Towards an Interoperable Corpus of Austrian Historical Newspapers: The case of PressMint-AT

Tanja Wissik, Jona Hassenbach, Hannes Pirker, Claudia Resch, Stefan Resch

Austrian Centre for Digital Humanities
Austrian Academy of Sciences, Vienna, Austria
{Tanja.Wissik, JonaMarie.Hassenbach, Hannes.Pirker, Claudia.Resch, Stefan.Resch}@oeaw.ac.at

Abstract

In this paper the PressMint-AT project is presented, which aims to create a historical newspaper corpus based on the *Wiener Abendpost*. The quality of automatic text recognition (ATR) is a key factor in creating historical newspaper corpora. Therefore, the performance of established ATR tools, multimodal large language models (LLMS), and existing full-text transcriptions provided by the Austrian National Library via ANNO is evaluated in order identify the most suitable approach for the *PressMint-AT* project. Even though recent research has demonstrated promising results for OCR tasks using multimodal LLMs, the experiments presented in this paper show, that PERO OCR achieves the best performance for the *PressMint-AT* dataset.

Keywords: historical newspaper corpora, OCR, multimodal LLMs

1. Introduction

Since the early 2000s, regional and national libraries, alongside transnational organisations and commercial providers, have invested substantially in the digitisation of historical newspapers converting newspaper content into machine-readable text by means of optical character recognition (OCR). As a result, millions of newspaper facsimiles, together with their corresponding textual transcriptions, have been produced at regional, national, and international scales (Ehrmann et al., 2023). These developments have led to numerous historical newspaper corpora (Fišer and Lenardič, 2018). Examples include corpora containing several newspapers such as *sPeriodika 1.0* (Dobranić et al., 2024) for Slovenia or the *Couranten Corpus 2.0* (2025) for Dutch, as well as corpora containing a specific newspaper such as the Corpus *Wienerisches Digitarium (Resch and Kampkaspar 2020*). However, several challenges remain regarding historical newspaper corpora: first, the varying quality of OCR-generated full texts provided by libraries, and second, the lack of interoperability between corpora resulting from differing encoding standards.

The CLARIN funded project *PressMint*[1], aims to address these issues by compiling a multilingual, comparable, annotated, translated and interoperable set of corpora of European historical newspapers from around the start of the 20th century. Within the Austrian sub-project, *PressMint-AT*, a corpus of the *Wiener Abendpost,* a supplement of the *Wiener Zeitung,* published from 1863 to 1921, will be created.

This paper examines and discusses different approaches (including the use of multimodal LLMs) that can contribute to producing automatic high-quality transcripts of Fraktur typeface. These transcriptions form the basis for subsequent processing steps within the *PressMint* project, such as part-of-speech tagging, named entity recognition (NER), entity linking, and TEI encoding.

2. Related Work on Austrian Historical Newspapers

In recent years, archives and libraries in Austria have made substantial progress in the digitisation of historical newspapers, providing full-text search functionality based on textual transcriptions generated through OCR. For example, the Austrian National Library is providing access to the full texts of 27 million pages from historical newspapers and magazines.[2] In practice, however, the quality of these full texts varies considerably, ranging from low-quality or “noisy” OCR output to manually corrected OCR results. Therefore, several research projects, working with historical newspapers, have created new full texts by applying different OCR approaches: The full text digitization project of the *Wiener Zeitung* between 1703 and 1798 (Resch, 2023), which trained a dedicated Transkribus model for Fraktur typeface (Resch and Kampkaspar, 2020). The CLARIAH-AT funded Esperanto Newspaper Excerpt project used the open-source software Tesseract OCR in order to create full texts of digitized newspaper excerpts about Esperanto, which are preserved in the Department of Planned Languages and Esperanto Museum of the Austrian National Library (Mayer, 2024). Also, the JobAds Project produced OCR and corrected several thousand job advertisements from 14 different newspapers between 1850 and1950 provided by ANNO (Venglarova et al., 2024). For the OCR task Tesseract with the frak2021 model (M. U. Library, 2021) was used and then the

[1] https://www.clarin.eu/pressmint

[2] https://www.onb.ac.at/en/

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 34–39

automatic OCR output was manually corrected within Transkribus.

As these examples demonstrate, most earlier projects used well-established tools or platforms for the transcription and processing of historical documents such as Transkribus or Tesseract. But recently, with the rise of multimodal LLMs, LLMs are also used for OCR tasks with promising results (Greif et al., 2025). Early research found that LLM-based text recognition often outperforms state-of-the-art pipelines. Multimodal LLMs often transcribe unseen manuscripts zero-shot for printed and even handwritten documents with better results than well-established tools like Transkribus (Humphries et al., 2024) or Tesseract (Kim et al., 2025).

However, using multimodal LLMs for automatic text recognition also has limitations, and poses challenges such as hallucinations (Li, 2024; Boros et al., 2024) and that their performance for English texts is better than for other languages (Corsillia et al., 2025).

Accordingly, this paper evaluates both established automatic text recognition tools and multimodal LLMs for OCR tasks using *PressMint-AT* data.

3. Source Material

As source material for the *PressMint-AT* corpus the *Wiener Abendpost* was selected. It was a newspaper supplement of the *Wiener Zeitung*, published and printed by the *Wiener Zeitung* between 1st July 1863 and 31st Dezember 1921. The *Wiener Abendpost* was published 6 days a week (except Sundays) and had on average between 4 and 8 pages. It could be subscribed to separately or together with the Wiener Zeitung. In terms of content, the *Wiener Abendpost* contained a daily news section, a feuilleton section and advertisements. The newspaper had a three-column layout and was written in Fraktur typeface (see Figure 1). The *Wiener Abendpost* is accessible via ANNO (AustriaN Newspaper Online)[3], a virtual reading room for digitized newspapers maintained by the Austrian National Library. There, the scanned images, as well as full-text transcriptions are available, with all the limitations of automatic OCR, created some years ago without manual corrections (Resch & Kampkaspar, 2019). Since the *Wiener Abendpost* was digitized as a supplement of the *Wiener Zeitung*, there are no separate entries in ANNO, the *Wiener Abendpost* pages are just part of the *Wiener Zeitung*. Consequently, no dedicated metadata for the *Wiener Abendpost* is available. Therefore, a separate workflow to extract the *Wiener Abendpost* pages, needed to be set up (see section 4.2). For the *PressMint-AT* Corpus more than 17,450 newspaper issues will be processed containing more than 90,300 pages.


Wiener Abendpost.

1917. — Nr. 296. Beilage zur Wiener Zeitung. Freitag, 28. Dezember.

Figure 1: First page of the *Wiener Abendpost* issue 296 from 28. December 1917 (ANNO/Österreichische Nationalbibliothek) made available by the Austrian National Library via ANNO[4].

4. Data Preparation

4.1 Ground Truth

The ground truth data consist of four issues of the *Wiener Abendpost*, published between 1914 and 1918, totalling 20 pages. Given the limited variation in print quality and newspaper layout across the source period, issues were selected largely based on their historical significance, while allowing for the inclusion of some lower-quality pages (e.g. a line through the text as shown in Figure 1).

The ground truth was created using Transkribus (READ-COOP SCE, n.d.). Layout and text regions were manually annotated, followed by automatic transcription using Transkribus' *Text Titan I* text recognition model. There was no additional fine-tuning of models, neither for layout nor for text recognition. Lastly, extensive manual correction was applied, in accordance with the following guidelines: The transcription aims to reproduce the original print as faithfully as possible in terms of pages, lines, and characters. The texts have undergone multiple rounds of collation, have been carefully checked for quality, and reflect the historical state of the language without alteration. The original typography has largely been preserved; that is, *u* and *v* remain unchanged, as does the alternation between Fraktur and Antiqua typefaces, as well as the use of bold and italic print - except for the distinction between round s and long s, and for ligatures for which no Unicode representation exists. Abbreviations in the text have not been expanded. Original line-end hyphenation was preserved. The ground truth is thus aligned at the line level, with each text line transcribed individually within

[3] https://anno.onb.ac.at/node/15

[4] https://anno.onb.ac.at/cgi-content/annoshow?call=wrz|19171228|17|100.0|0

annotated text regions. While higher-level layout elements (e.g., titles, paragraphs, headers) were annotated, they were not considered in the computational evaluation. The ground truth data have been made available in the project's GitHub repository[5].

4.2 Identification of relevant pages

The *Wiener Abendpost* was published as a supplement to the *Wiener Zeitung,* which is accessible via ANNO, but the metadata on ANNO lacks information on the exact location of the *Wiener Abendpost* within each issue of the *Wiener Zeitung* (see Section 3). Therefore, it is necessary to provide a method for reliably spotting the *Wiener Abendpos*t by identifying the pages, which mark the starts and end of the supplement.

A first attempt consisted of identifying the *Wiener Abendpost* on a textual basis, i.e. by searching the transcriptions of the *Wiener Zeitung* provided by ANNO for stable indicators. This approach was abandoned because the quality of the existing OCR was deemed too low.

Instead, a visual approach using image recognition and classification was taken. For a human observer it is simple to "spot" the relevant supplement by viewing the thumbnail representation of a complete *Wiener Zeitung* issue. The first page of the *Wiener Abendpost* is indicated by its prominent title area (see Figure 1). The last page of the *Wiener Abendpost* is either indicated by the title page of another supplement called *Amtsblatt* (official gazette), or just by the end of the whole issue. This capability was emulated by training two classifier models: one for spotting the title page for the *Wiener Abendpost* and one for the *Amtsblatt*. Transfer learning was applied using a pretrained ResNet-18 model from PyTorch's torchvision library, fine-tuned for binary page classification by replacing the final layer with a single sigmoid output. Training used stratified splits, BCEWithLogitsLoss with optional class-weight balancing, and early stopping via an Adam optimizer with adaptive learning rate scheduling. To provide the necessary training samples for this supervised method, a simple browser plugin was created which allows to manually point out the title pages of the relevant supplements directly on ANNO's webpage (see Figure 2). The same GUI is used for visualising classifier predictions and enable immediate correction of faulty results. The overall process performed very satisfactorily. With an initial training set of 50 positive examples, the error rate was already below 10%. With the visualisation and correction GUI it took minimal effort to increase the amount of training data. The classifiers reached 0% error rate with only 500 positive samples.

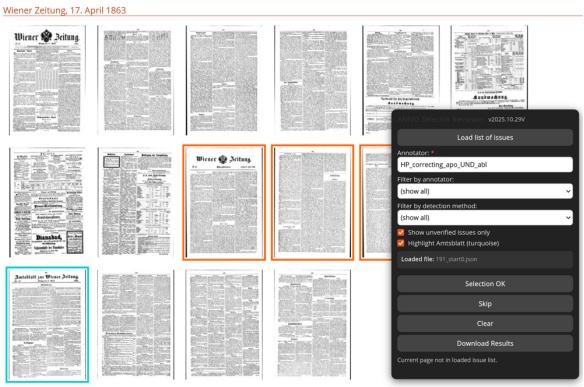

Figure 2: Screenshot of ANNO with a complete issue of the *Wiener Zeitung* and the *Wiener Abendpos*t Annotation plugin in place. Orange frames indicate the extent of the *Wiener Abendpos*t supplement, the turquoise frame depicts the title page of the *Amtsblatt* supplement.

5. Experiments

Different automatic text recognition tools, as well as several large language models (LLMs), were compared to evaluate their suitability for historical newspaper transcription. The systems evaluated include the OCR tools Google Cloud Vision OCR, Churro OCR, (Semnani et al., 2025), dots.ocr (Li et al, 2025), and PERO OCR (Kodym & Hradiš, 2021) using the German Fraktur modell as well as LLMs such as Anthropic's Claude Sonnet 4, DeepSeek, OpenAI's GPT-4o, and Google Gemini. A zero-shot procedure was employed, where the entire page image was provided as a single input. This approach maximises the context information but risks hallucination (Levchenko, 2025). The original, uncorrected Transkribus output obtained during ground truth creation (see Section 4.1), and the transcriptions provided by ANNO, the digital newspaper database of the Austrian National Library, were likewise included in the comparison.

Multiple prompts were tested, including system prompts, if available, (e.g. for dots.ocr or Churro OCR), and longer, hand-crafted prompts in both German (e.g. "Das ist ein Scan einer deutschen historischen Zeitung aus dem frühen 20. Jahrhundert. Bitte führe OCR darauf aus, also extrahiere den Text und behalte dabei die Leserichtung bei. Beachte auch, dass die Schrift in Fraktur gehalten ist. Der Output soll nur der Text alleine sein"), and in English ("This is a scan of a historic german newspaper from the early 20th century. Please do OCR on it, extract all the text and keep the reading order. Also keep in mind that the writing is in german 'Fraktur'. The output should only be text."), as the language of the prompt might have an influence on model performance (Kmainasi et al. 2025). All different

[5] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation/tree/main/data/texts/transkribus_corrected

prompts used are documented in the open GitHub repository[6].

For evaluation, standard OCR metrics were employed, namely Character Error Rate (CER) and Word Error Rate (WER), both of which are based on the Levenshtein distance. To complement these metrics, the similarity between continuous text sequences was assessed using the Ratcliff/Obershelp pattern recognition algorithm (Ratcliff and Metzener, 1988), here referred to as DIFFLIB. Quantitative evaluation was further supplemented by qualitative error analysis focusing on recurring challenges in Fraktur script.

SYSTEM	WER	CER	DIFFLIB
PERO	0.13	0.09	0.94
DOTS.OCR_1	0.38	0.11	0.91
DOTS.OCR_4	0.4	0.14	0.89
ANNO	0.35	0.17	0.88
TRANSKRIBUS	0.37	0.25	0.85

Table 1: Scores per workflow, ranked by mean performance across all three evaluation metrics. Highest score per metric is displayed in bold.

Results for the five highest-performing systems are presented in Table 1. All system–prompt combinations with a CER above 0.25 were excluded from further investigation. However, the complete set of results is documented in the GitHub repository[7]. The remaining scores include PERO OCR, dots.ocr with different system prompts (system prompt 1 and system prompt 4), the existing ANNO transcriptions, and the automatically generated Transkribus output. Across all three metrics, PERO OCR achieved the highest performance, followed closely by both dots.ocr systems, the existing transcriptions published in ANNO, and the automatically generated Transkribus output from the automatic text recognition feature (without manual corrections). All remaining systems underperformed relative to these results, exhibiting widely varying levels of error. Consistent with these findings, the four top-performing outputs showed low variance across all metrics, indicating stable and reliable performance at their respective levels.

Additional qualitative analysis of the data obtained through PERO OCR, dots.ocr, ANNO, and Transkribus further corroborated the observed pattern. In the case of Transkribus, the primary source of error consists of column crossings, i.e., instances in which the system fails to follow the correct reading order and instead merges lines from different columns into a single line. This issue was not observed in the other three highest-performing systems. In these cases, differences in performance are almost entirely attributable to word- and character-level OCR quality: PERO OCR produced the cleanest transcription; dots.ocr contained some misspellings typical of OCR for Fraktur texts, such as confusion between *f* and the Fraktur long *s*, or the faulty transcription of *tz* as *ß*, *b*, *z*, *ft*, *szt*, or *gt*; the transcriptions provided by ANNO contained numerous non-systematic misspellings, often involving non-alphabetic characters (e.g., #, », «, ‘), rendering some words unrecognizable.

In conclusion, its superior performance, together with its computational efficiency, renders PERO OCR in combination with the German Fraktur model the most suitable system for the task at hand; it was therefore selected for further experimentation.

6. Future Work

The next step will involve generating automatic full-text transcriptions for all *Wiener Abendpost* issues using PERO OCR, followed by TEI encoding in accordance with the *PressMint* schema and subsequent data processing (e.g., POS tagging and semantic enrichment) as outlined in the *PressMint* project work plan.

7. Conclusion

In this paper we described the *PressMint-AT* project, the Austrian subproject within *PressMint,* that aims at creating an Austrian historial newaspaper corpus for the *Wiener Abendpost*. The submission discusses and evaluates different approaches for creating full-text transcription via ORC, including methods based on LLMs. We compared several OCR tools such as Google Cloud Vision OCR, Churro OCR, dots.ocr, PERO OCR, Transkribus (without manual corrections) as well as generic LLMs such as Anthropic’s Claude Sonnet 4, DeepSeek, OpenAI’s GPT-4o, and Google Gemini alongside the already existing transcriptions provided by ANNO. Among these approaches, PERO OCR achieved the best results for our dataset. Considering both transcription quality and computational efficiency, PERO OCR appears to be the most suitable system for the OCR task within the *PressMint-AT* project.

8. Acknowledgments

The submission was supported by the PressMint CLARIN Flagship Project and CLARIAH-AT.

[6] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation

[7] https://github.com/acdh-oeaw/pressmint-OCR-AI-evaluation/tree/main?tab=readme-ov-file#comparison-results-plot

9. Bibliographical References

Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S. and Kaplan, F. (2024). Postcorrection 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. https://doi.org/10.18653/v1/2024.latechclfl-1.14.

Crosilla, G., Klic, L. and Colavizza, G. (2025). Benchmarking large language models for handwritten text recognition. https://arxiv.org/pdf/2503.15195.

Ehrmann, M, Bunout, E. and Clavert, F. (2023). "Digitised Historical Newspapers: A Changing Research Landscape". Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology, edited by Estelle Bunout, Maud Ehrmann and Frédéric Clavert, Berlin, Boston: De Gruyter Oldenbourg, 1-22. https://doi.org/10.1515/9783110729214-001.

Fišer, D. and Lenardič, J. (2018). CLARIN Resources Families / Newspaper Corpora. https://www.clarin.eu/resource-families/newspaper-corpora.

Humphries, M., Leddy, L. C., Downton, Q., Legace, M., McConnell, J., Murray, I. and Spence, E. (2024). Unlocking the archives: Using large language models to transcribe handwritten historical documents.

Kmainasi, M.B., Khan, R., Shahroor, A.E., Bendou, B., Hasanain, M. and Alam, F. (2025). Native vs Non-native Language Prompting: A Comparative Analysis. In: Barhamgi, M., Wang, H., Wang, X. (eds) Web Information Systems Engineering – WISE 2024. WISE 2024. Lecture Notes in Computer Science, vol 15440. Singapore: Springer. https://doi.org/10.1007/978-981-96-0576-7_30.

Kim, S., Baudru, J., Ryckbosch, W., Bersini, H. and Ginis, V. (2025). Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records. https://arxiv.org/abs/2501.11623.

Li, L. (2024). Handwriting Recognition in Historical Documents with Multimodal LLM. In CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark https://arxiv.org/html/2410.24034v1.

Li, Y., Yang, G., Liu, H., Wang, B. and Zhang, C. (2025). *dots.ocr: Multilingual document layout parsing in a single vision-language model*. arXiv. https://arxiv.org/abs/2512.02498.

Mayer, S. (2024). Reviving History: Reconstructing Esperanto Newspaper Excerpts from the Hachette Collection (1898-1915). ESF Connected https://esfconnected.org/2024/10/14/reviving-history-reconstructing-esperanto-newspaper-excerpts-from-the-hachette-collection-1898-1915/.

Ratcliff, John W. and Metzener, D. (1988). "Pattern Matching: The Gestalt Approach". Dr. Dobb's Journal (46).

READ-COOP SCE. (n.d.) Transkribus. Innsbruck: READ-COOP SCE, [30.07.2025]. https://transkribus.eu/.

Resch, C. (2023). Volltextoptimierung für die historische Wiener Zeitung Mit einem Anwendungsszenario aus der germanistischen Sprachgeschichte. In Bunout, Ehrmann, M., Clavert, F. (Eds.), Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology. Berlin, Boston: De Gruyter, 89-111. https://doi.org/10.1515/9783110729214.

Resch, C. and Kampkaspar D. (2019). DIGITARIUM – Unlocking the Treasure Trove of 18th-Century Newspapers for Digital Times. In: Wallnig, T, Romberg M. and Weis J. (Eds.): Digital Eighteenth Century: Central European Perspectives. Wien / Köln / Weimar: Böhlau Verlag, 49-64.

Semnani, S., Han Zhang, H., Xinyan He, X., Tekgurler, M. and Lam, M. (2025). CHURRO: Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34777–34824, Suzhou, China. Association for Computational Linguistics.

Kodym, O. and Hradiš, M. (2021). Page Layout Analysis System for Unconstrained Historic Documents. ICDAR, https://arxiv.org/abs/2102.11838.

Greif, G., Griesshaber, N. and Greif, R. (2025). Multimodal LLMs for OCR, OCR Post-Correction, and Named Entity Recognition in Historical Documents. https://arxiv.org/abs/2504.00414

Venglarova, K., Adam, R., Mölzer, W., Balasubramanian, S., Kleinert, J., Füllsack, M. and Vogeler, G. (2024). Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads. In: Proceedings of the Computational Humanities Research Conference 2024. Aarhus, Denmark. CEUR. 2024. 788-801.

10. Language Resource References

Couranten Corpus (version 2.0) (2025) [Online Service]. Available at the Dutch Language Institute: https://hdl.handle.net/10032/tm-a3-c2.

Dobranić, F., Evkovski, B. and Ljubešić, N. (2024). A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection. LREC-COLING 2024.

Resch, C. and Kampkaspar, D. (Eds) (2020). Wienerisches DIGITARIUM - Digitale Ausgabe des "Wien[n]erischen Diarium" (322 Ausgaben im Volltext).
Resch, C. and Kampkaspar, D. (Eds) (2022). German Fraktur 18th Century – WrDiarium_M9. https://www.transkribus.org/models/german-fraktur-18th-century.
Weil, S. (2021). Tesseract OCR models for historic prints based on Latin script. Zenodo. https://doi.org/10.5281/zenodo.10125246.