Data Matters: Looking for High-Quality Corpora to Build Robust and Reliable Models for Humanists

Jaione Macicior-Mitxelena, Ana Garcia-Serrano
Universidad Pública de Navarra (UPNA), ETSI Informática (UNED)
jaione.macicior@unavarra.es, agarcia@lsi.uned.es

Abstract

The digitization of Spanish historical newspapers poses significant challenges due to low scan quality, typographical diversity, complex layouts and linguistic variation from contemporary Spanish. While advances in Optical Character Recognition (OCR) and layout-aware models offer promising results, their effectiveness strongly depends on the quality and consistency of the underlying training corpora. This work focuses on corpus construction and evaluation for historical document processing. Two experiments were conducted. In the first corpus *los101* was used, a manually curated and structurally annotated subcorpus derived from historical Spanish newspapers, designed to ensure coherent ground truth under heterogeneous real-world conditions. This corpus enables systematic experimentation across OCR and document layout analysis tasks. In a second experimental phase, we apply an additional layout-focused corpus characterized by structural regularity and consistent page organization, allowing us to isolate the impact of layout homogeneity on segmentation performance. State-of-the-art OCR models and a layout detection model are evaluated as validation instruments to assess corpus adequacy rather than as primary contributions. Quantitative and qualitative analyses based on (1) relationship between annotation quality, (2) structural variability, and (3) model behavior, show that heterogeneous corpora challenge both transcription and segmentation stability, while layout-consistent data significantly improves structural detection reliability.

Keywords: Digital Humanities, Corpus Construction, OCR, Layout Analysis, Historical Newspapers

1. Introduction

Before the digital era, research on historical newspapers required physical access to archives and printed collections, resulting in labour-intensive and geographically constrained workflows (Tworek, 2024). Since the early 2000s, systematic digitization initiatives, such as those led by institutions like the National Library of Spain (BNE), have enabled remote access to large newspaper repositories and facilitated computational analysis. In parallel, advances in Artificial Intelligence and Natural Language Processing have expanded research possibilities in Digital Humanities, with Large Language Models supporting large-scale processing of unstructured historical texts and enabling applications such as conversational heritage interfaces (Sergeev et al., 2025), sentiment analysis (Jaber et al., 2025), and automated image description (Garcia-Arias and Garcia-Serrano, 2025), as discussed in recent work (Simons et al., 2025; Lastra-Díaz et al., 2021).

However, the effectiveness of these technologies depends fundamentally on the quality and consistency of the underlying corpora. Historical Spanish newspapers present specific challenges: degraded scans, typographical variation, non-standard orthography, multi-column layouts, advertisements, marginal notes, and irregular page structures. In such contexts, model performance is often constrained not only by architectural design but by the reliability and internal coherence of the annotated data used for training and evaluation.

The main objective of this work is to examine how corpus design, particularly layout variability, affects the performance of OCR and layout analysis systems in historical newspapers. To this end, we conduct two experimental phases: first, using a heterogeneous manually curated corpus (*los101*) for joint OCR and layout experimentation; and second, introducing a structurally homogeneous corpus dedicated to layout analysis in order to isolate the impact of layout regularity on segmentation performance. *los101* (Miguez Lamanuzzi et al. (2025)) is a manually curated corpus guided by philological and structural annotation criteria to ensure ground-truth consistency and reproducibility (Torterolo-Orta et al., 2025). Rather than prioritizing model comparison, OCR and layout architectures are employed as diagnostic instruments to examine how corpus design, annotation coherence, and layout variability affect computational performance. Key factors that influence system behaviour are systematically analyzed. Moreover, domain adaptation is limited by the modest size of the annotated dataset, typographical heterogeneity and variable scan quality. Together, both corpora provide complementary experimental conditions that allow us to analyze the relationship between annotation quality, structural variability, and model behavior.

The remainder is organized as follows. Section 2 reviews previous work on OCR and layout analysis for historical document processing. Section 3 details the motivation and construction of the corpora. Section 4 describes the experimental setup and models used and subsequent subsections present

Proceedings of the First Workshop on Creating Interoperable Corpora of Historical Newspapers (PressMint 2026), pages 82–91

preprocessing strategies and evaluation metrics while section 5 discusses the experimental results. Section 6 concludes with implications for corpus design and future research directions.

2. Related Work

Optical Character Recognition (OCR) converts scanned images into machine-readable text and enables access to digital archives for research and retrieval (Benavent et al., 2010). The accuracy of OCR impacts research outcomes like authorship attribution (Hill and Hengchen, 2019; Garcia Serrano and Menta Garuz, 2022) and user satisfaction in historical documents search (Kettunen et al., 2022). Recent neural approaches have improved transcription quality through automatic correction and normalization (Fleischhacker et al., 2025). Collaborative tools like OCR4all and Impresso further aid this by allowing human-machine interaction for refinement (Düring et al., 2024). Transformer architectures, such as TrOCR, have revolutionized OCR by offering end-to-end transcription with a visual encoder and textual decoder, achieving state-of-the-art results on various datasets (Li et al., 2023; Moreno-Sandoval et al., 2024; Bengio et al., 2013).

Digitizing historical Spanish newspapers poses unique challenges due to varying scan quality, typographical diversity, and complex layouts (Sánchez-Salido et al., 2023; Liebl and Burghardt, 2021). While traditional OCR struggles with these issues, deep learning (CNNs, RNNs) has enhanced robustness by learning visual patterns, though they have limitations with long sequences and parallel computing (Vaswani et al., 2017; Cho, 2014).

Alongside OCR, layout analysis ensures correct reading order and semantic coherence by segmenting elements like columns and titles (Rezanezhad et al., 2023). Early methods improved accuracy by preserving page structure (Chen et al., 2017; Alberti et al., 2017; Zhu et al., 2022). More recently, object-detection frameworks like YOLO, specifically DocLayoutYOLO models, treat structural regions as visual objects, particularly for complex newspaper layouts (Zhao et al., 2024; Santos Júnior et al., 2025; Shen et al., 2021).

Several institutional digitization pipelines have adopted a segmentation-first strategy, where layout boxes are detected and normalized prior to OCR transcription. For example, the Austrian National Library Labs project Esperanto Newspaper Excerpts implements a workflow in which document regions are first identified using object-detection models before text recognition is applied, highlighting the importance of structural preprocessing for historical newspapers (Austrian National Library Labs, 2024).

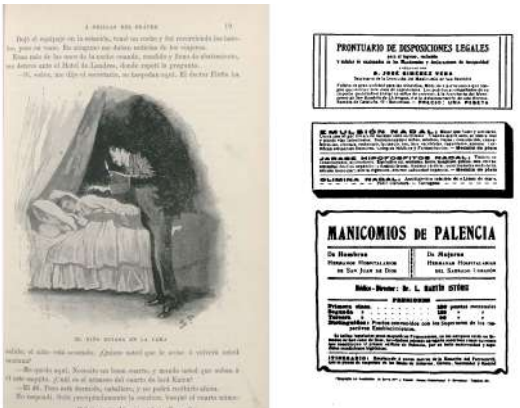
These advancements show a convergence of visual and linguistic modeling in document digitization. However, their effectiveness hinges on high-quality annotated corpora, which are lacking for Spanish historical materials, motivating the creation of the *los101* corpus for this study's experiments.

Nevertheless, while numerous studies evaluate OCR and layout models, fewer works explicitly analyze how corpus structural variability conditions model behavior across tasks. This gap motivates the dual-phase experimental design used in this study.

3. The Need for a Quality Corpus

The experimentation is organized into two phases as introduced in the following paragraphs.

Phase 1: OCR and Layout Experiments with a Heterogeneous Corpus The first phase of this study focused on evaluating OCR systems on historical Spanish newspapers, using a classical approach with Tesseract, an open-source OCR engine maintained by Google. Experiments were conducted within the framework of the PastReader 2025 shared task (IberLEF)[1], which addressed automatic transcription of historical newspapers. The PastReader corpus Montejo-Ráez et al. (2025) consists of over 12,000 scanned pages from eight heterogeneous publications encompassing cultural, scientific, satirical, and literary genres. These materials exhibit diverse typographies, layouts, and states of preservation, presenting a demanding benchmark for OCR systems (view Figure 1).



(a) (b)

Figure 1: Sample pages from the PastReader corpus.

Two configurations of Tesseract were evaluated:

1. **Baseline model (spa):** the default Spanish model (*spa* v5.3), used as an out-of-the-box

[1] https://sites.google.com/view/pastreader2025

reference.

2. **Fine-tuned model (los101):** adapted to the PastReader training set using the `tesstrain` utility. Fine-tuning follows Tesseract's LSTM-based pipeline, requiring aligned pairs of text line images and corresponding UTF-8 transcriptions. Each training sample was segmented line by line, that means that every line needs to be defined by its line coordinates.

Evaluation used standard OCR metrics, including *Character Error Rate (CER)*, *Word Error Rate (WER)*, and semantic measures (*BLEU* and *ROUGE-L*) (some details at subsection 4.1). Results showed that fine-tuning did not improve performance: CER slightly increased from 0.36 to 0.38, and BLEU decreased marginally. Error analysis revealed that the heterogeneous annotations and complex multi-column layouts hindered convergence, particularly for pages with irregular typography or line segmentation inconsistencies.

To overcome these limitations, a dedicated subcorpus with coherent annotations and controlled ground truth were created, ensuring high-quality data for subsequent experiments. From the PastReader dataset, 101 documents were carefully selected based on qualitative analyses and manually transcribed using the Transkribus platform, which supports both textual transcription and structural markup through regions, labels, and inter-region relationships. The corpus (*los101*) was split into 80% training, 10% validation, and 10% test sets Miguez Lamanuzzi et al. (2025).

Annotations followed a unified transcription guide with a *literal modernized* approach, preserving orthography while omitting purely paleographic features (Miguez Lamanuzzi and García Serrano, 2026). In other words, normalization targets graphical variability without altering the underlying linguistic content. Guidelines covered illegible text, marginalia, footnotes, and structural segmentation into paragraphs, headings, and other boxes. Inter-region relationships, such as linking titles to text bodies or images to captions, were included to capture reading order and document layout, supporting multimodal OCR models (Miguez Lamanuzzi et al., 2026).

Phase 2: Layout Analysis with a Homogeneous Corpus Despite its high-quality annotations, *los101* exhibits considerable layout heterogeneity, including variations in column structure, typography, spacing, and overall visual organization.

To investigate the impact of layout regularity, a second corpus (*layout-homogeneous*) was constructed specifically for layout analysis. This corpus consists of digitized pages from nine different newspapers, spanning diverse editorial styles (see Figure 2). It was designed to exhibit more homogeneous page structures, with more consistent column configurations and layout patterns across samples Obispo et al. (2026). It contains layout annotations only, as its purpose is to isolate the effect of structural consistency on segmentation performance rather than to evaluate OCR quality.

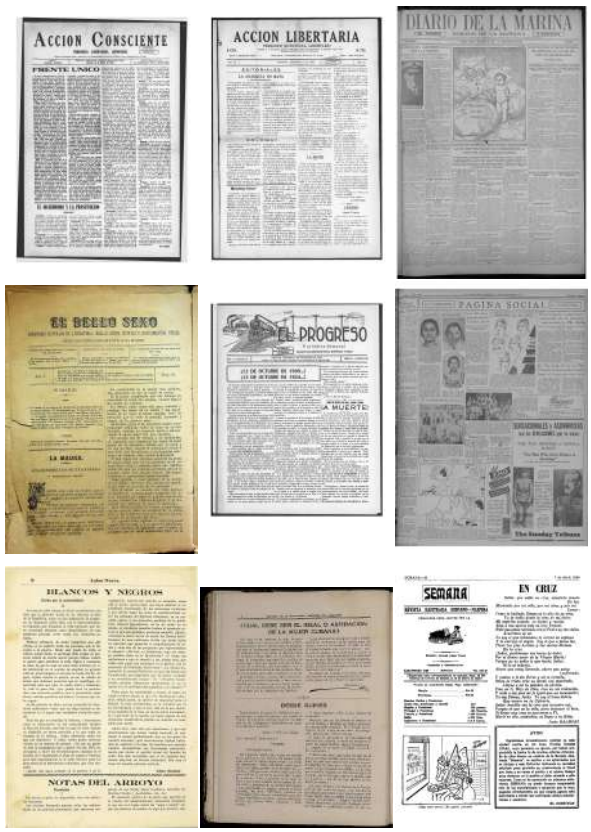


Figure 2: Sample pages from the corpus. Each image is from a different newspaper included in the dataset.

4. The Benchmark for OCR and Layout Tasks

To evaluate the adequacy of the constructed corpora, representative OCR and layout models were employed as diagnostic tools rather than as primary research contributions. The selected systems, Tesseract, TrOCR, Granite, and DocLayoutYOLO, cover classical OCR, transformer-based recognition, multimodal transcription, and object-detection-based layout segmentation. This diversity allows us to observe how corpus characteristics influence performance across different architectural paradigms.

Experiments were conducted in two phases reflecting the corpus design:

- **Phase 1:** OCR and layout experiments on the

heterogeneous *los101* corpus, containing both text and structural annotations.

- **Phase 2:** Layout-segmentation experiments conducted independently on both corpus, using a specialized training configuration focused solely on layout detection to isolate the impact of structural regularity on segmentation performance.

Each model requires a preprocessing pipeline adapted to its architectural assumptions and input constraints. Tesseract, as a classical OCR engine, operates on TIFF images paired with character-level annotations and generates intermediate training representations internally. TrOCR, a transformer-based OCR model, processes JPEG images aligned with their corresponding textual transcriptions, following a sequence-to-sequence learning paradigm. Granite, designed as a multimodal OCR system, requires images to be resized proportionally and embedded within a fixed-size canvas to preserve the aspect ratio and ensure uniform input dimensions. In contrast, DocLayoutYOLO addresses layout segmentation and therefore processes document images together with YOLO-format structural annotations that encode bounding box coordinates and region classes.

Although these preprocessing workflows differ according to architectural design, all models are integrated into a unified experimental framework to guarantee methodological consistency and fair comparison.

4.1. Evaluation Metrics

To assess model behavior, we considered several task-appropriate evaluation metrics. A broader set of measures was explored in (Macicior Mitxelena, 2025) and the code was published on GitHub (Jaione Macicior Mitxelena and Ana Garcia-Serrano., 2026)); however, for the sake of clarity and conciseness, this paper reports and discusses only the most representative ones.

Regarding **OCR subtask** metrics, Character Error Rate (CER) (lower is better), complemented by semantic metrics (BLEU and ROUGE-L) are used. The metrics used for **Layout Analysis subtask** has to take into account segmentation and classification accuracies. Segmentation accuracy is measured using Intersection over Union (IoU), which quantifies the overlap between predicted and ground-truth bounding boxes:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}.$$

Classification performance is evaluated using Precision and Recall, in order to get their harmonic mean (F1-score) computed per region type.

In addition, Mean Average Precision at IoU threshold 0.5 (mAP@0.5) and mAP@0.5:0.95 which averages the performance across multiple IoU thresholds (0.5–0.95) are considered. These metrics provide a standard region-detection evaluation framework that jointly captures localization and classification quality under varying levels.

Finally, to explicitly balance spatial and categorical accuracy, we define a new metric: the **composite layout score**:

$$\text{Score}_{\text{layout}} = \alpha \cdot \text{IoU} + (1 - \alpha) \cdot \text{F1-score},$$

where $\alpha \in [0, 1]$ controls the relative weight of localization versus classification. In this study, $\alpha = 0.75$, prioritizing accurate spatial placement while still accounting for correct region labeling.

4.2. Phase 1: Assesment of corpus adaptation

To ensure methodological consistency, the same experimental protocol was applied to the three OCR models (Tesseract, TrOCR, and Granite). Following the preprocessing procedures described, each baseline model was fine-tuned using the training split of the *los101* corpus. Performance was then evaluated on the corresponding train, val and test sets, and results from the fine-tuned models were systematically compared against their respective baselines.

This controlled setup allows us to isolate and measure the specific impact of corpus characteristics on transcription performance. By maintaining identical training and evaluation splits across models, any observed variation between baseline and fine-tuned configurations can be attributed to the influence of the constructed corpus rather than to differences in experimental design. This comparison serves as a diagnostic mechanism to assess how well the corpus supports adaptation across distinct OCR architectures, including classical OCR (Tesseract), transformer-based sequence-to-sequence recognition (TrOCR), and multimodal transcription (Granite).

The initial layout detection and box classification experiments were conducted on the heterogeneous *los101* corpus used a joint configuration that combined layout detection and box classification. (view Table 2).

Fine-tuning on the model DocLayoutYOLO model was performed using a minimal configuration: 30 epochs, batch size of 2, no explicit data augmentation, no layer freezing, and no regularization beyond default optimization parameters. Mixed precision was disabled and validation was enabled, but the training regime did not include any kind of data augmentation or controlled convergence strategies.

Under this setup, the model was optimized jointly for layout detection and box classification, thereby increasing task complexity while operating on a structurally heterogeneous corpus. The limited number of epochs and absence of regularization mechanisms likely exacerbated instability, particularly in the presence of inconsistent page structures.

4.3. Phase 2: Assesment of layout segmentation

As commented previously, in the second phase, a controlled fine-tuning strategy on the model DocLayoutYOLO was applied separately to both corpora, *los101* and *layout-homogeneous*, obtaining the models *los101_augment* and *per_augment* respectively.

To enhance the model's robustness against digitization artifacts, a **geometric data augmentation strategy** was applied, simulating real-world document variability. This included small rotations ($\pm 2.0°$) and lateral shearing ($2.0°$) to replicate page skew and binder-induced curvature, alongside translation (0.1) and scaling (0.3) to account for inconsistent margins. While mosaic augmentation was utilized to improve image feature extraction across multiple scales, it was strategically disabled during the final 15 epochs. This de-augmentation phase allowed the model to refine bounding box precision on the original document distribution, ensuring the final weights were optimized for the undistorted layouts of the target corpora.

The configuration was explicitly adapted to small-data regimes and to isolate the effect of corpus structure: training was restricted to layout detection only (box classification removed), the first 10 backbone layers were frozen to preserve low-level visual priors, perspective distortion and flips were disabled, and Mosaic Augmentation was enabled for most epochs. Dropout (0.15), a learning rate of 0.001 with a 3-epoch warmup, and early stopping (patience = 20) were also employed.

Early stopping was determined based on the detection metrics mAP@0.5 (mean Average Precision at IoU threshold 0.5) and mAP@0.5:0.95 (average mAP across multiple IoU thresholds from 0.5 to 0.95 in steps of 0.05). These metrics reflect the model's detection precision and localization accuracy: mAP@0.5 emphasizes correct object detection with moderate overlap, while mAP@0.5:0.95 provides a stricter evaluation by averaging performance over increasingly demanding IoU thresholds. Incorporating both metrics ensures the model balances accurate box placement with reliable detection confidence.

Despite the initially configured maximum of 300 training epochs, early stopping (patience = 20) was triggered independently for each corpus, indicating that no further performance improvements were observed beyond a certain number of epochs. The stopping behavior suggests convergence of the optimization process and stabilization of detection performance. For *los101*, training stopped at epoch 138, with best performance at epoch 66: $mAP@0.5 = 0.832$, $mAP@0.5 : 0.95 = 0.424$, while for the second homogeneous corpus, training stopped at epoch 138, with the best performance at epoch 118: $mAP@0.5 = 0.768$, $mAP@0.5 : 0.95 = 0.469$.

5. Results

Analysis results are described below.

5.1. Phase 1 results

Table 1 summarizes the results of the OCR experiments on the heterogeneous *los101* corpus. The outcomes highlight the interplay between corpus properties, model architecture, and fine-tuning strategies. Tesseract achieves the lowest baseline error rates (CER = 0.108), confirming its robustness in out-of-the-box scenarios. Interestingly, fine-tuning Tesseract on *los101* leads to a substantial deterioration in performance, with CER more than doubling. This negative impact reflects the influence of heterogeneity in the training corpus: inconsistent line segmentation, variable fonts, and multiple column layouts introduce noise that disrupts the LSTM training process. These results emphasize that classical OCR engines rely heavily on consistent annotations, and their performance can degrade when exposed to heterogeneous historical data, even if the corpus is relatively large.

Transformer-based TrOCR behaves differently. Its baseline performance is comparatively poor given its high error rate (CER = 0.996), largely due to the small size of the training corpus relative to the model's capacity. Fine-tuning on *los101* reduces CER to 0.906, demonstrating that domain adaptation can partially offset the initial lack of specialization. Although the absolute error rates remain high, the positive delta shows that transformer-based models benefit from exposure to domain-specific samples, provided they are cleanly annotated. This indicates that architectural flexibility allows adaptation to historical document variability, albeit limited by the dataset size.

Granite, the multimodal OCR model, maintains intermediate error rates (CER $\approx$ 0.29–0.30) across baseline and fine-tuned versions. Fine-tuning slightly increases CER, suggesting mild overfitting to the small training set. Nevertheless, Granite demonstrates relative stability compared with Tesseract and TrOCR, likely due to its ability to combine visual and textual cues, which makes it more

Model	Approach	CER	△CER	BLEU	ROUGE-L
Tesseract	spa	**0.1080**	–	**0.5998**	**0.8696**
	los101	0.2454	-127%	0.0505	0.4364
TrOCR	baseline	0.9961	–	0.0000	0.0017
	los101	0.9060	**+9.0%**	0.0000	0.0437
Granite	baseline	0.2897	–	0.2577	0.6557
	los101	0.3033	-4.7%	0.2495	0.6390

Table 1: OCR performance on the first phase with corpus *los101* (test set). Relative changes (△) indicate performance variation after fine-tuning.

Configuration	IoU	F1	Composite
Baseline	**0.3099**	0.0000	**0.2324**
Fine-tuned	0.0000	0.0000	0.0000

Table 2: Layout segmentation and box classification results of in Phase 1 on *los101* corpus on test set using DocLayoutYOLO model.

resilient to typographic and layout variation.

Overall, the results in Table 1 highlight that corpus heterogeneity strongly affects OCR performance. Classical engines are highly sensitive to inconsistent annotations, while transformer-based models benefit from fine-tuning only when sufficient clean data is available. Semantic metrics (BLEU and ROUGE-L) reflect the same trends: Tesseract declines sharply after fine-tuning, TrOCR shows modest gains, and Granite remains relatively stable, demonstrating its resilience to layout and typographic variability. In summary, the *los101* corpus itself limits OCR quality, underscoring the need for consistent annotations and controlled layouts. This motivates the second experiment using a new corpora of pages with more homogeneous features.

As far as the layout segmentation subtask is concerned, the model failed to converge to stable spatial representations (view Table 2). Intersection over Union (IoU) and F1 scores collapsed during validation and testing, in some cases approaching zero. Rather than improving segmentation quality, fine-tuning amplified instability, suggesting that structural variability, multiple editorial formats, inconsistent column structures and heterogeneous box organizations introduced excessive noise into the optimization process. These results indicate that the interaction between corpus heterogeneity and multi-task training limits reliable layout learning.

To complement the quantitative evaluation, a qualitative analysis was conducted on the validation set, focusing on files that consistently ranked among the lowest-performing across multiple metrics. The results show that errors are not driven by a single dominant factor, but correlate with several recurring document characteristics. In particular, performance degrades on low-quality scans (e.g., low contrast), non-standard layouts such as diagrams, and pages containing decorative or complex typographic elements, all of which interfere with text segmentation and recognition. Additionally, pages with little or no textual content tend to produce unstable outputs. These factors often co-occur, making error attribution non-trivial. While both baseline and fine-tuned models are affected, the fine-tuned model shows increased sensitivity to visually complex inputs. Overall, these findings indicate that OCR performance is strongly conditioned by visual and structural variability, which is not captured by aggregate evaluation metrics.

5.2. Phase 2 results

The Phase 2 evaluation highlights the impact of augmentation strategies (related to the image-preprocessing) on layout detection performance across the two corpora. The `los101_augment` and `per_augment` consist of models obtained by fine-tuning the baseline model with the *los101* and the *layout-homogeneus* corpus respectively.

For the *los101* corpus (Table 3), `los101_augment` consistently achieves the highest mean IoU across all splits, indicating improved spatial alignment of predicted boxes with the ground truth. It also attains the best mAP50 and mAP50:95 in most splits, suggesting that corpus-specific augmentations help the model better detect and localize layout elements with higher precision. The `per_augment` model shows moderate improvement over the baseline in detection metrics, but its mean IoU remains slightly lower, implying that augmentation strategies designed for general layouts may not fully capture corpus-specific structural patterns. The baseline model maintains reasonable spatial overlap but suffers from lower detection precision and recall, as reflected by its lower mAP50 and mAP50:95.

For the *layout-homogeneous* corpus (Table 4), `per_augment` clearly outperforms other models

	Train			Val			Test		
Model	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95
baseline	0.773	0.231	0.134	**0.788**	0.162	0.105	**0.773**	0.270	0.171
per_augment	0.720	0.293	0.136	0.722	0.283	0.130	0.726	0.362	0.172
los101_augment	**0.799**	**0.696**	**0.464**	0.745	**0.596**	**0.361**	0.747	**0.545**	**0.305**

Table 3: Phase 2 evaluation metrics for the *los101* corpus.

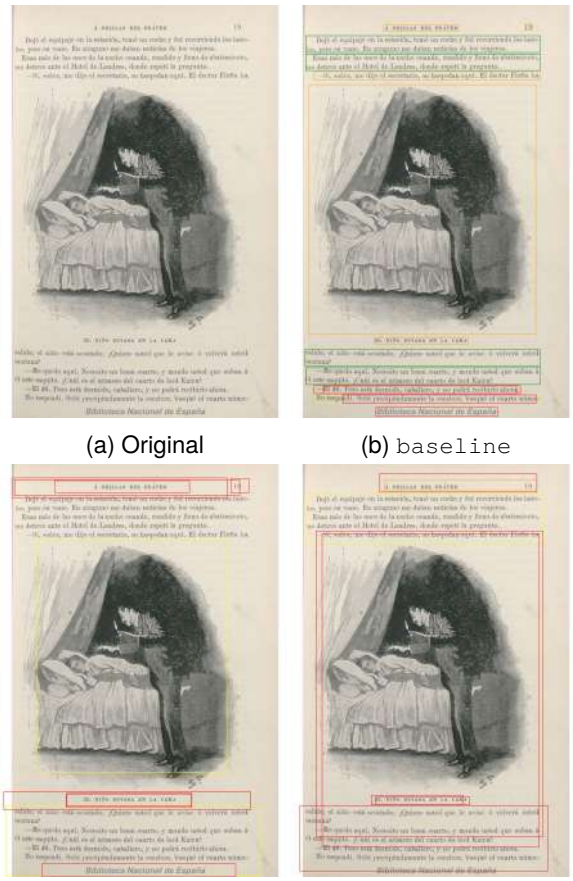
	Train			Val			Test		
Model	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95	mean IoU	mAP50	mAP50:95
baseline	0.730	0.182	0.098	0.740	0.170	0.097	0.734	0.196	0.106
per_augment	**0.841**	**0.755**	**0.560**	**0.810**	**0.652**	**0.436**	**0.804**	0.554	**0.370**
los101_augment	0.749	0.589	0.331	0.769	0.447	0.261	0.741	**0.592**	0.324

Table 4: Phase 2 evaluation metrics for the *layout-homogeneous* corpus.

in all splits and metrics, achieving the highest mean IoU, mAP50, and mAP50:95. This result demonstrates that layout-aware augmentations effectively handle heterogeneous and complex newspaper page structures, improving both box localization and detection confidence. The `los101_augment` model performs well but slightly lags behind `per_augment`, particularly in mAP50:95, suggesting that augmentations optimized for *los101* do not generalize perfectly to different editorial formats. The baseline model remains the weakest across metrics, highlighting the necessity of data augmentation for robust layout learning in highly variable corpora.

To complement the quantitative evaluation, we visualize model predictions in Figures 3 and 4. Bounding boxes are colored according to detection confidence, allowing a quick assessment of model certainty: green indicates high-confidence predictions ($conf > 0.9$), orange represents moderately high confidence ($0.75 < conf \leq 0.9$), yellow corresponds to medium confidence ($0.5 < conf \leq 0.75$), and red highlights low-confidence predictions ($0.2 < conf \leq 0.5$). This visual coding facilitates the inspection of both spatial alignment and model certainty across different corpora and augmentation strategies.

It can be observed that the baseline model generally produces fewer boxes with lower confidence variation, often appearing predominantly green, reflecting a more conservative but consistent detection behavior. In contrast, the fine-tuned models (`per_augment` and `los101_augment` show improved detection coverage and spatial alignment, but their predictions include a higher proportion of orange, yellow, and red boxes. This indicates that while the fine-tuned models are not inherently worse, their confidence is more distributed, reflecting increased sensitivity to diverse page structures and the presence of more challenging layout elements.



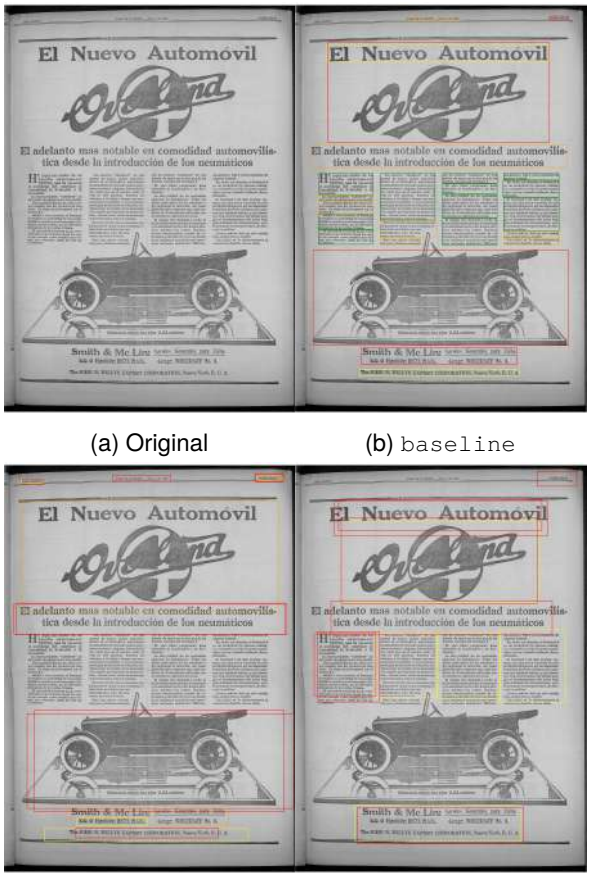
(a) Original (b) `baseline`

(c) `per_augment` (d) `los101_augment`

Figure 3: Visualization of layout detection results in an example of the corpus *los101* corpus.

5.3. Corpus adaptation discussion

Several key trends emerge from the evaluation across both corpora, highlighting the impact of augmentation and the nuances of detection performance. Primarily, the use of augmentation strategies leads to significant improvements in both mean IoU and mAP metrics. This demonstrates that such techniques are highly effective in stabilizing layout learning, likely by providing the model with a more



(a) Original (b) `baseline`

(c) `per_augment` (d) `los101_augment`

Figure 4: Visualization of layout detection results in an example of the new corpus.

diverse and robust set of spatial variations during training.

The results also reveal a trade-off between specialization and generalization. Models optimized for a specific corpus, such as `los101_augment` for the *los101* dataset, consistently achieve the highest performance on their native data. However, these gains do not always translate to other corpora, suggesting that while corpus-specific optimization maximizes local accuracy, it may limit the model's ability to generalize across different layout styles.

Furthermore, the performance discrepancy between mAP50 and mAP50:95 illustrates the complexity of precise localization. The mAP50:95 values are consistently lower across all tests, reflecting the increased difficulty of maintaining high precision across stricter overlap thresholds. By evaluating both spatial alignment through mean IoU and detection precision through mAP metrics, a more comprehensive view of model performance is achieved. This combined approach successfully reveals the inherent trade-offs between achieving accurate box placement and maintaining high overall detection precision.

6. Conclusions and Future Work

In this study of historical document processing, the quality and structural regularity of the underlying corpus are as critical as the choice of model architecture. Our experiments with *los101* reveal that high-quality, manually curated annotations alone are insufficient to overcome extreme layout heterogeneity in classical OCR engines like Tesseract, which suffered a performance drop of over 120% in CER after fine-tuning.

Conversely, the second phase results indicate that structural homogeneity significantly stabilizes layout analysis. By using a corpus with consistent column patterns and refined augmentation strategies, we achieved a marked improvement in spatial localization (mean IoU) and detection precision (mAP50). For the humanist researcher, this implies that a smaller, structurally consistent dataset may be more valuable for model training than a larger, noisier, and highly variable collection. Finally, our results suggest that multimodal models like Granite offer a more resilient middle ground, balancing visual and textual cues to navigate the "noise" of historical digitizations.

Future research will follow three primary directions. First, we intend to explore hybrid training regimes that combine the structural regularity of our second corpus with the linguistic richness of *los101*, starting with simple layouts and gradually introducing complexity. Second, we aim to implement automated layout normalization techniques as a preprocessing step to reduce the "visual noise" before it reaches the OCR engine.

Finally, we plan to expand the *los101* corpus to include a broader range of 19th-century scientific journals, testing the generalizability of our 'literal modernized" transcription approach. This will also involve evaluating the impact of OCR errors on downstream NLP tasks, such as Named Entity Recognition (NER) and Topic Modeling, to quantify exactly how much "annotation matters" for the final historical analysis.

7. Acknowledgements

This work is partially supported by the Ministerio de Ciencia e Innovación/AEI within the framework of the coordinated Spanish National project GRESEL UNED (PID2023-151280OB-C22).

8. Bibliographical References

Michele Alberti, Mathias Seuret, Vinaychandran Pondenkandath, Rolf Ingold, and Marcus Liwicki. 2017. Historical document image segmentation with lda-initialized deep neural networks. In *Proceedings of the 4th international workshop on historical document imaging and processing*, pages 95–100.

Austrian National Library Labs. 2024. Esperanto newspaper excerpts. GitLab repository. Accessed 2026-02-25.

Joan Benavent, Xaro Benavent, Esther de Ves, Ruben Granados, and Ana García-Serrano. 2010. Experiences at imageclef 2010 using CBIR and TBIR mixing information approaches. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, volume 1176 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. 2017. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 965–970. IEEE.

Kyunghyun Cho. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Marten Düring, Estelle Bunout, and Daniele Guido. 2024. Transparent generosity. introducing the impresso interface for the exploration of semantically enriched historical newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 57(1):20–40.

David Fleischhacker, Roman Kern, and Wolfgang Göderle. 2025. Enhancing ocr in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries*, 26(1):3.

Enrique Garcia-Arias and Ana Garcia-Serrano. 2025. Creación de un modelo de descripciones de imágenes especializado en arqueología griega (pending edit). *Procesamiento del Lenguaje Natural*, 75(0).

Ana Garcia Serrano and Antonio Menta Garuz. 2022. La inteligencia artificial en las humanidades digitales: dos experiencias con corpus digitales. *Revista de Humanidades Digitales*, 7:19–39.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Areej Jaber, Israa Bahati, and Paloma Martínez. 2025. Leveraging pre-trained embeddings in an ensemble machine learning approach for arabic sentiment analysis. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Jaione Macicior Mitxelena and Ana Garcia-Serrano. 2026. From Paper To Pixel: Experimental Framework for Access to Historical Spanish Documents. `https://github.com/jaionemacicior/from-paper-to-pixel`. Software/Code.

Kimmo Kettunen, Heikki Keskustalo, Sanna Kumpulainen, Tuula Pääkkönen, and Juha Rautiainen. 2022. Ocr quality affects perceived usefulness of historical newspaper clippings–a user study. *arXiv preprint arXiv:2203.03557*.

Juan J. Lastra-Díaz, Josu Goikoetxea, Mohamed Ali Hadj Taieb, Ana Garcia-Serrano, Mohamed Ben Aouicha, Eneko Agirre, and David Sánchez. 2021. A large reproducible benchmark of ontology-based methods and word embeddings for word similarity. *Information Systems*, 96:101636.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13094–13102.

Bernhard Liebl and Manuel Burghardt. 2021. An evaluation of dnn architectures for page segmentation of historical newspapers. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5153–5160. IEEE.

Jaione Macicior Mitxelena. 2025. Del papel al pixel: Experimentos para la digitalización de documentos históricos españoles. Master's thesis, UNED: Universidad Nacional de Educación a Distancia, Madrid, Spain, September. Directed by Ana Garcia-Serrano. Máster en Tecnologías del Lenguaje. Grade: Sobresaliente (9).

M. Miguez Lamanuzzi and A. García Serrano. 2026. Annotation of historical texts for automatic processing in the gresel-uned project. In *Congreso Internacional de Lingüística de Corpus (CILC 2026)*, Madrid. UAM. Aceptada.

M. Miguez Lamanuzzi, J. Macicior Mitxelena, Y. Torterolo, R. Ortuño Casanova, and A. García Serrano. 2026. Guía de transcripción y anotación para prensa histórica. https://doi.org/10.5281/zenodo.19187624.

Antonio Moreno-Sandoval, Leonardo Campillos-Llanos, and Ana García-Serrano. 2024. Language resources in spanish for automatic text simplification across domains.

Vahid Rezanezhad, Konstantin Baierer, Mike Gerber, Kai Labusch, and Clemens Neudecker. 2023. Document layout analysis with deep learning and heuristics. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, pages 73–78.

Eva Sánchez-Salido, Antonio Menta, and Ana García-Serrano. 2023. Seeking information in spanish historical newspapers: The case of diario de madrid (18th and 19th centuries). *DHQ: Digital Humanities Quarterly*, (4).

Eder Silva dos Santos Júnior, Thuanne Paixão, and Ana Beatriz Alvarez. 2025. Comparative performance of yolov8, yolov9, yolov10, and yolov11 for layout analysis of historical documents images. *Applied Sciences*, 15(6):3164.

Alexander Sergeev, Valeriya Goloviznina, Mikhail Melnichenko, and Evgeny Kotelnikov. 2025. Talking to data: Designing smart assistants for humanities databases.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*.

Arno Simons, Michael Zichert, and Adrian Wüthrich. 2025. Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives.

Yanco Amor Torterolo-Orta, Jaione Macicior-Mitxelena, Marina Miguez-Lamanuzzi, and Ana García-Serrano. 2025. Transcribing spanish texts from the past: Experiments with transkribus, tesseract and granite.

Heidi JS Tworek. 2024. Digitized newspapers and the hidden transformation of history. *The American Historical Review*, 129(1):143–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*.

Wenzhen Zhu, Negin Sokhandan, Guang Yang, Sujitha Martin, and Suchitra Sathyanarayana. 2022. Docbed: A multi-stage ocr solution for documents with complex layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12643–12649.

9. Language Resource References

M. Miguez Lamanuzzi, A. García Serrano, Jaione Macicior Mitxelena, and Yanco Torterolo. 2025. Los101. [Data set].

A. Montejo-Ráez, E. Sánchez Nogales, G. Expósito Álvarez, A. Ureña López, M. T. Martín-Valdivia, J. Collado-Montañez, I. Cabrera de Castro, M. V. Cantero Romero, A. García Serrano, R. Ortuño Casanova, and Y. A. Torterolo Orta. 2025. Pastreader 2025. https://doi.org/10.5281/zenodo.15084265. [Data set].

F. Obispo, R. Ortuño Casanova, L. Garçon, Y. Seyeux, E. Sinardet, D. Villanueva Romero, and E. Vivó Capdevila. 2026. Corpus de prensa histórica con el layout marcado en pagexml. https://doi.org/10.5281/zenodo.18774961. [Data set].