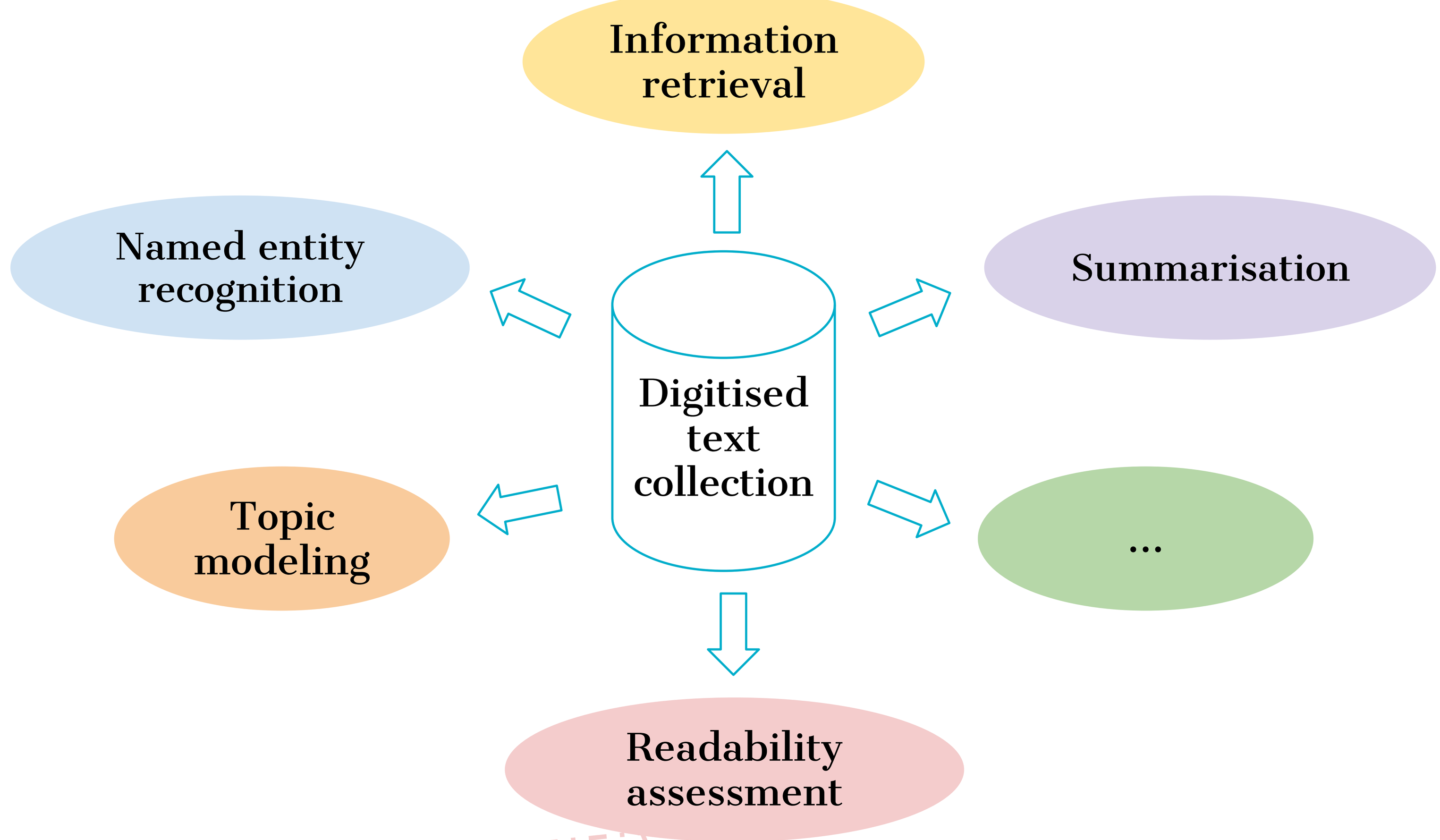
ReadOCR: A Novel Dataset and Readability Assessment of OCRed Texts

Hai Thi Tuyet Nguyen[1], Adam Jatowt[2], Mickael Coustaty[3], and Antoine Doucet[3]
tuyethai@ptithcm.edu.vn, , adam.jatowt@uibk.ac.at, {mickael.coustaty, antoine.doucet}@univ-lr.fr

[1] Posts and Telecommunications Institute of Technology, Ho Chi Minh, Vietnam
[2] Department of Computer Science, University of Innsbruck, Innsbruck, Austria
[3] L3i, La Rochelle University, La Rochelle, France

1. Motivation



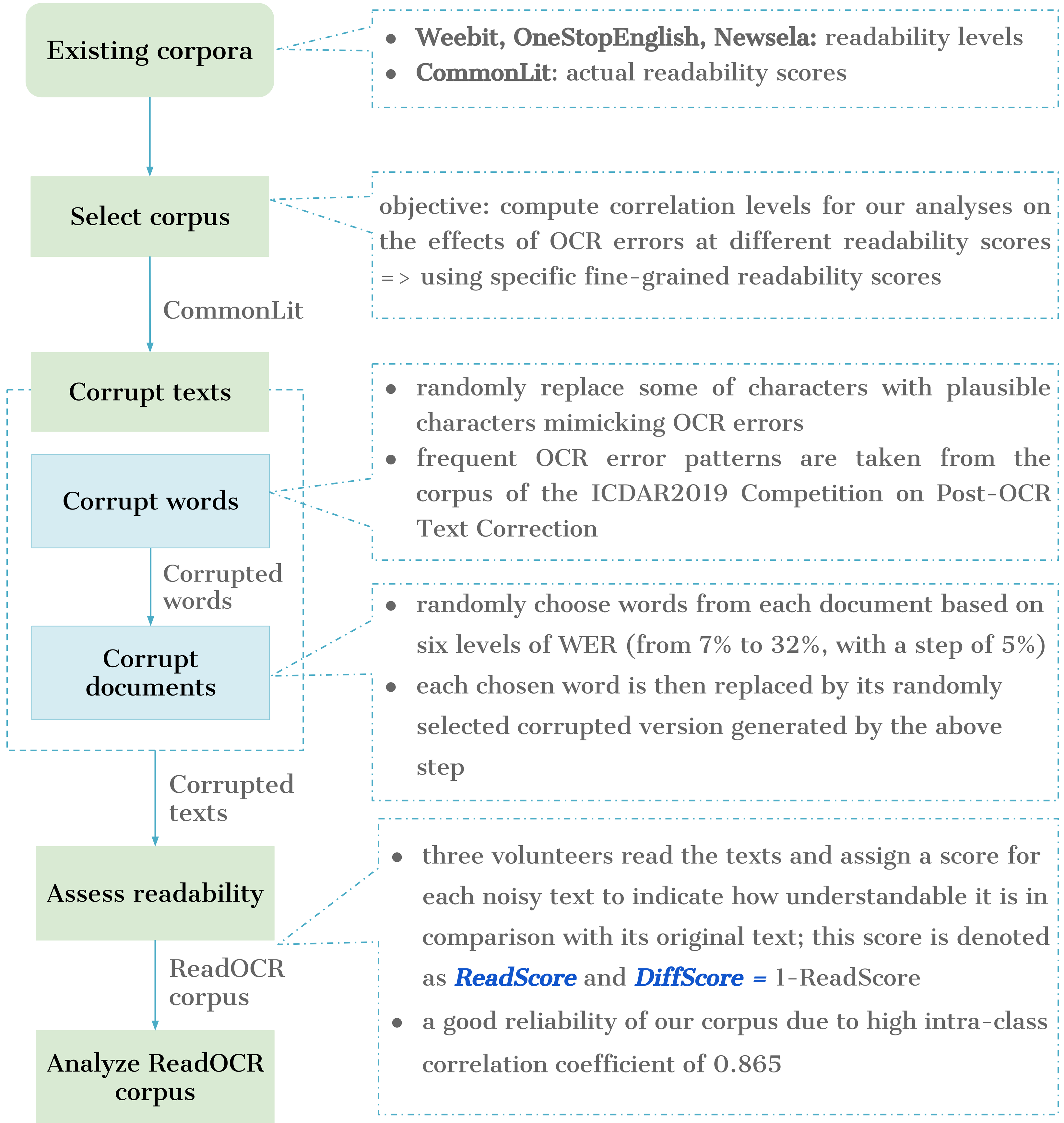
Our contributions:

- proposing a novel dataset for readability assessment of OCRed texts
- studying relations between readability reduction and other measures
- applying state-of-the-art methods for readability assessment

Text	WER	CER	Readability reduction
Radiosurgery is surgery using radiation, that is, the deitruction of precisely selected areas of lissue using ionizing radiation ra1her than excision with a blade.	0.048	0.008	0.023
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.	0.259	0.041	0.48
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.	0.2	0.034	0.368

Table 1: Examples of texts at different WERs along with readability reductions.

2. Proposed dataset



3. Dataset Analysis

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

Table 1: Statistics on the constructed corpus and its split parts.

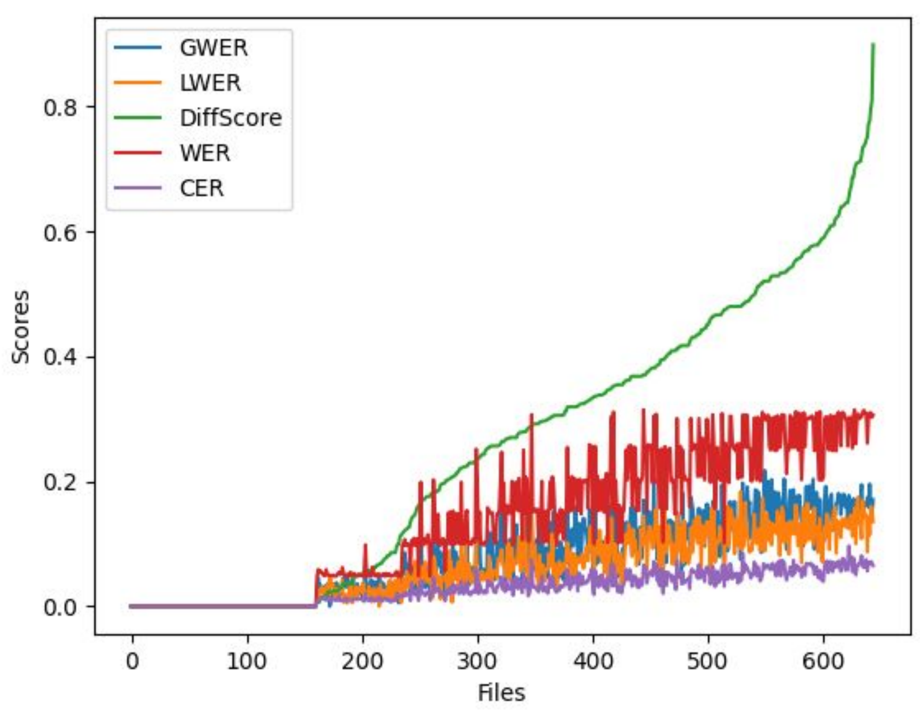


Fig. 1: Grammatical word error rate (GWER), lexical word error rate (LWER), WER, CER, and the *DiffScore* of the whole corpus whose documents are ordered on X-axis by their *DiffScores*. Pearson correlation coefficients between the other metrics and the *DiffScore* are 0.902, 0.910, 0.941, and 0.931,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

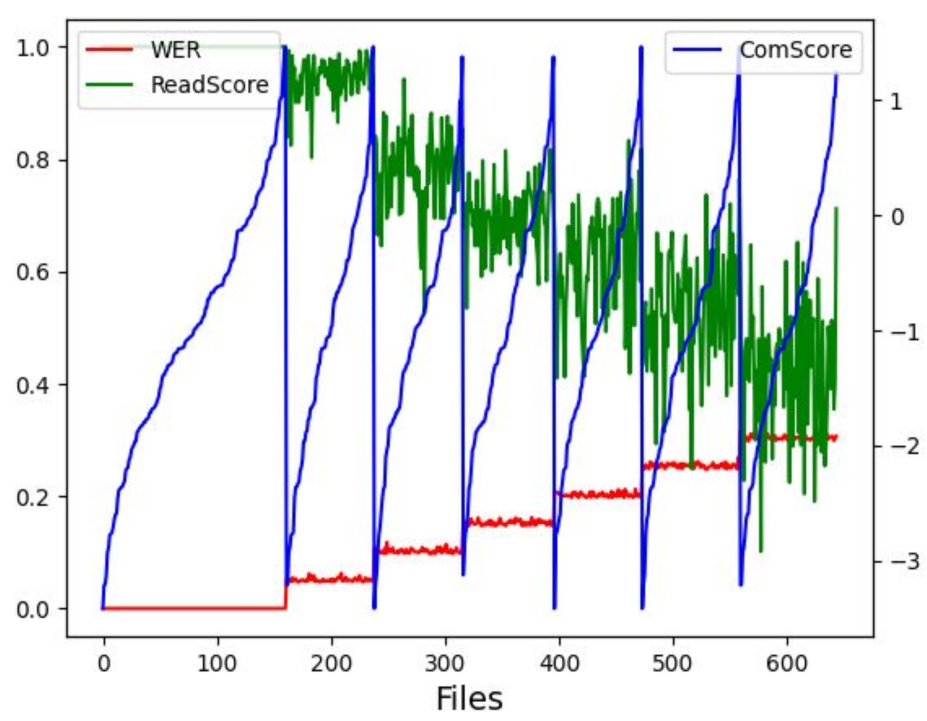
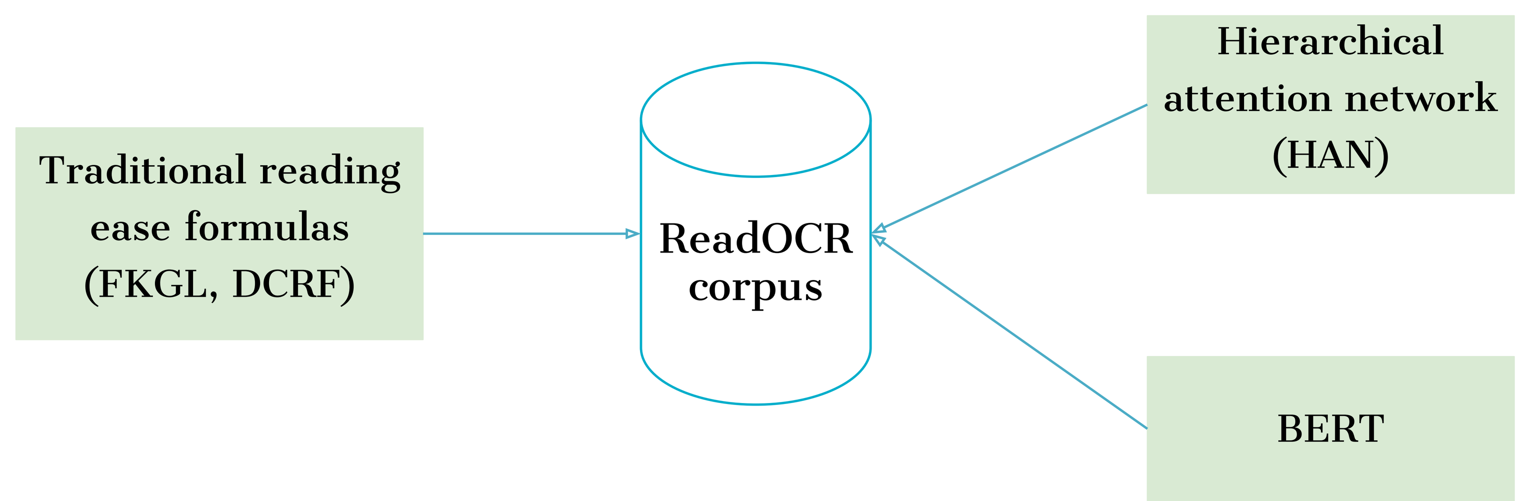


Fig. 2: *ComScores* and *ReadScores* of the whole corpus. *ComScores* is the readability scores of the original CommonLit texts. The left Y axis shows *ReadScores* and WER, the right Y axis indicates *ComScores*. These scores are grouped according to all WER levels.

- 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 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.

4. Readability Assessment



Method	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

Table 2: MSE and correlations between the *DiffScore* and DCRF reduction (i.e., DCRFRed), FKGL reduction (i.e., FKGLRed), BERT's prediction, HAN's prediction, CER, and WER on the test data.

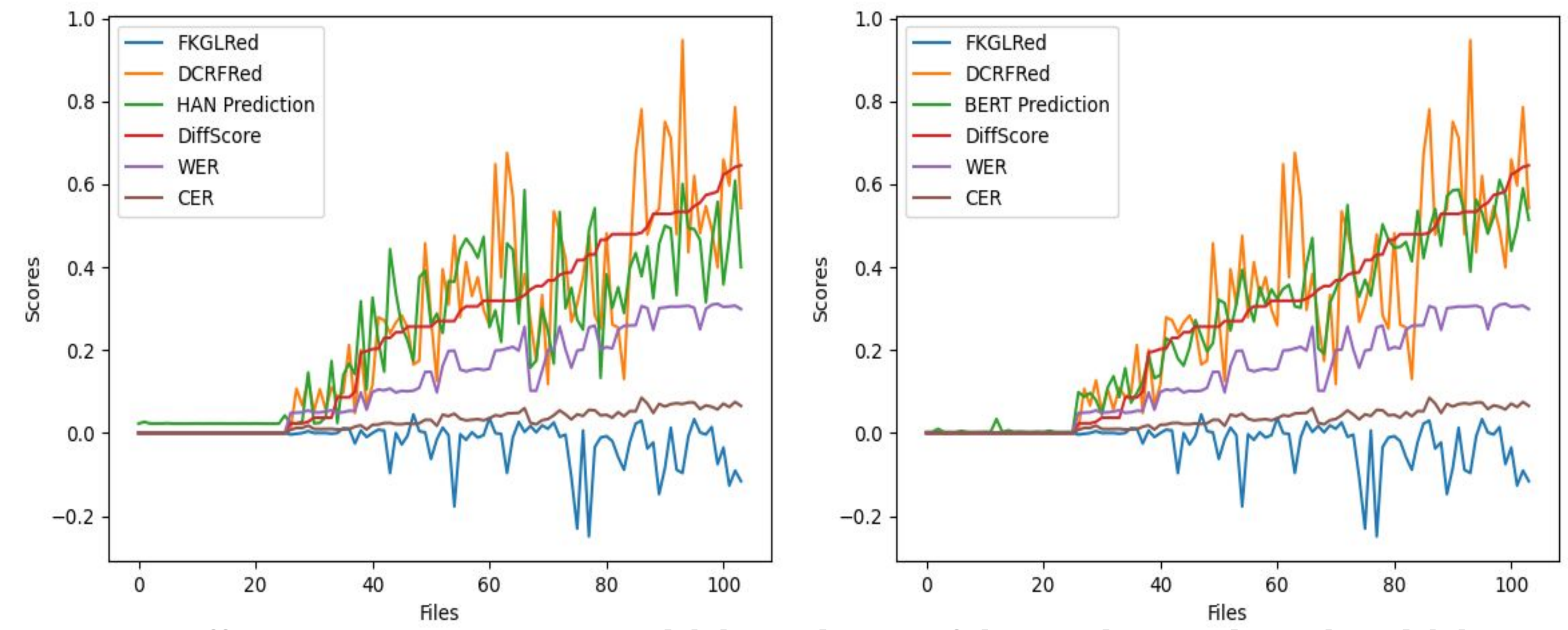


Fig. 3: Different scores in assessing readability reduction of the test data: traditional readability scores (FKGLRed as FKGL reduction, DCRFRed as DCRF reduction); error rates (WER, CER); reading difficulty or reduction as *DiffScore*; predictions of HAN and BERT models denoted as HAN prediction and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

5. Conclusion

- It is the pilot work on the topic of readability assessment of OCRed texts.
- We provide a novel dataset, analyze the impact of OCR errors on readability, test two traditional measures and two SOTA baselines on our ReadOCR corpus.
- Whereas WER highly correlates with the reading difficulty, the best BERT model has a smaller MSE and its prediction is much closer to the *DiffScore* than WER.
- The impact of the corrupted lexical words has been found to be not much higher than that of corrupted grammatical words.

References

[1] Dale,E.,Chall,J.S.:A formula for predicting readability:Instructions.Educational research bulletin pp. 37–54 (1948)
[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] Martinc, M., Pollak, S., Robnik-Sˇikonja, M.: Supervised and unsupervised neural approaches to text readability. Computational Linguistics 47(1), 141–179 (2021)
[4] 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.

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