From Manuscript to Model: Developing HTR for Medieval Greek

Nicklas Sindlev Andersen[1], **Byron MacDougall**[1], **Tariq Yousef**[1], **Aglae Pizzone**[1]
[1] University of Southern Denmark, Odense, Denmark
sindlev@imada.sdu.dk, bmacdoug@risd.edu, yousef@imada.sdu.dk, pizzone@sdu.dk

Abstract

We develop and evaluate manuscript-specific text line detection (TLD) and handwritten text recognition (HTR) models for two 14th-century Medieval Greek manuscripts, *Vat. gr. 2228* and *Phil. gr. 130*, comprising 1,356 handwritten pages. From these, we curate and document 36 pages with complete, manually curated text line annotations, together with 10 additional pages with layout annotations only for TLD, forming two manuscript-specific ground truth (GT) datasets. To ensure representative evaluation despite limited annotations, validation splits are optimized for character coverage and distributional similarity using Jensen-Shannon divergence. Using the Transkribus platform, we train manuscript-specific TLD models from scratch and manuscript-specific HTR models, comparing HTR training from scratch with fine-tuning of a publicly available Medieval Greek base model. TLD achieves validation pixel-wise misclassification rates of 5.42% for Vat. gr. 2228 and 8.76% for the more layout-variable Phil. gr. 130. For HTR, fine-tuning consistently outperforms training from scratch. On validation pages with manually curated text line annotations, Vat. gr. 2228 reaches 5.13% character error rate (CER) and 23.66% word error rate (WER), while Phil. gr. 130 reaches 27.13% CER and 65.72% WER after continued training. A supplementary held-out evaluation on Vat. gr. 2228 shows that the fine-tuned model reaches 5.97% CER and 23.52% WER on test pages with manually corrected line polygons, degrading to 12.12% CER and 44.21% WER under automatic TLD-based segmentation. The study also provides a reproducible workflow and evaluation protocol for Medieval Greek HTR under low-resource conditions.

Keywords: Medieval Greek, Handwritten Text Recognition, HTR, Cultural Heritage Digitization, Datasets

1. Introduction

The digitization and transcription of historical documents have become increasingly important for both preserving cultural heritage and enabling new research. By converting physical materials into digital formats, these efforts not only safeguard unique historical sources but also broaden access. Manual transcription, however, is a time-consuming and highly specialized task, making automated approaches, particularly handwritten text recognition (HTR), increasingly valuable.

Medieval Greek manuscripts, in particular those featuring minuscule scripts penned by informal scholarly hands, present a unique set of difficulties. On top of the rich system of diacritics characterizing ancient and medieval Greek, those scripts, designed for rapid execution, present a high degree of cursivity, frequent ligatures, and extensive use of tachygraphic symbols and abbreviations. They also exhibit considerable variation in scribal execution and, depending on the textual typology, complex layouts. Such features further increase transcription complexity. While HTR systems have advanced considerably in recent years, progress has been uneven across languages and scripts. Compared to Latin-based languages, non-Latin scripts such as Medieval Greek still lag behind in terms of available datasets, models, and tools, often requiring researchers to develop resources in relative isolation.

Recently, however, the situation has begun to change. Collaborative communities have started to emerge, bringing together datasets, annotation guidelines, and trained models for Medieval Greek transcription (see, e.g., Table 1) (Verstraete et al., 2025). At the same time, user-friendly platforms such as Transkribus (READ-COOP SCE, 2025) have democratized access to advanced machine learning techniques, enabling broader participation in the transcription process. These developments are transforming archives that were previously hardly accessible to the non-specialist into analyzable data and streamlining research workflows.

In this work, we (1) curate and document two Medieval Greek ground truth (GT) datasets consisting of annotated manuscript pages with line-level text line annotations (used here as a catch-all term for baselines, enclosing line polygons, and transcriptions), covering pages from *Vat. gr. 2228*[1] and *Phil. gr. 130*[2]; (2) develop and evaluate manuscript-specific text line detection (TLD) models trained from scratch and manuscript-specific HTR models in Transkribus, comparing HTR training from scratch with fine-tuning of a publicly available Medieval Greek base model (including continued training where applicable), and reporting Transkribus validation metrics for TLD and externally computed recognition scores, namely character error rate (CER) and word error rate (WER), for HTR; and

[1] Vatican Library: Vat. gr. 2228, part 1 + part 2
[2] Austrian National Library: Phil. gr. 130

Proceedings of LT4HALA 2026 @ LREC 2026, pages 139–151.
11 May 2026.

Table 1: Datasets and models for Ancient and Medieval Greek HTR, including the two datasets presented in this paper (Vat. gr. 2228 and Phil. gr. 130). In the table, D = dataset, and M = model. For Vat. gr. 2228, numbers in parentheses indicate additional pages and lines annotated only for TLD, not transcription.

Name	Type	Period	Pages	Lines	Characters (total / unique)
Zenon Papyri (Marthot-Santaniello and Hodel, 2022)	D	3rd BCE	27	321	6,059 / 47
Méléagre (Guénette et al., 2024)	D + M	10th CE	70	3374	108,231 / 108
Chrysostomicus I (Perdiki, 2022)	M	10-14th CE	–	–	– / –
Stavronikita (no. 53) (Pratikakis et al., 2021b)	D	14th CE	54	1038	63,228 / 147
Stavronikita (no. 114) (Pratikakis et al., 2021a)	D	15th CE	44	1006	66,086 / 159
Eparchos (Papazoglou et al., 2020)	D	16th CE	120	2272	203,338 / 156
Stavronikita (no. 79) (Pratikakis et al., 2021c)	D	16th CE	40	803	50,422 / 133
HPGTR (Platanou et al., 2024)	D	8-16th CE	70	1698	66,739 / 159
Vat. gr. 2228	D	14th CE	18 (+10)	731 (+439)	55,637 / 144
Phil. gr. 130	D	14th CE	18	847	123,908 / 143

(3) systematically document the end-to-end workflow from GT preparation to evaluation to support reproducibility and comparative analysis in future studies.

2. Related Work

2.1. HTR Platforms

HTR platforms integrate the core stages of the HTR workflow, typically combining layout analysis, often focusing on text line detection (TLD), with handwritten text recognition. Many platforms also support model training and fine-tuning through graphical or web-based interfaces (Boros et al., 2024).

These systems differ primarily in their trade-off between usability and experimental control. Hosted, closed platforms reduce setup overhead but may limit automation, transparency, and reproducibility. In contrast, open-source systems with application programming interface (API) access enable customizable and fully automated pipelines, at the cost of local infrastructure requirements and technical effort.

In this work, we use Transkribus (READ-COOP SCE, 2025), a widely adopted hosted platform whose HTR stack is based on the open-source PyLaia engine (Teklia, 2022; Tarride et al., 2024). While Transkribus facilitates rapid iteration and collaborative annotation, its closed environment constrains hyperparameter control and full reproducibility. We therefore account for these limitations in our evaluation protocol by performing repeated training runs and carefully documenting configuration settings.

Open-source alternatives such as eScriptorium (Kraken) (eScriptorium, 2025; Kiessling et al., 2019), Arkindex (PyLaia) (Teklia, 2025), and OCR4all (Centre for Philology and Digitality, 2025) provide greater transparency and automation, but typically require institutional hosting or dedicated compute resources for training at scale.

Together, these platforms form the technical foundation upon which most recent work in Ancient and Medieval Greek HTR is built. Understanding their respective strengths and limitations helps contextualize the methodological choices in the current study.

2.2. Ancient & Medieval Greek HTR

Recent efforts in Ancient and Medieval Greek HTR have produced a range of datasets and trained models across different platforms and architectural approaches.

Platanou et al. (2024) introduced the HPGTR dataset, constructed from digitized images of Bodleian Library Greek manuscripts dating from the 10th to the 16th century, and trained recognition models using Transkribus to assess HTR performance. Based on their findings, they argue that paleographic style and manuscript dating strongly influence recognition performance.

In particular, they correlated variation to chronology: manuscripts from the 10th-13th centuries achieved lower character error rates, whereas later manuscripts (14th-16th CE) proved more challenging due to an increasingly cursive style, ligatures, and stylistic complexity. These findings seem to underscore the importance of accounting for temporal and stylistic variation when developing manuscript-specific HTR models, although the question remains whether variations due to genres and uses might not be equally, if not more important.

Addressing scalability, Perdiki (2023) investigated how to minimize training requirements for large collections. Focusing on John Chrysostom's manuscript tradition, they showed that usable Transkribus models can be trained with as little as 1,000 words of GT. Their results highlight the value of careful manuscript selection and transfer learning strategies, particularly when annotated data are limited.

At the level of individual manuscripts, Guénette et al. (2024) studied Codex Palatinus Graecus 23 using eScriptorium. They trained recognition models both from scratch and via fine-tuning (initializing from the CREMMA Medieval Latin model (Pinche, 2022)), achieving approximately 90-91% accuracy. Their work demonstrates both the feasibility of

cross-script transfer learning and the importance of capturing scribal diversity within a manuscript.

From a methodological perspective, Markou et al. (2021) proposed a CRNN-FCNN architecture and introduced the EPARCHOS dataset (16th CE), which captures key challenges of Greek manuscripts, including abbreviations, floating characters, and polytonic orthography. Their ablation study showed that data augmentation and regularization significantly improved recognition performance. Building on convolutional-recurrent architectures, Tsochatzidis et al. (2021) introduced an OctCNN-BGRU model and released additional benchmark datasets from the Stavronikita Monastery (nos. 53, 79, and 114).

Extending beyond medieval manuscripts, and moving backward in time, the Zenon Papyri dataset developed by Marthot-Santaniello and Hodel (2022) within Transkribus as part of the D-Scribes project (D-Scribes Project, 2023) represents a key contribution to Ancient Greek HTR. As one of the few publicly available resources for Greek scripts preserved on papyri, it provides a foundation for the study of significantly earlier material.

Taken together, these studies highlight three recurring observations: (i) recognition performance is strongly influenced by paleographic and layout variation; (ii) transfer learning and fine-tuning consistently outperform training from scratch, particularly when annotated data are limited; and (iii) carefully curated, manuscript-specific GT subsets can be sufficient to train usable HTR models.

Our work builds directly on these observations by adopting a manuscript-specific modeling strategy, comparing training from scratch with fine-tuning of a publicly available Medieval Greek base model, and relying on carefully curated but relatively small annotated subsets.

3. Data

3.1. Source Manuscripts

We develop workflows and models for an overall corpus of approximately 1,356 handwritten pages drawn from two 14th-century Medieval Greek manuscripts: Vatican, Biblioteca Apostolica Vaticana, Vat. gr. 2228 and Vienna, Österreichische Nationalbibliothek, Phil. gr. 130 (see Figure 1). Within the broader manuscript project motivating this study, however, the current phase of the HTR work reported here focuses on specific folio ranges within these manuscripts.

Vat. gr. 2228 is a paper manuscript measuring 231×156 mm and consisting of 508 folia (1,016 pages), currently divided into two volumes. Each folio, with the exception of folia 1r-v and 16r-v, contains 35 to 45 lines of writing on a writing surface of mostly 174×113.118 mm or 166×114.115 mm, according to the full description provided by Salvator Lilla in the catalog of the Greek Vatican manuscripts (Lilla, 1985, pp. 307–313). The layout follows the ruling cataloged by Leroy as 20 D 1 (Leroy, 1976, p. 6) (see Figure 2 for the ruling pattern). The manuscript, copied by nine different hands, offers a selection of rhetorical treatises featuring prominently the Graeco-Roman handbooks authored by Hermogenes of Tarsus (2nd-3rd century CE) and his medieval exegeses. For the current phase, our primary focus is folia 194–313, copied by Hand 2 (according to Lilla's description) and containing the 11th-century commentary by John Doxapatres on Hermogenes' treatise *On Invention*, a commentary that remains unpublished to date. At the same time, the availability of published sections from an earlier portion of the manuscript makes it possible to construct additional GT for model development before targeting this main range of interest.

Phil. gr. 130 is also a paper manuscript dated to the first half of the 14th century. It consists of 170 folia (340 pages) measuring 240/245×155/160 mm. Each folio contains 38 to 56 or 55 to 62 lines of writing, depending on the dimensions of the folio. The manuscript is described by Herbert Hunger (Hunger, 1961, pp. 238–239) in the catalogue of the Greek manuscripts preserved by the Austrian National Library. Like the Vatican manuscript, the Viennese manuscript contains Graeco-Roman rhetorical treatises accompanied by medieval Byzantine commentaries. For the purpose of this study, we focused on folia 85v-170v, containing Hermogenes' treatise *On Issues*, with the exegesis of John Doxapatres. The commentary material is organized in a frame layout on the entirety of the written surface, whereby the commentary forms an open frame around small blocks of main text (Maniaci, 2022).

The particular reproductions of Vat. gr. 2228 we have been annotating are high-resolution grayscale images obtained from the Vatican Library, each measuring 2174×2996 px, while for Phil. gr. 130, the corresponding reproductions consist of high-resolution color images, ranging from 4032-4179 px in width and 6040-6086 px in height.

The broader underlying aim of this work is to facilitate access to the currently unpublished commentary of John Doxapatres, relevant to the history of medieval rhetoric. The material considered here reflects variation in textual density, layout complexity, and handwriting. In particular, Phil. gr. 130 exhibits greater variability in both layout and handwriting, whereas Vat. gr. 2228 is comparatively consistent and visually regular. This difference between the two manuscripts, despite their closeness in time, is reflected in model performance, as discussed in Section 5.

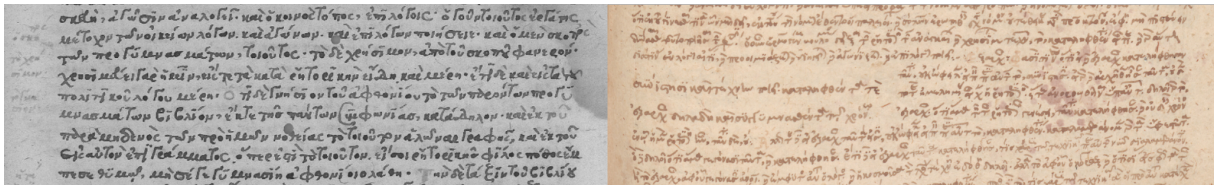

Figure 1: Excerpts from Vat. gr. 2228 (left) and Phil. gr. 130 (right). The image from Vat. gr. 2228 is generally representative due to the manuscript's consistency, while Phil. gr. 130 exhibits more variability.

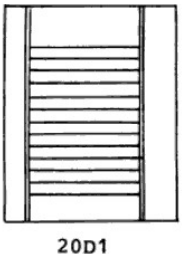

Figure 2: The ruling pattern of Vat. gr. 2228.

3.2. The Ground Truth & Splits

The two GT datasets consist of manuscript pages with line-level text line annotations.

Pages selected for annotation were chosen based on editorial needs (for Phil. gr. 130) and the availability of published sections of the manuscript text (for Vat. gr. 2228).

The GT datasets are distributed as PAGE-XML and split-definition files[3] with no manuscript images, since image rights remain with the holding institutions. In total, the GT release contains 46 PAGE-XML files: 18 for Phil. gr. 130 and 28 for Vat. gr. 2228.

The GT for Vat. gr. 2228 covers folia 12r–20v and 194r–199v. Of these, 18 unique pages are fully annotated at the line level. The GT was established by comparing folia 12r–20v with the existing edition by Christian Walz (Walz, 1835) and folia 194r-v by applying the workflow described in Section 4.1. An additional 10 pages (folia 195r–199v) contain layout annotations only (baselines and enclosing line polygons) and are used exclusively for TLD, contributing 439 lines but no transcriptions (these are shown in parentheses in Table 1).

For Phil. gr. 130, the GT covers folia 86r–94v, all fully annotated from scratch at the line level.

Because annotated data are limited, we partition the pages used for model development into a training set (≈ 90%) and a validation set (≈ 10%). For Phil. gr. 130, this corresponds to 16 training pages and 2 validation pages; for Vat. gr. 2228, 14 training pages and 2 validation pages are used, with 2 additional pages held out for test evaluation. For Vat. gr. 2228, these held-out test pages became available only in a later iteration of the workflow described in Section 4.1. They were therefore excluded when selecting the training/validation split. Accordingly, all splits are performed at the page level to prevent leakage across partitions.

All transcriptions are normalized prior to defining the character vocabulary used for training and evaluation. Normalization removes zero-width characters, maps Unicode whitespace to a single ASCII space, strips leading and trailing whitespace, and collapses consecutive spaces. In this context, let $\mathcal{C}$ denote the resulting normalized character vocabulary of a manuscript corpus, i.e., the set of characters occurring in the GT transcriptions.

Given the small number of annotated pages, evaluation may vary depending on the specific pages selected for validation. To mitigate this effect and promote representativeness, validation pages are selected according to two complementary criteria:

Character coverage. The validation set should contain as many unique characters present in the corpus as possible. For a subset S of pages, let $\mathcal{C}(S) \subseteq \mathcal{C}$ denote the set of characters occurring in S. We define the coverage of S as

$$\text{Coverage}(S) = \frac{|\mathcal{C}(S)|}{|\mathcal{C}|}.$$

A value of 1 indicates that all character types in $\mathcal{C}$ are represented in the validation set.

Distributional similarity. Beyond presence or absence, the relative character frequencies in the validation set should approximate those of the full corpus to avoid biased evaluation. Let P and Q denote the empirical character distributions over $\mathcal{C}$ for the subset and the full corpus, respectively. We then measure similarity using the symmetric Jensen-Shannon divergence:

$$D_{\text{JS}}(P \,\|\, Q) = \tfrac{1}{2} D_{\text{KL}}(P \,\|\, M) + \tfrac{1}{2} D_{\text{KL}}(Q \,\|\, M),$$

where D_{KL} is the Kullback-Leibler divergence (measuring the difference between two probability distributions) and $M = \frac{1}{2}(P + Q)$ is their pointwise average. Smaller D_{JS} values indicate closer agreement between the character frequency distributions of the validation set and the full corpus.

[3] https://doi.org/10.5281/zenodo.19425687.

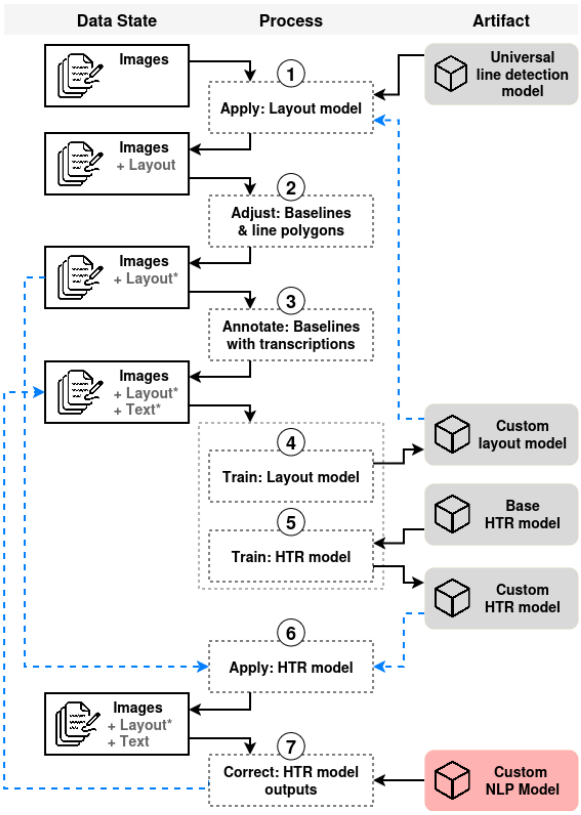

Figure 3: The workflow applied to Vat. gr. 2228 and Phil. gr. 130. Left column: an evolving GT dataset. Middle column: the processing steps. Right column: the resulting artifacts (applied or newly trained models). In the diagram, +Layout/+Layout* denote unadjusted/adjusted baselines and line polygons, and +Text/+Text* unadjusted/adjusted transcriptions. Blue dashed arrows indicate the iterative refinement cycle, where adjusted outputs are reintegrated into a GT dataset and models retrained.

Validation sets are selected using a two-step procedure. We first identify page combinations that maximize character coverage; among those, we select the subset minimizing D_{JS}. This approach yields validation sets that include rare characters while maintaining a character frequency distribution closely aligned with the overall corpus.

4. Workflow & Methods

4.1. Workflow in Transkribus

The workflow in Transkribus is illustrated in Figure 3. We initially employed the workflow consisting of steps ①-⑤, comprising data preparation and model development. The process began with uploading digitized manuscript images to the Transkribus web interface. A critical early step was layout analysis, here focusing specifically on TLD ①, which provided the structural basis for subsequent HTR processing.

For initial TLD, we applied the Transkribus *Universal Lines* model, which predicts baselines and corresponding line polygons of approximately rectangular shape. While generally effective, the automatically generated line polygons occasionally truncated characters, as letters in both manuscripts frequently extended beyond predicted boundaries. We therefore manually adjusted these annotations ②, producing high-quality layout annotations suitable for TLD training and for subsequent pairing with transcriptions in HTR training.

Expert transcriptions were then created on a per-line basis, pairing each annotated text line with its corresponding machine-readable transcription ③. This pairing of precise segmentation and diplomatically accurate transcription formed the foundation of the manuscript-specific GT.

GT preparation was carried out collaboratively by a domain expert in Medieval Greek and a data scientist. The domain expert ensured accurate transcription, including diacritical marks, resolutions of tachygraphic signs, ligatures, floating characters, and other paleographic features, while refining the associated layout annotations. The data scientist oversaw model application, dataset organization, and quality control within Transkribus in preparation for model training ④-⑤. All annotated pages were cross-checked to ensure transcription accuracy and annotation consistency.

In addition to steps ①-⑤, we employed steps ⑥ and ⑦, applying trained TLD and HTR models to previously unannotated pages. Together, steps ①-⑦ form an iterative learning cycle (illustrated in Figure 3 by blue dashed arrows), particularly for Vat. gr. 2228. Automatically segmented and transcribed pages were manually corrected and incorporated into the respective manuscript-specific GT dataset, after which updated models were retrained. This iterative refinement was applied independently to each manuscript, progressively improving TLD and recognition quality while expanding the annotated material, until converging on final manuscript-specific GT datasets (Section 3.2).

4.2. TLD Model Training

In parallel with HTR model training ⑤, we trained manuscript-specific TLD models ④ for each manuscript.

TLD models were trained using all default settings in Transkribus and using the same training and validation partitions as the HTR models (Section 3.2). Their primary purpose was to enable automated segmentation of the remaining unannotated manuscript pages.

Accurate TLD is a prerequisite for reliable HTR performance, as misplaced or truncated line polygons result in incomplete character capture and

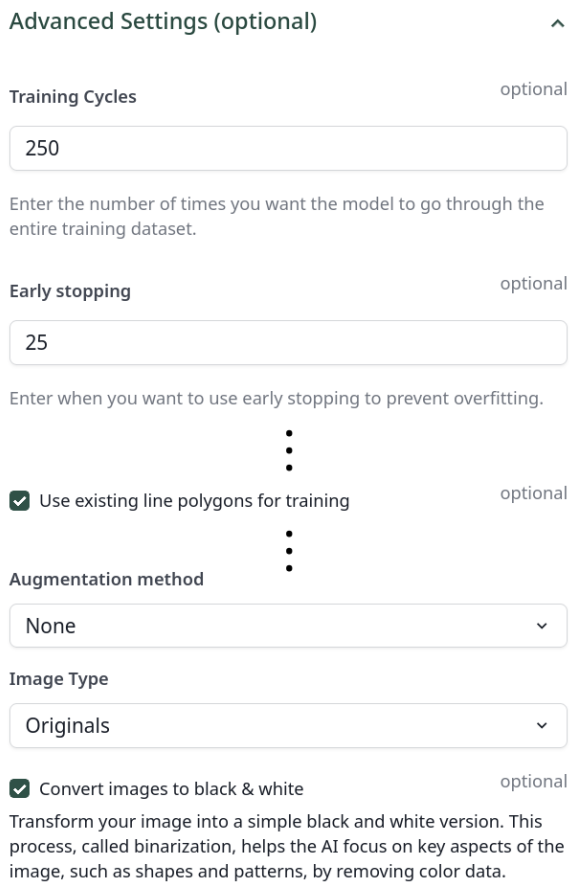

Figure 4: Screenshot of Transkribus HTR "Advanced Settings" used in the various experiments.

subsequent recognition errors (Fizaine et al., 2024). Manuscript-specific training therefore aimed to reduce systematic segmentation artifacts introduced by generic models (Universal Lines) available in Transkribus.

4.3. HTR Model Training

HTR model training in Transkribus ⑤ was carried out using the PyLaia engine, which implements a convolutional-recurrent neural network (CRNN) architecture. The model consists of a convolutional neural network (CNN) backbone for spatial feature extraction, followed by stacked bidirectional long short-term memory (BiLSTM) layers for sequence modeling. Recognition is optimized using the connectionist temporal classification (CTC) loss function, which resolves alignment between image features and character sequences without requiring pre-segmented character boundaries.

For each manuscript, we compared two training strategies: (i) training from scratch using only the manuscript-specific GT, and (ii) fine-tuning a publicly available Medieval Greek base model (transfer learning) (Perdiki, 2022). Where applicable, we further explored continued training to assess any further possible performance gains.

Unless otherwise noted, each training strategy was configured for up to 250 epochs with early stopping after 25 epochs without improvement on the validation set (see Figure 4). Advanced settings included using existing layout annotations during training and converting images to black and white. By binarizing both manuscripts, we aimed to enforce a consistent starting point for model training and evaluation, avoiding differences in color information as an additional source of variation across pipelines. Validation followed the page-level 90/10 split described in Section 3.2, ensuring separation between training and validation pages.

As additional corrected pages were incorporated into the GT datasets, models were retrained to reflect the expanded and refined training data. The results reported in Section 5 correspond to a final training phase conducted after this iterative refinement process had converged on the current GT datasets, described in Section 3.2.

4.4. Evaluation Protocol & Error Metrics

Because Transkribus training is non-deterministic, each training strategy was repeated three times. Evaluation results reported in Tables 2 and 3 therefore correspond to mean and standard deviation across three independent runs, training a model on the training set and evaluating it on the validation set.

TLD Evaluation. TLD performance is reported as the pixel-wise misclassification rate provided by Transkribus and retrieved directly from the platform's web interface. Lower values indicate better segmentation performance[4].

HTR Evaluation. Evaluation was performed externally to Transkribus to ensure reproducible scoring and to enable future, more extensive, line- and character-level error analysis. GT and predicted outputs were exported as PAGE-XML and aligned one-to-one at the text-line level. For all evaluated pages, the number of GT and predicted lines matched exactly.

Let g_i and h_i denote the normalized GT and predicted strings for line $i = 1, \ldots, N$. Before scoring, GT and predictions were normalized as described in Section 3.2, ensuring consistency with the vocabulary construction and dataset splits.

Recognition performance was quantified using character error rate (CER) and word error rate (WER), computed as micro-averaged Levenshtein

[4] https://help.transkribus.org/baselines-models

Table 2: TLD performance reported as a pixel-wise misclassification rate. Results are reported as mean $\pm$ std over three training runs.

Manuscript	Miscl. Rate	
	Train	Val
Vat. gr. 2228	6.36% $\pm$ 0.07%	5.42% $\pm$ 0.21%
Phil. gr. 130	4.86% $\pm$ 0.01%	8.76% $\pm$ 0.30%

distance with unit insertion, deletion, and substitution costs (no transpositions):

$$\mathrm{CER} = \frac{\sum_{i=1}^{N} d_{\mathrm{Lev}}(g_i, h_i)}{\sum_{i=1}^{N} |g_i|},$$
$$\mathrm{WER} = \frac{\sum_{i=1}^{N} d_{\mathrm{Lev}}(\tau(g_i), \tau(h_i))}{\sum_{i=1}^{N} |\tau(g_i)|}.$$

Here, $|g_i|$ denotes the character length of g_i, $\tau(\cdot)$ denotes whitespace tokenization, and micro-averaging means that edit operations are aggregated across all lines before normalization by total reference length.

In addition to the validation-based comparison used for model development, we report a supplementary held-out evaluation on unseen Vat. gr. 2228 test pages. We restrict this additional analysis to Vat. gr. 2228 because held-out test pages became available for this manuscript through an additional iteration of the workflow described in Section 4.1, and recognition quality had reached a level that made assessment under practical segmentation conditions meaningful. The same three trained HTR models per strategy were reused. Each model was then evaluated under two segmentation settings: manually corrected line polygons, representing a near-best-case recognition condition, and automatically detected line polygons produced by the manuscript-specific TLD model with the lowest validation misclassification rate, representing a more realistic end-to-end condition. Scoring followed exactly the same external PAGE-XML protocol as for validation.

5. Results

5.1. TLD Performance

Table 2 reports overall TLD results, while per-run results are provided in the Appendix (Section 11).

For Vat. gr. 2228, training and validation misclassification rates are comparable (6.36% vs. 5.42%), suggesting similar performance across the selected training and validation pages. This pattern is consistent with the manuscript's comparatively regular layout and homogeneous line structure.

In contrast, Phil. gr. 130 shows a larger gap between training and validation error (4.86% vs. 8.76%), suggesting reduced generalization. This behavior aligns with the manuscript's greater layout variability and scribal complexity.

On qualitative inspection of unseen pages, the manuscript-specific models generally capture the main body of text lines in both manuscripts. Residual errors appear primarily boundary-related, including partial exclusion of diacritics and occasional overlap with adjacent lines. Ultimately, automatically generated layout annotations remain below the quality of the manual layout annotations.

5.2. HTR Performance

Table 3 reports the overall results, while per-run results can be found in the Appendix (Section 11).

For Vat. gr. 2228, fine-tuning the publicly available Medieval Greek base model yields the best validation performance (5.13% CER, 23.66% WER), improving over training from scratch (6.23% CER, 28.26% WER). Training error remains low overall, though the from-scratch runs show noticeably larger variance than the fine-tuned runs.

Recognition of Phil. gr. 130 is substantially more challenging. Training from scratch results in 38.20% CER and 77.95% WER. Fine-tuning reduces validation CER to 28.62% (WER: 68.18%), and continued training further lowers it to 27.13% (WER: 65.72%). Although overall error remains high, transfer learning consistently improves performance, with the largest absolute gains observed for the more complex manuscript, indicating that a pretrained model can provide a beneficial initialization even under paleographic and structural variation.

Table 4 extends the validation results with a supplementary held-out evaluation for Vat. gr. 2228 under two segmentation conditions. On the held-out test pages with manually corrected line polygons, the fine-tuned base model achieved 5.97% CER and 23.52% WER, compared with 8.77% CER and 31.14% WER for training from scratch. Under automatically detected line polygons, recognition performance dropped to 12.12% CER and 44.21% WER for the base model, and to 18.47% CER and 54.81% WER for training from scratch. The relative ranking observed on validation is therefore preserved on unseen material, while the end-to-end setting makes explicit the expected propagation of segmentation errors into downstream HTR.

6. Discussion & Limitations

Because annotated material is limited, we use a fixed page-level split optimized for character coverage and distributional similarity. While this ensures textual representativeness, it does not explicitly account for visual variability. Reported validation scores should therefore be interpreted as estimates

Table 3: HTR performance reported as micro-averaged CER/WER. Mean ± std over three training runs.

Manuscript	Strategy	CER		WER	
		Train	Val	Train	Val
Vat. gr. 2228	From Scratch	0.68% ± 0.55%	6.23% ± 0.22%	3.91% ± 2.84%	28.26% ± 1.41%
	Base Model	0.02% ± 0.01%	5.13% ± 0.09%	0.15% ± 0.07%	23.66% ± 0.60%
Phil. gr. 130	From Scratch	37.42% ± 3.45%	38.20% ± 1.98%	66.80% ± 4.01%	77.95% ± 2.78%
	Base Model	32.86% ± 0.53%	28.62% ± 0.33%	57.28% ± 0.41%	68.18% ± 0.77%
	Base Model (Cont. Train.)	27.95% ± 0.05%	27.13% ± 0.43%	47.69% ± 0.13%	65.72% ± 1.64%

Table 4: Supplementary held-out evaluation of HTR performance for Vat. gr. 2228 on two test settings: manually corrected line polygons and automatically detected line polygons from the manuscript-specific TLD model with the lowest validation misclassification rate. Results are reported as micro-averaged CER/WER, mean ± std over three training runs.

Manuscript	Strategy	CER		WER	
		Manual	Automatic (TLD)	Manual	Automatic (TLD)
Vat. gr. 2228	From Scratch	8.77% ± 0.45%	18.47% ± 0.83%	31.14% ± 0.47%	54.81% ± 0.29%
	Base Model	5.97% ± 0.16%	12.12% ± 0.64%	23.52% ± 0.68%	44.21% ± 2.50%

under this controlled setting rather than guarantees for all pages within each manuscript.

A second limitation concerns experimental control in Transkribus. Model training is non-deterministic because random seeds cannot be fixed. Although we repeat each strategy three times and report mean and standard deviation, run-to-run variance remains. In addition, a single fixed validation split increases sensitivity to page selection and does not capture variability that cross-validation would reveal.

Finally, most of the recognition results are computed on validation pages with manually curated text line annotations. We supplement these with a held-out evaluation for Vat. gr. 2228 under both manual and automatic segmentation conditions, but this more realistic end-to-end setting is so far assessed only for one manuscript and on a small number of test pages. Extending and expanding this type of end-to-end evaluation therefore remains a natural next step, in particular for Phil. gr. 130, where layout and handwriting are more variable.

7. Conclusion & Future Work

We documented two manuscript-specific Medieval Greek GT datasets, including fully annotated pages with manually curated text line annotations and additional layout-only pages for TLD, trained manuscript-specific TLD and HTR models in Transkribus, and described an iterative workflow and evaluation protocol for Medieval Greek HTR under low-resource conditions. Across both manuscripts, fine-tuning a publicly available Medieval Greek base model consistently improved recognition performance over training from scratch, with the largest gains observed for the more complex Phil. gr. 130. TLD results likewise reflected the contrast between the two manuscripts, with more comparable training and validation performance for Vat. gr. 2228 than for the more layout-variable Phil. gr. 130.

Future work will focus on four directions. (1) Data and imaging quality: extend the annotated corpus, especially for Phil. gr. 130, and assess the effect of higher-quality scans on both segmentation and recognition. (2) End-to-end evaluation: extend the supplementary held-out evaluation reported here by quantifying the interaction between TLD quality and HTR performance on automatically segmented pages beyond the current Vat. gr. 2228 test-set analysis. (3) Training improvements: explore data augmentation to improve robustness to layout and handwriting variation, like in Markou et al. (2021). (4) Post-correction: benchmark post-correction models that treat error correction as sequence-to-sequence translation. Byte-level models such as ByT5 (Xue et al., 2022) have shown robustness to noise and rare characters in related historical settings (Löfgren and Dannélls, 2024; Momtaz et al., 2025), and recent work suggests similar potential for Medieval Greek (Pavlopoulos et al., 2024, 2023) and for large language model (LLM)-based correction of Medieval Greek HTR output (Evangelatos et al., 2026). We aim to evaluate similar approaches on Vat. gr. 2228 and Phil. gr. 130 to reduce remaining errors.

8. Acknowledgements

The work presented here is supported by the Carlsberg Foundation, grant CF24-1999.

9. Bibliographical References

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)*, page 133–159. Association for Computational Linguistics.

Centre for Philology and Digitality. 2025. Ocr4all. URL: https://www.ocr4all.org. Accessed: 2025-09-02.

D-Scribes Project. 2023. | Home | Ancient History | D-Scribes | University of Basel. URL: https://d-scribes.philhist.unibas.ch/en. Accessed: 2025-08-25.

eScriptorium. 2025. escriptorium. URL: https://escriptorium.rich.ru.nl. Accessed: 2025-09-02.

Andreas Evangelatos, Konstantinos Palaiologos, Basilis Gatos, Panagiotis Kaddas, Aikaterini Christopoulou, Vassilis Katsouros, and Andreas Kakridis. 2026. Using llms for improving the ocr accuracy of old greek handwritten documents. In *New Trends in Theory and Practice of Digital Libraries*, pages 100–109, Cham. Springer Nature Switzerland.

Florian Côme Fizaine, Patrick Bard, Michel Paindavoine, Cécile Robin, Edouard Bouyé, Raphaël Lefèvre, and Annie Vinter. 2024. Historical text line segmentation using deep learning algorithms: Mask-rcnn against u-net networks. *Journal of Imaging*, 10(3).

Maxime Guénette, Mathilde Verstraete, Alix Chagué, and Marcello Vitali-Rosati. 2024. Codex palatinus graecus 23 - Ground Truth Dataset Medieval Greek Manuscripts. URL: https://gitlab.huma-num.fr/ecrinum/anthologia/htr_cpgr23. Accessed: 2025-08-25.

Herbert Hunger. 1961. *Katalog der griechischen Handschriften der Österreichischen Nationalbibliothek*, volume 1. Prachner–Hollinek, Wien.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stokl Ben Ezra. 2019. escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19, Piscataway. IEEE.

Julien Leroy. 1976. *Les types de réglures des manuscrits grecs*. Centre national de la recherche scientifique, Paris.

Salvatore Lilla. 1985. *Codices Vaticani Graeci*. Bibliotheca Apostolica Vaticana, Vatican City.

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.

Marilena Maniaci. 2022. *Words within Words : Layout Strategies in Some Glossed Manuscripts of the Iliad*, pages 575–598. De Gruyter, Berlin, Boston.

K. Markou, L. Tsochatzidis, K. Zagoris, A. Papazoglou, X. Karagiannis, S. Symeonidis, and I. Pratikakis. 2021. A convolutional recurrent neural network for the handwritten text recognition of historical greek manuscripts. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 249–262, Cham. Springer International Publishing.

Isabelle Marthot-Santaniello and Tobias Hodel. 2022. Ground-truthed data set of zenon papyri for handwritten text recognition.

Yahya Momtaz, Lorenza Laccetti, and Guido Russo. 2025. Modular pipeline for text recognition in early printed books using kraken and byt5. *Electronics*, 14(15).

Aleksandros Papazoglou, Ioannis Pratikakis, Kleopatra Markou, and Lazaros Tsochatzidis. 2020. Eparchos - historical greek handwritten document dataset. Zenodo.

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. Association for Computational Linguistics.

John Pavlopoulos, Vasiliki Kougia, Paraskevi Platanou, and Holger Essler. 2023. Detecting erroneously recognized handwritten byzantine text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7818–7828, Singapore. Association for Computational Linguistics.

Elpida Perdiki. 2022. Chrysostomicus i | transkibus models. URL: https://app.transkribus.org/models/public/text/chrysostomicus-i. Accessed: 2025-08-25.

Elpida Perdiki. 2023. Preparing Big Manuscript Data for Hierarchical Clustering with Minimal HTR Training. *Journal of Data Mining and Digital Humanities*, Historical Documents and automatic text recognition.

Ariane Pinche. 2022. Cremma Medieval. URL: https://github.com/HTR-United/cremma-medieval.

Paraskevi Platanou, John Pavlopoulos, and Georgios Papaioannou. 2024. HPGTR: Handwritten Paleographic Greek Text Recognition dataset. URL: https://github.com/vivianpl/hpgtr. Accessed: 2025-08-25.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021a. Stavronikita monastery greek handwritten document collection no.114.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021b. Stavronikita monastery greek handwritten document collection no.53.

Ioannis Pratikakis, Aleksandros Papazoglou, Symeon Symeonidis, and Lazaros Tsochatzidis. 2021c. Stavronikita monastery greek handwritten document collection no.79.

READ-COOP SCE. 2025. Transkribus. URL: https://www.transkribus.org. Accessed: 2025-09-02.

Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. 2024. Improving automatic text recognition with language models in the pylaia open-source library. In *Document Analysis and Recognition - ICDAR 2024*, pages 387–404, Cham. Springer Nature Switzerland.

Teklia. 2022. Pylaia. URL: https://gitlab.teklia.com/atr/pylaia. Accessed: 2025-08-25.

Teklia. 2025. Arkindex: A document processing platform. URL: https://doc.teklia.com/arkindex. Accessed: 2025-09-02.

Lazaros Tsochatzidis, Symeon Symeonidis, Alexandros Papazoglou, and Ioannis Pratikakis. 2021. Htr for greek historical handwritten documents. *Journal of Imaging*, 7(12).

Mathilde Verstraete, Maxime Guénette, Malamatenia Vlachou-Efstathiou, Marianne Reboul, and Marcello Vitali-Rosati. 2025. Harmonizing Guidelines for Handwritten Text Recognition of Ancient Greek. URL: https://dhtr25.anthologiagraeca.org. Accessed: 2025-08-25.

Christian Walz. 1835. *Rhetores Graeci*, volume 2. Cotta.

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.

10. Appendix A: Data Splits

Figure 5 and Figure 6 illustrate the resulting character frequency distributions for the full corpus and the corresponding training and validation splits for Vat. gr. 2228 and Phil. gr. 130, respectively. In each figure, characters are ordered by frequency in the full corpus and split into the 50% most frequent and 50% least frequent characters. Relative frequencies are computed within each subset, enabling direct comparison independent of corpus size. For both manuscripts, the training distribution closely follows the full corpus distribution, while the validation set shows slightly weaker distributional alignment but still achieves good coverage across the full set of characters, including less frequent ones.

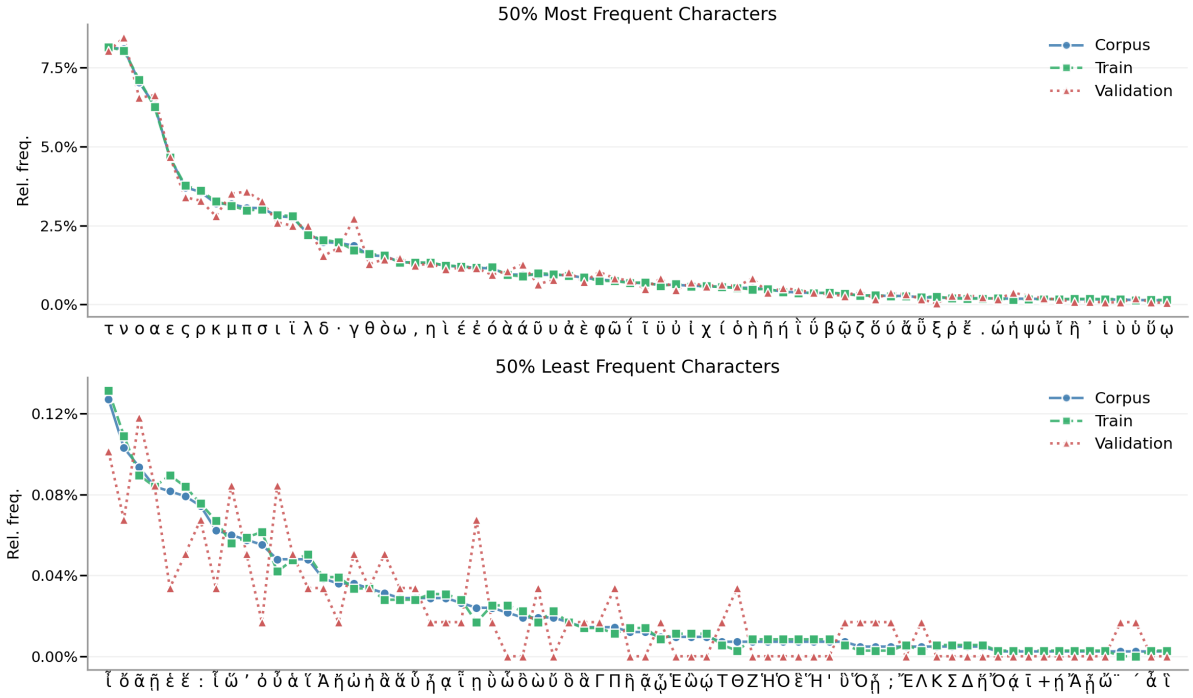

Figure 5: Character frequency distribution in Vat. gr. 2228, split into the 50% most frequent and 50% least frequent characters.

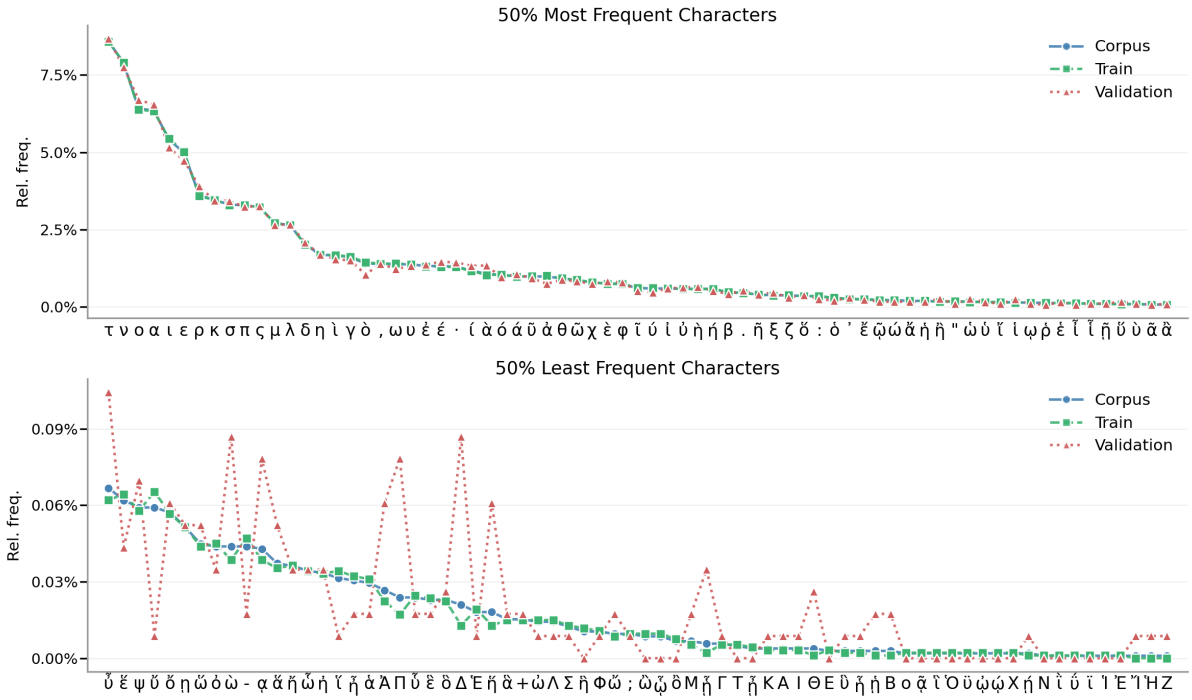

Figure 6: Character frequency distribution in Phil. gr. 130, split into the 50% most frequent and 50% least frequent characters.

11. Appendix B: Per-run Results

This appendix reports per-run results underlying the aggregated metrics presented in Section 5. We present per-run metrics for TLD, HTR on the training and validation splits, and the Vat. gr. 2228 HTR test-set evaluations. The tables reported here make more explicit the run-to-run variability arising from the non-deterministic training process.

Table 5: Per-run TLD performance reported as pixel-wise misclassification rate. Lower values indicate better segmentation performance.

Manuscript	Run	Miscl. Rate	
		Train	Val
Vat. gr. 2228	Run 1	6.32%	5.19%
	Run 2	6.44%	5.59%
	Run 3	6.32%	5.50%
Phil. gr. 130	Run 1	4.87%	8.93%
	Run 2	4.85%	8.41%
	Run 3	4.87%	8.93%

Table 6: Per-run HTR performance on the training and validation splits, reported as micro-averaged CER/WER.

Manuscript	Strategy	Run	CER		WER	
			Train	Val	Train	Val
Vat. gr. 2228	From Scratch	Run 1	0.363%	6.48%	2.21%	29.82%
		Run 2	1.320%	6.07%	7.19%	27.88%
		Run 3	0.367%	6.15%	2.33%	27.08%
	Base Model	Run 1	0.012%	5.05%	0.10%	23.45%
		Run 2	0.026%	5.22%	0.23%	24.34%
		Run 3	0.014%	5.12%	0.11%	23.19%
Phil. gr. 130	From Scratch	Run 1	36.59%	38.17%	65.75%	76.61%
		Run 2	34.46%	36.24%	63.43%	76.10%
		Run 3	41.21%	40.20%	71.23%	81.15%
	Base Model	Run 1	33.48%	28.88%	57.43%	68.69%
		Run 2	32.55%	28.74%	57.58%	68.55%
		Run 3	32.56%	28.25%	56.81%	67.30%
	Base Model (Cont. Train.)	Run 1	28.01%	26.63%	47.63%	64.47%
		Run 2	27.94%	27.41%	47.60%	65.12%
		Run 3	27.90%	27.34%	47.84%	67.58%

Table 7: Per-run HTR test-set performance for Vat. gr. 2228, reported as micro-averaged CER/WER for manually corrected line polygons and automatically detected line polygons from TLD.

Strategy	Run	CER		WER	
		Manual	Automatic (TLD)	Manual	Automatic (TLD)
From Scratch	Run 1	9.27%	19.09%	31.61%	54.75%
	Run 2	8.43%	18.78%	31.14%	54.56%
	Run 3	8.60%	17.53%	30.67%	55.13%
Base Model	Run 1	6.16%	12.27%	24.27%	44.50%
	Run 2	5.86%	12.67%	23.33%	46.57%
	Run 3	5.90%	11.42%	22.95%	41.58%