Semi-automatic pipeline for constructing HTR corpora from Ukrainian-language historical documents*

Andrii Ivasechko[1,*,†] and Khrystyna Lipianina-Honcharenko[1,†]

[1] *Faculty of Computer Information Technologies, West Ukrainian National University, 46000 Ternopil, Ukraine*

Abstract

Recognition of historical manuscripts is a challenging task due to multilingualism, non-standard spelling, and variability in writing styles, which limits the effectiveness of traditional OCR systems, especially for low-resource languages. This study presents an integrated system for creating a dataset for deep learning that combines automated preprocessing, character segmentation, and a collaborative interface for annotation. Based on documents from the State Archive of Khmelnytskyi Region, 1684 validated entries with 215 unique symbols in two languages were created. The platform proved user-friendly for untrained users — 48 students participated in collaborative labeling, ensuring high annotation quality. The challenges of segmentation, data variability, and the prospects for expanding the corpus and implementing active learning are discussed.

Keywords

historical manuscripts, handwritten text recognition, dataset creation, image segmentation, character annotation, cultural heritage

1. Introduction

This article presents a semi-automatic pipeline for constructing HTR corpora from Ukrainian-language historical documents. Chapter 2 provides a review of related work. Chapter 3 outlines the system architecture, including the dataset loader, pre-processing procedures, labeling interface, quality aggregation algorithm, and evaluation metrics. Chapter 4 presents the testing results, describes the dataset, and reports on annotation quality based on the defined metrics.

Automatic Handwritten Text Recognition (HTR) is a key component in the digital transformation of historical documents and cultural heritage sources. This issue is particularly relevant in the context of the humanities, where the preservation, analysis, and reuse of archival manuscripts are fundamental tasks. Unlike modern printed texts, manuscripts from the 14th to 19th centuries feature high variability in handwriting, multilingualism, non-standard spelling, rare fonts, and significant degradation of the media. These factors significantly limit the effectiveness of traditional Optical Character Recognition (OCR) systems designed for modern printed texts.

One of the major challenges in the field of deep learning for HTR is the limited availability of high-quality training data, particularly for low-resource languages and historical fonts. Creating such corpora requires substantial human resources and time. To address this issue, recent studies propose methods such as semi-automated labeling, active learning, synthetic data generation, and transformer architectures that rely less on large labeled datasets. At the same time, character segmentation in irregular historical handwriting remains a critical step that requires further improvements[1].

This study presents the development of an interactive system for constructing handwritten text corpora, combining automated image preprocessing, segmentation, a collective character labeling interface, and annotation storage in a format suitable for further training deep learning models.

* *BAITmp'2025: The 2nd International Workshop on "Bioinformatics and Applied Information Technologies for medical purpose", November 12-13, 2025, Ben Guerir, Morocco*

[1*] Corresponding author.

[†] These authors contributed equally.

andrewivasechko@gmail.com@tuni.fi (A. Ivasechko); kh.lipianina@wunu.edu.ua (K. Lipianina-Honcharenko)

0009-0002-8623-7828 (A. Ivasechko); 0000-0002-2441-6292 (K. Lipianina-Honcharenko)

The system was tested on digitized multilingual documents from the State Archive of Khmelnytskyi Region, resulting in over 1,600 validated characters and creating a foundation for future text recognition.

In this context, the paper analyzes the dataset creation process, the structure of the annotation interface, the challenges of segmentation quality, and the organization of group collaboration to enhance labeling reliability. The proposed approach can be scaled for other historical corpora and serves as a valuable tool for researchers in digital humanities, automated manuscript recognition, and the creation of intelligent access systems for cultural heritage.

This article presents a semi-automatic pipeline for constructing HTR corpora from Ukrainian-language historical documents. Chapter 2 provides a review of related work. Chapter 3 outlines the system architecture, including the dataset loader, preprocessing procedures, labeling interface, quality aggregation algorithm, and evaluation metrics. Chapter 4 presents the testing results, describes the dataset, and reports on annotation quality based on the defined metrics.

2. Related works

The goal of this study is to develop an interactive system for constructing handwritten text corpora, which combines automated image preprocessing, segmentation, a collective character labeling interface, and annotation storage in a format suitable for further training deep learning models. To justify the design of such a system and identify knowledge gaps, this section summarizes related works in six directions: (i) the availability and quality of multilingual resources and the consequences of their scarcity for low-resource languages, including Ukrainian [1]; (ii) end-to-end HTR pipelines for historical documents (layout recognition, segmentation, transcription) [2]; (iii) synthetic data generation and augmentation to enhance models in the absence of labeling [3–7]; (iv) interactive and semi-supervised approaches (active/self-training) to reduce the cost of manual annotation [8–9]; (v) the capabilities and limitations of modern MLLMs and foundation models for manuscripts and related tasks [10–14]; (vi) methods for targeted collection, structuring, and tool support for annotations compatible with subsequent ML training [15–20]. This systematization serves as the methodological foundation for making engineering decisions in the proposed system (automation of preprocessing and segmentation, "human-in-the-loop" for symbols, interoperable storage formats), aimed at rapidly building high-quality corpora for further deep learning.

Yu et al. [1] conduct a quantitative and qualitative analysis of multilingual NLP resources, covering 156 public datasets, manually annotating text sources and annotations, creation tools, tasks, and motivations. The researchers show that simply counting datasets is misleading due to the predominance of automatically generated and English-translated corpora, and they identify a correlation between experts' and crowdworkers' assessments of data availability and the actual existence of resources. The authors' crowdsourcing experiments lead to practical recommendations for collecting high-quality multilingual data for languages with limited corpora. In their classification, Ukrainian is categorized as a low-resource language, providing a methodological basis for justifying the scarcity of Ukrainian-language corpora and planning data collection, particularly for tasks like fake news detection.

The iForal study [2] focuses on automating the transcription of historical manuscripts to facilitate access to cultural heritage. The developed system includes layout recognition, segmentation, and transcription of text. A corpus of 67 Portuguese charters was used for training. The system achieved an accuracy of 0.98 mAP@0.50 for layout, 0.91 mAP@0.50 for segmentation, and 8.1% CER. It reduces the need for expert intervention and supports adaptation to other writing styles through transfer learning. The dataset has been published in HTR United, and it is also made available in the HTR United catalog for reuse.

The study by Lisa Koopmans [3] is dedicated to the automated dating of historical manuscripts using SVM analysis of texture and grapheme features. To address the issue of limited labeled data, data augmentation was applied, resulting in a 1–3% increase in accuracy. The models were tested on several corpora, including the Medieval Paleographical Scale. The authors note the potential for adaptation to specific handwriting styles to improve accuracy.

Lars Vögtlin's work [4] presents a framework for generating synthetic historical documents with accurate labeling based on unlabeled images. A two-step approach is proposed: creating templates with controlled content and transferring the style of historical scans. This method ensures realism and accuracy without expert involvement. Pre-training on generated data outperforms baseline models, opening the way for creating scalable datasets for low-resource languages and rare writing systems.

Wei Chen [5] proposes the Fine-grained Automatic Augmentation (FgAA) method to improve handwritten text recognition for languages with limited sample data. Unlike traditional approaches that perform global transformations on words, FgAA operates at the level of individual strokes: each word is segmented into strokes, approximated by Bézier curves, and local transformations are applied. Optimal augmentation parameters are automatically selected using Bayesian optimization.

In the work by Arthur Flor de Souza Neto [6], a systematic review of data augmentation methods for offline Handwritten Text Recognition (HTR) systems is presented, which are crucial for improving model quality when labeled data is limited. The review analyzes 32 relevant studies from 976 found in databases from 2012 to 2023. The authors note that traditional Digital Image Processing (DIP) is still widely used, although recent interest in Generative Adversarial Networks (GANs) is increasing, allowing for the synthesis of handwritten text with arbitrary style and content. The paper also discusses the datasets used and recognition levels in the studies.

In Yahia Hamdi's article [7], four data augmentation strategies to improve online handwritten text recognition (OHR) with small datasets are presented. The proposed approaches include: (1) geometric transformations (italic angle, scale, tilt), (2) frequency processing of trajectories, (3) beta-elliptical modeling of writing dynamics, and (4) a hybrid combination of all strategies. The system was tested on multilingual datasets (Arabic: ADAB, ALTEC-OnDB, Online_KHATT; Latin: UNIPEN) using CNN architecture. Results demonstrate significant improvements in recognition accuracy compared to baseline and contemporary approaches.

In the study by Alejandro Héctor Toselli [8], an interactive system for transcribing historical manuscripts is proposed, which continuously fine-tunes based on user-validated results. The goal is to reduce user interactions while improving their efficiency. Three approaches are considered: adaptation through semi-supervised learning, active learning for selecting ambiguous examples, and error probability assessment for regulating user intervention. Experiments on two historical documents confirm the effectiveness of the approach.

In Fabian Wolf's study [9], a self-learning method for Handwritten Text Recognition (HTR) and word search is proposed, which eliminates the need for manual annotation. The baseline model is trained on synthetic data, after which it generates pseudo-labels for real images and is further trained on them in a semi-supervised manner. To improve accuracy, mechanisms for filtering unreliable pseudo-annotations are applied. The proposed approach demonstrates better accuracy and stability compared to other annotation-free methods.

In Shukang Yin's article [10], a review of the development of multimodal large language models (MLLMs), specifically GPT-4V, is presented, which combine language processing with image, video, and text analysis. The paper discusses architectures, training strategies, types of data, and evaluation metrics, as well as challenges, including the issue of multimodal hallucinations. The authors analyze the prospects for expanding MLLMs to new modalities, languages, and application scenarios, particularly through M-ICL, M-CoT, and LAVR. The review is accompanied by an open GitHub repository with current research.

In the study by Jacob Murel and David Smith [11], a method for improving the detection of handwritten annotations in early printed books based on visual similarity between text samples is proposed. The authors explore the impact of pseudo-labeled page images on the performance of manuscript localization models, using pages from copies of Shakespeare's "First Folio." Self-learning and active learning approaches with pseudo-labels for both positive and negative examples are compared. The results show a 15% improvement in average accuracy for individual copies, although the effectiveness on collections from multiple sources was less conclusive.

Carina Geldhauser and Konstantin A. Malyshev [12] introduced a prototype for integrating Handwritten Text Recognition (HTR) and semi-automated annotation of textual features in the graphical interface eScriptorium. The solution is aimed at humanists, particularly researchers of ancient Greek texts (majuscules) who are creating critical editions or digital collections. The prototype allows for simultaneous transcription and annotation, an important step in reconstructing textual variants and facilitating the analysis of manuscript traditions, such as in the study of Homeric or biblical texts.

Giorgia Crosilla, Lukas Klic, and Giovanni Colavizza [13] compare the capabilities of multimodal large language models (MLLMs), such as Claude 3.5 Sonnet, with traditional HTR systems like Transkribus for handwritten text recognition. Unlike classical models, which require significant manual annotation, MLLMs can recognize various handwriting styles without specialized training. Experiments cover modern and historical texts in four languages (English, French, German, Italian). The results demonstrated the advantages of proprietary models in a zero-shot setting, particularly for English, but also revealed the limitations of LLMs in independently correcting transcriptions.

In the study by Li Y. et al. [14], the efficient use of Vision Transformer (ViT) for handwritten text recognition tasks under limited data conditions is proposed. The authors replace the standard patch representation in ViT with features extracted using a convolutional neural network (CNN) and apply the Sharpness-Aware Minimization (SAM) optimizer to achieve stable and generalized loss minimization. The method also introduces span masking, a regularizer that masks related areas in the feature map. The proposed approach demonstrates competitive results on small IAM and READ2016 datasets and sets a new benchmark on the largest LAM dataset (19,830 text lines).

One key study [15] analyzes data collection for machine learning, emphasizing the need for large volumes of annotated data for deep models. The entire data collection cycle is discussed, from acquisition to dataset enhancement, with a focus on integrating Big Data and AI.

The study [16] examines automated data collection (ADC) in the context of Industry 4.0, where IoT devices are used. ADC reduces the workload for users by applying AI to identify relevant data.

The study [17] proposes formalizing the data collection process as an optimization task that minimizes costs while maintaining a balance between the amount of data and model accuracy, especially under semi-supervised learning conditions.

In the study [18], a model for recognizing irregular text in images is proposed, combining ResNet-31, an LSTM encoder-decoder, and an attention mechanism to reduce data preparation costs. This solution demonstrates high effectiveness on test datasets.

The research [19] focuses on the use of synthetic data to reduce the costs of dataset creation. The generation of synthetic images with automatic labeling showed that they can provide performance comparable to real data.

In the work [20], a system for creating annotated datasets based on historical manuscripts is described. The system combines semi-automated character labeling and multi-level verification to enhance data quality. It supports Cyrillic, Latin, and Arabic scripts and utilizes a "human-in-the-loop" approach.

The study [21] proposes an intelligent system for online product promotion, which includes keyword generation, product catalog creation, advertisement content generation, and targeting. The experiment confirmed that the system improves advertising effectiveness by 125% while reducing costs by 87%.

The paper [22] compares three neural networks — ResNet, EfficientNet, and Xception. The models were evaluated for accuracy, sensitivity, specificity, and F1-score. The Xception model achieved the highest accuracy (87.7%), EfficientNet showed high efficiency under limited resources, while ResNet faced challenges with classifying underrepresented classes, highlighting the importance of data balance and training methods.

In the study [23], a method for segmentation of atmospheric cloud images obtained via remote sensing was presented. The authors developed an algorithm to isolate cloud structures in satellite images, demonstrating how classical computer-vision techniques can effectively separate complex visual objects. This approach can be adapted for preprocessing and segmenting handwritten manuscript images.

The paper [24] introduces a model for classifying information objects by combining neural networks with fuzzy logic. This hybrid approach improves classification accuracy in conditions of uncertainty and data heterogeneity. Such techniques can be leveraged to classify symbols or identify languages in multilingual historical documents.

The review revealed that most approaches either rely on large, carefully annotated corpora or are narrowly focused on specific languages/scripts; at the same time, there is a noticeable lack of open, symbol-level standardized resources for Cyrillic and mixed scripts [1–2]. Synthetic data and augmentation can partially compensate for the data shortage [3–7], while interactive/self-learning methods reduce annotation costs [8–9]. However, there is a lack of practical, reproducible solutions that integrate these approaches into a unified workflow with transparent data quality and interoperable formats [10–14, 15–21]. The proposed system directly addresses these gaps: automated preprocessing and segmentation reduce input costs, collective labeling ensures symbol-level quality control, and standardized annotation storage makes the data ready for training deep models and further expansion with synthetic/semi-supervised samples. A pilot test on digitized multilingual documents from the State Archive of Khmelnytskyi Region resulted in over 1,600 validated symbols, confirming the viability of the approach and creating a foundation for the next stage — building and evaluating handwritten text recognition models for Ukrainian and mixed corpora. Thus, the results of the review directly inform the requirements for the system's architecture and functionality, focused on effectively solving the task at hand.

3. System architecture

A system for forming a training dataset, focused on handwritten text recognition tasks from historical manuscripts, has been developed, along with a specialized application and conducted test labeling. The approach includes sequential stages of preprocessing, segmentation, and annotation of character images, ensuring high-quality preparation of input data for training optical character recognition models under conditions of variability and degradation of manuscript sources. The system is schematically presented in Figure 1.

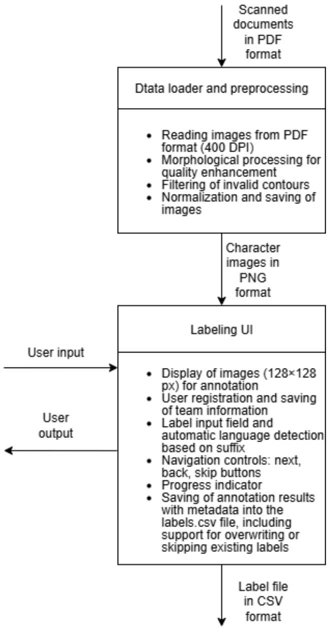

Scanned documents in PDF format
Dtata loader and preprocessing
Reading images from PDF format (400 DPI)
Morphological processing for quality enhancement
Filtering of invalid contours
Normalization and saving of images
Character images in PNG format
Labeling UI
User input
User output
Display of images (128×128 px) for annotation
User registration and saving of team information
Label input field and automatic language detection based on suffix
Navigation controls: next, back, skip buttons
Progress indicator
Saving of annotation results with metadata into the labels.csv file, including support for overwriting or skipping existing labels
Label file in CSV format

Figure 1: System scheme.

3.1. Dataset Loader & Pre-processing

The study used documents obtained from the State Archive of Khmelnytskyi Region. The sample included digitized descriptions of materials from several fonds: Fond 442 (Kamianets-Podilskyi County Treasury, 1861–1913), Fond 507 (Office of the Head of the Southwestern Customs District, 1907–1913), Fond 596 (Podillya Branch of the Princess Tatiana Nikolaevna Committee for Assisting War Victims, 1914–1915), Fond 598 (Judicial Investigation Department of Kamianets-Podilskyi County, 1875–1880), Fond 616 (Military Affairs of Kamianets-Podilskyi County, 1884–1919), and Fond 309 (Isakovets Customs, 1915–1931). The documents are written in Ukrainian and Russian, with a total of 80 pages.

Each page of the PDF document is rendered in RGB format at a density of 400 DPI using PyMuPDF. The images are converted to grayscale and binarized using Otsu's method for character segmentation. Morphological opening is applied to remove noise and separate fused components. Next, contours are detected using the cv2.findContours algorithm, which highlights only the outer boundaries of the characters. The contours are analyzed based on several criteria: minimum character area (less than 80 pixels), aspect ratio ($0.2 \leq w/h \leq 4.0$), and fill factor ($0.1 \leq extent \leq 0.25$). Valid contours are normalized to a size of 64×64 pixels and saved in PNG format. This approach improves the quality of the sample and the preparation of data for training text recognition models. As a result of the processing, 7464 segmented character images were obtained, an example of which is shown in Figure 2.

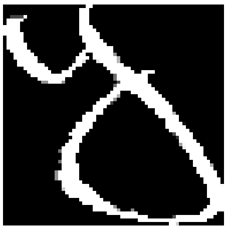


Figure 2: Example of an image after segmentation.

3.2. Labeling UI

A local application based on Gradio has been developed for manual character annotation, which accelerates the labeling process for images obtained at the preprocessing stage. The main goal is to involve experts in entering labels (annotations) for the images.

Authorization Section: At the beginning, the user enters their name, group, and archive team number, which allows identifying participants during the subsequent analysis of the collected labels.

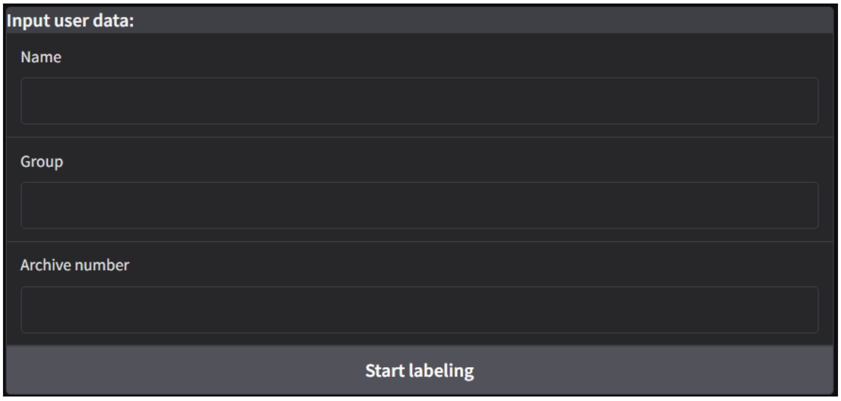
Main Labeling Interface: The symbol is displayed, the language is indicated (automatically determined from the filename), a field for entering the label, and buttons for navigation (save, skip, return to the previous symbol) are provided. The language of the symbol is automatically detected from part of the filename (_ukr, _eng, _pol, _rus), which can be used for further classification.

Each label is saved in a CSV file, which includes:

1. label — the entered label (symbol),
2. language — the language, determined in advance and indicated as a suffix in the filename (e.g., "Manuscripts/230-1-1_ukr.pdf"),
3. image_path — the path to the image,
4. user_name, user_group, team_number — metadata about the annotator.

The indexing update mechanism allows for correcting labels in case of skipped or revisited images. The application works locally, without the need to connect to external servers, making it convenient for handling confidential data or working in areas with limited network access. The simplicity of the interface allows even untrained users to participate in character annotation.

The annotation application's interface is implemented as a local web application using the Gradio library. After launching the application, the user is directed to the registration page, an example of which is shown in Figure 3. The user enters their personal data: name, surname, group number, and archive identifier (set number of 200 images). This ensures user identification and control over the quality of the entered labels.



Input user data:

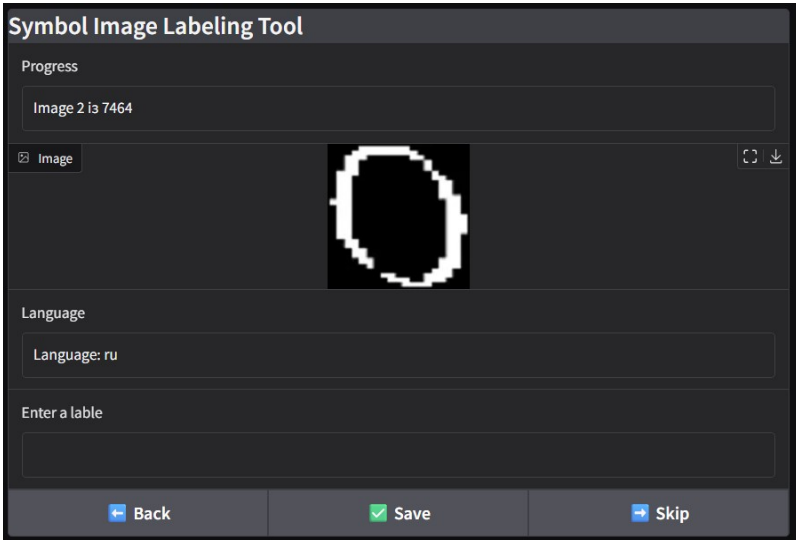
Name

Group

Archive number

Figure 3: User registration page.

After confirming the data by clicking the "Start labeling" button, the system redirects the user to the main character annotation page (see Figure 4). The interface displays a progress indicator showing the number of processed and remaining images. The symbols are displayed at a fixed size of 128×128 pixels, with the option to zoom in or download. The language of the symbol is automatically determined by the filename.



Symbol Image Labeling Tool

Progress

Image 2 із 7464

Language

Language: ru

Enter a lable

Figure 4: Labeling page.

In the central part of the interface, there is a text field for entering the label. Below, there are navigation buttons: to move to the next or previous image, as well as a "Save" button that records the label in the CSV file and initiates the transition to the next symbol. If necessary, the user can skip an image or return to the previous one, with the label being updated accordingly.

3.3. Quality aggregation

To merge annotation data obtained from multiple user teams in the form of CSV files, a software module was developed that ensures standardized merging of the labels based on the majority vote principle. The code is implemented in Python using the pandas library.

At the first stage, the module searches and reads all available CSV files in a specified directory (csv_inputs). Empty files or files with reading errors are ignored, which enhances the processing robustness.

For each record containing metadata about the language (language), image path (image_path), label (label), as well as user and team information (user_name, user_group, team_number), a check is performed to ensure compliance with the required column set. After that, the language names are normalized using a pre-defined mapping dictionary (LANG_MAP), which ensures consistent representation of labels like "ua", "ukr", "uk" to a single standard "uk".

The main aggregation operation is performed at the record grouping level by the key (team_number, image_path). For each group, the agreed-upon values for the fields label and language are determined using majority voting (majority_vote). If there are multiple values with the same number of votes, the one that appears first alphabetically is selected, ensuring result stability.

The summarized results are stored in the final file merged_labels.csv with UTF-8-SIG encoding for compatibility with local processing systems.

3.4. Metrics

To evaluate the quality of annotations, two inter-rater agreement metrics were chosen — Krippendorff's α [25] and Fleiss' κ [26].

Krippendorff's α was calculated on the nominal scale of measurement, which corresponds to the nature of the character classification task. Formally, the coefficient α is defined as:

$$\alpha = 1 - \frac{D_0}{D_e} \tag{1}$$

where D_0 is the observed variance (the number of disagreements between annotators), and $D_e$is the expected variance under random distribution of labels. In our case, α=0.637\alpha = 0.637α=0.637, which indicates an acceptable level of agreement between annotators, sufficient for analytical conclusions.

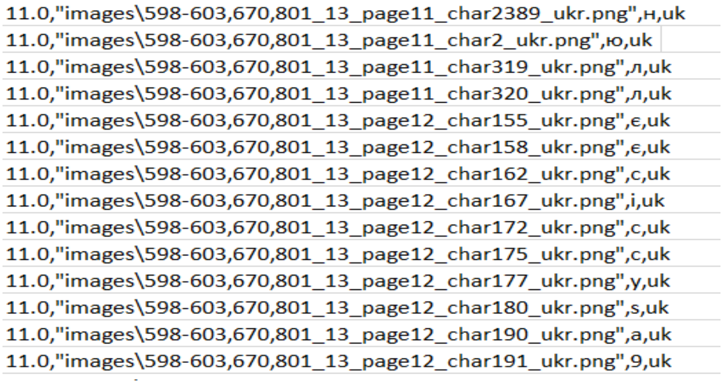
Fleiss' κ was applied to the subset of images that were annotated by exactly three annotators (n = 3), which meets the conditions for applying the metric. The formula for Fleiss' κ is as follows:

$$k = \frac{\overline{P} - \overline{P}_e}{1 - \overline{\overline{P}}_e} \tag{2}$$

where $\overline{P}$ is the average agreement proportion between annotators for all objects, and $\overline{P}_e$is the expected proportion of agreement under random label assignment. In this study, the value of Fleiss' κ is 0.562, which also indicates a moderate, but acceptable level of agreement.

4. Result

As a result of the semi-automated system, a corpus of 1684 handwritten character images was created. Of these, 215 are unique in terms of content and label, indicating a certain level of redundancy (≈87% repeated entries), which was intentionally built in to allow for label aggregation based on the majority voting principle. Example records can be seen in Figure 5.



```
11.0,"images\598-603,670,801_13_page11_char2389_ukr.png",н,uk
11.0,"images\598-603,670,801_13_page11_char2_ukr.png",ю,uk
11.0,"images\598-603,670,801_13_page11_char319_ukr.png",л,uk
11.0,"images\598-603,670,801_13_page11_char320_ukr.png",л,uk
11.0,"images\598-603,670,801_13_page12_char155_ukr.png",є,uk
11.0,"images\598-603,670,801_13_page12_char158_ukr.png",є,uk
11.0,"images\598-603,670,801_13_page12_char162_ukr.png",с,uk
11.0,"images\598-603,670,801_13_page12_char167_ukr.png",і,uk
11.0,"images\598-603,670,801_13_page12_char172_ukr.png",с,uk
11.0,"images\598-603,670,801_13_page12_char175_ukr.png",с,uk
11.0,"images\598-603,670,801_13_page12_char177_ukr.png",у,uk
11.0,"images\598-603,670,801_13_page12_char180_ukr.png",s,uk
11.0,"images\598-603,670,801_13_page12_char190_ukr.png",а,uk
11.0,"images\598-603,670,801_13_page12_char191_ukr.png",9,uk
```

Figure 5: Labels file.

The corpus covers two language domains — Ukrainian and Russian — which are identified based on the suffixes in the original PDF document filenames. All images were pre-normalized to a size of 64×64 pixels in grayscale, maintaining the proportions of the characters and with additional padding to avoid losing context.

Table 1
Dataset statistic

Parametr	**Value**
Total number of images	1684
Number of unique records	215
Number of languages	2 (Ukrainian, russian)
Image format	PNG, 64×64 px, binarised
Average character area	112.6
Average aspect ratio	0.73
Percentage of images with noise	≈ 8.2%
Typical PDF page size	A4, 400 DPI
Average number of characters per page	120–170

Figure 6 shows the frequency graph of characters in the collected corpus. The most common character was "C" — with over 140 occurrences, significantly surpassing the frequency of other characters. Other frequent graphemes include "o" (≈100 occurrences) and the symbol "/".

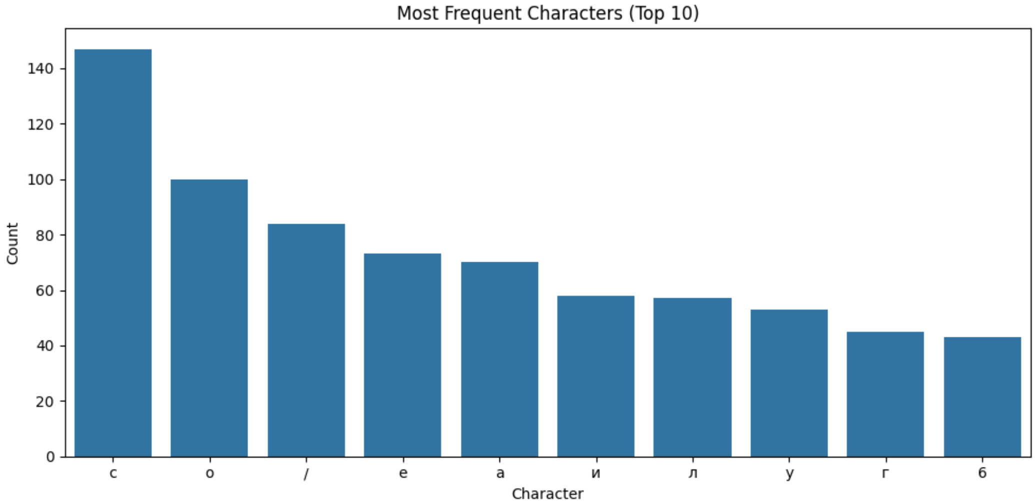


Figure 6: Top 10 characters by frequency.

This uneven frequency distribution is caused both by the characteristics of the linguistic material (e.g., the frequency of the letter "o" in Ukrainian and Russian texts) and by segmentation errors. Specifically, the high frequency of the symbol "/" indicates misclassification of fragments or line breaks as separate symbols.

Manual review revealed that some of the input images labeled as "c", "o", and "/" correspond to incomplete symbols rather than full characters, or artifacts from the images. This confirms the need for improvement in preprocessing modules and the implementation of filtering based on context evaluation.

To assess the quality of annotations, two inter-rater agreement metrics were used — Krippendorff's α and Fleiss' κ. Krippendorff's α (nominal scale) was 0.637, indicating an acceptable level of agreement between participants. Fleiss' κ was calculated only for images with the same number of annotations (n = 3) and was 0.562, which also suggests moderate but acceptable agreement. The results confirm the sufficient quality of annotations for further analysis.

To evaluate the effectiveness of the labeling module, an analysis was conducted on the time spent annotating characters as well as preprocessing the images. During the annotation of 100 characters, the total time recorded was 4 minutes and 45 seconds, corresponding to an average time of 2.85 seconds per character.

The effectiveness of the image segmentation module was also analyzed. A total of 1866 characters from handwritten text pages were processed, with an average processing time of 1.98 seconds per page and a total processing time of 39.62 seconds for the entire dataset. It is important to note that the processing time remained stable and was practically independent of the number of segmented characters on the page. For example, when 8 characters were preserved from page 1, the processing time was 2.07 seconds, while for 179 characters on page 18, it was only 2.00 seconds.

5. Conclusions

A prototype of an interactive system for creating handwritten character corpora has been developed and tested. The system implements a full pipeline: preprocessing, segmentation, collective labeling, quality aggregation, and export to a standardized format for HTR tasks from historical documents. A total of 80 pages of multilingual materials were processed. 7,464 character crops (PNG, 64×64 px) were generated. As a result of labeling, 1,684 validated examples were obtained, with 215 being unique. The redundancy was approximately 87%, which was deliberately incorporated for majority voting. The annotation consistency is Krippendorff's α = 0.637 and Fleiss' κ = 0.562 (n = 3). This corresponds to a moderate and practically sufficient level of agreement. The average labeling time is 2.85 seconds per symbol. The average segmentation time is 1.98 seconds per page, and it remains stable across a range of 8–179 symbols per page. The total time for the analyzed subset is 39.62 seconds. The frequency analysis reveals the dominance of the symbol "C" (>140 occurrences) and "o" (~100 occurrences). The increased frequency of "/" indicates segmentation artifacts.

The volume of validated data is currently limited: 1,684 examples, including 215 unique ones. The language coverage includes only Ukrainian and Russian. The distribution of graphemes is uneven. False positives, particularly for "/", are noted, caused by imperfections in preprocessing and segmentation. Inter-annotator agreement is moderate. The labeling is stored in CSV format at the symbol level. PAGE-XML/ALTO formats have not yet been integrated, making direct comparison with benchmarks difficult and excluding "line" and "word" levels.

The plan is to scale the corpus to at least 10,000 validated symbols. Language-script coverage will be expanded, and grapheme frequencies will be balanced. Segmentation will be improved through adaptive morphological filters, symbol/non-symbol classification, contextual filtering, and detectors based on Mask R-CNN or ViT. The goal is to reduce false positives for "/" by at least 50%. Active learning and self-training integration is planned, including Dawid–Skene and self-training with pseudo-labels, aiming to increase α to ≥0.75. The labeling will be converted to PAGE-XML/ALTO and COCO formats and supplemented with "line" and "word" levels. The final stage will involve benchmarking HTR models (CRNN+CTC and Transformer/ViT with SAM). The impact of synthetic data and fine-tuning on CER and WER will be evaluated, and "quality–labeling volume" curves will be plotted.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Yu X. V. Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources (Findings EMNLP 2022).

[2] Matos, A., Almeida, P., Correia, P. L., & Pacheco, O. (2025). iForal: Automated handwritten text transcription for historical medieval manuscripts. Journal of Imaging, 11(2), 36.

[3] Koopmans, L., Dhali, M. A., & Schomaker, L. (2023). The effects of character-level data augmentation on style-based dating of historical manuscripts. In M. De Marsico, G. Sanniti di Baja, & A. Fred (Eds.), Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023) (pp. 124–135).

[4] Vögtlin, L., Drazyk, M., Pondenkandath, V., Alberti, M., & Ingold, R. (2021). Generating synthetic handwritten historical documents with OCR constrained GANs. In J. Lladós, D. Lopresti, & S. Uchida (Eds.), Document Analysis and Recognition – ICDAR 2021 (Lecture Notes in Computer Science, Vol. 12823, pp. 610–625).

[5] Chen, W., Su, X., & Hou, H. (2025). Fine-grained automatic augmentation for handwritten character recognition. Pattern Recognition, 159, 111079.

[6] de Sousa Neto, A. F., Bezerra, B. L. D., de Moura, G. C. D., & Toselli, A. H. (2024). Data augmentation for offline handwritten text recognition: A systematic literature review. SN Computer Science, 5, 258. https://doi.org/10.1007/s42979-023-02583-6 SpringerLink

[7] Hamdi, Y., Boubaker, H., & Alimi, A. M. (2021). Data augmentation using geometric, frequency, and beta modeling approaches for improving multi-lingual online handwriting recognition. International Journal on Document Analysis and Recognition (IJDAR), 24, 283–298. https://doi.org/10.1007/s10032-021-00376-2 SpringerLink

[8] Toselli, A. H., Vidal, E., & Casacuberta, F. (2011). Active interaction and learning in handwritten text transcription. In A. H. Toselli, E. Vidal, & F. Casacuberta (Eds.), Multimodal interactive pattern recognition and applications (pp. 119–133). Springer. https://doi.org/10.1007/978-0-85729-479-1_5 SpringerLink

[9] Wolf, F., & Fink, G. A. (2024). Self-training for handwritten word recognition and retrieval. International Journal on Document Analysis and Recognition (IJDAR), 27, 225–244. https://doi.org/10.1007/s10032-024-00484-9 SpringerLink

[10] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. National Science Review, 11(12), nwae403. https://doi.org/10.1093/nsr/nwae403 Oxford Academic

[11] Murel J., Smith D. Self-training and Active Learning with Pseudo-relevance Feedback for Handwriting Detection in Historical Print. In: Proc. ICDAR 2024, LNCS 14967, pp. 305–324. Springer, 2024.

[12] Geldhauser C., Malyshev K. Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation. FedCSIS 2024 – AI in Digital Humanities (Comm. Papers), pp. 37–44, 2024. DOI: 10.15439/2024F1772. ACM Journal

[13] Crosilla G., Klic L., Colavizza G. Benchmarking Large Language Models for Handwritten Text Recognition. arXiv preprint arXiv:2503.15195, 2025. arXiv

[14] Li Y., Chen D., Tang T., Shen X. HTR-VT: Handwritten Text Recognition with Vision Transformer. Pattern Recognition, 158 (2024): 110967.

[15] Baek, J., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[16] Medvedeva, A., & Ponomarenko, T. (2023). Automating data collection process in Industry 4.0. AIP Conference Proceedings.

[17] Wang, P., Zhang, Y., Wang, Q., Yu, H., Xie, E., & Luo, P. (2022). Scene text recognition with permuted autoregressive sequence models. In Advances in Neural Information Processing Systems (NeurIPS).

[18] Shi, B., Bai, X., & Yao, C. (2019). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. Proceedings of the AAAI Conference on Artificial Intelligence.

[19] Rocco, C., de Mello, M. R., & Oliveira, L. (2022). Synthetic dataset creation for computer vision application: Pipeline proposal.

[20] Ivasechko, A. V., & Lipianina-Honcharenko, K. V. (2025). Architecture of a semi-automated annotation system for multilingual archival handwritten texts. Systems and Technologies.

[21] Lipianina-Honcharenko, K., Wolff, C., Sachenko, A., Desyatnyuk, O., Sachenko, S., & Kit, I. (2023). Intelligent information system for product promotion in internet market. Applied sciences, 13(17), 9585.

[22] Lipianina-Honcharenko, K., Telka, M., & Melnyk, N. (2024, December). Comparison of ResNet, EfficientNet, and Xception architectures for deepfake detection. In Proceedings of the 1st International Workshop on Advanced Applied Information Technologies CEUR-WS, Khmelnytskyi, Ukraine, Zilina, Slovakia (pp. 26-34).

[23] Rusyn,B et al (2018) Segmentation of atmospheric clouds images obtained by remote sensing.14th International Conrefences on Advanced Trends in Radioelectronics, Telecommunication and Computer Engineering,TCSET,2018,Proceeding,pp.213-216.

[24] Mukhin,V,et al (2025) A model for classifying information objects using neural networks and fuzzy logic.Scientific Reports,v.15,is.1,15904.

[25] Marzi G., Balzano M., Marchiori D. K-Alpha Calculator—A user-friendly tool for computing Krippendorff's Alpha. MethodsX, 12, 102545, 2024. DOI: 10.1016/j.mex.2023.102545. (Короткий огляд α, інтерпретація порогів і практичний інструмент.)

[26] Halpin S.N. Inter-Coder Agreement in Qualitative Coding: Considerations for its use. American Journal of Qualitative Research, 2024.