

# 000 001 002 003 004 005 HTR for Russian Empire Period Manuscripts: 006 A Two-Stage Framework with New Annotated Resources 007 008 009

010 **Anonymous authors**  
011

012 Paper under double-blind review  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030

## Abstract

031 Historical handwritten documents represent a valuable source of information about  
032 the language, culture, and society of earlier periods. In the context of global-  
033 ized scholarship, the development of automatic handwriting recognition tools for a  
034 wide range of languages has become increasingly important to ensure broader ac-  
035 cessibility to the cultural heritage of different nations. Pre-revolutionary Russian  
036 presents a particular challenge for such systems due to its significant orthographic  
037 differences from the modern language. This work introduces a universal tool for  
038 recognizing handwritten documents written in pre-revolutionary Russian orthog-  
039 raphy, dated from the 19<sup>th</sup> century to the early 20<sup>th</sup> century. We present a two-  
040 stage handwritten text recognition (HTR) system combining YOLOv8-based line  
041 segmentation with TrOCR<sub>pre</sub>, a transformer architecture pre-trained on Russian-  
042 language data. The system is performed on a manually annotated corpus of 38,501  
043 lines across three document types: Gubernatorial Reports (31,083 lines), Statu-  
044 tory Charters (5,868 lines), and Personal Diaries (1,550 lines), split into training,  
045 validation, and test sets. Our approach achieves a character error rate (CER) of  
046 8.5% and a word error rate (WER) of 29.1% overall, with performance varying  
047 by document type - ranging from 4.8% CER on formal administrative documents  
048 to 19.0% CER on informal personal writings. The transformer-based architecture  
049 demonstrates a 53.8% improvement over traditional CNN-RNN baselines (from  
050 18.4% to 8.5%), providing a practical tool for large-scale digitization of historical  
051 Russian archives. **Demo:** Interactive demo.

## 1 Introduction

032 Historical documents play a crucial role in preserving cultural heritage and providing the scholarly  
033 community with direct access to primary sources. These materials offer unique insights into the  
034 linguistic, cultural, and societal dynamics of earlier periods and serve as empirical foundations for  
035 studying long-term macroeconomic and political developments.

036 Despite their importance, the automated processing of such documents remains a substantial chal-  
037 lenge - particularly in the case of low-resource languages. Modern optical character recognition  
038 (OCR) systems, typically powered by machine learning techniques (Garrido-Munoz et al., 2025b;  
039 Romein et al., 2025; Zhu et al., 2025), require large volumes of annotated training data. This cre-  
040 ates a fundamental obstacle for historical languages or scripts with limited digital resources, where  
041 labeled corpora are scarce or altogether absent.

042 A notable case of this challenge is the recognition of handwritten Russian documents from the Rus-  
043 sian Empire period, spanning from the 17<sup>th</sup> century until the orthographic reform of 1918 (Council  
044 of People's Commissars of the RSFSR, 1918b). Unlike many Western European languages - whose  
045 orthographic systems remained relatively stable - Russian orthography underwent radical changes  
046 following the 1917 revolution. The 1918 reform eliminated several letters and altered spelling rules,  
047 creating a substantial divergence between pre- and post-reform Russian texts.

048 By contrast, English orthography began stabilizing in the 18<sup>th</sup> century, especially after the publica-  
049 tion of Samuel Johnson's dictionary in 1755 (Johnson, 1755), which contributed to enduring spelling  
050 conventions. As a result, modern OCR systems can recognize 18<sup>th</sup>-century English texts with min-  
051 imal adaptation (Edwards III, 2007; Garrido-Munoz et al., 2025a). The situation with Russian is

markedly different: the substantial orthographic transformation renders pre-reform Russian nearly a distinct variant from the perspective of automated language processing.

Our novelty lies in addressing a critical, previously unsolved gap rather than architectural innovation. Pre-revolutionary Russian (19th–early 20th century) represents a genuinely low-resource domain with no prior HTR systems or comprehensive datasets. Unlike English (stable since 18th century), Russian underwent radical 1918 orthographic reform eliminating obsolete characters (ѣ, ѿ, ѿ, ѵ) and spelling conventions, preventing direct application of modern Russian HTR models.

According to recent estimates (Degtareva, 2022), only about 5% of archival Russian documents requiring digitization have been converted into digital form. This severely limits scholarly access to these materials and justifies classifying pre-reform Russian as a low-resource language in the context of modern text processing technologies.

Existing solutions for historical Russian OCR are typically narrow in scope. For example, the system designed for early 18<sup>th</sup>-century Petrine-era cursive targets a highly specific domain (Potanin et al., 2021), while the project focused on A.S. Pushkin’s manuscripts is limited to a single individual’s handwriting style (Kokorin et al., 2025). Although effective in their respective contexts, such tools are not generalizable to a broader range of documents due to the specificity of their training data and temporal focus.

To address this gap, we present ***the first universal tool for handwritten text recognition (HTR) in historical Russian documents from the Russian Empire period*** (19th–early 20th centuries), capable of handling diverse handwriting styles and document types within this era.

Our approach makes three key contributions:

First, we develop **a novel two-stage architecture for low-resource handwritten text recognition (Russian)**, combining YOLOv8-based layout analysis for line segmentation (Yaseen, 2024) with a specialized TrOCR transformer architecture (Li et al., 2023) for text recognition.

Second, we curate and manually annotate **the largest dataset of pre-reform Russian manuscripts to date, comprising 38,501 lines spanning diverse handwriting styles from the late 19th to early 20th centuries**. This represents the first comprehensive dataset of its kind for this historical period and writing system.

Third, **our system demonstrates state-of-the-art performance** with a character error rate (CER) of 8.5% and word error rate (WER) of 29.1% on held-out test data, substantially outperforming all existing methods in this domain.

These contributions enable, for the first time, large-scale automated digitization of historical Russian archives, opening new possibilities for digital humanities research and cultural heritage preservation.

## 2 Related Works

### 2.1 HTR for Historical Documents

Recent surveys Garrido-Munoz et al. (2025b) document the shift from CNN-RNN to transformer and LMM architectures, though complex layouts and rare scripts remain challenging. Benchmarks Romein et al. (2025); Ghaboura et al. (2025) confirm state-of-the-art performance: Humphries et al. (2025); Kim et al. (2025) achieve CER below 6 % on 18th-19th-century Western manuscripts. Self-supervised pre-training (Penarrubia et al., 2025) and post-OCR correction (Beshirov et al., 2025) further improve results, yet accuracy degrades on severely degraded or non-Latin material.

### 2.2 HTR for Pre-Revolutionary Russian Orthography

Pre-1918 Russian employed four now-obsolete letters and mandated the hard sign at word-final position, producing an alphabetic gap that prevents direct transfer of modern Russian HTR models. Limited prior work addresses only narrow subsets: Potanin et al. (2021) developed *Digital Peter* (9k lines, early-18th-century Petrine cursive) showing CRNN baselines incur triple the error on Latin datasets; Kokorin et al. (2025) focused on single-author *Chronicle of Pushkin* via hybrid OCR-plus-

108 rule pipeline requiring heavy manual post-editing. Neither exploits recent transformer or LMM  
 109 advances.

110 Direct comparison with state-of-the-art models (LLMs, commercial OCR, DAN (Constum et al.,  
 111 2024; Coquenet, 2025)) is methodologically invalid: modern systems lack obsolete Cyrillic char-  
 112 acters (ѣ, ё, в, і); DAN requires page-level annotation for Latin scripts vs. our line-level archival  
 113 workflow; cross-period evaluation (Digital Peter’s 18th vs. our 19th–20th century corpus) would  
 114 confound results due to 150-year paleographic gap. We compare architectures trained on our corpus  
 115 - the only valid approach given domain-specific constraints.

116 Data scarcity and orthographic divergence remain principal bottlenecks. Our system is the first gen-  
 117 eralizable approach across multiple 19th–20th century document types, enabling access to approxi-  
 118 mately 95% of undigitized Russian archival materials (Degtareva, 2022).

### 120 3 Corpus Description

121 The source material for this study comprises high-resolution color scans - digital facsimiles of hand-  
 122 written documents preserved in the State Historical Archive of the Russian Federation (Russian State  
 123 Historical Archive, 2024) and the Presidential Library (Presidential Library named after B.N. Yeltsin,  
 124 2024).

125 The corpus is characterized by a high degree of heterogeneity across multiple dimensions. From a  
 126 palaeographic perspective, it features a wide range of handwriting styles and graphical conventions.  
 127 Linguistically, the documents contain archaic morphological forms and diachronic orthographic varia-  
 128 tion. The physical condition of the sources varies considerably, resulting in uneven image quality.  
 129 Thematically, the dataset spans a broad spectrum of genres, including private correspondence, diary  
 130 entries, legal documents, and official reports. Structural layouts range from continuous prose  
 131 to decorated headings and formal symbolic elements. These combined factors define a corpus of  
 132 significant historical value and substantial computational complexity.

133 With 38,501 annotated lines across three document types, this represents the largest annotated cor-  
 134 pus for pre-revolutionary Russian to date, featuring multi-scribe coverage spanning reports from  
 135 more than 20 provinces of the Russian Empire. We are finalizing licensing with archival institu-  
 136 tions. Immediate access is available to specialists upon request for non-commercial research via our  
 137 *Anonymous access request form*.

#### 138 3.1 Dataset Composition

139 The corpus (38,501 lines) is organized into three distinct datasets, each reflecting a different type of  
 140 historical document (see Fig. 1):

- 141 • **Gubernatorial Reports (19<sup>th</sup>-early 20<sup>th</sup> centuries)** – 31,083 *lines*. These are official re-  
 142 ports submitted by provincial administrations to the central government of the Russian Em-  
 143 pire between 1804 and 1914. The content includes statistical tables, descriptive narratives,  
 144 and incident records from various provinces (Razdorskii, 2011). The general records in-  
 145 clude text describing the economic and social situation of the province, as well as appendices  
 146 in the form of dozens of tables with data on population dynamics, crop yields, industrial ac-  
 147 tivity, education, etc. The documents exhibit a uniform clerical font, standardized terminol-  
 148 ogy, and well-structured tabular formatting, but the number and structure of tables may have  
 149 changed over time. Our sample includes reports from more than 20 provinces of the Rus-  
 150 sian Empire with different climatic and economic characteristics (including Arkhangelsk,  
 151 Astrakhan, Moscow and Tobolsk).
- 152 • **Statutory Charters (19<sup>th</sup> century)** – 5,868 *lines*. Statutory documents (charters) (19<sup>th</sup>  
 153 century) – 5868 lines . These legal documents were prescribed by the Emancipation Man-  
 154 ifesto of 1861 and record the transition of peasants from serfdom to a free state. Typically  
 155 featuring hybrid layouts with handwritten clauses and printed inserts, they detail financial  
 156 obligations, party lists, and administrative endorsements (Sofronenko, 1954). The stylistic  
 157 formality and juridical lexicon reflect their legislative function.
- 158 • **Konstantin P. Pobedonostsev Personal Papers Collection (Diaries) (19<sup>th</sup>–20<sup>th</sup> cen-  
 159 turies)** – 1,550 *lines*. A collection of private reflections by the statesman and legal scholar

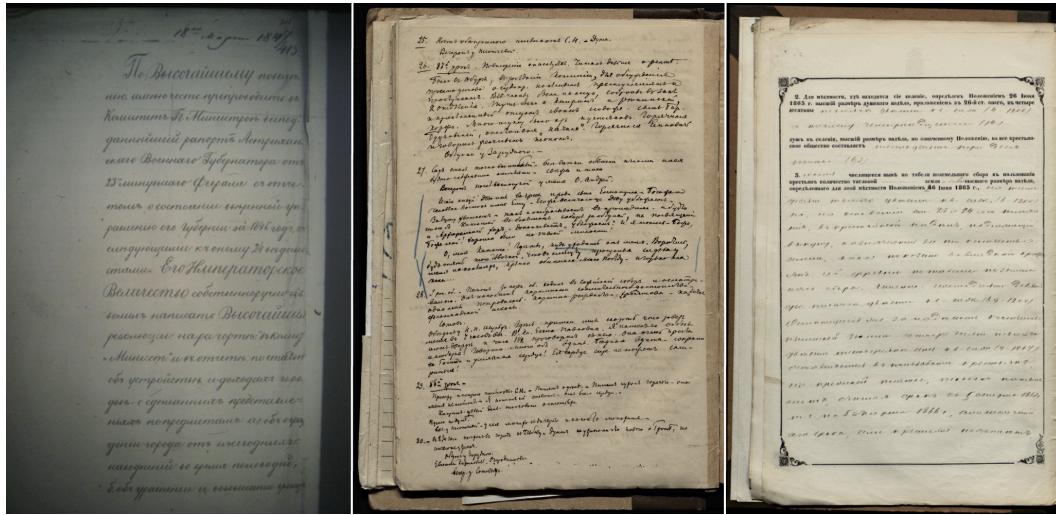


Figure 1: Example manuscript pages from the three datasets used in this study: Gubernatorial Reports (left), Personal Diaries of Konstantin P. Pobedonostsev (center), and Statutory Charters (right). The examples illustrate the diversity of handwriting styles, document structure, and historical orthography encountered across different sources.

who oversaw educational and ecclesiastical reforms. The texts offer insights into everyday life, courtly affairs, and personal impressions. Written in a freeform style, they exhibit emotional interjections, nonstandard constructions, orthographic irregularities, and extensive manual corrections.

While the first two datasets reflect official documentation in polished chancery handwriting, the third represents informal, idiosyncratic entries marked by irregular lineation, overwrites, and expressive variation. The complete dataset is currently being prepared for public release to support future research in historical document digitization.

### 3.2 Linguistic Specificity and Representativeness

All three datasets belong to a highly specific linguistic class characterized by pre-reform Russian orthography, archaisms, and palaeographic idiosyncrasies. These factors impose additional demands on HTR systems and necessitate domain expertise during annotation (see Appendix A).

The corpus includes both formal administrative genres and spontaneous private writing, thus capturing a broad spectrum of textual production practices in Imperial Russia. Despite being geographically limited to archival collections, the documents span a representative array of Eastern European written forms from the 19<sup>th</sup> to early 20<sup>th</sup> centuries. Stylistic and graphic features include syntactic complexity, orthographic variability, nonstandard abbreviations, and a wide range of handwriting - from chancery calligraphy to diary informality.

In total, over 38,000 lines were transcribed by domain experts, including historians trained in Old Church Slavonic and pre-revolutionary Russian script. This expert-driven annotation is critical given the rare and often ambiguous glyphs present in the sources.

### 3.3 Annotation Format and Procedure

The dataset partitioning follows standard machine learning practices with training, validation, and test splits designed to ensure robust model evaluation:

*Gubernatorial Reports (19th-early 20th centuries) – 31,083 lines:*

- Train: 28,671 lines (92.2%)
- Validation: 1,191 lines (3.8%)
- Test: 1,221 lines (3.9%)

216 *Statutory Charters (19th century)* – 5,868 lines:

217

- 218 • Train: 4,441 lines (75.7%)
- 219 • Validation: 662 lines (11.3%)
- 220 • Test: 765 lines (13.0%)

222 *Konstantin P. Pobedonostsev Personal Papers Collection (Diaries) (19th–20th centuries)* – 1,550  
223 lines:

224

- 225 • Train: 960 lines (61.9%)
- 226 • Validation: 264 lines (17.0%)
- 227 • Test: 326 lines (21.0%)

229 The training sets were used for model parameter optimization, validation sets for hyperparameter  
230 tuning and early stopping criteria, and test sets for final performance evaluation. The varying split  
231 ratios reflect the different dataset sizes, with smaller collections allocated proportionally larger val-  
232 idation and test portions to ensure statistically meaningful evaluation metrics.

233 The annotation process was conducted using the LabelStudio (Heartex, 2020) platform and followed  
234 a two-stage pipeline: (1) line localization and (2) content transcription.

235 **Stage 1: Line Localization.** Two complementary techniques were employed depending on the  
236 structural complexity of the source material:

237

- 238 • **Bounding-box detection** was applied to documents with regular layouts and well-separated  
239 lines, specifically the *Gubernatorial Reports* and *Statutory Charters*. Each text line was  
240 enclosed in an axis-aligned rectangle.
- 241 • **Segmentation masks** were used for the *Personal Diaries*, which often featured overlap-  
242 ping, skewed, or irregularly spaced lines. This approach enabled finer-grained delineation  
243 of visually entangled handwriting.

245 **Stage 2: Line Transcription.** Each localized line - regardless of whether it was detected via bound-  
246 ing box or segmentation - was manually transcribed by expert annotators. The final annotated unit  
247 thus consists of a paired representation: a cropped image of the handwritten line and its correspond-  
248 ing textual content in UTF-8 format.

## 249 4 Architecture of the Recognition System

250 Our system adopts a two-stage architecture comprising YOLOv8-based line segmentation (Yaseen,  
251 2024) and TrOCR transformer-based text recognition (Li et al., 2023), trained on three manually an-  
252 notated datasets of late 19<sup>th</sup> to early 20<sup>th</sup> century manuscripts (totaling 38,501 lines). The complete  
253 pipeline implements a four-stage process: image preprocessing, text line localization, optical char-  
254 acter recognition, and postprocessing. This modular design ensures high accuracy under challenging  
255 paleographic conditions.

### 256 4.1 Image Preprocessing

257 Each image undergoes a series of enhancement operations to optimize the quality of the input prior  
258 to recognition. The preprocessing stage includes:

259

- 260 • noise suppression to remove background artifacts;
- 261 • contrast enhancement to improve text-to-background visibility.

262 Empirical evaluations confirm consistent accuracy gains across all tested architectures. While mod-  
263 ern HTR models may be less sensitive to preprocessing on high-quality contemporary documents,  
264 our degraded historical manuscripts-exhibiting ink deterioration, paper discoloration, and variable  
265 contrast-demonstrably benefit from these enhancement operations. Preprocessing is therefore ap-  
266 plied both during model training and inference.

270  
271

## 4.2 Text Line Localization

272  
273  
274  
275  
276  
277  
278  
279

A modified YOLOv8 (Yaseen, 2024) performs joint detection-and-segmentation for localizing hand-written lines, enabling precise identification of overlapping or distorted text. Trained on challenging configurations (overlapping, slanted, curved lines), YOLOv8 employs mask-based cropping for complex layouts: an axis-aligned bounding box is extracted with pixels outside the segmentation mask set to neutral background, preserving original geometry without explicit warping. Selected after benchmarking against alternative solutions (Tesseract OCR, EasyOCR, Craft Text Detector) that proved unreliable on degraded manuscripts, YOLOv8 achieves high localization accuracy (mAP = 0.98) on the validation set.

280  
281

## 4.3 Text Recognition

282  
283  
284  
285  
286  
287

Following localization, cropped line images are passed to the OCR component. We evaluate five architectures: *three traditional baselines* (VGG-CTC (Simonyan & Zisserman, 2015; Graves et al., 2006), CRNN (Shi et al., 2015), ResNet-BiLSTM-CTC (Shonenkov et al., 2021)) and *two transformer-based models* (TrOCR, TrOCR<sub>pre</sub>) (Li et al., 2021). Among these, TrOCR<sub>pre</sub> consistently delivered the highest accuracy and was selected for the final pipeline due to its robust handling of non-horizontal text through grid-like feature maps (detailed evaluation in Section 6.1).

288  
289  
290  
291  
292  
293

Among these, TrOCR<sub>pre</sub> consistently delivered the highest character-level accuracy and was selected for downstream integration. The transformer architecture's grid-like feature maps enable robust decoding of non-horizontal text. Through exposure to diverse line geometries during fine-tuning—including slanted, curved, and irregular alignments prevalent in Personal Diaries: TrOCR<sub>pre</sub> learns to decode text along varied spatial trajectories without requiring explicit line straightening.

294  
295

## 4.4 Postprocessing

296  
297

To improve transcription quality, two postprocessing strategies were evaluated:

298  
299  
300  
301  
302  
3031. **Dictionary-based correction:**

- a lexicon was constructed from the training corpus;
- Levenshtein-distance matching was used to suggest corrections;
- improvements were observed only on subsets with initially low accuracy, showing limited overall benefit.

304  
305  
306  
307  
308  
309  
310  
311  
3122. **LLM-assisted correction:**

- the model received the erroneous OCR output and reference context;
- an LLM generated a corrected version based on semantic coherence;
- performance consistently degraded due to inappropriate historical substitutions: modern LLMs lack training data in pre-reform orthography; their tokenizers and corpora do not contain obsolete characters (ѣ, ё, в, і) or archaic spelling conventions, leading to undesirable modernization of historical text. Integrated language modeling with historical Russian corpora represents a promising direction for future work.

313  
314  
315

## 5 Training and Inference Procedure

316  
317

## 5.1 Training Pipeline

318  
319

The training process followed a sequential pipeline composed of the following stages:

320  
321  
322  
323

1. **Image preprocessing:** All images were enhanced using the standard preprocessing routine described in Section 3 prior to any downstream training.
2. **Data partitioning:** The dataset was split into three subsets - train, validation, and test - at the document level. Pages originating from the same source document were never assigned to multiple subsets, thereby preventing content leakage.

324        3. **Detector and segmenter training:** The YOLOv8 architecture was trained on annotated  
 325        data using both bounding-box and segmentation-mask formats. This dual-format annotation  
 326        enabled the model to generalize across both well-structured and visually entangled layouts.  
 327

328        4. **Text recognition training:**

329            • Line images were extracted from the localized regions.  
 330            • Each cropped line was paired with its corresponding transcription and used to train  
 331              OCR models, as detailed in Section 3.3.

332        5. **Postprocessing optimization:** For each postprocessing strategy, hyperparameters were  
 333              tuned to maximize evaluation metrics on the validation set. This included thresholding  
 334              for dictionary-based correction and alignment strategies for candidate selection.

335

336        5.2 Inference Workflow

337

338        During inference, the system operates as a unified pipeline:

339

340        1. Each input page is passed through the image preprocessing module.  
 341        2. YOLOv8 performs line localization using either bounding boxes or segmentation masks,  
 342              depending on the document type.  
 343        3. Localized line images are processed by the chosen OCR model (e.g.,  $\text{TrOCR}_{\text{pre}}$ ).  
 344        4. Optional postprocessing is applied to refine the textual output.

345

346        5.3 Model Adaptation and Fine-Tuning

347

348        To tailor the recognition pipeline to the characteristics of pre-revolutionary Russian manuscripts,  
 349        both detection and recognition components were adapted:

350

351            • **YOLOv8** was fine-tuned on historical documents annotated to capture the specificities of  
 352              early Cyrillic script, including characters with diacritical marks and obsolete graphemes.  
 353            • **TrOCR<sub>pre</sub>** was initialized from pretrained weights and further fine-tuned on the task-  
 354              specific corpus. The adaptation included additional exposure to synthetic Russian strings.

355

356        6 Experiments and Evaluation

357

358        6.1 Comparison of HTR Architectures

359

360        To evaluate handwriting recognition performance, we conducted experiments with five model archi-  
 361        tectures:

362

363        1. **VGG-CTC**: a CNN architecture with CTC decoding (based on PyLaia);  
 364        2. **CRNN**: a hybrid model combining CNN and RNN layers with a CTC decoder (adapted  
 365              from wronnyhuang/htr);  
 366        3. **ResNet-BiLSTM-CTC**: deep convolutional layers followed by bidirectional LSTMs (from  
 367              the indic-htr project);  
 368        4. **TrOCR**: a transformer-based model with a Vision Transformer encoder, pre-trained on En-  
 369              glish handwriting corpora;  
 370        5. **TrOCR<sub>pre</sub>**: the same architecture as TrOCR, further pre-trained on synthetic Russian lines  
 371              and fine-tuned on real handwritten Russian and Kazakh data.

372

373        We do not include general-purpose OCR engines as recognition baselines. Their default Russian  
 374        models target modern orthography and printed text and lack coverage of obsolete graphemes; with-  
 375        out retraining or custom alphabets they produce inconsistent or modernized outputs on pre-reform  
 376        handwriting, making results non-comparable. We therefore report only HTR architectures trained  
 377        on our corpus; off-the-shelf OCR was used, if at all, only for line-detection sanity checks.

378 The experimental results, summarized in Table 1, demonstrate a clear advantage of transformer-based  
 379 models. The best overall performance was achieved by the **YOLOv8 + TrOCR<sub>pre</sub>** configuration,  
 380 reaching a CER of 8.5% across the full corpus. More granular metrics per document type are pre-  
 381 sented in Table 2, where the same configuration achieved a CER of 4.8% on the *Gubernatorial*  
 382 *Reports* subset. Examples of system usage can be found in Appendix B.

383

## 384 6.2 Component-wise Evaluation

385

### 386 6.2.1 Image Preprocessing

387 In ablations, image preprocessing improved both line localization and transcription accuracy. Be-  
 388 cause filtering operations (noise suppression, contrast enhancement) yielded consistent gains on the  
 389 validation set, we include them in the final pipeline.

390

### 391 6.2.2 Line Detection and Segmentation

392 We compared our detection module against standard tools (Tesseract OCR, EasyOCR, CraftTextDe-  
 393 tector). These baseline tools proved unstable when applied to historical manuscripts. In contrast,  
 394 the fine-tuned YOLOv8 model unified detection and segmentation stages and achieved high local-  
 395 ization accuracy (mean average precision = 0.98). Additional evaluation on the test set confirms  
 396 YOLOv8’s robustness, with mean IoU of 97.2%, indicating precise line boundary delineation even  
 397 in challenging layouts.

398 Cross-archive out-of-distribution evaluation is currently impossible: this represents the first publicly  
 399 annotated dataset for pre-revolutionary Russian from this historical period. However, we demon-  
 400 strate multi-scribe generalization capability on the Gubernatorial Reports subset, which contains  
 401 documents from more than 20 provinces of the Russian Empire, each written by different scribes  
 402 with varying handwriting styles. We are actively collaborating with archival institutions to create  
 403 annotated manuscripts from other archives, which will enable rigorous cross-archive benchmarks in  
 404 future work.

405 To isolate the contribution of each pipeline component, we conducted experiments using manually  
 406 annotated ground-truth bounding boxes and segmentation masks. Under these ideal localization  
 407 conditions, TrOCR<sub>pre</sub> achieved 7.2% CER compared to 8.5% CER in the full end-to-end pipeline,  
 408 yielding a segmentation-to-recognition error gap of only 1.3 percentage points. This indicates that  
 409 while line detection errors do propagate to the final output, the recognition architecture remains the  
 410 dominant factor in overall performance.

411 Furthermore, TrOCR<sub>pre</sub> demonstrates robustness to non-horizontal text geometries without requir-  
 412 ing explicit line straightening or geometric warping. Through exposure to diverse line orientations  
 413 during fine-tuning—including slanted, curved, and irregular alignments prevalent in the Personal Di-  
 414 aries subset—the model learns to decode text along varied spatial trajectories. The achievement of  
 415 19.0% CER on the most challenging Personal Diaries corpus, despite its extreme geometric irregu-  
 416 larity, empirically validates this geometric robustness.

417

### 418 6.2.3 Postprocessing

419 Postprocessing experiments yielded the following observations:

420

- 421 • Dictionary-based correction helped in cases with low initial recognition quality but slightly  
 422 degraded performance when the base output was already accurate.
- 423 • Integration of large language models (LLMs) in the current setup led to distortion of the  
 424 original historical content and requires further investigation.

425

## 426 6.3 Final Results

427 The final CER and WER metrics reflect the cumulative error of the two-stage pipeline: line local-  
 428 ization and text recognition. As shown in Table 1, the **YOLOv8 + TrOCR<sub>pre</sub>** system achieved:

429

- 430 • CER = 8.5% and WER = 29.1% across the full corpus test-subset (2,312 lines), as detailed  
 431 in Table 1,

432 • CER = 4.8% and WER = 21.5% on the Gubernatorial Reports test-subset (1,221 lines), as  
 433 detailed in Table 2.  
 434

435 To evaluate the recognition stage independently from detection quality, we conducted additional  
 436 experiments using manually annotated bounding boxes and segmentation masks. Under these ideal  
 437 conditions,  $\text{TrOCR}_{\text{pre}}$  achieved a CER of 7.2% and WER of 24.0%, highlighting the architecture’s  
 438 strong potential when accurate localization is provided.

440 Table 1: Overall CER and WER on the full corpus test-subset (2,312 samples)

Model	CER (%)	WER (%)
VGG-CTC	18.4	50.3
CRNN	12.8	39.8
ResNet-BiLSTM-CTC	9.8	32.2
YOLOv8 + TrOCR	9.3	30.2
YOLOv8 + $\text{TrOCR}_{\text{pre}}$	<b>8.5</b>	<b>29.1</b>

447 Table 2: CER and WER across individual test-subsets: *Gubernatorial Reports*, *Statutory Charters*,  
 448 and *Personal Diaries*

Model	Gubernatorial Reports		Statutory Charters		Personal Diaries	
	CER	WER	CER	WER	CER	WER
VGG-CTC	10.4	36.2	20.2	54.4	28.8	72.4
CRNN	7.9	30.1	13.1	42.3	24.5	63.2
ResNet-BiLSTM-CTC	5.5	24.1	11.2	33.7	20.5	54.2
YOLOv8 + TrOCR	5.2	23.5	10.5	32.6	20.1	53.5
YOLOv8 + $\text{TrOCR}_{\text{pre}}$	<b>4.8</b>	<b>21.5</b>	<b>9.4</b>	<b>30.8</b>	<b>19.0</b>	<b>50.0</b>

## 456 7 Results and Discussion

457 Our evaluation on 2,312 test samples shows that transformer-based models consistently outperform  
 458 traditional CNN–RNN architectures for historical Russian HTR. The YOLOv8 +  $\text{TrOCR}_{\text{pre}}$ , config-  
 459 uration achieved the lowest error rates (CER: 8.5%, WER: 29.1%), representing a 53.8% reduction  
 460 in CER compared to the VGG-CTC baseline (18.4% CER) as shown in Table 1.

461 While a WER of 29.1% overall (50% on diaries) indicates that manual correction remains neces-  
 462 sary, this represents meaningful progress for an extremely challenging domain. Importantly, error  
 463 analysis reveals that many mistakes affect functional words or minor orthographic variants while  
 464 preserving key entities and overall semantic structure. This characteristic significantly reduces the  
 465 amount of meaningful human correction needed compared to the baseline errors produced by tradi-  
 466 tional architectures, making the output practically useful for search, reading, and further scholarly  
 467 analysis.

468 The performance hierarchy follows a clear pattern: VGG-CTC < CRNN < ResNet-BiLSTM-CTC  
 469 < TrOCR <  $\text{TrOCR}_{\text{pre}}$ , with each architectural advancement yielding measurable improvements.  
 470 Notably, language-specific pretraining ( $\text{TrOCR}_{\text{pre}}$  vs. TrOCR) provides consistent gains across all  
 471 document types, with improvements of 0.4–1.1% CER as evidenced in Table 2.

### 474 7.1 Document Type Impact on Recognition Accuracy

475 Recognition performance varies systematically across document types, correlating with their struc-  
 476 tural and linguistic characteristics (Table 2). Gubernatorial Reports achieve the highest accuracy  
 477 (4.8% CER, 21.5% WER), followed by Statutory Charters (9.4% CER, 30.8% WER), and Personal  
 478 Diaries (19.0% CER, 50.0% WER) using the best-performing configuration.

479 Performance correlates strongly with writing formality: 4.8% CER on formal administrative reports  
 480 versus 19.0% CER on informal personal diaries. This fourfold accuracy gap reflects not only the  
 481 inherent task difficulty posed by irregular handwriting styles but is also compounded by severe data  
 482 scarcity: Personal Diaries comprise only 960 training samples compared to 28,671 samples for Gu-  
 483 bernatorial Reports. The CER-to-WER ratio also differs by type: 1:4.5 for formal documents vs.  
 484 1:2.6 for personal diaries—consistent with shorter word lengths and greater lexical diversity in infor-  
 485 mal writing.

486  
487

## 7.2 Training Data Requirements

488  
489  
490  
491  
492  
493  
494

Dataset size directly impacts model performance, particularly for complex document types. Personal Diaries, with only 960 training samples, show the highest error rates (19.0% CER) despite representing the most challenging recognition task. In contrast, Gubernatorial Reports with 28,671 training samples achieve optimal performance (4.8% CER). This performance gap reflects the inherent scarcity of preserved informal historical manuscripts from the Russian Empire period rather than methodological limitations in dataset construction, as data scarcity compounds the inherent difficulty of irregular handwriting styles.

495  
496

## 7.3 Component Contribution Analysis

497  
498  
499  
500  
501  
502

Ablation studies reveal that accurate line localization is critical for overall system performance. When evaluated with manually annotated bounding boxes,  $\text{TrOCR}_{\text{pre}}$  achieves 7.2% CER compared to 8.5% CER in the full pipeline, indicating that detection errors contribute 1.3 percentage points to the total error rate. On the validation set, the YOLOv8 model attains  $\text{mAP} = 0.98$ , indicating high detection performance on this corpus.

503  
504  
505  
506  
507  
508  
509

Image preprocessing provides consistent improvements across all architectures, while postprocessing methods show mixed results. Dictionary-based correction helps only when base accuracy is low, and LLM-based correction degrades performance due to inappropriate historical text modifications.

This analysis establishes quantitative baselines for historical Russian HTR and shows that transformer architectures with language-specific pretraining achieve the lowest CER/WER among the evaluated models.

510  
511

## 8 Conclusion

512  
513  
514  
515  
516  
517  
518  
519

This study demonstrates that modern computer vision and natural language processing techniques can be successfully adapted for handwritten text recognition (HTR) in historical documents of varying types. We acknowledge using established components (YOLOv8 +  $\text{TrOCR}_{\text{pre}}$ ). However, the contribution lies in the successful adaptation and comprehensive evaluation of these architectures for this unique domain-pre-revolutionary Russian manuscripts from the 19th to early 20th centuries. Our system demonstrates a 53.8% CER improvement over traditional CNN-RNN baselines (from 18.4% to 8.5%), establishing the first viable automated processing pipeline for this previously inaccessible historical corpus.

520  
521  
522

The best results were obtained using transformer-based architectures, particularly when combined with language-specific pretraining and accurate line localization.

523  
524  
525  
526  
527

Recognition quality varied with the typology of the historical material. Official documents written in calligraphic script achieved high accuracy (CER as low as 4.8%, WER 21.5%), whereas Personal Diaries with individual handwriting styles were more challenging (CER up to 19%, WER 50.0%). These findings highlight the need for a differentiated processing approach tailored to document type, as well as the importance of constructing representative and typologically diverse training corpora.

528  
529  
530  
531

The resulting system achieves a character error rate (CER) of 8.5% and a word error rate (WER) of 29.1% on a held-out test set, consistently outperforming other approaches across all three document types (Gubernatorial Reports, Statutory Charters, and Personal Papers) in this domain and enabling large-scale digitization of historical Russian archives.

532  
533  
534  
535

The results establish a solid foundation for advancing automated processing of historical archival materials. Furthermore, they open new opportunities for digital humanities research by enabling scalable access to large volumes of previously unstructured handwritten sources.

536  
537

## References

538  
539

Angel Beshirov, Milena Dobreva, Dimitar Dimitrov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. Post-ocr text correction for bulgarian historical documents. *International Journal on Digital Libraries*, 26(1):4, 2025.

540 Thomas Constum, Pierrick Tranouez, and Thierry Paquet. Daniel: A fast document attention  
 541 network for information extraction and labelling of handwritten documents. *arXiv preprint*  
 542 *arXiv:2407.09103*, 2024. URL <https://arxiv.org/abs/2407.09103>.

543

544 Denis Coquenet. Meta-dan: towards an efficient prediction strategy for page-level handwritten  
 545 text recognition. *arXiv preprint arXiv:2504.03349*, April 2025. URL <https://arxiv.org/abs/2504.03349>.

546

547 Council of People's Commissars of the RSFSR. Dekret Soveta Narodnykh Kommissarov o Vvedenii  
 548 Novoi Orfografii, 1918a. URL <https://docs.cntd.ru/document/901855491>.

549

550 Council of People's Commissars of the RSFSR. Dekret Soveta Narodnykh Kommissarov o Vve-  
 551 denii Novoi Orfografii [decree of the council of people's commissioners on the introduction of new  
 552 spelling], 1918b. <https://docs.cntd.ru/document/901855491>. Accessed 15 Jan 2024.

553

554 Irina A. Degtareva. Otsifrovka Arkhivnykh Dokumentov: Reshennaya Zadacha ili Zadacha  
 555 Budushchego [digitization of archival documents: A solved problem or a future challenge]. *Is-  
 556 toriya i Arkhivy [History and Archives]*, (4):122–135, 2022.

557

558 John Alexander Edwards III. *Easily adaptable handwriting recognition in historical manuscripts*.  
 559 University of California, Berkeley, 2007.

560

561 Carlos Garrido-Munoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. Handwritten text recognition:  
 562 A survey. *arXiv preprint arXiv:2502.08417*, 2025a.

563

564 Carlos Garrido-Munoz, Antonio Rios-Vila, and Jorge Calvo-Zaragoza. Handwritten text recognition:  
 565 A survey, 2025b. URL <https://arxiv.org/abs/2502.08417>.

566

567 Sara Ghaboura, Ketan More, Ritesh Thawkar, Wafa Alghallabi, Omkar Thawakar, Fahad Shah-  
 568 baz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. Time travel: A  
 569 comprehensive benchmark to evaluate lmms on historical and cultural artifacts. *arXiv preprint*  
 570 *arXiv:2502.14865*, 2025.

571

572 Alex Graves, Santiago Fernandez, and Jürgen Schmidhuber. Connectionist temporal classification:  
 573 Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd*  
 574 *International Conference on Machine Learning*, pp. 369–376, 2006. URL <https://www.cs.toronto.edu/~graves/nc-2006.pdf>.

575

576 Heartex. Label studio: Data labeling software. <https://github.com/heartexlabs/label-studio>, 2020.  
 577 Open source data labeling tool.

578

579 Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella  
 580 Murray, and Elizabeth Spence. Unlocking the archives: Using large language models to transcribe  
 581 handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisci-  
 582 plinary History*, pp. 1–19, 2025.

583

584 Samuel Johnson. *A Dictionary of the English Language*. Printed by W. Strahan for J. and P. Knapton;  
 585 T. and T. Longman; C. Hitch and L. Hawes; A. Millar; and R. and J. Dodsley, London, 1755. Folio  
 586 edition; originally published April 15, 1755.

587

588 E. V. Kharitonova. Formation of the spelling of compound nouns with an initial imperative in the  
 589 history of russian writing. *Prepodavatel XXI Vek*. URL <http://prepodavatel-xxi.ru/sites/default/files/408417.pdf>.

590

591 Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. Early evidence  
 592 of how llms outperform traditional systems on ocr/htr tasks for historical records. *arXiv preprint*  
 593 *arXiv:2501.11623*, 2025.

594

595 Pavel P Kokorin, Aleksandr A Kotov, Sergey V Kuleshov, and Aleksandra A Zaytseva. Extracting  
 596 structured data from the chronicle of the life and work of as pushkin: Hybrid approach. *Software  
 597 & Systems*, 38(1):39–46, 2025.

598

599 V. Kuznetsov. On language, political power and the regulation of russian orthography. *David Pub-  
 600 lisher*. URL <http://www.davidpublisher.org/index.php/Home/Article/index?id=34882.html>.

594 Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun  
 595 Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models.  
 596 *arXiv preprint arXiv:2109.10282*, 2021. Version 5, published 6 Sep 2022.  
 597

598 Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun  
 599 Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models.  
 600 In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13094–13102,  
 601 2023.

602 Carlos Penarrubia, Carlos Garrido-Munoz, Jose J Valero-Mas, and Jorge Calvo-Zaragoza. Spatial  
 603 context-based self-supervised learning for handwritten text recognition. *Pattern Recognition Let-  
 604 ters*, 2025.

605 Mark Potanin, Denis Dimitrov, Alex Shonenkov, Vladimir Bataev, Denis Karachev, Maxim  
 606 Novopoltsev, and Andrey Chertok. Digital peter: New dataset, competition and handwriting  
 607 recognition methods. In *Proceedings of the 6th International Workshop on Historical Document  
 608 Imaging and Processing*, pp. 43–48, 2021.

609

610 Presidential Library named after B.N. Yeltsin. Prezidentskaia Biblioteka imeni B.N. E'Istina [presi-  
 611 dential library named after b.n. yeltsin]. <https://www.prlib.ru/>, 2024. [Official Website].

612 Aleksei I. Razdorskii. *Obzory Gubernii, Oblastei i Gradoначальств Rossiiskoi Imperii (Prilozheniya k Vsepoddanneishim Otchetam Gubernatorov, Nachal'nikov Oblastei i Gradoначальникov) 1870–1916* [Surveys of Governorates, Regions, and City Administrators of the Russian Empire (Appendices to the Most Humble Reports of Governors, Regional Heads, and City Administrators) 1870–1916]. Limited Liability Company Dmitry Bulanin Publishing House, 2011.

613

614 CA Romein, Achim Rabus, Gundram Leifert, and Phillip Benjamin Ströbel. Assessing advanced  
 615 handwritten text recognition engines for digitizing historical documents. *International Journal of  
 616 Digital Humanities*, pp. 1–20, 2025.

617

618 Russian State Historical Archive. Rossiiskii Gosudarstvennyi Arkhiv [russian state historical  
 619 archive]. <https://fgurgia.ru/>, 2024. [Official Website].

620

621 A. V. Shapovalova. On language, political power and the regulation of russian orthography. *Jour-  
 622 nal of Literature and Art Studies*. URL <http://www.davidpublisher.org/index.php/Home/Article/index?id=34882.html>.

623

624 Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-  
 625 based sequence recognition and its application to scene text recognition. *arXiv preprint  
 626 arXiv:1507.05717*, 2015. URL <https://arxiv.org/abs/1507.05717>.

627

628 Alex Shonenkov, Denis Karachev, Max Novopoltsev, Mark Potanin, Denis Dimitrov, and Andrey  
 629 Chertok. Handwritten text generation and strikethrough characters augmentation. *arXiv preprint  
 630 arXiv:2112.07395*, 2021. arXiv:2112.07395v1 [cs.CV].

631

632 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image  
 633 recognition. In *International Conference on Learning Representations (ICLR)*, 2015. URL <https://arxiv.org/abs/1409.1556>.

634

635 Kseniia A. Sofronenko. *Krest'ianskaia Reforma v Rossii 1861 Goda: Sbornik Zakonodatel'nykh  
 636 Aktov* [The Peasants' Emancipation Reform in Russia 1861: Collection of Legislative Acts]. State  
 637 Publishing House of Legal Literature, 1954.

638

639 Muhammad Yaseen. What is yolov8: An in-depth exploration of the internal features of the next-  
 640 generation object detector. *arXiv preprint arXiv:2408.15857*, August 2024. arXiv:2408.15857v1  
 641 [cs.CV] 28 Aug 2024.

642

643 Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. Xunzi-mllm: a  
 644 multimodal large language model for ancient text and image recognition. *Digital Scholarship in  
 645 the Humanities*, pp. fqaf026, 2025.

648 A Orthographic Differences in Pre- and Post-Reform Russian  
649650 The 1918 Russian orthography reform was a pivotal event that fundamentally reshaped written Russian,  
651 creating distinct visual and linguistic characteristics between pre-reform (prior to 1918) and  
652 post-reform (after 1918) texts (Council of People’s Commissars of the RSFSR, 1918a; Kuznetsov;  
653 Kharitonova; Shapovalova). These changes are crucial for understanding the challenges in historical  
654 Handwritten Text Recognition (HTR). Key Differences:  
655

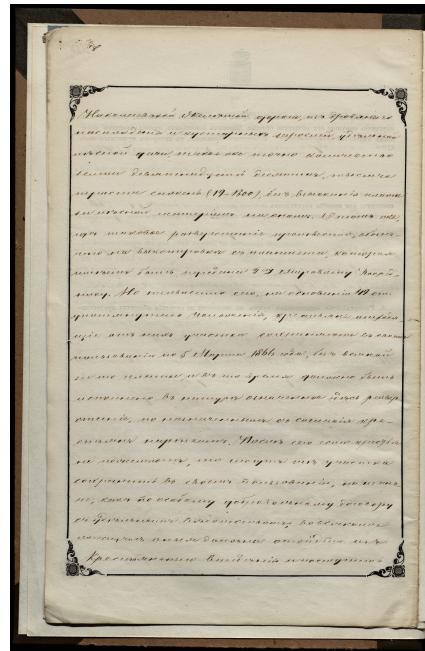
1. **Elimination of Obsolete Letters:** Four letters, primarily serving etymological rather than phonetic purposes, were removed from the alphabet:
  - Ѳ (Yat): Replaced by E. Its usage often required memorization as its pronunciation had largely merged with 'E'.
  - i (Decimal I): Replaced by И. Used before vowels, the hard sign, and the letter Й (e.g., *история*).
  - ѧ (Fita): Replaced by Ф. Primarily used in words of Greek origin (e.g., *театръ*).
  - ѵ (Izhitsa): Replaced by И or В. Used very rarely, mainly in certain Greek borrowings (e.g., *муро*).
2. **Abolition of the Hard Sign (Ђ) at the End of Words:** Prior to the reform, the hard sign Ђ was obligatorily written at the end of every word ending in a hard consonant, with no phonetic value (e.g., *домъ*, *городъ*). The reform eliminated this practice, retaining Ђ only as a separator between a prefix and a root (e.g., *объяснить*, *подъём*). This significantly altered the visual density and appearance of texts.
3. **Simplification of Prefix Spelling:** Rules for prefixes ending in -3/-C before voiced/voiceless consonants were largely streamlined. For instance, prefixes like *без-*, *воз-*, *из-*, *раз-* consistently ended in 3 before voiced consonants and vowels, and in C before voiceless consonants. While these rules existed before, the reform clarified and solidified their application.
4. **Other Minor Changes:** The reform also included adjustments to the spelling of compound nouns, certain grammatical endings, and a general move towards phonetic principles over etymological ones.

679 These fundamental changes resulted in distinct visual forms and vocabulary, making pre-reform  
680 Russian texts a highly specialized domain for HTR. Models trained on modern Russian or other  
681 languages without adaptation for these obsolete characters and spelling rules are largely ineffective,  
682 underscoring the necessity and novelty of our dedicated approach.  
683684 B Recognition Results  
685686 Here we can look at examples of documents and their transcripts in text form.  
687688 Listing 1: Results of text recognition on Figure 2  
689

690 Николаевской

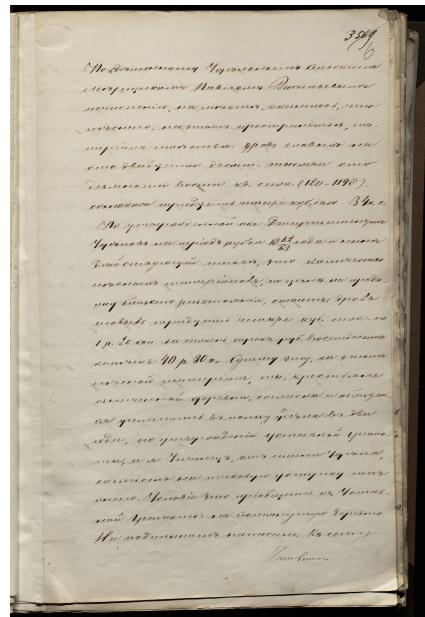
691 Железней дороги, изъ дровянского насажденія и кустарныхъ зарослей удельной лесной  
692 дачи, такое же точно количество величи девятнадцать бесятинъ, тысяча шириста  
693 сажень (19-1300), безъ высоканія планы ки лесной материале на ономъ. О том, те, гдѣ  
694 таковое реверстане произвести, оболее ного на выкотировке съ планшета, которая  
695 бомъ быть передана 8.7 Мировому Посре нику. Но независимо сего, на основани 49  
696 ст. упомянутаго Положения, крестьяне открыя ще отъ нихъ участки степаняются въ  
697 своего пользовании по 5 Марта 1866 года, безъ всякой пото платы, а въ те времена  
698 должно быть исполнено въ натуре означенное здесь развер ствене, по назначеннымъ  
699 до согласия крестьянъ наризкамъ. После сего ссаже предел не пожелокъ, то смотре изъ  
700 участки сохранить въ своемъ пользовани, на не ино не, какъ по особому добровольномъ  
701 того дору съ удельнымъ Солдатствомъ, вовсякими жащихъ оныя должны отойдели изъ  
Крестьянского визделя и поступить

702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725



726 Figure 2: Document example.  
727  
728  
729  
730  
731  
732

733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755



753 Figure 3: Document example.  
754  
755

756

## Listing 2: Results of text recognition on Figure 3

757

По  
делебному Удельнымъ Семъскимъ мерщикомъ Павломъ Васильевымъ начислению, на месте,  
оказалось, что леснаго, надимыхъ пространствъ, ма териала имеется: дровъ словыхъ на  
сто двадцати десят. тысяча ста девяносто восми кв. саж. (120-1198) сенокоса тридцать  
четыре куб. саж..34 к. ю. по утвержденной же Департаментомъ Уделовъ на периодъ  
рубки 1863 года и ныне действующий таксъ, что количество лесныхъ материаловъ, по  
цене на прода ону близкаго разстоянія, места. дровъ словыхъ тридцать четыре куб.  
сесок. то 1 р. 20 коп. на сажень сорокъ руб. восемьдесятъ количенъ 40 р. 80 к. Одину,  
на этомъ лесной материалѣ, мы, крестьяне означенной деревни, согласны и обязуемъ ся  
уплатить въ пользу удела въ два года, по утверждении уставной грамоты, а я Уманецъ,  
отъ имени Удела, согласенъ на таковую уступку имъ онаго. Условие это приобщить къ  
уставной грамоте на помянутую деревню На подлинномъ написано: Къ сему

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

щеннаго  
крестьянина Корчевскаго Уезда, деревни Третиши Степана Андреева проживающаго въ  
семь Ель цахъ Уставная Грамота прости тана была отъ слова до соловыхъ надлежащими  
поясненіями, всѣхъ пунктовъ грамоты крестьяне ни какихъ возраженій непредъявили,  
на все статьи грамоты изъявили согласіе и подпись ева по дѣтвердили, а потому  
Постановлено грамоту на основании 99 ст. положенія о крестья нахъ водворенныхъ на  
земляхъ Государевъ хъ Дворцовиа и Удельныхъ считать окончательно повереною и  
составителяхъ. и на основании 99 стати того же положенія утвердить подлин ный  
подпись Кондидатъ Ми роваго Посредника Краморевъ.

802

803

804

805

806

807

808

809

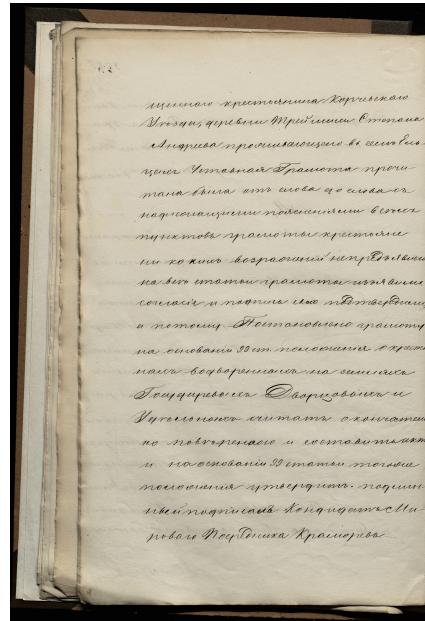


Figure 4: Document example.

## Listing 3: Results of text recognition on Figure 4