Exploring Dataset Size and Diversity for OCR Post-Correction with hmByT5 Models

Anonymous ACL submission

Abstract

This study explores the application of the hmByT5 model for Optical Character Recognition (OCR) post-correction, focusing on historical German job advertisements. Two versions of the model—standard and fine-tuned on the ICDAR-2019 dataset—were evaluated across subsets of the JobAds dataset. The effects of dataset size and OCR model diversity on post-correction performance were analyzed. Results show that larger training datasets improve performance, but with diminishing returns, suggesting an optimal balance between annotation effort and model effectiveness. Training on outputs from multiple OCR systems enhances generalization with limited data but may introduce conflicting patterns in larger datasets. Fine-tuning on unrelated datasets, such as ICDAR, reduced performance, underscoring the importance of domain alignment in pre-training.

1 Introduction

Accessing data from historical newspapers in the machine-readable form offers unique opportunities for scholars of many disciplines. Indeed, several present or past projects focused on digitized newspapers, e.g. (Doucet et al., 2020; Ehrmann et al., 2020; Manrique-Gomez et al., 2024) with the aim to develop tools or create large collections. However, challenges in Optical Character Recognition (OCR) still remain far from being solved (Jarlbrink & Snickars, 2017; Late & Kumpulainen, 2021; Torget, 2023; Wevers, 2023) and pose problems to keyword search or further automated processing, as OCRed text often contains 'character recognition mistakes, formatting issues, and hyphenation problems' (Guan & Greene, 2024).

Through the process of post-correction, the OCRed text quality can be significantly improved. While a lot of work has been done in this regard and the progress and interest of the community can be documented e.g. by the ICDAR competitions (Rigaud et al., 2019; Chiron et al., 2017), many problems still remain unsolved and in specific languages and domains, custom post-correction models have to be trained or fine-tuned.

While the influence of the amount of training data on the model performance was evaluated on synthetical data (Guan & Greene, 2024), finding the least amount of annotation necessary to get a good model is a pressing question. Also, generalization of the post-correction models remains a question. Exploring performance of post-correction models trained on data coming from various OCR models would complement studies like (Dannélls & Persson, 2020) who explore the role of an OCR system for post-correction on Swedish data using as SVM model and find that in most cases, the performance of the post-correction decreases when the model is trained on mixed data from different OCR systems.

In this paper, we present a series of evaluations with a fine-tuned Byte-to-Byte Text-to-Text Transfer Transformer (ByT5) model for historical German on the corpus of job advertisements from digitized newspapers. We focus on how the number of annotations affects the performance of the model. We also compare the effect of different OCR models used and whether a model trained on outputs of more OCR models reaches better generalization ability.

2 Related Work

In a survey, Nguyen et al. (2022)distinguish between manual approaches that mostly benefit from crowdsourcing, and (semi-)automatic approaches. The latter are further divided into

approaches working with isolated words, e.g. using a dictionary or merging outputs from several OCR systems, and context-dependent, that benefit from information gained from the context. These include language models, feature-based machine learning (ML) models, and sequence-to-sequence models which conceive post-correction as a machine-translation task.

As the approaches using isolated words often face challenges when containing proper nouns, words of specific domains, historical orthographic variations (Nguyen et al., 2022), but also abbreviation, which is the case of our data, we further focus only on the contemporary approaches considering context.

Guan and Greene (2024) explored generating synthetic data for post-OCR corrections, including based on glyph similarity. Using dataset of 8 languages and several models, including mT5 (Xue et al., 2020), mBART (Tang et al., 2020) or ByT5 (Xue et al., 2022), they explore the impact of data volume and various methods of data generation on the model's performance. Finding that ByT5 performs the best for the post-OCR correction, they also find that for all languages, the results are the best when using the whole dataset instead of its fraction for training, and that the data augmentation x4 reaches generally the best improvement.

A ByT5 model was also used with very good results in (Maheshwari et al., 2022) on Sanskrit or in (Löfgren & Dannélls, 2024) for Swedish newspapers, encouraging further use of this model. On the other hand, (Debaene et al., 2025) report that in the post-correction of early modern Dutch theatre plays, the ByT5 model unnecessarily modifies sequences of the OCRed text.

Soper et al. (2021) fine-tuned BART for post-OCR correction and demonstrated its usefulness for this task. Later, Thomas et al. (2024) compare Llama and BART models on the task of post-correction of historical newspapers from the 19th century. This corpus is characterized by abbreviations and spelling variations over time. Authors report that Llama significantly outperforms BART, however, as it was trained predominantly on English data, its adaptation to other languages might prove a challenge. They also compare the performance with the amount of training data and show that while BART improves significantly with the increased amount of training data, Llama performs very well already from the outset.

3 Dataset

In our experiments, we use two datasets: Part of the ICDAR 2019-POCR (Rigaud et al., 2019) and the JobAds dataset created in the course of our project.

From the ICDAR 2019-POCR dataset, the DE1 (Ströbel & Clematide, 2019) DE2, DE3 and DE7 (Springmann et al., 2018) were used. They include frontpages of newspapers and literary works written in the Fraktur font in the form of images together with their textual ground-truth, as well as in the early modern latin font. We used 7896 files with an average length of 1569 (±666) bytes from those written in the Fraktur font. This dataset is used to create a fine-tuned hmByt5 model to investigate whether fine-tuning on unrelated OCR data can reduce the ground truth needed to create a suitable OCR post-correction model. The ICDAR data shows a SacreBLEU (Post, 2018) score of 11,87. The fine-tuned model raises this to 79,39.

JobAds dataset was created from ANNO corpus (Österreichische Nationalbibliothek, 2021), a collection of digitized historical newspapers predominantly in German. From 29 different newspaper titles, we picked issues from 1850-1950. From this corpus, a random sample of 1 page per year per newspaper was made. On these pages, all job advertisements were manually annotated using doccano (Nakayama et al., 2018). The annotated ads were OCRed with the frak2021 model (Mannheim University Library, 2021) and manually corrected using the Transkribus platform (Kahle et al., 2017). This resulted in 9680 machine-readable, proof-read job advertisements consisting of on average 196 (±159) bytes. Additional OCR text for each job ad was created with the german_print (Mannheim University Library, 2024) and austrian_newspaper (Mannheim University Library, 2023) models. As frak2021 is the main model used in our project, the evaluation dataset contains only OCR text from this model. The evaluation dataset has a SacreBLEU score of 68,46.

Because ByT5 encodes text on the byte level, encoding an entire job ad is computationally infeasible. Therefore, both datasets are split into segments containing a maximum of 150 bytes.

4 Methods

4.1 hmByT5

ByT5 is a variant of the Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) model

OCR Variant	Model	% of training set				
		10%	25%	50%	75%	100%
Single	Base	65.18	**73.39**	73.72	74.20	**74.60**
	ICDAR	65.14	70.36	71.75	72.29	72.96
Multi	Base	**72.09**	73.31	**74.05**	**74.22**	74.42
	ICDAR	68.82	70.91	72.15	72.54	7317

Table 1: SacreBLEU scores of the single-OCR and multi-OCR training variants on the base hmByT5 model and the model fine-tuned on the ICDAR dataset.

designed to process text at the byte level rather than relying on subword tokenization. This architecture enables ByT5 to handle diverse languages, character sets, and noisy text more effectively. By operating directly on raw byte sequences, the model avoids tokenization biases and can better capture fine-grained character-level patterns. While this approach increases computational complexity, it enhances robustness in tasks involving text with unconventional structures, misspellings, or rare word forms. ByT5 has demonstrated strong performance across a variety of natural language processing applications, particularly in scenarios where traditional tokenization methods struggle.

hmByT5[1] are a set of multi-lingual Byt5 models fine-tuned on historical data in English, German, French, Finnish, Swedish and Dutch. Here, we use byt5-small-historic-multilingual-span20-flax as it shows the best performance on German.

4.2 Training

In this study, we investigate the performance of two different versions of the hmByt5 model for OCR post-correction: The standard hmByt5 model and a fine-tuned version trained on a subset of the ICDAR-2019 dataset. The training data for our experiments is derived from the Anno-dataset, which consists of historical OCR outputs and their corresponding corrected versions by creating a randomized 80-20 split. To assess the impact of dataset size on model performance, we create five different subsets containing 10%, 25%, 50%, 75%, and 100% of the available training data. Additionally, we introduce two variations for each subset: A single-OCR variant, which consists of OCR text produced exclusively by the frak2021 model and a multi-OCR variant, which contains text generated by the three OCR models mentioned in the previous section.

Each hmByt5 variant is trained separately on all dataset configurations. We use the Adafactor optimizer for training, as this was used in the original ByT5 model. Training is conducted for five epochs and evaluated periodically with SacreBLEU.

5 Results

In this section, we are presenting SacreBLEU scores for the two models trained on different subsets of our training data. Table 1: SacreBLEU scores of the single-OCR and multi-OCR training variants on the base hmByT5 model and the model fine-tuned on the ICDAR dataset. shows how well each model performed for five different dataset sizes when trained on either single- or multi-OCR versions of the JobAds data.

6 Discussion

The results of this study reveal several notable trends and provide valuable insights into the factors influencing OCR post-correction performance. The analysis of dataset size and OCR model diversity highlights critical aspects of model generalization and domain adaptation. Additionally, the relatively lower performance of the fine-tuned model raises important questions about the suitability of pre-training datasets and their impact on downstream tasks.

The observed improvements in SacreBLEU scores with increasing dataset size are consistent with established findings in machine learning, where larger training datasets enable models to capture more diverse patterns and relationships. The most significant performance gain occurs between using 10% and 25% of the training data. Beyond this point, further improvements become more moderate across all model configurations. For example, the base hmByT5 model trained on the single-OCR variant achieves a 7.2% increase in SacreBLEU when moving from 10% to 25% of the training data, while using the entire dataset yields a total improvement of 9.0%. Considering the

[1] https://github.com/stefan-it/hmByT5

substantial manual effort required to create four times as much training data, the trade-off of slightly lower performance might be acceptable in real-world applications.

Fine-tuning the hmByT5 model on the ICDAR dataset consistently lowers performance across all training sizes. Despite the ICDAR dataset's focus on OCR error correction, it was created with older OCR models and features lower OCR quality. This suggests that error correction does not generalize well between datasets with significantly different qualities or originating from different OCR models. Another key factor may be the ICDAR dataset's larger size relative to the JobAds dataset, as it consists of longer texts. This likely results in the model overfitting to the specific error patterns in the ICDAR dataset. This overfitting could explain why SacreBLEU scores for the fine-tuned model exhibit higher absolute changes as additional training data is introduced, progressively mitigating this overtuned correction behavior. Notably, this effect is further supported by the fact that the multi-OCR training variant has a more pronounced impact on the fine-tuned ICDAR model, highlighting its sensitivity to training diversity.

In contrast, the multi-OCR training variant produces mixed results for the base hmByT5 model. When using only 10% of the training data, the model's performance shifts from a 4.8% reduction in SacreBLEU to a 5.3% improvement. However, when trained on the entire dataset, the multi-OCR variant underperforms compared to the single-OCR variant. This suggests that while multi-OCR training can improve generalizability when data is limited, it may lead to reduced performance on larger datasets, possibly due to conflicting error patterns among different OCR systems. This is also in agreement with older findings comparing post-correction performance across different OCR systems (Dannélls & Persson, 2020).

7 Conclusion

This study highlights key factors influencing OCR post-correction performance, including dataset size, OCR model diversity, and pre-training strategies. Larger training datasets improve model performance, though gains diminish as dataset size increases, emphasizing the need for balancing annotation effort and performance benefits.

Training on multi-OCR outputs enhances generalization, especially with smaller datasets, but can introduce conflicting patterns in larger datasets. Fine-tuning on the ICDAR dataset consistently reduced performance, likely due to domain mismatch and overfitting, underscoring the importance of dataset alignment in pre-training.

These findings stress the need for careful dataset design and training strategies. Future work could explore domain-adaptive fine-tuning, synthetic data augmentation, and more nuanced evaluation metrics to develop robust OCR post-correction systems, improving access to digitized historical texts.

8 Limitations

A significant limitation of this study is its focus solely on correcting errors in texts that have already undergone OCR processing. The performance of OCR models is heavily influenced by the preceding layout analysis step, which determines the structure and organization of the document before text extraction (Liebl & Burghardt, 2021). The generalizability of post-correction methods may vary depending on the specific techniques employed during layout analysis, even when the same OCR model is used. This aspect warrants further investigation to better understand its impact on OCR post-correction.

Another limitation lies in the size of the JobAds dataset used in this study. While the findings illustrate the relationship between dataset size and improvements in OCR quality through post-correction, the observed performance decline of multi-OCR models with increasing amounts of training data requires validation on a larger dataset. A more extensive dataset would allow for a deeper exploration of these trends and provide more robust conclusions about the scalability of multi-OCR training approaches.

Acknowledgments

References

C. Rigaud, A. Doucet, M. Coustaty, & J. -P. Moreux. (2019). ICDAR 2019 Competition on Post-OCR Text Correction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1588–1593. https://doi.org/10.1109/ICDAR.2019.00255

Chiron, G., Doucet, A., Coustaty, M., & Moreux, J.-P. (2017). ICDAR2017 Competition on Post-OCR Text Correction. *2017 14th IAPR International Conference*

on Document Analysis and Recognition (ICDAR), 1423–1428. https://doi.org/10.1109/ICDAR.2017.232

Dannélls, D., & Persson, S. (2020). Supervised OCR Post-Correction of Historical Swedish Texts: What Role Does the OCR System Play? *Digital Humanities in the Nordic and Baltic Countries Publications*, *3*(1), 24–37. https://doi.org/10.5617/dhnbpub.11176

Debaene, F., Maladry, A., Lefever, E., & Hoste, V. (2025). Evaluating Transformers for OCR Post-Correction in Early Modern Dutch Theatre. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Hrsg.), *Proceedings of the 31st International Conference on Computational Linguistics* (S. 10367–10374). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.690/

Doucet, A., Gasteiner, M., Granroth-Wilding, M., Kaiser, M., Kaukonen, M., Labahn, R., Moreux, J.-P., Muehlberger, G., Pfanzelter, E., Therenty, M.-E., Toivonen, H., & Tolonen, M. (2020, Juni). *NewsEye: A digital investigator for historical newspapers*. Digital Humanities 2020 (DH 2020), Ottawa, Canada, 20-25 July 2020. https://doi.org/10.5281/zenodo.3895269

Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P. B., & Barman, R. (2020). Language Resources for Historical Newspapers: The Impresso Collection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 958–968. https://doi.org/10.5167/uzh-191270

Guan, S., & Greene, D. (2024). Advancing Post-OCR Correction: A Comparative Study of Synthetic Data. *Findings of the Association for Computational Linguistics ACL 2024*, 6036–6047. https://doi.org/10.18653/v1/2024.findings-acl.361

Jarlbrink, J., & Snickars, P. (2017). Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *Journal of Documentation*, *73*(6), 1228–1243. https://doi.org/10.1108/JD-09-2016-0106

Kahle, P., Colluto, S., Hackl, G., & Mühlberger, G. (2017). Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *04*, 19–24. https://doi.org/10.1109/ICDAR.2017.307

Late, E., & Kumpulainen, S. (2021). Interacting with digitised historical newspapers: Understanding the use of digital surrogates as primary sources. *Journal of Documentation*, *78*(7), 106–124. https://doi.org/10.1108/JD-04-2021-0078

Liebl, B., & Burghardt, M. (2021). An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers. *2020 25th International Conference on Pattern Recognition (ICPR)*, 5153–5160. https://doi.org/10.1109/ICPR48806.2021.9412571

Löfgren, V., & Dannélls, D. (2024). Post-OCR Correction of Digitized Swedish Newspapers with ByT5. In Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, & S. Szpakowicz (Hrsg.), *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)* (S. 237–242). Association for Computational Linguistics. https://aclanthology.org/2024.latechclfl-1.23/

Maheshwari, A., Singh, N., Krishna, A., & Ramakrishnan, G. (2022). A Benchmark and Dataset for Post-OCR text correction in Sanskrit. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6258–6265. https://doi.org/10.18653/v1/2022.findings-emnlp.466

Mannheim University Library. (2021). *Frak2021* (Version frak2021-0.905) [Software]. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/frak2021-0.905.traineddata

Mannheim University Library. (2023). *Austrian_newspapers* [Software]. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/tesseract/austriannewspapers/

Mannheim University Library. (2024). *German_print* [Software]. https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/german_print/

Manrique-Gomez, L., Montes, T., Rodriguez Herrera, A., & Manrique, R. (2024). Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction. In M. Hämäläinen, E. Öhman, S. Miyagawa, K. Alnajjar, & Y. Bizzoni (Hrsg.), *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (S. 132–139). Association for Computational Linguistics. https://aclanthology.org/2024.nlp4dh-1.13

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). *doccano: Text Annotation Tool for Human*. https://github.com/doccano/doccano

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2022). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, *54*(6), 1–37. https://doi.org/10.1145/3453476

Österreichische Nationalbibliothek. (2021). *ANNO Historische Zeitungen und Zeitschriften*. https://anno.onb.ac.at/

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. https://doi.org/10.18653/v1/W18-6319

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(1), 140:5485-140:5551.

Soper, E., Fujimoto, S., & Yu, Y.-Y. (2021). BART for Post-Correction of OCR Newspaper Text. *Proceedings of the Seventh Workshop on Noisy User-Generated Text (W-NUT 2021)*, 284–290. https://doi.org/10.18653/v1/2021.wnut-1.31

Springmann, U., Reul, C., Dipper, S., & Baiter, J. (2018). *Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin*. arXiv. https://doi.org/10.48550/ARXIV.1809.05501

Ströbel, P., & Clematide, S. (2019). *Ground truth for Neue Zürcher Zeitung black letter period* [Dataset]. figshare. https://doi.org/10.6084/M9.FIGSHARE.8864510.V1

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*. arXiv. https://doi.org/10.48550/ARXIV.2008.00401

Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for Post-OCR Correction of Historical Newspapers. In R. Sprugnoli & M. Passarotti (Hrsg.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024* (S. 116–121). ELRA and ICCL. https://aclanthology.org/2024.lt4hala-1.14/

Torget, A. (2023). Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections. In E. Bunout, M. Ehrmann, & F. Clavert (Hrsg.), *Digitised Newspapers – A New Eldorado for Historians?* (S. 47–66). https://doi.org/10.1515/9783110729214-003

Wevers, M. (2023). Mining Historical Advertisements in Digitised Newspapers. In E. Bunout, M. Ehrmann, & F. Clavert (Hrsg.), *Digitised Newspapers – A New Eldorado for Historians?* (S. 227–252). https://doi.org/10.1515/9783110729214-011

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, *10*, 291–306. https://doi.org/10.1162/tacl_a_00461

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). *mT5: A massively multilingual pre-trained text-to-text transformer*. arXiv. https://doi.org/10.48550/ARXIV.2010.11934