electronics

MDPI

Article

Modular Pipeline for Text Recognition in Early Printed Books Using Kraken and ByT5

Yahya Momtaz *, Lorenza Laccetti † and Guido Russo †

Department of Physics "Ettore Pancini", University of Naples Federico II, 80126 Naples, Italy
* Correspondence: yahya.momtaz@unina.it
† These authors contributed equally to this work.

Abstract

Early printed books, particularly incunabula, are invaluable archives of the beginnings of modern educational systems. However, their complex layouts, antique typefaces, and page degradation caused by bleed-through and ink fading pose significant challenges for automatic transcription. In this work, we present a modular pipeline that addresses these problems by combining modern layout analysis and language modeling techniques. The pipeline begins with historical layout-aware text segmentation using Kraken, a neural network-based tool tailored for early typographic structures. Initial optical character recognition (OCR) is then performed with Kraken's recognition engine, followed by post-correction using a fine-tuned ByT5 transformer model trained on manually aligned line-level data. By learning to map noisy OCR outputs to verified transcriptions, the model substantially improves recognition quality. The pipeline also integrates a preprocessing stage based on our previous work on bleed-through removal using robust statistical filters, including non-local means, Gaussian mixtures, biweight estimation, and Gaussian blur. This step enhances the legibility of degraded pages prior to OCR. The entire solution is open, modular, and scalable, supporting long-term preservation and improved accessibility of cultural heritage materials. Experimental results on 15th-century incunabula show a reduction in the Character Error Rate (CER) from around 38% to around 15% and an increase in the Bilingual Evaluation Understudy (BLEU) score from 22 to 44, confirming the effectiveness of our approach. This work demonstrates the potential of integrating transformer-based correction with layout-aware segmentation to enhance OCR accuracy in digital humanities applications.

Keywords: OCR post-correction; historical documents; early printed books; incunabula; ByT5; sequence-to-sequence learning; Kraken OCR; digital humanities; text recognition; OCR noise reduction; manuscript processing; cultural heritage digitization

Academic Editor: Xenophon Zabulis

Received: 29 June 2025
Revised: 26 July 2025
Accepted: 29 July 2025
Published: 1 August 2025

Citation: Momtaz, Y.; Laccetti, L.; Russo, G. Modular Pipeline for Text Recognition in Early Printed Books Using Kraken and ByT5. *Electronics* **2025**, *14*, 3083. https://doi.org/10.3390/electronics14153083

1. Introduction

The digital preservation of historical manuscripts, especially early printed books like incunabula, is growing in the digital humanities and archive sciences. These books represent the earliest instances of books printed with movable type and were written during the mid-15th and early 16th centuries. They are of cultural, linguistic, and historical importance and provide insight into early modern philosophy, literature, and scientific exploration. So digitizing these historic manuscripts is crucial for retaining cultural heritage, increasing access to historical texts, and protecting them for future generations. In recent decades,

libraries and research institutes throughout Europe and beyond have started extensive digitization projects to keep these artifacts and enhance accessibility for scholars and the public. Given these actions, the automatic transcription of early written texts continues to be a challenge. Multiple reasons contribute to this complexity: diverse typefaces, non-standard orthography, ornamental embellishments, deteriorated paper, bleed-through effects from acidic inks, and non-linear layouts. Typical Optical Character Recognition (OCR) systems struggle with these problems, which result in error transcriptions characterized by mistake rates and restrictions for future study. To address these limitations, we propose a *modular pipeline* for text recognition tailored to early printed books. The pipeline integrates three key stages: (1) *Layout-aware text segmentation using Kraken*, which is well suited for historical documents with irregular structures; (2) *OCR generation* using a conventional recognition model; and (3) *OCR correction using ByT5*, a multilingual byte-level transformer model that we fine-tune to correct the noisy outputs generated by OCR. This structure allows each component to be independently adapted or replaced as new tools emerge, making the pipeline scalable and flexible for diverse corpora. This work builds on the objectives of the *MAGIC project* [1], which started in 2023 to enable digital access to ancient books [2]. In previous work, we developed a robust *bleed-through removal system* [3] based on four complementary approaches: non-local means (NLM), Gaussian mixture models (GMMs), biweight estimation, and Gaussian blur. This preprocessing stage significantly improves the visual quality of degraded images and prepares them for OCR. This study focuses on the post-OCR correction phase, when we train a ByT5 model to map OCR-generated hypotheses with precise ground-truth transcriptions. Our training dataset comprises segmented lines from early printed books, along with their manually corrected transcriptions. This fine-tuning method enables the model to recognize mistake patterns characteristic of early print OCR and to produce enhanced text output. This document outlines the dataset preparation, model architecture, and training methodology. Next, we assess the system's performance based on the character error rate (CER) and examine its ramifications for the wider domain of historical document processing and digital humanities. The code is available on GitHub (https://github.com/yahyamomtaz/medieval-ocr-pipeline, accessed on 28 July 2025).

2. Related Works and Literature Review

Optical character recognition (OCR) in the context of early printed books and manuscripts presents unique challenges due to non-standard orthography, degraded page quality, and complex historical typefaces. While the history of OCR research has seen considerable advancements since the early 1990s [4], traditional systems like Tesseract [5] perform well on modern documents but often struggle with these irregularities because of their reliance on rigid character templates and limited layout adaptability. To address these limitations, OCR engines like Kraken [6] have adopted recurrent neural network architectures trained with Connectionist Temporal Classification (CTC) loss, enabling more flexible, line-based recognition. Kraken's support for custom training makes it particularly suitable for early printed texts with specialized fonts or ligatures. Similarly, the Transkribus platform [7] integrates HTR+ (Handwritten Text Recognition) technology and offers collaborative model training for historical handwriting, although its proprietary nature may limit interoperability with open-source pipelines. Beyond traditional OCR engines, recent research has explored hybrid or neural approaches to post-correction. Models such as ByT5 [8] offer byte-level, token-free architectures capable of correcting noisy OCR outputs without language-specific preprocessing. These models have proven effective in handling fine-grained character-level noise, particularly in historical corpora. Concurrently, large language models (LLMs) are being explored for tasks related to cultural heritage

processing [9–12], offering new possibilities for semantic enrichment, classification, and user-facing interfaces in digital humanities platforms.

2.1. Post-OCR Correction Techniques

While early OCR efforts relied on rigid rule-based logic and predefined fonts [13], modern systems like Kraken and Tesseract have broadened their capabilities. However, historical texts still present structural and orthographic inconsistencies that challenge these tools. Kraken's ability to fine-tune recognition models on custom datasets makes it ideal for early printed books, whereas platforms like Transkribus remain popular in handwritten document analysis despite their closed-source limitations.

2.2. Post-OCR Correction and Transformer-Based Neural Models

Postprocessing remains one of the most persistent challenges in OCR pipelines, particularly when working with degraded or noisy inputs such as bleed-through, faded ink, or non-standard typography. As discussed by [14], even advanced OCR systems frequently introduce errors that necessitate specialized correction strategies. Traditional post-OCR correction methods have relied heavily on rule-based or dictionary-based techniques, typically minimizing the edit distance to fix out-of-vocabulary or malformed words. While useful in many modern contexts, these methods often underperform when faced with compounded OCR errors or obsolete orthographies—frequent in historical printed materials [15]. To overcome these limitations, recent advances have shifted toward neural sequence-to-sequence (seq2seq) models, which treat post-OCR correction as a translation problem by learning to map noisy input to accurate transcriptions. Transformer-based architectures have become the standard in this domain. In particular, ByT5—a byte-level variant of the T5 (Text-to-Text Transfer Transformer) model introduced by [8]—offers significant advantages for historical OCR correction. Unlike token-based models, ByT5 operates directly on raw byte sequences, making it robust to spelling inconsistencies, typographic noise, and rare characters—qualities particularly useful for historical documents. While its effectiveness has been demonstrated in post-OCR correction of Swedish newspapers [16], our work adapts it to the domain of early printed Latin texts.

2.3. Layout Analysis and Digitization Challenges

Accurate layout analysis and text line segmentation are prerequisites for OCR or text correction to be applied successfully to digitized historical materials [17]. Various methods have been proposed to tackle these challenges, including the Labeling, Cutting, and Grouping approach by Alberti et al. [18], which presents an efficient technique for segmenting text lines in medieval manuscripts using document-specific heuristics and machine learning. The diversity of historical document layouts, including marginalia, multi-column structures, ornamentation, and irregular line spacing poses significant hurdles to automated systems [19]. Recent advances in neural network-based methods underpin specialized tools such as Kraken, which supports region-based segmentation and XML-based PAGE format annotation, enabling nuanced representations of lines, areas, and metadata [20]. The quality of segmentation is a core determinant of OCR output reliability: poorly segmented lines often cascade into recognition and correction errors throughout the pipeline [21]. While additional tools such as OCR-D, Calamari, and Tesseract have begun integrating advanced segmentation modules, Kraken stands out for its flexibility in custom training for historical layouts and its compatibility with modular pipeline architectures.

2.4. Digitization Challenges: Bleed-Through Removal

OCR systems for historical documents are often compromised by bleed-through—ink permeating the paper, visible from the reverse side. This issue degrades image quality and

confounds text recognition, especially in aged manuscripts [13]. Recent research integrates both traditional image processing techniques and modern machine learning methods to mitigate bleed-through. Notable approaches include non-local means filtering, Gaussian mixture models, the use of biweight estimators, Gaussian blur, conditional random fields (CRFs), and various deep learning models [Le Deunf, Julian, et al.]. These techniques are selected and tuned based on the unique features of each manuscript, and their effectiveness is amplified when coupled with robust post-OCR correction pipelines. Modular solutions that treat bleed-through removal as a dedicated preprocessing stage greatly enhance the subsequent accuracy of OCR and postprocessing [22].

2.5. Digital Humanities Integration Pipelines

The digital humanities community increasingly adopts modular, AI-driven pipelines for digitization, transcription [23], and scholarly text analysis. Full-stack platforms—such as Transkribus, IMPRESSO, and eScriptorium—integrate segmentation, OCR/HTR (Handwritten Text Recognition), and metadata enrichment. However, research increasingly favors modular and interoperable architectures over monolithic systems to enhance repeatability and flexibility. By combining high-precision segmentation (e.g., with Kraken) and byte-level transformer-based correction (e.g., fine-tuned ByT5), the proposed approach aligns with best practices in digital preservation. Such modular designs facilitate adaptation across varied historical document types, languages, and typographic systems, and support integration with broader semantic enrichment workflows. Alignment with FAIR (Findable, Accessible, Interoperable, and Reusable) principles and open-source tools helps ensure the repeatability and sustainability of these digital infrastructures. Our modular pipeline supports interoperability and long-term sustainability, making the corrected texts suitable for further indexing, metadata generation, and historical search systems [24].

3. Methodology

The whole architecture and implementation specifics of our modular text recognition and OCR correction system for early printed books are discussed in this part. Three main stages comprise the pipeline: segmentation, OCR, and OCR post-correction with a fine-tuned ByT5 model. Every stage is made to operate autonomously and is compatible using common formats.

3.1. Summary of the Pipeline

We introduce a modular OCR correction pipeline tailored for medieval manuscripts, as illustrated in Figure 1. The workflow consists of two main phases: model fine-tuning and automated inference.

In the fine-tuning phase, we construct a training dataset by aligning raw OCR outputs from medieval manuscript lines with their manually verified ground-truth transcriptions. These aligned pairs are used to fine-tune a ByT5 transformer model, enabling it to learn the specific error patterns and correction strategies characteristic of historical texts. During inference, the pipeline processes new manuscript images in three steps. First, Kraken performs line segmentation, extracting individual text lines from full-page images. Next, the TrOCR model—pretrained for historical scripts—is used to generate initial OCR predictions for each line. Specifically, we adopt the `medieval-data/trocr-medieval-print` model, which has been fine-tuned on early printed texts for improved accuracy on historical typefaces [25]. Finally, the fine-tuned ByT5 model corrects these raw OCR outputs, producing high-accuracy transcriptions.

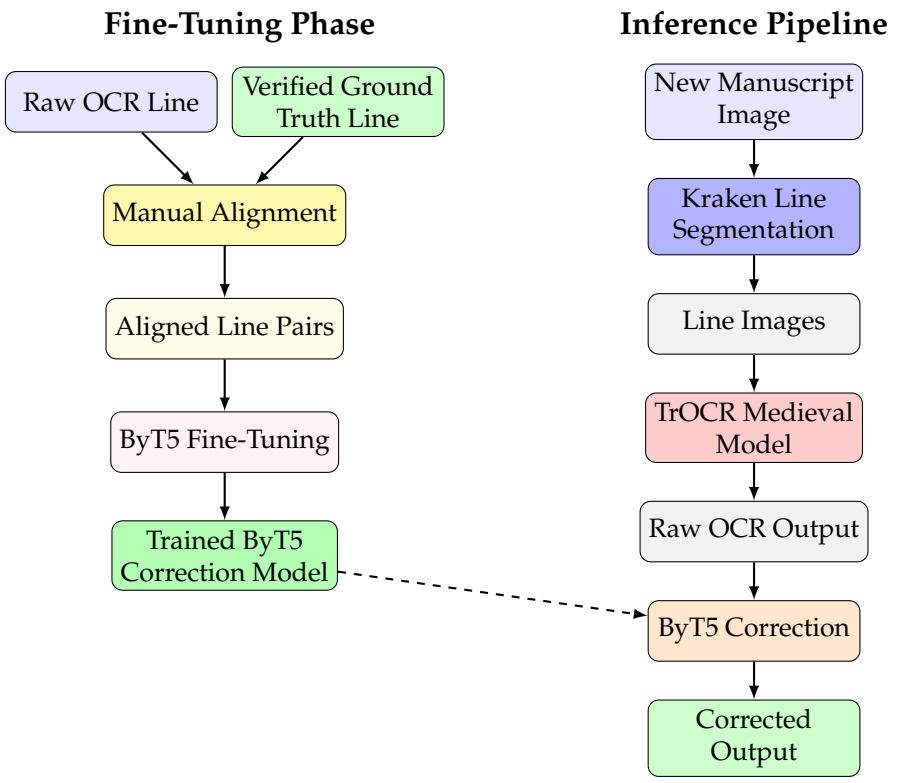

Figure 1. End-to-end workflow showing the ByT5 fine-tuning process and its integration into the automated OCR inference pipeline.

The entire pipeline is designed for interoperability and reproducibility, utilizing standard data formats (such as CSV and JSON) and modular components. This approach enables seamless integration with existing digital humanities tools and supports robust evaluation across diverse manuscript collections. Our results demonstrate a substantial reduction in the character error rate, highlighting the effectiveness of combining neural OCR with transformer-based post-correction for challenging medieval sources.

3.2. The Usage of Kraken

Kraken is a key part of both layout segmentation and the first OCR phase in our modular pipeline. While OCR systems such as Tesseract [5] and OCR4all [6] provide modular workflows, they often struggle with historical variability, requiring the integration of specialized segmentation and correction models. We used a Kraken model for segmentation using a carefully chosen set of manually annotated PAGE XML files. These files had regions, baselines, and individual text lines marked with great care. This step is very important because the accuracy of line-level segmentation directly affects the quality of the OCR. The training set included a wide range of page layouts and typographical structures from the incunabula, which ensured that the model could work with different types of historical source. We checked the accuracy of the segmentation both qualitatively (by looking at it) and quantitatively (by looking at the pixel-level overlap between the model output and the ground-truth annotations). After extracting the correct line regions, they were sent to Kraken's default recognition model for OCR. This model is made to handle the strange things that happen with early printed books, like characters that are not shaped like they should be, ligatures, and long s forms. Although the first transcriptions had a lot of structure, they often kept historical spelling differences and character misclassifications, especially with glyphs that were faded or damaged. These inaccurate predictions showed the flaws of a purely image-based OCR approach and required a neural post-correction component. We used a fine-tuned ByT5 transformer model to add this feature. This second

stage allowed us to normalize spelling variations and fix common OCR-specific errors, greatly improving the accuracy of the final output.

Figure 2 shows that the Kraken segmentation output agrees well with the text lines on the first printed page. Most lines are correctly bounded, but there are still some problems due to degraded glyphs, bleed-through, or uneven line spacing. These flaws show how important it is to have strong post-correction strategies in downstream OCR processing.

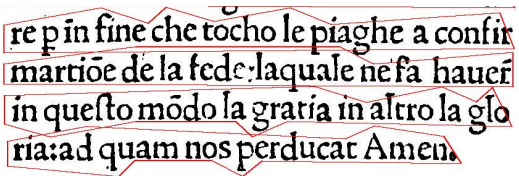

Figure 2. Segmented lines using Kraken.

3.3. OCR Correction with ByT5

Recent advances in language modeling have shown promise for historical text correction, particularly using models such as ByT5 [8,26] and large language models tailored to cultural heritage [9–11]. We employed the `google/byt5-small` model in a sequence-to-sequence framework to correct OCR errors in early printed Latin texts. The training data consisted of line-level pairs where the input was the noisy OCR output, and the target was its manually verified transcription. This supervised learning approach which is shown in Figure 3 allowed the model to learn transformations from erroneous text to corrected output, capturing both character-level noise and orthographic variation.

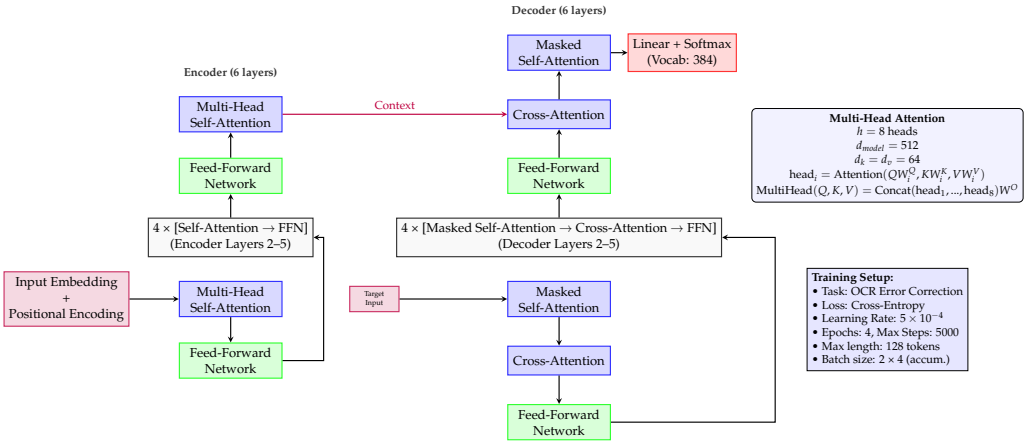

Figure 3. Simplified ByT5 architecture for OCR correction. The model uses a 6-layer encoder–decoder transformer with multi-head attention and feed-forward networks. Input text is processed through byte-level embeddings, encoded bidirectionally, and decoded autoregressively with cross-attention to generate corrected output.

3.3.1. Neural Architecture

ByT5-small is a transformer-based encoder–decoder model with six encoder and six decoder layers. Each layer includes multi-head attention with eight heads (64-dimensional per head, given the 512 hidden size) and a 2048-dimensional feed-forward subnetwork, connected through residual connections and layer normalization. The encoder applies byte-level embeddings plus positional encodings to generate context-rich representations of OCR inputs. The decoder mirrors this structure while introducing masked self-attention and encoder–decoder cross-attention for autoregressive generation.

3.3.2. Training Strategy

We used the Hugging Face Transformers library to fine-tune ByT5 on historical OCR data. The model was trained with cross-entropy loss using teacher forcing, with targets shifted as decoder input. To ensure generalization, we applied early stopping and checkpointing based on the validation Character Error Rate (CER), a metric widely adopted in OCR evaluation [7]. The maximum sequence length was 128, and we used gradient accumulation with an effective batch size of 8. The model's byte-level nature allows it to learn fine-grained transformations, which is critical for correcting early print inconsistencies such as long-s/s substitutions, ligatures, or variant spellings.

The integration of ByT5 substantially improved the quality of the OCR output, reducing error rates and producing more readable, analyzable text. This postcorrection step is essential for downstream digital humanities tasks, including indexing, search, and semantic analysis.

3.4. Evaluation Metrics

To evaluate the effectiveness of the OCR postcorrection module, we employed two widely recognized metrics: *Character Error Rate (CER)* and *BLEU (Bilingual Evaluation Understudy)*.

CER quantifies the number of insertions, deletions, and substitutions required to transform the predicted sequence into the reference transcription. It is formally defined as:

$$\text{CER} = \frac{D(p, g)}{|g|} \tag{1}$$

where $D(p, g)$ denotes the Levenshtein edit distance between the predicted string p and the ground-truth string g, and $|g|$ is the total number of characters in the ground-truth. *CER* is particularly suited for OCR tasks due to its sensitivity to character-level distortions, which are common in historical document transcription [7]. The *BLEU* score, originally introduced for machine translation, measures the overlap of n-grams between the system output and the reference transcription. It is calculated using the following formulation:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2}$$

where p_n is the modified precision for n-grams, w_n are weights (typically uniform), and BP is the brevity penalty that penalizes overly short translations. While *CER* captures the accuracy at the character level, *BLEU* complements it by assessing the fluency and sequence-level similarity between predicted and ground-truth lines [27].

We computed both *CER* and *BLEU* on a held-out validation set composed of historical line images not used during training. Metrics were evaluated after each training checkpoint to monitor the convergence and generalization of the model. The combination of character- and sequence-level metrics ensures a balanced assessment of both transcription accuracy and linguistic quality.

3.5. Dataset

The early printed books used for OCR and postcorrection tasks originate from the MAGIC project website https://www.magic.unina.it (accessed on 1 June 2025), which provides open access to digitized manuscripts. Our training data were created by aligning OCR outputs with manually verified transcriptions derived from these sources. These books, representative of early modern printing, exhibit a wide range of typographic features, printing techniques, and non-standardized orthographic conventions typical of incunabula. The presence of varied typefaces, frequent abbreviations, and inconsistent spelling patterns poses significant challenges for both optical character recognition and subsequent text

correction. Figure 4 shows several examples of cropped text lines extracted from full-page scans using Kraken's polygonal segmentation. These line images constitute the core input to the OCR model and the ByT5-based correction stage. The line-level granularity ensures more consistent recognition and correction, particularly in documents with complex layouts and significant typographic variation.

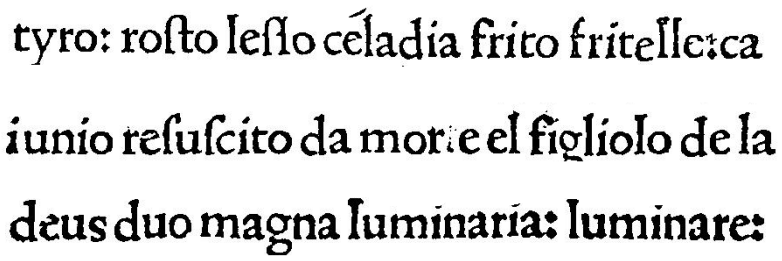

Figure 4. Examples of cropped line images used as inputs for OCR and correction.

The dataset was structured as a CSV file, with each row representing one line of text taken from a scanned page. A segment of the dataset is shown in Table 1, Each entry includes a unique **ID**, like **0031_031_line_10**, that shows where the line is in the book. The second column has the raw OCR output from Kraken's recognition model, which is the noisy input that needs to be fixed. The third column has the transcription of the line that has been manually checked and made standard. This is the ground truth. The dataset was divided into three parts for training and testing the model: 80% for training, 10% for validation, and 10% for testing. This structure makes sure that the model's learning process is strong and can be used in many different situations. It also gives a good way to judge how well it performs on data it has not seen before.

Table 1. Segment of the dataset used for fine-tuning ByT5.

ID	OCR Prediction	Ground Truth
0031_031_line_01	est in principio erat verbum	est in principio erat verbum
0031_031_line_02	et verbum erat apud deũ	et verbum erat apud deum
0031_031_line_03	et deus erat verbum	et deus erat verbum

3.5.1. Preprocessing

Before training, a number of preprocessing steps were taken to ensure that the data were consistent and of high quality. First, a basic but effective preprocessing step was applied to all scanned page images: grayscale inversion, which is shown on Figure 5. This process enhances the visual contrast between the text and the background, which is especially important in early printed books where ink bleed-through, faded characters, or uneven page coloration often hinder accurate recognition. Using OpenCV, each RGB image was first converted to a single-channel grayscale image and then inverted using a pixel-wise transformation. Given a grayscale image I defined over pixel coordinates (x, y) with intensity values in the range $[0, 255]$, the inverted image I' is computed as

$$I'(x, y) = 255 - I(x, y) \tag{3}$$

This operation effectively reverses the luminance values: dark regions (e.g., text strokes) become bright, and light regions (e.g., parchment background) become dark. This inversion enhances edge detection and region proposal accuracy during Kraken's segmentation stage.

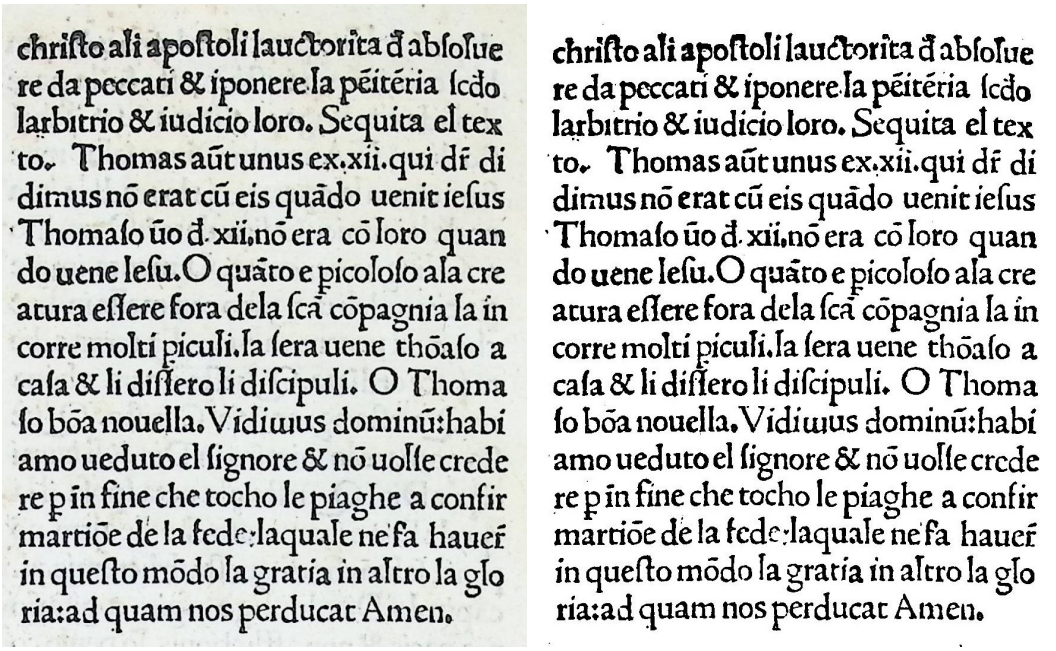

Figure 5. Comparison of an early printed page before and after grayscale inversion. The inverted image improves contrast for segmentation and OCR. (Page 64v of *Sermones Quadragesimales* by Roberto Caracciolo, printed in Venice by Thomaso de Alexandria in 1485; copy held at the Biblioteca dei Girolamini, Naples, shelfmark 4B1).

The inverted images were subsequently used for both layout analysis and line extraction. The improvement in segmentation consistency and OCR accuracy resulting from this simple transformation highlights the importance of preprocessing in document image pipelines, especially when dealing with degraded or historically complex materials. Then, all text data were normalized to Unicode to make sure that all characters were encoded the same way. This eliminated any inconsistencies caused by OCR artifacts or different character forms. This step was very important for getting historical texts, which often have diacritics and old glyphs, into a clean and consistent format. Next, a filtering process was used to remove lines that were too short, incomplete, or corrupted, which could have added noise to model training. The ByT5 model works at the byte level and does not need traditional tokenization, but it was still important to make sure that all text samples had the same and correct character encoding. These steps helped to make the training process more stable and effective, especially when working with the noisy and uneven text of early printed books.

3.5.2. Dataset Statistics

The dataset, summarized in Table 2, captures many of the distinctive challenges associated with OCR in early printed texts. Comprising over 10,000 line pairs, it offers a substantial volume of training and evaluation material for fine-tuning transformer-based correction models. Each line contains an average of 55 characters, providing sufficient sequence length to model typical errors, including character insertions, deletions, and substitutions that arise from OCR misinterpretation of degraded or ornate typography. With a UTF-8 vocabulary size of 256 bytes, the dataset supports fine-grained modeling at the byte level, which is particularly advantageous when using ByT5, as it enables the model to handle a wide range of typographic and orthographic variations without relying on token-based preprocessing. The corpus includes both Latin and early Italian texts, reflecting the multilingual nature of early modern European printing. This linguistic diversity adds an additional layer of complexity, making the dataset a robust and realistic

benchmark for evaluating OCR correction systems. Moreover, it includes typical issues such as ligatures, irregular spacing, ink bleed-through, and archaic spelling, all of which demand sophisticated correction strategies. As such, the dataset not only supports rigorous quantitative evaluation but also enables the development of OCR correction tools that generalize well to a variety of historical texts.

Table 2. Dataset.

Attribute	Value
Total line pairs	10,000+
Average characters per line	55
Vocabulary size (UTF-8)	256 (bytes)
Languages	Latin, Italian

4. Experiments and Results

Using the dataset described above, we fine-tuned the `google/byt5-small` transformer model to learn a sequence-to-sequence mapping from noisy OCR predictions to accurate ground-truth transcriptions. All experiments were conducted on a workstation, with components sourced from various manufacturers and assembled in our laboratory in Naples, Italy, equipped with an Intel Core i9 processor, 64 GB of RAM, and an NVIDIA RTX A4000 GPU (16 GB VRAM), running under a Linux-based environment. This setup provided sufficient computational resources to train the model efficiently without requiring distributed or cloud-based infrastructure.

4.1. Training Configuration

We used the Hugging Face Transformers and Datasets libraries to fine-tune the ByT5-small model for the OCR post-correction task. The goal of the training was to reduce the token-level cross-entropy loss between the model output and the correct transcription. Using the AdamW optimizer with a learning rate of 5×10^{-5}, the training was carried out. We trained the model for about five epochs, using early stopping based on validation loss to keep it from overfitting. To find a good balance between speed and stability, a batch size of 16 was used. This setup was chosen because it is simple and works well on small- to medium-sized datasets with a lot of noise at the character level, like those found in early printed texts. At regular intervals, training and validation losses were documented (see Table 3). To evaluate performance on unseen data, we also calculated evaluation metrics such as the BLEU score and Character Error Rate (CER) (Table 4).

Table 3. Training curves.

Step	Training Loss	Validation Loss
250	73%	68%
1000	43%	54%
2000	36%	49%
3000	29%	48%
4000	21%	50%
5000	18%	52%

Table 4. Metrics.

Timestamp	CER	BLEU
Start	19.9%	30.6%
Midpoint	16.0%	40.8%
Final	15.1%	44.8%

4.2. Training Curves

The model was trained in 5000 steps using a dataset made up of pairs of line-level images and transcriptions from early printed books. To keep an eye on convergence and possible overfitting, both training and validation losses were recorded at regular intervals during training.

During the first 3000 steps, the model's training and validation loss both went down steadily and clearly, which meant that it was learning and generalizing well. The validation loss started to level off slightly after 3000 steps, but the training loss continued to decrease. This behavior is commonly observed and indicates that the model has reached the limits of what it can learn from the current dataset. The validation loss stayed between 49% and 52% from step 3000 to step 5000, which is interesting because it means that the model did not overfit even though it was training. This is especially important because early printed book scans are often very complicated and noisy.

4.3. Evaluation

We used both qualitative and quantitative methods to assess the effectiveness of the OCR correction model. Qualitative analysis showed that the ByT5 model corrected a wide range of OCR errors, making the text more readable and semantically accurate. Table 5 shows that many corrections addressed characters distorted over time or outdated spellings, and the model was able to restore these with high fidelity. This level of correction is essential to ensure that digital transcriptions remain accessible and useful for research and long-term preservation. The model's final Character Error Rate (CER) was 15.1%, which is considered strong performance given the degraded print quality and irregular typography of early printed materials. During the training, the BLEU score increased steadily, reaching 44.8%. This indicates improved alignment between predicted outputs and ground-truth transcriptions. These results are particularly noteworthy, as no explicit rule-based or hand-crafted post-processing was used; improvements stem entirely from learned byte-level corrections. This confirms the potential of our fine-tuned ByT5 model as an effective OCR correction module within document processing pipelines for historical texts.

The final results demonstrate a substantial correction of the OCR errors. Given the challenges posed by early printed texts, These values represent a strong outcome, especially given the typographic and orthographic variability of early printed materials.

Table 5. Correction examples.

Ground Truth	OCR Prediction	Corrected Output
impressionem	fmpenssorüel'	impressionem
denunciabat	do&nunçabat	denunciabat
disciplinae	di€$ciplinae	disciplinae

4.4. Dataset Description

The dataset used in this investigation is derived from high-resolution scans of incunabula, books printed before 1501, curated within the MAGIC Project's digital archive. Kraken, an OCR toolkit that can segment lines, was used to process each page. We manually

checked the segmentation to make sure that each line of text was correct. Then, we used those lines to make our training and evaluation datasets. We collected three things for each line in the dataset: a cropped grayscale image at 300–600 DPI, the OCR output of Kraken's default pretrained recognition model, and a transcription that had been manually checked for accuracy. This structure helped us think of the task as a sequence-to-sequence correction problem, where the model learns how to turn inaccurate OCR outputs into correct transcriptions. This study used two sets of data. The first dataset, which was used in the first experiments, had about 2000 samples. After getting good results, a bigger dataset with about 5000 annotated lines was put together and used to train and test the final model. The dataset has a lot of problems that are specific to its field. These include things like non-standard spelling, like Latin abbreviations and ligatures; degradation phenomena, like ink bleed-through and faded strokes; and historical typographic conventions, like the long "s" that looks a lot like the modern "f." To reduce alignment errors, all OCR outputs were checked and cleaned by hand to make sure they matched the ground truth.

4.5. Comparative Evaluation with Existing Methods

We compare our proposed pipeline to two recent high-performing approaches designed specifically for OCR on early printed materials. Table 6 presents a summary of the results from Reul et al. [21] and Seuret et al. [28], based on published Character Error Rates (CER) and degree of automation.

Table 6. Comparison of OCR methods on early printed books.

Method	Dataset Type	CER	Automated
Reul et al. (2018) [21]	Early Latin/German/Dutch	1.1–2.5%	No
Seuret et al. (2023) [28]	Early modern German	2.4–5.6%	No
Ours (Kraken + ByT5)	Latin incunabula (15th c.)	15.1%	Yes

Reul et al. [21] achieved outstanding OCR accuracy on early printed books using a labor-intensive pipeline involving pretraining on 1000 manually transcribed lines per book, five-fold model training, and ensemble voting with active learning. Their experiments span historical texts in Latin, German, and Dutch, and report Character Error Rates (CERs) as low as 1.1%. However, their approach relies heavily on curated ground truth and repeated training runs, which make it less accessible for small-scale or low-resource projects. Seuret et al. [28] proposed an ensemble method that combines outputs from Kraken, Calamari, Ocropus, and Tesseract using ISRI voting to improve OCR robustness on early modern German documents. Their evaluation was conducted on three datasets: Confessionale (1118 lines), Praktik (377 lines), and Augsburger Postzeitung (537 lines), for a total of over 2000 ground-truth lines. Their ensemble achieved CERs ranging from 2.4% to 5.6%, depending on the combination and alignment of models. While effective, this strategy requires managing multiple OCR systems, performing output alignment, and using voting mechanisms that add engineering complexity.

In contrast, our pipeline integrates Kraken-based layout-aware segmentation with a lightweight ByT5 correction model in a fully automated manner following minimal fine-tuning on fewer than 500 manually aligned line-level samples. Applied to Latin incunabula—arguably more degraded and typographically irregular than the sources used in the above studies—our system achieved a CER of 15%. Although our absolute accuracy is lower, the method is significantly more lightweight, scalable, and reproducible. It does not require ensemble modeling, voting mechanisms, or access to thousands of ground-truth lines. This makes our approach particularly well suited for small-scale digitization efforts,

especially in the cultural heritage domain, where ground truth is scarce and fast, and automatic deployment is often prioritized over absolute error minimization.

5. Discussion

Our results show that ByT5, a byte-level transformer model, is very good at fixing OCR errors in early printed books, especially when combined with Kraken for layout segmentation and initial recognition in a modular pipeline. The character error rate (CER) was significantly reduced after correction, from more than 35% in the baseline OCR to about 15%. This shows that the model can learn intricate mappings between accurate transcriptions and noisy OCR output. Given the physical deterioration, outdated typefaces, and orthographic irregularities that characterize the incunabula, this improvement is especially pertinent. The idea that sequence-to-sequence models like ByT5 can work well as "denoisers" for historically degraded text is supported by the concurrent rise in BLEU scores, which further suggests improved sequence-level fluency and coherence. Although the model performs well on the available dataset, it may not be able to generalize to previously unseen content, such as books with radically different fonts, layouts, or languages. Our dataset may not fully capture the typographic diversity of early print, despite capturing a representative sample of incunabula. Furthermore, because it takes a lot of time and effort to create manually verified ground-truth transcriptions, this approach presents scalability issues. To lessen this dependence, future research may investigate unsupervised or semi-supervised learning techniques. Because our pipeline is modular, each step, segmentation, recognition, and correction, can be changed or enhanced separately, making it suitable for a variety of digitization projects and OCR engines. By using byte-level correction instead of word- or character-level tokenization, the pipeline is more applicable to multilingual and morphologically rich corpora and requires less language-specific preprocessing. In the end, incorporating precise post-OCR correction into historical digitization processes is essential to improve access, searchability, and scholarly analysis of cultural heritage materials, in addition to improving the quality of textual data for subsequent natural language processing. Our method contributes significantly to the infrastructure of digital humanities and the long-term preservation of early printed texts by bridging the gap between raw scanned documents and usable digital editions.

While the overall results are promising, we acknowledge that the model's effectiveness can vary depending on the book's layout, typographic style, and degradation level. As observed in our sample outputs, some corrections were partial or inaccurate, indicating that further refinements—such as book-specific fine-tuning or more robust pretraining strategies—may be necessary for consistent performance across diverse early printed materials. However, this variability does not diminish the significance of our findings: the proposed pipeline operates fully automatically, requires only a small number of aligned lines for training, and achieves a meaningful reduction in CER even under challenging conditions. We believe that this trade-off between automation and absolute accuracy is appropriate for low-resource digitization workflows, and that future extensions, including document adaptation and self-supervised learning, may further improve robustness. In this regard, our method constitutes an important step toward scalable OCR correction infrastructure for historical texts. Despite some variability across documents, our pipeline demonstrates that even modest ground truth and automated correction can yield usable outputs, supporting real-world applications in historical OCR without requiring extensive manual intervention.

6. Conclusions and Future Work

In this study, we introduced a modular pipeline for text recognition in early printed books, integrating Kraken for segmentation and OCR with ByT5 for the rectification of OCR inaccuracies. Our findings indicate that the ByT5 model, meticulously calibrated on byte-level input and output, can acquire intricate corrections despite the interference of inaccurate OCR predictions from historical documents.

The pipeline made big strides in both Character Error Rate (CER) and BLEU scores, with a final CER of about 15% and a BLEU score of over 44%. This shows that the model is strong and useful in real life. These results are especially important because early printed texts had problems like typographical variation, bleed-through, and degraded glyphs.

Our contribution provides a comprehensive solution designed for incorporation into extensive digitization workflows. It improves the quality of transcriptions used in digital libraries, search engines, and scholarly editions. This helps both digital humanities research and the preservation of cultural heritage. There are several directions in which this work can be extended: Multilingual and Cross-Domain Generalization: ByT5's usefulness and generalizability could be improved by fine-tuning and testing it on datasets from other historical domains, languages, or typographic styles. Pretraining with Synthetic Noise: Adding a lot of synthetically generated OCR errors to old texts could make the model more robust, especially when there are not a lot of resources available. Interactive Correction Interfaces: By adding the correction pipeline to transcription tools that are easy to use, scholars and volunteers could work together to improve OCR outputs over time. Hybrid Correction Systems: Looking into how to combine statistical and neural methods might help, especially when dealing with old or unusual glyphs. Our goal is to support scalable and accurate digitization efforts by continuing to improve and expand this pipeline. This will make it easier for more people to access the rich textual legacy of early printed books. In summary, the proposed pipeline achieved a relative reduction in CER of over 50% compared to the raw OCR output and nearly doubled the BLEU score, demonstrating the effectiveness of combining layout-aware segmentation and byte-level transformer correction. These improvements confirm the feasibility of applying modern sequence-to-sequence models to historical OCR correction, especially in the context of early printed books.

Author Contributions: Y.M., L.L. and G.R. contributed equally to all scientific, technical, design and intellectual aspects of the work and in the preparation, editing and drafting of the article. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the MAGIC project , funded by the Ministry of Enterprise and Made in Italy, for the creation of a prototype of a service center for technologies applied to cultural heritage. Specifically, it is designed for the treatment of manuscripts, documents, ancient and printed texts that aim at the protection and conservation of cultural heritage, as well as its enhancement and use. The project was funded by the Ministry of Enterprise and Made in Italy, (code n. F/130093/03/X38 and CUP: B69J23000560005) and by the Ministry of University and Research (code n. PIR01_00011, CUP: I66C18000100006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used and referenced in this work are publicly available on the official website of the MAGIC project: https://www.magic.unina.it (accessed on 1 June 2025).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Russo, G.; Aiosa, L.; Alfano, G.; Chianese, A.; Cornevilli, F.; Di Domenico, G.M.; Maddalena, P.; Mazzucchi, A.; Muraglia, C.; Russillo, F.; et al. MAGIC: Manuscripts of Girolamini in Cloud. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *949*, 012081. [CrossRef]
2. Conte, S.; Maddalena, P.M.; Mazzucchi, A.; Merola, L.; Russo, G.; Trombetti, G. The Role of Project MA.G.I.C. in the Context of the European Strategies for the Digitization of the Library and Archival Heritage. In *Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage*; Bucciero, A., Fanini, B., Graf, H., Pescarin, S., Rizvic, S., Eds.; The Eurographics Association: Lecce, Italy, 2023. [CrossRef]
3. Ettari, A.; Brescia, M.; Conte, S.; Momtaz, Y.; Russo, G. Minimizing Bleed-Through Effect in Medieval Manuscripts with Machine Learning and Robust Statistics. *J. Imaging* **2025**, *11*, 136. [CrossRef] [PubMed]
4. Mori, S.; Suen, C.Y.; Yamamoto, K. Historical Review of OCR Research and Development. *Proc. IEEE* **1992**, *80*, 1029–1058. [CrossRef]
5. Smith, R. An Overview of the Tesseract OCR Engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil, 23–26 September 2007; pp. 629–633. [CrossRef]
6. Reul, C.; Christ, D.; Hartelt, A.; Balbach, N.; Wehner, M.; Springmann, U.; Wick, C.; Grundig, C.; Büttner, A.; Puppe, F. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Preprints* **2019**. [CrossRef]
7. Springmann, U.; Reul, C.; Dipper, S.; Baiter, J. Ground Truth for Training OCR Engines on Historical Documents in German Fraktur and Early Modern Latin. *arXiv* **2018**, arXiv:1809.05501. [CrossRef]
8. Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; Raffel, C. ByT5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models. *arXiv* **2022**, arXiv:2105.13626. [CrossRef]
9. Bu, F.; Wang, Z.; Wang, S.; Liu, Z. An Investigation into Value Misalignment in LLM-Generated Texts for Cultural Heritage. *arXiv* **2025**, arXiv:2501.02039. [CrossRef]
10. Trichopoulos, G. Large Language Models for Cultural Heritage. In *CHIGREECE 2023: Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter, Athens, Greece, 27–28 September 2023*; ACM: New York, NY, USA, 2023; pp. 1–5. [CrossRef]
11. Hwang, H.; Park, C.W.; Kim, H.K.; Lee, J.-H. CATS: Cultural-Heritage Classification using LLMs and Distribute Model. *npj Herit. Sci.* **2025**, *13*, 76. [CrossRef]
12. Cossatin, A.G.; Mauro, N.; Ferrero, F.; Ardissono, L. Tell Me More: Integrating LLMs in a Cultural Heritage Website for Advanced Information Exploration Support. *Inf. Technol. Tour.* **2025**, *27*, 385–416. [CrossRef]
13. Mantas, J. An Overview of Character Recognition Methodologies. 1986. Available online: https://www.sciencedirect.com/science/article/pii/0031320386900403 (accessed on 10 June 2025).
14. Nguyen, T.T.H.; Jatowt, A.; Coustaty, M.; Doucet, A. Survey of Post-OCR Processing Approaches. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 124. [CrossRef]
15. Li, Q.; An, W.; Zhou, A.; Ma, L. Recognition of Offline Handwritten Chinese Characters Using the Tesseract Open Source OCR Engine. In Proceedings of the 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 27–28 August 2016. [CrossRef]
16. Löfgren, V.; Dannélls, D. Post-OCR Correction of Digitized Swedish Newspapers with ByT5. 2024. Available online: https://aclanthology.org/2024.latechclfl-1.23/ (accessed on 5 June 2025).
17. Reul, C.; Dittrich, M.; Gruner, M. Case Study of a Highly Automated Layout Analysis and OCR of an Incunabulum: 'Der Heiligen Leben' (1488). In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2017), Göttingen, Germany, 1–2 June 2017*; ACM: New York, NY, USA, 2017; pp. 155–160. [CrossRef]
18. Alberti, M.; Vögtlin, L.; Pondenkandath, V.; Seuret, M.; Ingold, R.; Liwicki, M. Labeling, Cutting, Grouping: An Efficient Text Line Segmentation Method for Medieval Manuscripts. *arXiv* **2019**, arXiv:1906.11894
19. Binmakhashen, G.M.; Mahmoud, S.A. Document Layout Analysis: A Comprehensive Survey. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 109. [CrossRef]
20. Benjamin Kiessling. Kraken—A Universal Text Recognizer for the Humanities. In Proceedings of the Digital Humanities 2019, Utrecht, The Netherlands, 9–12 July 2019. [CrossRef]
21. Reul, C.; Springmann, U.; Wick, C.; Puppe, F. Improving OCR Accuracy on Early Printed Books by Combining Pretraining, Voting, and Active Learning. *arXiv* **2018**, arXiv:1802.10038. [CrossRef]
22. Sun, B.; Li, S.; Sun, J. Blind bleed-through removal for scanned historical document image with conditional random fields. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015. [CrossRef]
23. Muenster, S. Digital 3D Technologies for Humanities Research and Education: An Overview. *Appl. Sci.* **2022**, *12*, 2426. [CrossRef]
24. Leydier, Y.; LeBourgeois, F.; Emptoz, H. Textual Indexation of Ancient Documents. In *Proceedings of the 2005 ACM Symposium on Document Engineering (DocEng'05), Bristol, UK, 2–4 November 2005*; ACM: New York, NY, USA, 2005; pp. 111–117. [CrossRef]
25. Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Li, Z.; Wei, F. TrOCR: Transformer-Based Optical Character Recognition with Pretrained Models. 2022. Available online: https://arxiv.org/abs/2109.10282 (accessed on 18 July 2025).

26. Garst, P.; Ingle, R.; Fujii, Y. OCR Language Models with Custom Vocabularies. *arXiv* **2023**, arXiv:2308.09671. Available online: https://arxiv.org/abs/2308.09671 (accessed on 20 July 2025). [CrossRef]
27. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *ACL'02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002. [CrossRef]
28. Seuret, M.; van der Loop, J.; Weichselbaumer, N.; Mayr, M.; Molnar, J.; Hass, T.; Christlein, V. Combining OCR Models for Reading Early Modern Books. In *Document Analysis and Recognition—ICDAR 2023. ICDAR 2023*; Fink, G.A., Jain, R., Kise, K., Zanibbi, R., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 14191. [CrossRef]