the annotators attempted to understand the texts by guessing the meaning of errors or sometimes by simply ignoring them. In general, the scores were dependent on how hard it was to guess the meaning of the erroneous words, how these affected the annotators' overall understanding of the texts, the number of these errors, and how much the errors disturbed the reading flow.

[9] The three annotators are sophomores, two of them are law students, and one is an information technology student.

Table 3: Pearson correlation coefficients between *ComScores* and *ReadScores* according to WER levels that are illustrated in Figure 2 (except level 0).

WER	0.07	0.12	0.17	0.22	0.27	0.32
Correlation	0.204	0.080	0.182	0.319	0.267	0.003

2.3 Dataset Analysis

To answer our leading research question on the relation of readability and error rates, we used both the given scores (denoted as *ReadScore*) and the reading difficulty or reduction (denoted as *DiffScore*) computed as the difference between 100 and *ReadScore*. Min-max normalization was applied on the scores given by each annotator to ensure score comparability among the annotators.

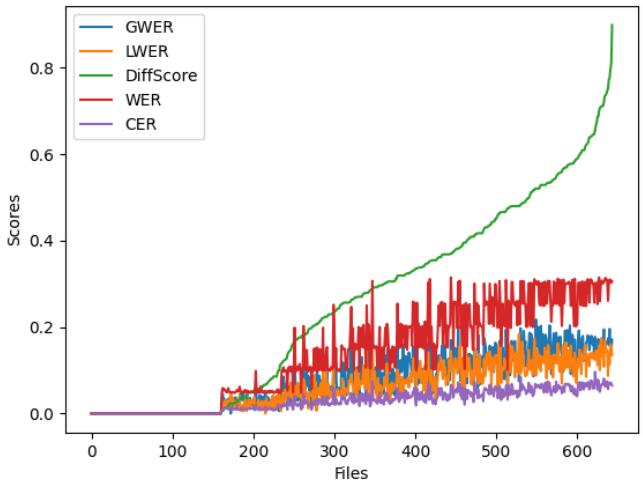


Fig. 1: Grammatical word error rate (GWER), lexical word error rate (LWER), WER, CER, and the *DiffScore* of the whole corpus whose documents are ordered on X-axis by their *DiffScores*. Pearson correlation coefficients between the other metrics and the *DiffScore* are 0.902, 0.910, 0.941, and 0.931, respectively.

Our analysis on the corrupted corpus shows that the *DiffScore* of noisy texts correlates well with the common error rates. However, this is not true for all WER values. For example, Figure 1 shows that WER remains around 0.32 even when the *DiffScore* increases further.

Regarding the relation between the *ReadScore* and the original reading ease CommonLit scores (denoted as *ComScore*), we examined it with respect to six levels of WER: 0.07, 0.12, 0.17, 0.22, 0.27, and 0.32. Figure 2 illustrates the

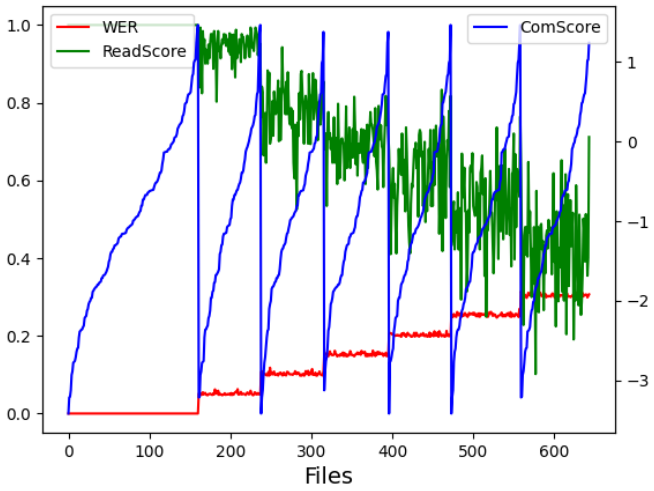


Fig. 2: *ComScores* and *ReadScores* of the whole corpus. The left Y axis shows *ReadScores* and WER, the right Y axis indicates *ComScores*. These scores are grouped according to all WER levels.

relations between *ComScores* and *ReadScores* which are grouped by all WER levels. The detailed correlation values of the examined relations according to six levels of WER are presented in Table 3. We expect that the more readable the original text is, the more readable its noisy version should be. In other words, the higher the *ComScore* is, the larger the *ReadScore* should be, at a given WER. Yet, in contrast to our expectation, the correlations between these scores fluctuate along with WER values. When texts are relatively noisy (e.g., WER of 0.32), it is probably difficult to understand them regardless of the reading ease scores of their non-corrupted (original) versions. The highest WER for our dataset is 0.32, however in reality OCRed texts (especially historical ones) could be even noisier. Further study is needed to investigate higher corruption levels.

Another analysis on the proposed corpus is to study the impact of the corrupted lexical and grammatical words on readability. Lexical words include nouns, verbs, and adjectives, while grammatical words (typically considered as stop words) consist of articles, pronouns, and conjunctions. We expect that corrupted lexical words have a high effect on readability than corrupted grammatical words. The correlation between the *DiffScore* and the error rate of the lexical words is a bit higher than the one for grammatical words, with 0.910 and 0.902, respectively. The difference between these two correlations is not so high since the readability of noisy texts relies not only on vocabulary but also on the reading flow and overall text understanding.

The next study on the ReadOCR corpus is to observe the effect of two common types of OCR errors, namely, *real-word* and *non-word* errors, on readability. Whereas *non-word* errors do not exist in a dictionary, *real-word* errors are dictionary entries but they are used in the incorrect context. With the above properties, *real-word* errors often mislead readers. However, according to our statistics, the rate of *real-word* errors correlates less with the *DiffScore* than that of *non-word* errors, with correlation values of 0.871 and 0.926, respectively. A possible reason for this lower correlation is that around a quarter of *real-word* errors in our corpus are stop words which are less important than lexical words.

3 Readability Assessment

In addition to the studies on relations between WER, CER, and the *ReadScore* or the *DiffScore*, we apply several readability assessment methods to predict the readability reduction and compare them against the results obtained so far.

3.1 Methods

Text readability scores can be assessed by traditional readability formulas and machine learning techniques. We utilize both of them to compute the readability reduction. The traditional readability formulas are used to assess the readability scores of corrupted and non-corrupted texts, and then to compute readability reduction. These formulas mainly focus on some lexical and syntactic information such as word length, sentence length, and word difficulty. As for machine learning techniques, we apply two state-of-the-art approaches to predict readability reduction of corrupted texts. One approach relies on a hierarchical attention mechanism to capture syntactic and structural information, while the other transfers knowledge of pre-trained models to predict the target scores.

Traditional reading formulas We measured the readability of documents in our corpus using two popular reading formulas, Flesch-Kincaid grade level (FKGL) [9] and Dale-Chall readability formula (DCRF) [7]. FKGL and DCRF represent, respectively, the number of years of education or the educational grade levels generally required to understand a given text. Whereas FKGL depends on a sentence's length and a number of syllables per word, DCRF relies on a sentence's length and a list of 3,000 words that fourth-grade American students could reliably understand. A word that does not exist in the list is deemed as a difficult word. The formulas for these metrics are given in Equations (1) and (2). Using FKGL and DCRF, we first calculated the readability of each original text and its noisy versions. These scores were then used to compute the readability reduction as in Equation (3).

$$FKGL = 0.39(\frac{totalWords}{totalSentences}) + 11.8(\frac{totalSyllables}{totalWords}) - 15.59 \quad (1)$$

$$DCRF = 15.79(\frac{difficultWords}{totalWords}) + 0.0496(\frac{totalWords}{totalSentences}) \quad (2)$$

$$reduction = \frac{originalRead - noisyRead}{originalRead} \quad (3)$$

Hierarchical attention network (HAN) HAN is one of the best performing approaches for readability assessment. In our experiment, we utilize it to predict readability reduction of texts.

HAN's architecture [22] contains 4 parts: a word encoder, a word-level attention layer, a sentence encoder, and a sentence-level attention layer. Words of each sentence are embedded into word vectors through an embedding matrix. Bidirectional GRU is then applied to encode contextual information from both directions of words. Since each word may contribute differently to the representation of the sentence, the word attention mechanism is used to extract informative words and aggregate their representation to form a sentence vector. Given the sentence vectors, they can be used to represent a document vector in a similar way. The encoded sentence is the bidirectional hidden state generated by bidirectional GRU. The attention mechanism is used again to compute the importance of the sentences in the document. The resulting vector becomes the final representation of the document and can be used as features for the target task. The hierarchical attention mechanism is expected to better capture the document structure and lead to accurate predictions of readability reduction.

For the experiments, we split documents into sentences and tokenize each sentence using NLTK library [3]. Each document contains up to 80 sentences, each of which has a maximum of 70 words. The model uses GoogleNews word2vec as word embeddings and is trained with a batch size of 16 and 50 epochs.

Transformer Martinc *et al.* [11] reported positive results by transferring the knowledge of pre-trained BERT models [8] for predicting text readability of multiple corpora. This approach leverages the pretrained neural language model for the prediction task. In particular, a fully connected layer is put on the top of the pretrained model. The whole model is fine-tuned on a new data to predict readability reduction. An obvious shortcoming of fine-tuning BERT model is that the model cannot handle long documents as the maximum sequence length is limited to 512 tokens. Fortunately, the texts in our corpus are relatively short, therefore, they are not affected by this issue.

We utilized the pretrained BERT model with 12 layers of size 768 and 12 self-attention heads, i.e., bert-base-uncased model. The model was then fine-tuned with a linear layer, as well as with the same batch size, and the same number of epochs as the HAN model.

3.2 Experimental results

We use our corpus to analyze the performance of the above methods. Since the corpus size is limited, as mentioned in Section 2.2, we split the dataset into

Table 4: MSE and correlations between the *DiffScore* and DCRF reduction (i.e., DCRFRed), FKGL reduction (i.e., FKGLRed), BERT's prediction, HAN's prediction, CER, WER on the test data.

Scores	**MSE**	**Pearson**
DCRFRed	0.014	0.863
FKGLRed	0.129	-0.380
BERT Prediction	**0.003**	0.960
HAN Prediction	0.012	0.854
CER	0.085	0.945
WER	0.026	**0.967**

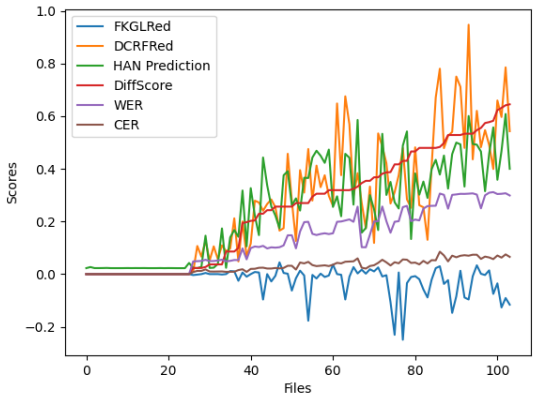
two parts: train and test. Then k-fold cross validation is applied ($k = 5$ in our experiment) on the training set. The best model of each method is evaluated on the same test part. More details about the data splits are given in Table 2.

Mean Squared Error (MSE) is a common metric used for regression. We utilize it to validate our models by computing MSE between the *DiffScore* and the predicted values. Besides MSE, we plot the predicted values together with the reading difficulty scores (Figures 3a and 3b) and compute the correlations shown in Table 4.

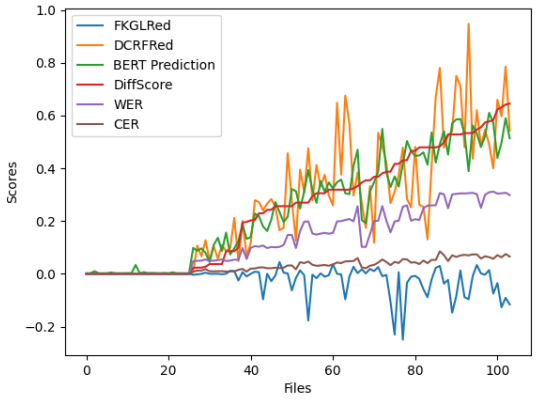
We notice that the BERT model is better than the HAN model in both MSE and correlation when comparing them on the same test data. The BERT model has a smaller MSE than the HAN model (0.003 vs. 0.012). The *DiffScore* has also a stronger correlation with the BERT model's prediction than the HAN model's (0.960 vs. 0.854). This is expected, as HAN is good at capturing document structure whereas BERT is a language model and focuses on text semantics. The sub-word segmentation of the BERT tokenizer also helps alleviating the out-of-vocabulary problem in corrupted texts. Moreover, BERT is more robust to noise when fine-tuning on noisy corpus.

MSE, correlation, and line plots are also used to compare performance of other traditional readability scores (i.e., FKGL, DCRF) as well as error rates (i.e., WER, CER) in readability assessment on corrupted texts. In particular, we computed MSE and correlation between the *DiffScore* and the other scores, which are reported in Table 4. All the computed correlation coefficients in this table are statistically significant with p-value under 0.05. Moreover, we plotted all scores according to HAN and BERT models, i.e., Figures 3a and 3b, respectively. The results reveal that WER has the strongest correlation with the *DiffScore* whereas both MSE and line plots support the best performance of BERT model. In fact, MSE of WER and the *DiffScore* is approximately 9 times higher than that of BERT predictions and the *DiffScore*. Figures 3a and 3b also indicate that the resulting predictions of BERT model are much closer to the *DiffScore* than the error rate lines.

Regarding two traditional readability scores, FKGL reduction seems to be ineffective with a high MSE and negative correlation, DCRF reduction has comparable MSE to WER but its correlation is much lower than other scores. In



(a) Predictions of HAN model.



(b) Predictions of BERT model.

Fig. 3: Different scores in assessing readability reduction of the test data: traditional readability scores (FKGLRed as FKGL reduction, DCRFRed as DCRF reduction); error rates (WER, CER); reading difficulty or reduction as *DiffScore*; predictions of HAN and BERT models denoted as HAN prediction and BERT prediction, respectively.

terms of DCRF, it depends on a list of easy words used (effectively assuming all other words as difficult). Since noisy texts contain many such words, their DCRFs are higher than DCRFs of their original texts. In terms of FKGL, corrupted texts often contain more words than their original version since OCR easily recognizes noise as punctuation, thus a tokenizer separates them into some words. When a number of syllables remains almost unchanged while the number of words increases, FKGL of a corrupted word becomes often lower than that of its original one. It should be noted that the increase of the total number of words is lower than that of difficult words.

Some examples of readability scores on different corruption levels (different error rates) of the same original text are illustrated in Table 5. Regarding two conventional scores, i.e, FKGL and DCRF, FKGL does not work effectively with the reduction as it obtains negative value due to the above-mentioned reasons (e.g, noisy texts have more words than their original one); DCRF exhibits some positive correlation with the *DiffScore* since it considers number of difficult words, and noisy words are often difficult ones. Similar to DCRF, error rates correlate well with the *DiffScore*; Nonetheless, readability is affected by not only the number of difficult words, the number of errors but also by other factors. BERT model implicitly takes into account many of such factors and hence gives good scores, which are closer to the *DiffScore*.

Table 5: Examples of different readability scores for the different corrupted versions of the same text.

Text	Radiosurgery is surgery using radiation, that is, the deitruction of precisely selected areas of lissue using ionizing radiation ra1her than excision with a blade.	Radiosurgery uts sur ery using radiation, that is, the des1ruction of precisély selected areat of tissul using ionizing rndiation rather than excision with n blade.	Radiosurgery is surgery using radiation, ihat is, •he destruction of precisely select d areas ol tissue using ionizing radiation rather than excision with a blade.
FKGL Reduction	-0.004	-0.089	0.01
DCRF Reduction	0.01	0.13	0.119
HAN Prediction	0.023	0.289	0.246
BERT Prediction	0.098	0.414	0.347
WER	0.048	0.259	0.2
CER	0.008	0.041	0.034
DiffScore	0.023	0.48	0.368

4 Conclusions

Our paper is the first work on the topic of readability assessment of OCRed texts. We provide a novel dataset and analyze the impact of OCR errors on readability as well as test two traditional measures and two SOTA baselines on our dataset. Among interesting findings, we observe that WER highly correlates with the reading difficulty. Nonetheless, the best BERT model has a smaller MSE and its prediction is much closer to the *DiffScore* than WER. By using this kind of model, one could potentially estimate the readability reduction of OCRed texts without having access to the original versions of those texts. Furthermore, the impact of the corrupted lexical words has been found to be not much higher than that of corrupted grammatical words since readability of noisy texts is additionally affected by other factors. Besides mis-recognized characters, layout errors may cause serious problems on text readability. In the next step of our future work, we will consider such errors.

Acknowledgements

This work has been supported by the "ANNA" and "Au-delà des Pyrénées" projects funded by the Nouvelle-Aquitaine Region.

References

1. Abdulkader, A., Casey, M.R.: Low cost correction of OCR errors using learning in a multi-engine environment. In: 10th International Conference on Document Analysis and Recognition, ICDAR 2009. pp. 576–580. IEEE Computer Society (2009)
2. Bazzo, G.T., Lorentz, G.A., Vargas, D.S., Moreira, V.P.: Assessing the impact of OCR errors in information retrieval. In: Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020. vol. 12036, pp. 102–109. Springer (2020)
3. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. pp. 69–72 (2006)
4. Boros, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., Doucet, A.: Alleviating digitization errors in named entity recognition for historical documents. In: Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020. pp. 431–441. Association for Computational Linguistics (2020)
5. Chiron, G., Doucet, A., Coustaty, M., Moreux, J.P.: Icdar2017 competition on post-ocr text correction. In: 14th IAPR International Conference on Document Analysis and Recognition. pp. 1423–1428. IEEE (2017)
6. Crossley, S.A., Skalicky, S., Dascalu, M., McNamara, D.S., Kyle, K.: Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. Discourse Processes **54**(5-6), 340–359 (2017)
7. Dale, E., Chall, J.S.: A formula for predicting readability: Instructions. Educational research bulletin pp. 37–54 (1948)

8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019). pp. 4171–4186. Association for Computational Linguistics (2019)
9. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
10. Koo, T., Li, M.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine **15** (03 2016)
11. Martinc, M., Pollak, S., Robnik-Šikonja, M.: Supervised and unsupervised neural approaches to text readability. Computational Linguistics **47**(1), 141–179 (2021)
12. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-ocr processing approaches. ACM Comput. Surv. **54**(6) (jul 2021)
13. Nguyen, T., Jatowt, A., Coustaty, M., Nguyen, N., Doucet, A.: Deep statistical analysis of OCR errors for effective post-ocr processing. In: 19th ACM/IEEE Joint Conf. on Digital Libraries. pp. 29–38 (2019)
14. Pontes, E.L., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR quality on named entity linking. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) Digital Libraries at the Crossroads of Digital Information for the Future - 21st International Conference on Asia-Pacific Digital Libraries, ICADL 2019. Lecture Notes in Computer Science, vol. 11853, pp. 102–115. Springer (2019)
15. Ranganathan, P., Pramesh, C., Aggarwal, R.: Common pitfalls in statistical analysis: Measures of agreement. Perspectives in Clinical Research **8**, 187 (10 2017)
16. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)
17. 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, ICAART 2020. pp. 484–496. SCITEPRESS (2020)
18. Traub, M.C., Van Ossenbruggen, J., Hardman, L.: 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 (2015)
19. Vajjala, S., Lučić, I.: Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications. pp. 297–304 (2018)
20. Vajjala, S., Meurers, D.: On improving the accuracy of readability classification using insights from second language acquisition. In: Proceedings of the seventh workshop on building educational applications using NLP. pp. 163–173 (2012)
21. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Trans. Assoc. Comput. Linguistics **3**, 283–297 (2015)
22. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. pp. 1480–1489. Association for Computational Linguistics, San Diego, California (Jun 2016)