Miktub: A Manuscript Dataset of Historical Maltese for Handwritten Text Recognition

Thomas Koppens, Claudia Borg
Department of Artificial Intelligence, University of Malta
{thomas.koppens.22, claudia.borg}@um.edu.mt

Abstract

The digitisation of handwritten historical material is essential for preserving cultural heritage and enabling search and computational analysis. For Maltese, historical handwritten resources are scarce, and, to the best of current knowledge, no public handwritten text recognition (HTR) dataset for historical Maltese exists. We introduce a Manuscript Dataset of Historical Maltese (Miktub), collected from the University of Malta Library Special Collections: 35 scanned pages transcribed by specialists and converted into a line-level HTR dataset. A key challenge was robust line extraction from heterogeneous pages; fully automatic line segmentation was insufficient, so we developed a semi-automatic pipeline combining horizontal projection profiling with lightweight post-processing and manual refinement to maximise line fidelity. We provide two annotation variants, including a corrected/standardised version (Miktub-COR) designed to improve consistency, accessibility, and downstream learning stability. We benchmark two strong public HTR models, HTR-VT and VAN, and report the best test performance of 4.68% character error rate (CER) and 13.59% word error rate (WER) on Miktub-COR with VAN. We will release Miktub publicly upon acceptance, along with scripts and splits, to support historical Maltese-language technology research.

Keywords: Handwriting Recognition, Maltese Historical Manuscripts

1. Introduction

Large-scale digitisation initiatives increasingly rely on Handwritten Text Recognition (HTR) to reduce the cost of transcribing handwritten documents, enabling search and computational study of collections that are otherwise locked in image form. Yet HTR performance depends heavily on the availability of labelled data and careful preparation of line-level training material. HTR is often seen as an extension of Optical Character Recognition (OCR), in which scanned or photographed documents are converted into machine-readable form through stages such as text-region detection, line/character segmentation, and recognition. Whilst all these processes are present, HTR must cope with substantial intra- and inter-writer variation, inconsistent character shapes, and irregular spacing, which makes it a considerably more complex recognition problem and motivates the use of more expressive modelling approaches.

This study addresses HTR for Historical Maltese Manuscripts. While HTR is used in several applied domains (e.g., medical documents and cheques), historical manuscripts introduce additional challenges such as physical degradation and ink bleed-through, which can distort character appearance and line structure. In practice, digitising historical collections still often relies on manual line segmentation and transcription—procedures that are repetitive, time-consuming, and require trained personnel. HTR can substantially accelerate this workflow by automatically segmenting and transcribing pages, leaving humans primarily with post-hoc correction; correspondingly, reducing recognition errors directly decreases the manual effort required to clean transcriptions.

Maltese is a Semitic language, with influence from Italian and English (Borg and Gatt, 2017). It is written in the Latin script, with a small set of orthographic diacritics and language-specific letters (notably ċ, ġ, ħ, ż). For HTR, the recogniser must still capture fine-grained distinctions introduced by diacritics, which are particularly vulnerable to noise (e.g., fading, bleed-through) in historical scans. The problem offers an opportunity to leverage techniques and architectures developed for other Latin-script settings, while still confronting the low-resource and historical-domain constraints specific to Maltese collections. The motivation for this work is rooted in cultural preservation and access. Many historically significant Maltese documents remain available only in handwritten form, and the scale of unprocessed archival material far exceeds the human resources typically available for full manual transcription.

Recent advances in neural HTR have made large gains in automation, but these systems are predominantly optimised for high-resource languages and domains; consequently, their effectiveness is limited when applied to historical Maltese texts. This project is therefore driven by the need to bridge that gap by developing and evaluating an HTR pipeline for the accurate transcription of historical handwritten Maltese, supporting preservation, scholarly study, and the broader dissemination of Malta's written heritage.

Concretely, the work pursues a model that min-

Proceedings of LT4HALA 2026 @ LREC 2026, pages 380–388.

imises transcription error, especially character error rate (CER), to reduce manual intervention in downstream transcription workflows. Accordingly, this work makes three primary contributions:

1. Miktub, a line-level dataset derived from 35 historical handwritten pages curated for diversity in authorship and difficulty (e.g., faded/struck-through content), enabling research in low-resource historical HTR.

2. A practical semi-automatic line segmentation workflow that bridges the gap between unreliable full automation and expensive manual segmentation, designed for heterogeneous scans

3. Strong baseline results with two state-of-the-art public architectures (HTR-VT and VAN), including analysis of how transcription consistency affects training and error rates.

For Maltese, especially historical Maltese, resources are limited; as far as we are aware, no historical Maltese HTR dataset currently exists, motivating the creation and release of Miktub.[1]

2. Literature Review

Early automated handwriting recognition relied on template matching and manually engineered features, inspired by online HTR methods that reconstruct stroke trajectories (Plamondon and Privitera, 1999). Systems used Hidden Markov Models (HMMs) and handcrafted descriptors (Plamondon and Srihari, 2000). These pipelines typically required explicit segmentation into characters or words and manual design choices for aligning feature sequences to HMM states.

Deep learning shifted the field toward end-to-end training. A key step was the introduction of Connectionist Temporal Classification (CTC), which enables learning alignments between long input sequences and shorter transcriptions without requiring pre-segmented character boundaries (Graves et al., 2006). Coupled with Multidimensional LSTMs (MDLSTMs), this provided robustness to two-dimensional distortions (e.g., slant and local shifts) while allowing line-level transcription without explicit segmentation (Graves and Schmidhuber, 2008). The standard CTC mechanism—including an explicit blank symbol to separate repeated characters—remains influential due to its suitability for HTR's alignment problem. Puigcerver (2017); Bluche and Messina (2017) found that MDLSTMs often do not substantially outperform faster alternatives based on a CNN feature extractor + BLSTM + CTC, especially when combined with data augmentation; these 1D recurrent models achieve strong accuracy while offering improved computational efficiency.

Beyond purely recurrent/CTC pipelines, attention-based encoder–decoder models became prominent: convolutional backbones produce feature maps, while attentive decoders learn soft alignments over image regions, improving robustness to irregular spacing, cursive joins, and local degradations (Bluche et al., 2017; Wu et al., 2019).

Hybrid objectives combining attention-based cross-entropy with CTC have also been used to stabilise training and improve convergence in low-data regimes. Transformer-based approaches unify visual encoding and sequence generation. TrOCR (Li et al., 2023) exemplifies this direction by pairing a Vision Transformer (ViT) encoder with a Transformer decoder: the encoder represents a line image as a sequence of patch embeddings and applies self-attention to capture long-range dependencies, while the decoder generates text autoregressively via cross-attention. Pre-training on large printed/handwritten corpora is central to this paradigm, enabling strong transfer when fine-tuned on line-level benchmarks.

2.1. HTR-VT and VAN

Two strong public architectures align well with the constraints of historical, low-resource HTR. HTR-VT (Li et al., 2025) adapts the ViT idea to handwriting by replacing patch embedding with a lightweight CNN that yields dense, high-resolution token embeddings aligned with stroke patterns. It introduces span feature masking (masking contiguous token spans during training) to encourage contextual reasoning rather than overfitting to local stroke cues, and it combines this with Sharpness-Aware Minimization (SAM) to improve data efficiency. Decoding is performed with CTC, preserving the benefits of alignment-free training. HTR-VT is particularly relevant here because it achieves strong results on the LAM dataset (2.8% CER / 7.4% WER), suggesting it may transfer well to similar historical letter collections.

The second architecture, VAN (Coquenet et al., 2023), is motivated by a major practical bottleneck in historical collections: the fragility of explicit line segmentation. Instead, VAN processes a paragraph image using a lightweight convolutional backbone and applies a recurrent vertical attention mechanism that scans top-to-bottom, implicitly isolating lines without an external segmentation stage. It iterates attention steps until all lines are covered, enabling end-to-end paragraph transcription, and has reported strong performance on historical datasets such as READ 2016 and IAM with relatively few parameters. Importantly, VAN's para-

[1] Code and data available here: https://github.com/MLRS/Miktub/

graph decoding still depends on a line recogniser, so line-level training (and thus line-aligned supervision) remains necessary, and robustness to unseen handwriting typically requires further fine-tuning.

2.2. Post-processing for HTR

Many HTR systems treat recognition as a two-stage problem: an optical model produces raw text, followed by a language-informed post-processing module that corrects likely errors (Li et al., 2023; Bluche and Messina, 2017). Traditional post-processing commonly integrates lexicons and n-gram language models into decoding, thereby reducing out-of-vocabulary errors and improving plausibility (Povey et al., 2011). However, two limitations are particularly acute in low-resource settings: these methods require carefully constructed corpora/dictionaries for LM training, and n-grams capture only short-range dependencies, limiting context-sensitive correction (Chen and Goodman, 1999; Bengio et al., 2003).

Neural post-processing addresses these issues by learning a mapping from noisy HTR output to corrected text using Seq2Seq architectures with attention, borrowing from machine translation and grammatical error correction (Neto et al., 2020). Because these models are data-driven, they can adapt to the characteristic error patterns of a given optical model without relying on handcrafted lexicons. Neto et al. (2020) conducts a comparative study evaluating statistical and neural correction approaches across multiple HTR datasets and reports that neural methods substantially outperform statistical baselines; notably, a Luong-attention Seq2Seq model corrected a markedly larger share of erroneous sentences than a widely used HMM-LM approach. This motivates prioritising neural correction when sufficient paired data exists, while still retaining lightweight statistical correctors as practical baselines.

2.3. HTR datasets

High-performance HTR models can be pre-trained on large public datasets before fine-tuning on a target domain to improve error rates (Parres and Paredes, 2023). Widely used datasets such as IAM (Marti and Bunke, 2002) and RIMES (Grosicki and El-Abed, 2011) consist of more recent documents with contemporary handwriting styles and show few signs of wear, suggesting they may not share many characteristics with our data. On the other hand, many benchmark datasets and other popular historical datasets were considered for the purpose of pre-training models before being fine-tuned on our dataset, including READ 2016 (Sánchez et al., 2016), Bentham (Sánchez et al., 2015), LAM (Cascianelli et al., 2022).

The target Maltese collection is characterised as early 20th-century cursive with occasional block-letter titles, exhibiting stains and bleed-through and spanning multiple authors over decades—placing it between older historical corpora and modern handwriting datasets. On this basis, READ 2016 is treated as a weaker match due to its highly specialised script/language, and Bentham is excluded primarily because it is single-author and thus offers limited stylistic variability for robust pre-training. In contrast, LAM is argued to be the closest overall match: it is large, historically degraded, spans decades (capturing handwriting drift), and resembles a letter-style manuscript domain.

3. Methodology

3.1. Data Source and Scope

We construct the Miktub dataset, named after the Maltese word for 'written', from historical material held by the University of Malta Library Special Collections Department. They provided 35 scanned page images together with full textual transcriptions, with explicit selection criteria aimed at maximising diversity and practical difficulty: multiple authors over multiple decades (hence multiple handwriting styles as shown in Fig. 1) and the inclusion of phenomena common in archival documents, such as struck-through text, fading, and other degradations. The intent is to support HTR development in a realistic setting and to promote robustness when models are deployed on unseen authors and pages.

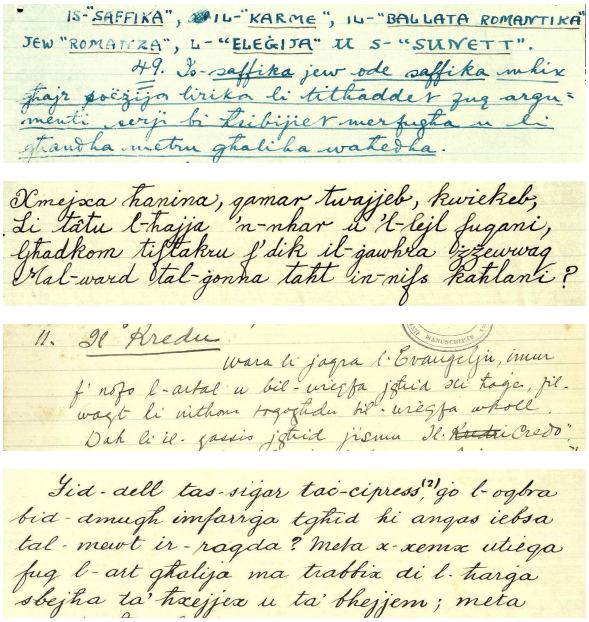

Figure 1: Four samples from different pages of the Miktub dataset illustrating the variety in conditions and handwriting.

3.2. Line-level dataset construction

Most contemporary HTR systems assume line-level inputs. Consequently, each page image and its transcription are converted into aligned line instances. Text splitting is straightforward: each page transcript is programmatically split into individual lines, stored as a single text file per line, grouped by page. Line image segmentation is substantially harder for heterogeneous historical pages. An initial approach was based on Horizontal Profile Projection (HPP) combined with A* pathfinding to refine line boundaries (Shernobyl, 2024; Muthu, 2022). However, this approach proved unsuitable: it frequently merged multiple lines, missed portions of lines, and was slow (minutes per page), which impeded iterative development. Multiple enhancements were explored, including peak clustering, line-height constraints, dilation strength, peak-detection thresholds, and cropping, yet the A* component remained both costly and unreliable, so it was removed.

The segmentation effort then focused on detecting robust line boundaries from HPP peaks. Two clustering strategies were tested (DBSCAN and adaptive gap thresholding), but both still produced closely spaced spurious boundaries and required sensitive tuning across pages. A simpler post-processing rule—peak filtering, i.e., rejecting any peak occurring within a fixed pixel radius of the previous accepted peak—provided a practical compromise: it enforces at most one boundary per line region and generalises across pages without per-page parameter search, at the cost of occasionally keeping the less accurate of two nearby candidate boundaries.

3.3. Semi-automatic segmentation workflow

A key empirical finding is that pages in the Miktub dataset are non-standardised, so parameters that work for one page often fail on others; some pages are not fully segmentable under a purely automatic regime, making full automation infeasible for building a clean training corpus. The adopted solution is a semi-automatic workflow: a fast automatic stage proposes approximate boundaries, followed by a manual adjustment stage that ensures each crop preserves the full vertical context of the handwriting (ascenders/descenders) while minimising overlap with neighbouring lines.

The automatic stage retains HPP but adds (i) optional auto-cropping to remove irrelevant margins (especially for small documents scanned on an A4 canvas) and (ii) the peak-filtering step. Most parameters are fixed globally (e.g., dilation and peak threshold), while cropping is enabled by default and disabled when it would clip text. This streamlined procedure runs in seconds and typically places boundaries slightly below baselines; however, because it effectively "slices" the page rather than explicitly isolating each line, it can cut off descenders and occasionally ascenders. Manual adjustment, therefore, remains necessary for high-quality line images.

Manual refinement is implemented as border editing (top/bottom and, optionally, left/right), rather than editing a single separating line, allowing a single line crop to be expanded without altering adjacent crops. This step is particularly important for interlinear corrections, where an author inserts text between two main lines (sometimes with arrows indicating intended reading order). Depending on length and layout, such corrections are either merged into the corrected line (when short and readable in-place) or treated as standalone lines (when written as a distinct intermediate line). Extreme cases may still force overlap between crops due to tight vertical spacing.

3.4. Transcriptions and dataset variants

During alignment, two transcription issues emerged. First, line-order mismatches occurred because the exported transcriptions supplied by the Data Provider are derived from line-bounding polygons; the textual line order may not correspond to the vertical order of line images. This problem is especially common when corrections are present, because a single correction line can be split into multiple annotated segments. Second, the transcriptions contained systematic inconsistencies and errors, including irregular spacing around punctuation, spacing near clitic articles, accent/diacritic substitutions or omissions, and case errors; additionally, some words were misspelt or omitted.

To handle these issues, we define two dataset variants. Miktub-ORG preserves the original transcriptions "as provided"; consequently, pages with annotation/image mismatches are excluded, reducing coverage. In contrast, Miktub-COR applies a standardisation pass with explicit guidelines for punctuation and stylistic consistency, correcting transcription errors and ensuring one-to-one alignment between each line image and its text. A log of edits is maintained for traceability. Miktub-COR standardisation prioritises consistent formatting to improve model training stability and usability for search and assistive technologies, while largely maintaining a diplomatic stance (transcribing what is written rather than "correcting" author misspellings). An exception is made when marks are clearly displaced (e.g., a delayed cross or dot shifting position), where the intended character may be transcribed to avoid systematic ambiguity.

As a result, Miktub-COR includes all 35 pages

and totals 1076 line images, with an 80/10/10 split into 860/107/109 lines for train/validation/test. Miktub-ORG is smaller: 21 pages, 668 lines total, split into 507/63/64. Agreement between transcription variants, measured at the page level on the full dataset to avoid alignment mismatches, was 3.76% CER and 16.67% WER.

For subsequent paragraph-level experiments, bounding-box metadata for line locations in the original page images is required; because this was not produced during initial dataset creation, it is generated retrospectively via template matching to recover line bounding boxes.

3.5. Out-of-distribution (OOD) test set

To assess generalisation to unseen handwriting, the Data Provider later provided additional pages from the letters collection of P.P. Saydon and Ġużè Aquilina (1932–1946). The two pages were sampled, segmented using the same semi-automatic colour workflow as Miktub-COR, and transcribed from scratch. The resulting Saydon–Ġużè Letters (SGL) dataset contains 55 line images, used exclusively as an OOD test set.

3.6. Experimental protocol and model training

We evaluate two open HTR architectures: HTR-VT and VAN. HTR-VT is used first because its published results on LAM enable replication and because it reports state-of-the-art performance on relevant benchmarks. During training, HTR-VT stores checkpoints for the best validation CER and WER. VAN supports selection by best CER or WER and also retains the latest checkpoint for resumability; following common HTR practice, we prioritise CER when selecting the best model.

Pre-training Before training on Miktub, HTR-VT is trained on the LAM dataset to reproduce reported performance and to obtain a strong initialisation for fine-tuning. Although an online link for pre-trained weights exists, only a READ2016 checkpoint is available. Thus, we pre-train HTR-VT on the LAM dataset and retain the same hyperparameters as in Li et al. (2025). The resulting model (HTR-VT-LAM) is used for further fine-tuning in the next stage.

HTR-VT on Miktub-ORG The initial Miktub experiments use a binary line image variant (Miktub-ORG-BINARY) with 668 lines, training both from scratch and via fine-tuning from HTR-VT-LAM. To support HTR-VT, the dataset is reformatted from a page-folder structure into a flat index of line images and annotations; train/val/test splits are stored as lists of line IDs (80/10/10). The character set size is adjusted for the dataset (92 classes for Miktub-ORG). Training instability near convergence motivates scrutiny of data quality; binary crops are found to be narrow and sometimes poorly binarised, reducing legibility. A colour variant (Miktub-ORG-COLOUR) is therefore created using the semi-automatic segmentation pipeline while retaining the original transcriptions.

Fine-tuning implementation and freezing Fine-tuning is implemented by loading pre-trained weights and freezing a configurable prefix of named layers; the classifier head is re-initialised when class counts differ. An ablation over freezing depth shows that freezing beyond the second embedding layer limits domain adaptation; the best setting freezes up to the second embedding layer and is used thereafter.

Pre-training dataset comparison To test whether LAM's domain similarity matters, a comparative fine-tuning experiment uses READ2016 as an alternative pre-training source. READ2016 is a large historical benchmark (10,527 pages; Early Modern German) with similar page degradations but markedly different script; available weights allow skipping full pre-training, enabling direct comparison under the same fine-tuning pipeline.

HTR-VT and VAN on Miktub-COR We consider two training modes for HTR-VT. The first is training HTR-VT from scratch on the Miktub-COR dataset. The second is to use HTR-VT-LAM (trained on LAM) and fine-tune this model on Miktub-COR. Because Miktub-COR expands coverage (all 35 pages) and modifies transcriptions, it also increases the character set to 97 classes. Empirical comparisons indicate that Miktub-COR is preferable to Miktub-ORG, and subsequent experiments focus on Miktub-COR. The best HTR-VT Miktub-COR model is additionally evaluated on the SGL OOD set.

VAN experiments start with line-level training on Miktub-COR to enable direct comparison with HTR-VT, and because VAN's paragraph-level pipeline typically relies on line-level weights. Dataset conversion is performed from the HTR-VT format into VAN's expected structure (split folders plus a pickle file containing annotations and charset). VAN is trained with default hyperparameters, selecting the best checkpoint based on validation CER.

4. Evaluation

We evaluate the raw outputs of the HTR systems (prior to any post-processing) on (i) the in-distribution test sets of each dataset variant and (ii) an out-of-distribution (OOD) test set written by

Experiment	CER (%)	WER (%)
ORG-BINARY	14.37	34.24
ORG-BINARY+LAM ≤ 3	12.42	29.86
ORG-BINARY+LAM ≤ 2	12.42	29.86
ORG-BINARY READ2016	14.95	37.86
ORG-COLOUR	10.86	28.43
ORG-COLOUR+LAM	10.23	24.12
COR	6.02	15.57
COR+LAM	**5.21**	**13.93**

Table 1: **Test Set Results of HTR-VT experiments** Experiments which freeze N layers during fine-tuning are shown as $\leq N$.

an unseen author. Reported metrics are CER and WER; for post-processing experiments, we also report SER (sequence/line error rate).

4.1. In-distribution HTR results

In-distribution results for all dataset variants using HTR-VT are summarised in Table 1.

Miktub-ORG-BINARY HTR-VT The aim of this initial experiment is to understand the limits of small/noisy data and the impact of freezing embedding layers. This Maltese baseline is trained directly on Miktub-ORG-BINARY and yields 14.37% CER / 34.24% WER, indicating that the small, noisy subset does not, by itself, support low error rates. Several fine-tuning strategies were then tested. Freezing a large portion of the network (up to the third embedding block) was strongly detrimental (26.44% CER / 56.26% WER), suggesting insufficient capacity remained to adapt to the target domain. In contrast, freezing only the first two embedding layers improved performance to 12.42% CER / 29.86% WER, showing that pre-training can help provided most layers remain trainable. Finally, replacing LAM with READ2016 pre-training did not help (14.95% CER / 37.86% WER), reinforcing that transfer effectiveness depends on the relevance of the source domain.

Miktub-ORG-COLOUR HTR-VT To address poor legibility caused by tight cropping and imperfect binarisation, line images were regenerated with the improved semi-automatic segmentation procedure and kept in colour (Miktub-ORG-COLOUR). This change alone reduced the HTR-VT baseline to 10.86% CER / 28.43% WER, a substantial absolute gain over the binary variant. Fine-tuning from the LAM checkpoint produced a small CER improvement but a larger WER improvement, reaching 10.23% CER / 24.12% WER. The evaluation

Experiment	CER (%)	WER (%)
COR	**4.68**	**13.59**

Table 2: **Test Set Results of VAN experiment**.

attributes this pattern to a ceiling effect at the character level once crops become clearer, while pre-training still provides useful priors that reduce word-level confusions.

Miktub-COR HTR-VT The final dataset iteration, Miktub-COR, uses improved line images alongside revised transcriptions, leading to a large jump in baseline accuracy: 6.02% CER / 15.57% WER without external pre-training. Interestingly, applying the earlier fine-tuning approach (with freezing) worsened results (7.10% CER / 18.40% WER), motivating a different strategy. Following findings that Transformer-based HTR benefits from adapting all layers, the model was fine-tuned from LAM with no freezing, yielding the best HTR-VT performance on Miktub-COR: 5.21% CER / 13.93% WER.

Overall, these results underline that (i) transcription quality and consistency can be as impactful as architectural choice, and (ii) for Transformer HTR, partial freezing can hinder domain adaptation when the target distribution differs materially.

Miktub-COR VAN Because Miktub-COR was the most reliable dataset variant, it was the only one used to train VAN. The resulting line-CTC baseline achieved 4.68% CER / 13.59% WER, outperforming the strongest HTR-VT configuration in this work and constituting the best "pure HTR" result reported in (Table 2).

4.2. Out-of-distribution evaluation (SGL)

To probe generalisation, the best HTR-VT models from each dataset variant were evaluated on the SGL OOD test set (two pages from an unseen author). Table 3 summarises how performance drops substantially across the board: 20.37% CER / 53.04% WER for Miktub-ORG-BINARY+LAM, 19.48% / 51.45% for Miktub-ORG-COLOUR+LAM, and 15.93% / 41.01% for the Miktub-COR model fine-tuned on all layers. Notably, the gap between binary and colour variants narrows in OOD, suggesting that once handwriting deviates significantly from training, gains from improved cropping alone are limited, and remaining dataset inaccuracies/inconsistencies become the dominant constraint. Although OOD error rates remain high, the best Miktub-COR model outputs are a useful starting point for assisted manual transcription, enabling a bootstrapping cycle in which new OOD labels can

be incorporated to improve robustness in subsequent iterations.

Experiment	CER (%)	WER (%)
ORG-BINARY+LAM ≤ 2	20.37	53.04
ORG-COLOUR+LAM	19.48	51.45
COR+LAM	15.93	41.01

Table 3: **Test Set Results HTR-VT experiments on the SGL OOD dataset** Experiments which freeze N layers during fine-tuning are shown as $\leq N$.

5. Conclusion

This work targeted accurate recognition of early 20th-century Maltese manuscripts and shows that—given careful data preparation—modern HTR systems can reach low error rates even on challenging historical material. The key driver was an iterative loop that jointly improved (i) line segmentation quality and (ii) transcription accuracy/consistency, complemented by the strategic use of a closely related pre-training dataset.

A central outcome is the construction of a progressively improved line-level dataset. Starting from the initially segmented Miktub-ORG-BINARY, the pipeline was refined to produce a higher-quality colour variant (Miktub-ORG-COLOUR) with better segmentation boundaries, and finally Miktub-COR, which incorporates updated manual transcriptions to improve alignment, accuracy, and consistency. Each iteration improved the legibility and reliability of training data and translated into stronger model performance.

On the modelling side, multiple experiments establish practical guidance for transfer to historical Maltese. For HTR-VT, the best-performing approach is to fine-tune all layers of a model pre-trained on LAM using the corrected Miktub-COR data, rather than relying on partial freezing strategies. In parallel, a single experiment with VAN achieved the best overall results reported in the work, reaching 4.7% CER and 13.6% WER, and thereby outperforming HTR-VT on the in-distribution evaluation and, by extension, on the OOD SGL setting reported in the study.

These findings jointly underscore (a) the value of domain-relevant pre-training and (b) the importance of transcription consistency when training high-capacity recognisers on low-resource historical data.

Despite meeting the primary objectives, limitations remain. The current dataset size (notably the 35-page Miktub-COR collection) constrains generalisation to unseen authors; both HTR-VT and VAN struggle with block letters and sequences of capital letters, and broader coverage (both in- and out-of-distribution) is expected to improve robustness.

Future work should therefore prioritise (i) expanding Miktub-COR with more transcribed pages and more handwriting-varied styles, (ii) incorporating expert-reviewed transcriptions to reduce annotation noise, and (iii) exploring larger or more recent pre-trained architectures (e.g., TrOCR) while recognising that data scale and quality are likely to remain the dominant bottlenecks. Post-processing remains open: one promising direction is fine-tuning language models for Maltese HTR correction or integrating them more tightly into decoding.

Overall, the project demonstrates that, even under limited resources, substantial accuracy is attainable for historical Maltese HTR, and that domain-specific datasets plus iterative refinement are foundational for progress in low-resource handwritten document understanding.

6. Acknowledgements

We acknowledge the collaboration and support of the University of Malta Library Special Collections, in particular Antida Mizza and Luke Joseph Brincat, for supplying the manuscripts, transcripts, annotations and feedback on this work.

7. Bibliographical References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Théodore Bluche, Jérôome Louradour, and Ronaldo Messina. 2017. Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1050–1055.

Théodore Bluche and Ronaldo Messina. 2017. Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 646–651.

Claudia Borg and Albert Gatt. 2017. Morphological Analysis for the Maltese Language: The challenges of a hybrid system. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 25–34, Valencia, Spain. Association for Computational Linguistics.

Silvia Cascianelli, Vittorio Pippi, Maarand Martin, Marcella Cornia, Lorenzo Baraldi, Kermorvant Christopher, and Rita Cucchiara. 2022. The lam dataset: A novel benchmark for line-level handwritten text recognition. In *International Conference on Pattern Recognition*.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Denis Coquenet, Clement Chatelain, and Thierry Paquet. 2023. End-to-End Handwritten Paragraph Text Recognition Using a Vertical Attention Network . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):508–524.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Alex Graves and Jürgen Schmidhuber. 2008. Offline handwriting recognition with multidimensional recurrent neural networks. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'08, page 545–552, Red Hook, NY, USA. Curran Associates Inc.

Emmanuele Grosicki and Haikal El-Abed. 2011. Icdar 2011 - french handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1459–1463.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13094–13102.

Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. 2025. Htr-vt: Handwritten text recognition with vision transformer. *Pattern Recognition*, 158:110967.

U.-V. Marti and H. Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

S. P. Muthu. 2022. Line segmentation in handwritten documents. https://github.com/muthuspark/line-segmentation-handwritten-doc. Last accessed: April 2026.

Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, and Alejandro Héctor Toselli. 2020. Towards the natural language processing as spelling correction for offline handwritten text recognition systems. *Applied Sciences*, 10(21):7711.

Daniel Parres and Roberto Paredes. 2023. Finetuning vision encoder–decoder transformers for handwriting text recognition on historical documents. In *Document Analysis and Recognition - ICDAR 2023*, pages 253–268, Cham. Springer Nature Switzerland.

R. Plamondon and C.M. Privitera. 1999. The segmentation of cursive handwriting: an approach based on off-line recovery of the motor-temporal information. *IEEE Transactions on Image Processing*, 8(1):80–91.

R. Plamondon and S.N. Srihari. 2000. Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, and Georg Stemmer. 2011. The kaldi speech recognition toolkit, workshop on automatic speech recognition and understanding. *US IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii*.

Joan Puigcerver. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition? In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 67–72.

Shernobyl. 2024. Writer identification system. https://github.com/Shernobyl/Writer-Identification-System. Last accessed: April 2026.

Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal. 2016. Icfhr2016 competition on handwritten text recognition on the read dataset. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 630–635.

Joan Andreu Sánchez, Alejandro H. Toselli, Verónica Romero, and Enrique Vidal. 2015. Icdar 2015 competition htrts: Handwritten text recognition on the transcriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1166–1170.

Long Wu, Ta Li, Li Wang, and Yonghong Yan. 2019. Improving hybrid ctc/attention architecture with time-restricted self-attention ctc for end-to-end speech recognition. *Applied Sciences*, 9(21):4639.