Evaluating Transformers for OCR Post-Correction in Early Modern Dutch Theatre

Florian Debaene, Aaron Maladry, Els Lefever, Véronique Hoste
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

This paper explores the effectiveness of two types of transformer models—large generative models and sequence-to-sequence models—for automatically post-correcting Optical Character Recognition (OCR) output in early modern Dutch plays. To address the need for optimally aligned data, we create a parallel dataset based on the OCRed and ground truth versions from the EmDComF corpus using state-of-the-art alignment techniques. By combining character-based and semantic methods, we design and release a qualitative OCR-to-gold parallel dataset, selecting the alignment with the lowest Character Error Rate (CER) for all alignment pairs. We then fine-tune and evaluate five generative models and four sequence-to-sequence models on the OCR post-correction dataset. Results show that sequence-to-sequence models generally outperform generative models in this task, correcting more OCR errors and overgenerating and undergenerating less, with mBART as the best performing system.

1 Motivation & Related Work

Inspired by the success of generative large language models in a variety of NLP tasks, their usefulness has recently also been explored to automatically correct the output of Optical Character Recognition (OCR) models for historical documents. Evaluating 14 foundation models in zero and few-shot settings against 8 OCR post-correction benchmarks for manuscripts, newspapers, literary commentaries and other historical documents in different languages, time periods and transcription quality, Boros et al. (2024) found that generative models did not improve faulty transcriptions in their applied experimental settings. Moreover, they often degraded the transcription quality of texts. Conversely, Thomas et al. (2024) compared generative models for OCR post-correction after supervised fine-tuning (SFT) to prevalent sequence-to-sequence (seq2seq) models for OCR post-correction on BLN600 (Booth et al., 2024), a dataset of 19th century British newspaper articles, and reported a Character Error Rate reduction of 54.51% after instruction-tuning generative models for a prompt-based approach to OCR post-correction.

Both Boros et al. (2024) and Thomas et al. (2024) recommend fine-tuning transformers on period- and genre-specific datasets to optimize OCR post-correction. Therefore, the models and results of their experiments do not directly transfer to datasets from other languages and periods. In the latter study, for instance, fine-tuned generative models were evaluated on 19th century English newspapers—a domain more contiguous with the predominantly English training data of most transformer models. In this paper, we extend this research by focusing on the automatic OCR post-correction of early modern Dutch texts from the EmDComF corpus (Debaene et al., 2024), which is a type of historical language that has no central position in the pre-training data of language models. This language variant presents significant challenges for OCR post-correction due to orthographic variations and substantial lexical and semantic shifts from modern Dutch, further complicating automatic text processing. Moreover, automatically transcribed texts in EmDComF have a reported Character Error Rate (CER) and Word Error Rate (WER) of 8.5% and 9.5% on average at the document level after pre-processing without post-correction. As nearly half of EmDComF consists of uniquely OCRed texts, this further underlines its need for additional processing. Our contributions to automatically processing historical OCRed texts include:

1. **Investigation of Alignment Methods**: We explore three alignment methods, one of which leverages layout features originating from book editions captured by OCR systems, such

as line breaks and formatting (Section 2.1).

2. **Parallel Corpus Creation**: We develop and release[1] a dataset for post-correcting OCRed early modern Dutch theatre texts, based on the three alignment methods (Section 2.2).

3. **Model Comparison**: We evaluate generative and seq2seq transformer models, showing that seq2seq models outperform generative models in this OCR post-correction task (Section 3).

2 Parallel Corpus Creation

The parallel data serving the task of post-correcting early modern Dutch OCR data comes from the EmDComF subset that has both gold standard texts, of which 92 originate from Census Nederlands Toneel (CENETON) and 34 texts from Digitale Bibliotheek voor de Nederlandse Letteren (DBNL), and their respective OCRed versions, originating from the automatic transcription of scanned plays on Google Books with Transkribus Print M1. This leaves us with a dataset of 126 parallel texts in total that we operationalize for OCR post-correction through automatic alignment.

2.1 Alignment Methods

2.1.1 RETAS Method

The RETAS method, a fast "recursive text alignment scheme" (Yalniz and Manmatha, 2011), aligns OCRed text with its ground truth by first using a longest common subsequence algorithm to match unique words as anchor points. These anchors divide the text into segments, which are recursively aligned by finding new shared words. This continues until the smaller segments can be aligned with an edit distance algorithm. This method follows a character-based genetic algorithm approach, where alignment evolves based on the recursive segmentation and alignment of text subsequences.

2.1.2 Baseline Semantic Search

As opposed to the first approach which predominantly relies on word and character matching, this second approach creates alignment based on semantic similarity. As EmDComF consists of unstructured .txt files for gold and OCR texts alike, the full-play text files for both OCR and gold text are split into sentences using nltk (Bird et al., 2009). After creating the two lists of sentences, we generate sentence embeddings for each gold and OCR sentence using all-mpnet-base-v2 (Reimers and Gurevych, 2019), a general-purpose sentence transformer. Consequently, we use these embeddings to calculate the cosine distance between each gold sentence compared to every OCR sentence, without considering the position of the candidate sentences. Note that OCR alignment candidates are limited to sentences corresponding to the same source text, i.e. the same play, to avoid cross-text mappings. From this semantic search, we select the OCR sentence with the lowest cosine distance to create aligned sentence pairs.

2.1.3 Improved Semantic Search

The third approach we employ is a variation on the aforementioned semantic search alignment method. After manual inspection of the OCR sentences, we found that many automatically detected sentences contained multiple sequences that were concatenated with newline characters, introduced by the positional encoding of the OCR system. As a result, the OCR sentences often contained the same sentence that was also present in the gold text, but also included preceding or consequent sequences, often indicating structural features such as acts, scenes, page numbers and titles that are frequently left out in ground truth versions of plays. To overcome this type of noise, we propose an improved semantic search approach. This approach involves creating additional alignment candidates by splitting the original OCR sentences based on newline characters. After splitting an OCR sentence into newline segments, we create up to 20 new candidates per sentence by recursively concatenating consequent segments. Note how in the example of Appendix A the combination of the final sentence chunk with none of the previous sequences creates an exact string match with the gold sequence. As this approach extends the set of alignable OCR candidates, new sentence embeddings are successively created and follow the previous semantic search method to create alignments.

2.2 Alignment Results

To evaluate the three alignment methods, we calculate CER and WER as proposed by Neudecker et al. (2021), and normalised Character Error Rate (nCER), or CER divided by the averaged length of both gold and OCR sequences. As shown in Table 1, the RETAS and improved semantic search (chunk) methods achieve equal error rates in almost 80% of all alignment pairs (120,829). While the im-

[1] floriandebaene/EmDComF_OCR_post-correction

proved semantic search approach yields the lowest error rates in 12% of the pairs, the RETAS algorithm finds the best results in 8%, and the baseline semantic approach does so in a very small fraction. This indicates that, whilst our improved semantic search approach is the best single approach, there are still cases where the other methods perform better. With these alignment methods, we aim to create an optimally aligned corpus, with the lowest number of wrong alignments, before building systems for OCR post-correction. We do so, because any wrong alignment could be detrimental to the training process and because wrong alignments in the test set would set unrealistic expectations for our OCR correction models, i.e. rewriting an OCR sentence as a completely different sentence (Lyu et al., 2021). To this end, we propose combining the three methods into a single alignment method (triadic) that selects the OCR sentence with the lowest CER out of any of the three approaches. We select CER as the defining metric in the creation of our parallel post-correction dataset, as it is frequently used to measure performance in OCR studies (Nguyen et al., 2021; Carlson et al., 2024; Pavlopoulos et al., 2024). As shown in Figure 1, our proposed triadic method consistently achieves substantially lower error rates across all metrics. We refer to Appendix B for the complete scores. We use this dataset in the experiments to post-correct early modern Dutch.

	CER	WER	nCER
Tie	78.91	79.20	78.70
Chunk	11.92	12.36	11.73
RETAS	8.80	6.91	9.10
Base	0.36	1.53	0.47

Table 1: Percentage of aligned pairs where each method delivered the best score, calculated per metric. Tie indicates when Chunk and RETAS scored equally.

3 OCR Post-Correction

Starting from 120,829 alignment pairs, we remove duplicate pairs to prevent having identical samples in both train and test set. Besides addressing data contamination, this minimizes the impact of much repeated and often exactly transcribed OCR sentences during training, like character names and structural indications such as acts and scenes. Then, we capitalize the first letter of each sentence, lowercase the rest and remove punctuation except for periods, commas, exclamation and question

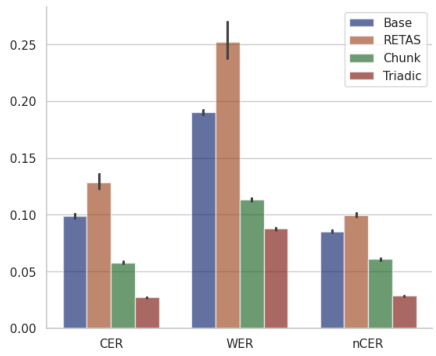

Figure 1: Average error rates (↓) per alignment method, with black lines indicating standard deviation.

marks. This results in a final OCR post-correction dataset of 79,481 alignment pairs, which we divide in a train (57,225), test (15,897) and development (6,359) set, stratified on the distribution of exact string matches at 44.78%.

3.1 Methodology

For the OCR post-correction task, we explore two distinct methodologies using transformer models. The first method involves fine-tuning a selection of pre-trained transformer models for seq2seq modelling, which is considered the current state-of-the-art (SOTA) approach for OCR post-correction. The second method encompasses fine-tuning of large generative models, a novel approach that is gaining ground as the new SOTA for many NLP problems. For seq2seq fine-tuning, we select mBART (Tang et al., 2020) and mT5 (Xue et al., 2021) for their multilingual capabilities, and two more recent T5 models, ByT5 (Xue et al., 2022) and Flan-T5 (Chung et al., 2024). These architectures have successfully been used in related work on OCR post-correction or language normalization in different languages and from different periods (Soper et al., 2021; Madarász et al., 2024; Löfgren and Dannélls, 2024; Wolters and Van Cranenburgh, 2024).

To investigate the large generative models, we employ parameter-efficient fine-tuning, with QLoRa (Dettmers et al., 2023), which updates an additional low-rank decomposition of the weight matrix instead of updating all model weights. We make use of the supervised fine-tuning with trl (von Werra et al., 2020) based on a prompt that combines the OCR text with the gold output (Appendix C). We fine-tune the following selection of instruction-tuned models. We begin with Llama 3, a robust multilingual model, and compare it to two similarly-sized, Dutch-specific models, GEITje

	Test Set (N=15,897)			
	CER	WER	nCER	match
Baseline	4.20	12.56	4.21	44.78
Fietje	12.61	19.72	9.75	40.05
GEITje	11.03	17.81	9.01	43.52
Reynaerde	10.84	17.50	8.39	43.93
Llama 3	10.64	15.15	8.15	50.96
OCRonos	7.94	15.31	5.26	44.95
mBART	**3.51**	**9.16**	**3.47**	**53.76**
Flan-T5	4.21	10.68	4.50	49.66
mT5	4.93	11.55	7.21	43.86
ByT5	9.98	15.51	14.55	33.69

Table 2: Error rates (↓) and matches (↑) on the test set.

(Vanroy, 2024) and Reynaerde, as well as Fietje, a Dutch-specific model with significantly fewer parameters. Finally, we also employ OCRonos, a generative model that is developed to correct OCR output for English. Since these generative models often provide additional examples and explanations, we employ a set of post-processing rules to reduce noise in the model outputs. These post-processing rules include the removal of the input of the prompt template, instruction-tuning tokens, and replacing outputs longer than the input text by 3 or more tokens with the baseline OCR text. For fair evaluation, the final rule is also applied to the output of the seq2seq models. We refer to Appendix E for the training and prediction parameters.

3.2 Performance Overview

Table 2 highlights the performance of the models on post-correcting the 15,897 samples in the test set. Among all models, mBART clearly outperformed all others, achieving the lowest error rates across all metrics and the highest match rate, indicating the percentage of exact string matches between system output and gold standard. Other seq2seq models such as Flan-T5 and mT5 showed moderate performance, with Flan-T5 standing as the second-best model after mBART. Generative models like Fietje, GEITje, and Reynaerde generally underperformed, showcasing error rates higher than the baseline. Among the generative models, Llama 3 and OCRonos performed best, even though they both still increased error rates on average. ByT5 had the weakest match rate, suggesting greater challenges in post-correcting OCR sentences. Finally, given the strong baseline, we recognize that Transkribus Print M1 performs well at automatic text transcription, even without post-correction. We note that this task seems particularly challenging for generative models, as they need to maintain a balance between correcting missing characters and leaving correct text untouched.

3.3 Error Analysis

Table 3 delves deeper into model performance by dividing the test set into its two subsets: one where post-correction was required (indicating room for improvement) and one where OCR outputs already matched the gold standard. This puts model performance into perspective by highlighting how challenging and straightforward cases were dealt with.

For the subset requiring post-correction, the patterns in model performance appear to be similar to the test set. Again, mBART stands out with the lowest error rates across the board, followed by the other seq2seq models, except for ByT5, and then the large generative models. The character and token difference scores (c_diff and tok_diff), which represent the difference between gold and model output lengths, further support this conclusion. Here, negative scores indicate that the generated output was longer than the gold standard. mBART's smaller length differences show that it overgenerated and undergenerated the least, producing more correct sequences (match). Concretely, this meant making OCR sequences longer, as the baseline indicates that on average OCR sequences needing post-correction were too short (nCER, c_diff, tok_diff). On the other hand, generative models like Fietje, GEITje and Reynaerde yield significantly higher error rates alongside their high and negative difference scores, which suggests overgeneration (CER, c_diff, tok_diff). Still, the nCER scores indicate that these models also suffered from undergeneration when post-correcting. Further, Llama 3 and OCRonos performed better than the other generative models. Llama 3 is most successful at completely correcting faulty OCR input (match). However, OCRonos achieves lower error rates by outputting less long sequences. OCRonos therefore seems to be more careful by making smaller adjustments that often don't completely correct the sequences (match), whilst also being susceptible to overgeneration (CER, c_diff, tok_diff). These results point at difficulties for generative models in maintaining the delicate balance of adding and removing characters or tokens required for OCR post-correction.

For the subset where no post-correction was needed, we observe negative length differences

	Room for Improvement (N=8,778)						No Room for Improvement (N=7,119)					
	CER	WER	nCER	c_diff	tok_diff	match	CER	WER	nCER	c_diff	tok_diff	match
Baseline	5.56	16.71	7.62	0.15	0.26	0	0	0	0	0	0	100
Fietje	15.72	24.54	14.80	-10.41	-1.90	9.89	3.04	5.12	3.52	-0.23	-0.09	77.23
GEITje	12.38	21.13	14.17	-7.70	-1.37	14.32	2.70	5.60	3.57	-0.48	-0.09	77.10
Reynaerde	13.65	21.76	12.93	-10.01	-1.76	14.87	2.17	4.59	2.80	-0.39	-0.08	79.76
Llama 3	13.52	19.40	12.96	-9.06	-1.71	18.50	1.79	2.28	2.23	-0.33	-0.08	90.98
OCRonos	10.10	19.18	8.51	-4.98	-0.87	13.83	1.30	3.59	1.26	0.19	0.02	83.33
mBART	**4.58**	**11.84**	**6.05**	-1.58	-0.22	**21.37**	**0.22**	**1.04**	**0.28**	-0.01	0	**93.71**
Flan-T5	5.42	13.71	6.77	-1.69	-0.24	15.71	0.51	1.48	1.71	-0.08	-0.02	91.52
mT5	5.75	14.02	8.32	-2.14	-0.31	15.08	2.42	4.06	5.85	-0.20	-0.02	79.34
ByT5	9.87	16.61	13.23	-1.54	-0.4	14.50	10.34	12.18	16.17	3.31	0.52	57.34

Table 3: Error rates (↓), matches (↑) and character and token difference scores (c_diff and tok_diff) on the subset which required post-correction (left) and on the subset which did not (right).

for all models, except for ByT5 and OCRonos, which again indicate general overgeneration, although these differences are overall much less significant than in the other subset. The performance of most models remains similar in this subset too, although ByT5 only maintains 57.34% of the originally matching OCR input sequences, suggesting that frequent unnecessary modifications and undergeneration (nCER, c_diff, tok_diff) heightened its error rates. Notably, the generative models attained less extreme error rates in this subset, with Llama 3 recognizing already correct sequences in 90% of the cases and with OCRonos having the least issues with overgeneration and undergeneration. Finally, see Appendix D for an example of a post-corrected OCR sentence per model.

4 Conclusion

With this work, we advance the processing of automatically digitised historical texts by exploring the impact of sentence alignment and OCR post-correction for early modern Dutch in the EmDComF corpus. First, we investigated different alignment approaches to build a parallel dataset for OCR post-correction. To this end, we explored a character-based genetic algorithm and two semantic search methods and found that, while semantic search methods attain lower error rates, all methods have their merits. Thus, we combined these three methods into our "triadic" alignment approach, reflecting their unequal contributions (see Table 1), to build the optimal dataset for OCR post-correction. Using this dataset, we evaluated a variety of models for OCR post-correction. This evaluation covers two distinct methodologies, both fine-tuning of (1) seq2seq transformer models, the traditional SOTA for this task, and (2) decoder-only generative models, more novel systems, that have shown promising results for historical English. Our experiments suggest that the excellent performance of generative models on historical English does not transfer to early modern Dutch, as seq2seq models generally perform better. Based on our results, mBART attains the best performance out of all models, correcting more OCR mistakes and leaving more already correct OCR sentences intact, followed by FlanT5. ByT5 seems to modify OCR sequences when it was unnecessary the most. Manual inspection reveals that many of its mistakes are caused by flooding and reduplication, which we also observe in the generative models. Further, the generative models generally suffered more from both overgeneration and undergeneration when there was room for improvement, but much less when there was none. One might expect Dutch-specific language models like Fietje, GEITje and Reynaerde to outperform the multilingual Llama 3, given that early modern Dutch is closer to modern Dutch than to English. However, this is not the case. Llama 3, with its extensive training data and instruction tuning, outperforms the Dutch-specific models and proves to be the best generative model to post-correct early modern Dutch after fine-tuning. Contrarily, OCRonos, a generative model specifically designed to correct English OCR, did not generalize as well to early modern Dutch and therefore could not beat Llama 3 in this context. In conclusion, we found that in our experimental settings seq2seq models learn more effectively when and how to post-correct, making them better suited for adapting to early modern Dutch. As a result, they outperform generative models in this domain-specific OCR post-correction task.

Limitations

Firstly, the proposed alignment methods to create a parallel dataset for OCR post-correction were only tested on a single early modern Dutch corpus pertaining to the dramatic genre. In future work, we will experiment with held-out train and test sets from different source databases, employing more out-of-distribution settings. Furthermore, the models are expected to be directly transferable only to other historical Dutch corpora, as their performance on distantly or unrelated languages will likely be significantly different. In our alignment methodology, we made use of sentence-level splitting of the texts. Conversely, it would also be possible to group the texts into dialogues or paragraphs to allow language processing in larger and (sometimes) more meaningful contexts, which is possible for structured drama corpora like those available on DraCor. The automatic insertion of this type of structural information into unstructured text is left for future work. Finally, we hypothesize that generative models might work better on larger sequences, when additional context can be provided beyond the sentence level.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. This work was supported by the Research Foundation Flanders (FWO) under grant G032123N and by Ghent University under grant BOF.24Y.2021.0019.01.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Callum William Booth, Alan Thomas, and Robert Gaizauskas. 2024. BLN600: A parallel corpus of machine/human transcribed nineteenth century newspaper texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2440–2446, Torino, Italia. ELRA and ICCL.

Emanuela Boros, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

Jacob Carlson, Tom Bryan, and Melissa Dell. 2024. Efficient OCR for building a diverse digital history. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8105–8115, Bangkok, Thailand. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Florian Debaene, Kornee van der Haven, and Veronique Hoste. 2024. Early Modern Dutch comedies and farces in the spotlight: Introducing EmDComF and its emotion framework. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 144–155, Torino, Italia. ELRA and ICCL.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Viktoria Löfgren and Dana Dannélls. 2024. Post-OCR correction of digitized Swedish newspapers with ByT5. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 237–242, St. Julians, Malta. Association for Computational Linguistics.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural OCR post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Gábor Madarász, Noémi Ligeti-Nagy, András Holl, and Tamás Váradi. 2024. Ocr cleaning of scientific texts with llms. In *International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs*, pages 49–58. Springer Nature Switzerland Cham.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, pages 13–18.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

John Pavlopoulos, Vasiliki Kougia, Esteban Garces Arias, Paraskevi Platanou, Stepan Shabalin, Konstantina Liagkou, Emmanouil Papadatos, Holger Essler, Jean-Baptiste Camps, and Franz Fischer. 2024. Challenging error correction in recognised byzantine Greek. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 1–12, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021. Bart for post-correction of ocr newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Bram Vanroy. 2024. Geitje 7b ultra: A conversational model for dutch. *Preprint*, arXiv:2412.04092.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Andre Wolters and Andreas Van Cranenburgh. 2024. Historical dutch spelling normalization with pre-trained language models. *Computational Linguistics in the Netherlands Journal*, 13:147–171.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ismet Zeki Yalniz and R. Manmatha. 2011. A fast alignment scheme for automatic ocr evaluation of books. In *2011 International Conference on Document Analysis and Recognition*, pages 754–758.

Appendix

A Improved Semantic Search

{gold} **En wyl ik dat rapsody ken,** ...

{baseline} gitized by Google \n 27 \n 28 \n BESLIKTE SWAANTJE \n **En wyl ik dat rapsody ken,** ...

{chunk} **En wyl ik dat rapsody ken,** ...

{translation} And while I know that rhapsody, ...

B Error Rates per Alignment Method

	CER	WER	nCER
Base	0.0987	0.1901	0.0853
Chunk	0.0577	0.1133	0.0607
RETAS	0.1286	0.2521	0.0997
Triadic	**0.0272**	**0.0876**	**0.0285**

C Prompt Template

{user} Correct the OCR errors in the provided text.
Not all texts contain errors.
Text: {INPUT_OCR}
{ass}### Corrected Text: {CORRECTED_OCR}

D Post-Correction Example

{gold} Of de redenen daar van?
{base} Ofde rede! nen daar van?
{mBART} Of de redenen daar van?
{Flan-T5} Of de redenen daar van?
{Reynaerde} Of de redenen daar van?
{mT5} De redenen daar van?
{ByT5} Ofde redenen daar van?
{GEITje} Of de redenen daarvan?
{Llama3} Of de reeden daar van?
{OCRonos} Ofde redden daar van?

{translation} Or the reasons thereof?

E Model Parameters

E.1 Seq2Seq Models

Training	
Max Sequence Length	1024
Learning Rate	2e-5
Batch Size	2
Training Epochs	2
Predicting	
Max Sequence Length	1024
Number of Beams	5

E.2 Generative Models

Training	
Max Sequence Length	1024
Lora Alpha	32
Ranks	128
Learning Rate	5e-5
Batch Size	4
Training Epochs	1
Predicting	
Max Sequence Length	150
Top K	4
Returned Sequences	1