Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 453-460 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of PastReader at IberLEF 2025: Transcribing Texts From the Past

Resumen de la Tarea PastReader en IberLEF 2025: Transcribiendo Textos del Pasado

Arturo Montejo-Ráez,[1] **Elena Sánchez-Nogales,**[2] **Gloria Expósito-Álvarez,**[2]
L. Alfonso Ureña-López,[1] **María Teresa Martín-Valdivia**,[1] **Jaime Collado-Montañez**,[1]
Manuel Carlos Díaz-Galiano,[1] **Isabel Cabrera-de Castro**,[1]
María Victoria Cantero-Romero,[1] **Rocío Ortuño-Casanova**[3]
[1]University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]Biblioteca Nacional de España
[3]Universidad Nacional de Educación a Distancia
[1]{amontejo, laurena, maite, jcollado, mcdiaz, iccastro, vcantero}@ujaen.es,
[2]{elena.sanchez,gloria.exposito}@bne.es
[3]rocio.ortuno@flog.uned.es

Abstract: The PastReader 2025 task, within the framework of IberLEF 2025, focuses on the automatic transcription of digitized Spanish historical press. It uses as a basis the Digital Newspaper Library of the National Library of Spain, a collection that is part of the Hispanic Digital Library project and that gathers millions of pages of newspapers and magazines representative of the thematic and stylistic diversity of the Hispanic press. Although the documents are available in PDF with OCR, the quality of the extracted texts is often poor due to deteriorated scans, irregular page structures, old spelling, and other visual problems. To further automate this process, the task proposes two challenges: the correction of OCR errors and the generation of curated texts from scanned images, applying multimodal models. The main objective is to reduce the need for human intervention in mass digitization processes, promoting systems capable of improving the accessibility, recovery, and preservation of Spanish newspaper heritage through robust and efficient technological solutions.
Keywords: OCR, historical press, automatic transcription, Digital Humanities.

Resumen: La tarea PastReader 2025, en el marco de IberLEF 2025, se centra en la transcripción automática de prensa histórica española digitalizada. Utiliza como base la Hemeroteca Digital de la Biblioteca Nacional de España, una colección que forma parte del proyecto Biblioteca Digital Hispánica y que reúne millones de páginas de periódicos y revistas representativas de la diversidad temática y estilística de la prensa hispánica. Aunque los documentos están disponibles en PDF con OCR, la calidad de los textos extraídos suele ser baja debido a escaneos deteriorados, estructuras de página irregulares, ortografía antigua y otros problemas visuales. Para avanzar en la automatización de este proceso, la tarea propone dos retos: la corrección de errores OCR y la generación de textos curados a partir de imágenes escaneadas, aplicando modelos multimodales. El objetivo principal es reducir la necesidad de intervención humana en los procesos de digitalización masiva, promoviendo sistemas capaces de mejorar la accesibilidad, recuperación y preservación del patrimonio hemerográfico español mediante soluciones tecnológicas robustas y eficientes.
Palabras clave: OCR, prensa histórica, trasncricpión automática, Humanidades Digitales.

ISSN 1135-5948 DOI 10.26342/2025-75-33

1 Introduction

The "Hemeroteca Digital" is part of the Hispanic Digital Library project, which aims to provide public consultation and dissemination via the Internet of the Spanish Bibliographic Heritage held by the Biblioteca Nacional de España.

The Hemeroteca was created in March 2007 to provide public access to the digital collection of Spanish historical press housed in the Library, with an initial collection of 143 press and magazine titles. Today, it contains millions of digitalised pages publicly available. This digital collection has been created to serve as a key reference for the study and consultation of Spanish historical press and magazines. In addition to offering access to the texts for reading and consultation, it also provides information on major digital newspaper collections, facilitating greater awareness and access to the still partly unexplored Spanish newspaper heritage.

The guiding criterion for curating this collection has been the selection of newspapers and magazines that are representative of their time and showcase the thematic diversity of Hispanic press publishing. Visitors to the newspaper library will thus find publications spanning politics, satire, humor, science, religion, illustration, entertainment, sports, art, literature, and more.

The digital publications are provided in PDF format with OCR, enabling users to search for any desired topic within the text. These advanced text search capabilities make the Digital Newspaper Library an invaluable tool for research purposes.

The process of generating a final transcription from scanned pages is a challenging task that, nowadays, requires a vast amount of human resources. The process involves not only a very performant OCR, but also a robust error-correction approach, as many pages are of bad quality or the OCR system is just unable to reproduce the original text (spots, stains, non-standard spelling used, and other issues).

In addition, newspapers have several OCR difficulties associated with their structure: they tend to be arranged in columns that are not always regular, they include different types of images with and without text, and often news items start on one page and continue on successive non-continuous pages, with the abbreviated title at the beginning of the news item and "continued on page xx" at the end. This task aims to advance the automation of this process.

This paper presents in detail the PastReader 2025 task developed in the framework of IberLEF 2025(González-Barba, Chiruzzo, and Jiménez-Zafra, 2025). First, the state of the art in automatic processing of digitized historical documents is presented, with emphasis on recent approaches such as pixel-based models (PHD) and large multimodal language models (mLLMs). Next, the two tasks posed in the competition - OCR error correction and end-to-end text extraction - are described, along with the evaluation metrics used to measure the performance and efficiency of the proposed solutions. Subsequently, the dataset, compiled from the Spanish National Library's Digital Newspaper Library, is presented, and the methodologies applied by the participating teams (OCRTIST, GRESEL 1, and GRESEL 2) are detailed, including an analysis of the environmental impact of their approaches. Finally, the results obtained are discussed, followed by a critical discussion and a presentation of the conclusions and lines of future work.

2 State of the Art

Processing digitised historical documents presents significant challenges, such as scan quality, archaic fonts and the loss of visual context that traditional Optical Character Recognition (OCR) systems often introduce(Borenstein et al., 2023), (Fleischhacker, Goederle, and Kern, 2024) (Fleischhacker, Kern, and Göderle, 2025). To address these limitations, innovative approaches have been developed:

PHD (Pixel-based model for Historical Documents) is a pixel-based language model that processes documents as images directly, avoiding conversion to text via OCR and associated noise (Borenstein et al., 2023). It has demonstrated a high ability to reconstruct masked image patches and remarkable language understanding, being useful for historical question and answer and semantic search tasks.

Another leading approach combines machine learning-based layout detection with an OCR engine such as Tesseract, which is matched with a custom font for the specific historical document (Fleischhacker,

Goederle, and Kern, 2024) (Fleischhacker, Kern, and Göderle, 2025). This method has achieved significant improvements in OCR quality, reducing character (CER) and word (WER) error rates in documents with complex layouts. Layout detection allows context preservation and correct text reordering, especially in documents with multiple columns.

More recently, Large Multimodal Language Models (mLLMs) have emerged as a promising solution. Models such as Gemini 2.0 Flash outperform conventional OCR in transcription tasks without the need for preprocessing, achieving very high accuracy (normalized CER of 1.27%) (Greif, Griesshaber, and Greif, 2025). In addition, these mLLMs are exceptionally effective for OCR post-correction by using both the original image and the noisy transcript as input, resulting in highly accurate transcripts ($< 1\%$ CER). They also prove to be an efficient solution for Named Entity Recognition (NER) and the extraction of structured information directly from historical documents.

3 Tasks Description

Two tasks are proposed and related to the basic workflow in a transcription process (see Figure 1): extraction of text from scanned documents (OCR) and curation of the extracted text to fix found errors. But, instead of dedicating a specific task to each step, we encourage participants to overcome the following tasks:

- Task 1 Error Correction: In this task, participants are provided with the output of an OCR system and are asked to generate clean and fixed versions of the extracted texts.
- Task 2 End-to-end Extration: Due to the advances in multimodal systems, this task aims to explore end-to-end approaches, using scanned pages as input and expecting to produce curated texts as output.

3.1 Metrics

We have multiple metrics to measure systems' behaviour arranged in two main categories: performance and efficiency.

3.1.1 Performance metrics

This type of metric is intended to measure how well systems achieve the proposed task in terms of quality. Several metrics are proposed, Levenshtein metric being the main one for the final ranking of systems. Here is a summarized list of the additional performance metrics that will be calculated for each submission:

- **Word Error Rate (WER)**: Measures errors at the word level, including insertions, deletions, and substitutions.
- **Levenshtein Distance (Lev.Dist.)**: Calculates the minimum number of single-character edits (insertions, deletions, substitutions) needed to transform one text into another.
- **Normalized Edit Distance (NED)**: Normalizes the Levenshtein Distance by the length of the ground truth text.
- **BLEU (Bilingual Evaluation Understudy)**: Measures the overlap of n-grams (e.g., unigrams, bigrams) between the transcribed text and the reference text, focusing on precision.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Measures content overlap between the transcribed text and the reference using n-grams (ROUGE-N), longest common subsequences (ROUGE-L), or skip-grams (ROUGE-S).

3.1.2 Efficiency metrics

Efficiency metrics are intended to measure the impact of the system in terms of resources needed and environmental issues. We want to recognize those systems that are able to perform the task with minimal demand for resources. This will allow us to, for instance, identify those technologies that could run on a mobile device or a personal computer, along with those with the lowest carbon footprint. To this end, each submission (each prediction sent to the server) must contain the following information:

- Total RAM needed.
- Total % of CPU usage.
- Floating Point Operations per Second (FLOPS).
- Total time to process (in milliseconds).
- Kg in CO2 emissions. For this, the Code Carbon tool (Courty et al., 2024) will be used.

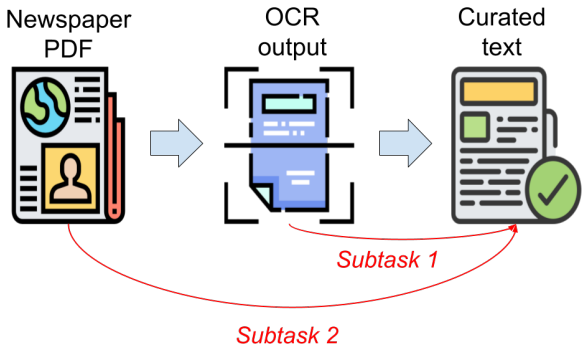


Figure 1: Workflow of transcription process and proposed tasks

A notebook with a sample code to collect this information was made available to participants[1].

4 Dataset

The corpus for this study comprises historical press publications from the public domain, digitized by the National Library of Spain (BNE) and freely accessible through Hemeroteca Digital. As of the date of this proposal, this corpus contains 298 press titles, 88,748 issues, and 8,302,407 pages. The collection continues to grow thanks to the BNE's ongoing digitization efforts.

This collection spans from the 17th to the 20th century and covers a wide range of topics. Content is accessible in PDF format via the Hemeroteca Digital viewer[2]. Additionally, the OCR text of specific digital objects can be accessed via a separate URL[3]. We provide an example for both the PDF and the OCR text in Figure 2.

The quality of OCR results varies depending on several factors, including: the date of digitization and available technology (BNE has been digitizing press materials since the late 1990s), the quality of the optical technology used, the state of preservation of the original documents, and the complexity of the text structure.

The BNE has undertaken efforts to improve the quality of these OCR-generated texts through various approaches. One of the most significant initiatives is open and collaborative OCR correction, among other projects, facilitated by the ComunidadBNE platform[4].

Split	Num. Samples	% of total
train	8959	73.46%
dev	500	4.10%
test	2736	22.43%

Table 1: Sample distribution per dataset split.

The manually corrected output serves as a valuable resource for testing and training technologies, particularly as 'ground truth' datasets. Project proposals for collaborative correction are selected based on criteria such as general interest (e.g., frequently consulted publications with poor OCR), the uniqueness of the publications, and the feasibility of an open correction process accessible to any user (i.e., avoiding overly complex orthotypographic or structural issues).

For this shared task, we compiled a corpus using the previously mentioned tools. This corpus was subsequently divided into train, development (dev), and test sets, as depicted in Table 1. This dataset has been made publicly available on Zenodo (Montejo-Ráez et al., 2025).

5 Participant Approaches

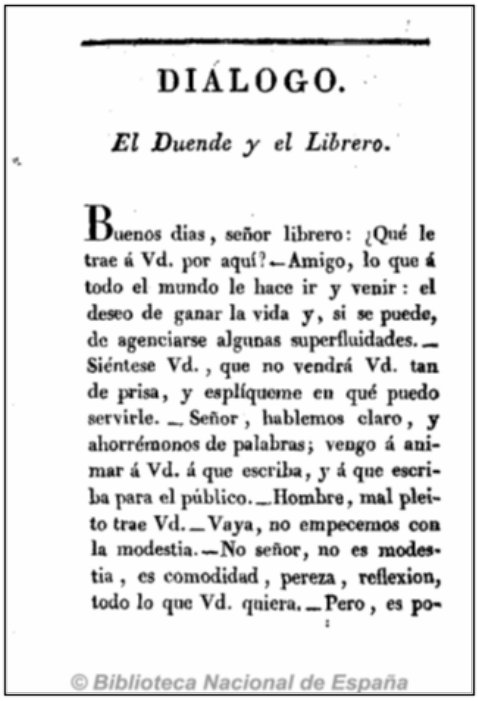
Of the two proposed tasks, only subtask 2 had participants, with three different teams: OCRTIST, GRESEL1, and GRESEL2. The systems developed by each of them and their emissions are presented below.

[1] https://colab.research.google.com/drive/1boavnGOir0urui8qktbZa0m0V2pS5cn6?usp=sharing

[2] https://hemerotecadigital.bne.es/hd/es/viewer?id=1de9fb49-08f1-43f7-a74e-28902d88bc93

[3] https://hemerotecadigital.bne.es/hd/es/text?id=1de9fb49-08f1-43f7-a74e-28902d88bc93

[4] https://comunidad.bne.es/



DIÁLOGO.

El Duende y el Librero.

Buenos dias, señor librero: ¿Qué le trae á Vd. por aquí?—Amigo, lo que á todo el mundo le hace ir y venir: el deseo de ganar la vida y, si se puede, de agenciarse algunas superfluidades.—Siéntese Vd., que no vendrá Vd. tan de prisa, y esplíqueme en qué puedo servirle.—Señor, hablemos claro, y ahorrémonos de palabras; vengo á animar á Vd. á que escriba, y á que escriba para el público.—Hombre, mal pleito trae Vd.—Vaya, no empecemos con la modestia.—No señor, no es modestia, es comodidad, pereza, reflexion, todo lo que Vd. quiera.—Pero, es po-

(a) PDF image.

```
DIALOGO.
El Duende y el Librero.

Xluenos días, señor librero: ¿Qué le trae á
Vd. por aquí?..-Amigo, lo que á todo el mundo
le hace ir y venir: el deseo de ganar la vida
y , si se puede, de agenciarse algunas
superfluidades.— Siéntese Vd. , que no vendrá
Vd. tan de f»-isa, y esplíqueme en qué puedo
servirle. _ Señor , hablemos claro, y
ahorrémonos de palabras j vengo á animar á
Vd. á que escriba, y á que escriba para el
público.—Hombre, mal pleito trae Vd._Vaya, no
empecemos con la modestia..»-No señor, no es
modeS'* tia , es comodidad , pereza ,
reflexión, todo lo que Vd. quiera. ^Pero , es
po-
```

(b) OCR text.

Figure 2: Example of a PDF and its corresponding OCR text.

5.1 Systems approaches

GRESEL1: Transcribing History with Tesseract (Macicior-Mitxelena, 2025) GRESEL1 explored a monomodal OCR strategy using the open-source Tesseract engine. The study evaluated both the out-of-the-box performance of Tesseract and the impact of domain-specific fine-tuning. Due to time limitations and the absence of structured layout data, the team focused exclusively on text-based processing.

The methodology included using Tesseract version 5.5.0, applying it first as a baseline and then with custom fine-tuning via the `tesstrain` toolkit. The dataset provided by the organizers included 8,959 training pages, 500 development pages, and 3,000 test pages, all sourced from the digitized archives of the Biblioteca Nacional de España (BNE). Two runs were submitted:

- **GRESEL1_run2 (Baseline):** Direct application of the default pre-trained Tesseract model on the test set. No data adaptation or model modification was conducted.
- **GRESEL1_run3 (Fine-tuned):** The base Spanish model was fine-tuned using the training set, involving image-to-TIFF conversion, box file creation, LSTM-compatible dataset preparation, and automated alignment with transcriptions.

Evaluation was conducted uniformly using high-resolution TIFF images.

GRESEL2: Fine-Tuning a Compact Multimodal Model on Consumer-Grade Hardware (Torterolo-Orta and Míguez-Lamanuzzi, 2025) GRESEL2 focused on developing a lightweight multimodal OCR system using IBM's Granite3.2-vision:2b model. Their work emphasized accessibility by executing the solution on consumer-grade hardware (an NVIDIA RTX 5080 with 16 GB VRAM).

The model was fine-tuned using a custom dataset format structured as chat-like interactions: a system prompt, a user prompt ("Please perform OCR on this Spanish document."), the document image in PNG format, and the expected OCR output. Images were resized to 414×585 pixels with padding to maintain aspect ratio, and the model was optimized via QLoRA (Quantized Low-Rank Adapter), enabling 4-bit quantization combined with LoRA adapters. Training employed a batch size of 1 with 8 gradient accumulation steps, bfloat16 precision, and the

AdamW optimizer.

OCRTIST: Beyond Traditional OCR – Leveraging LLMs in Document Processing (Narbona and Ros, 2025) OCRTIST presented a comparative study between two OCR pipelines: a multi-stage agent-based architecture and a minimalistic solution relying entirely on large language models (LLMs). Both aimed to transcribe Spanish historical documents with high fidelity while preserving orthographic authenticity.

The first architecture integrated three OCR engines (Surya, OlmOCR, Gemini 2.0) and fused their outputs using GPT-4o through Chain-of-Thought prompting. Prompts instructed the model to preserve historical spelling and formatting, correct OCR noise, and reconcile differences. The output was then refined through post-processing to remove artifacts without modernizing the language.

The second, and ultimately preferred, architecture used Gemini 2.5 PRO as a standalone system, with a single prompt performing OCR directly from scanned images. This setup required no preprocessing and achieved superior performance across all evaluation metrics.

Together, these three approaches illustrate a rich spectrum of strategies, from classic OCR systems to cutting-edge LLMs, highlighting the evolving landscape of document transcription in Digital Humanities.

5.2 Emissions

Table 2 shows the emissions of the different participant runs. In the case of the last three runs, the participants followed systems that did not allow them to extract this type of data.

6 Results

Table 3 shows the results obtained by the teams in the different runs they have sent for evaluation with the aforementioned metrics. As we can see, all three teams have developed systems capable of exceeding the proposed baseline, which was implemented using the Tesseract OCR Engine (Smith, 2007).

OCRTIST presented the system with the best overall results, outperforming the BASELINE on all indicators. Levenshtein Distance is the lowest among all participants with a value of 56.30, as well as 0.234 in the WER metric and 0.019 in NED. In the case of BLEU, it is observed that it obtains the highest values with 0.803. In the ROUGE metrics (1/2/L/LSUM), ROUGE2 with 0.806 and ROUGELSUM with 0.884 stand out.

GRESEL1 performed three different runs (GRESEL1_run1, GRESEL1_run2 and GRESEL1_run3). The first one stands out for having the best BLEU within the three runs, with a value of 0.686, and the best WER with 0.293, although it presents a higher Levenshtein with an index of 105.18. Run3 obtains the best Levenshtein and NED within all the runs performed, with values of 89.14 and 0.0302, respectively.

GRESEL2 carried out the execution of two runs: GRESEL2_run1 and GRESEL2_run2. If we look at the metrics obtained in the first of them, we can see the WER at 0.264, below the GRESEL1 runs, and high ROUGE and BLEU scores, especially in ROUGE2 with 0.805, very close to that obtained by OCRTIST. The second run presents a worse performance in WER (0.452) and Levenshtein (105.91), although with the highest BLEU of the team with 0.695.

7 Discussion

The results from the PastReader 2025 shared task reveal a clear trend in the evolution of historical document transcription, where large-scale multimodal models (LLMs) are outperforming traditional OCR architectures. The OCRTIST system, which employed an LLM like Gemini 2.5 Pro in an end-to-end approach, achieved the best results across all key performance metrics, including Levenshtein distance and WER. This finding is significant because it demonstrates that a single model, without the need for image preprocessing or a complex multi-engine architecture, can effectively handle the visual and textual irregularities of historical Spanish press. The simplicity and superiority of this minimalist approach call into question the long-term viability of systems that rely on separate components for layout detection, character recognition, and post-correction.

On the other hand, the approaches from the GRESEL teams show the value and limitations of more traditional strategies and compact models. GRESEL1, using Tesseract, confirmed that while fine-tuning improves performance over the baseline model, it still lags considerably behind LLM-based solu-

Team	Duration	Emissions	CPU_Energy	GPU_Energy	RAM_Energy	Enegy_Consumed	CPU_Count	GPU_Count	RAM_Total_Size
OCRTISTrun1	58,553.2138	0.1595	0.6910	0.2240	0.0015	0.9165	32	1	62.5033
GRESEL1run3	4,668.7926	0.0109	0.0551	0.0000	0.0076	0.0627	12	nan	15.6921
GRESEL1run2	2,283.7284	0.0053	0.0269	0.0000	0.0037	0.0307	12	nan	15.6921
GRESEL2run1	-	-	-	-	-	-	-	-	-
GRESEL1run1	-	-	-	-	-	-	-	-	-
GRESEL2run2	-	-	-	-	-	-	-	-	-

Table 2: Emissions by participating systems in Task 2.

Team	Lev. Dist	WER	NED	BLEU	ROUGE1	ROUGE2	ROUGEL	ROUGELSUM
OCRTISTrun1	56.3023	0.2344	0.0191	0.8035	0.8849	0.8065	0.8834	0.8843
GRESEL1run3	89.1427	0.3846	0.0302	0.6220	0.8232	0.6907	0.8180	0.8226
GRESEL1run2	93.3819	0.3650	0.0316	0.6229	0.8306	0.7114	0.8256	0.8299
GRESEL2run1	97.2399	0.2643	0.0330	0.6890	0.8841	0.8049	0.8806	0.8837
BASELINE	98.4497	0.3725	0.0334	0.6083	0.8244	0.7014	0.8190	0.8237
GRESEL1run1	105.1823	0.2933	0.0356	0.6863	0.8170	0.7110	0.8103	0.8167
GRESEL2run2	105.9065	0.4516	0.0359	0.6949	0.8686	0.7914	0.8626	0.8643

Table 3: Results achieved by participants in Task 2.

tions. Meanwhile, GRESEL2 demonstrated the feasibility of adapting a compact multimodal model for use on consumer-grade hardware, a crucial aspect for the accessibility and democratization of these technologies. Although its performance did not match that of OCRTIST, its results, particularly in the ROUGE2 metric, were competitive and far superior to the baseline, suggesting a promising balance between resource efficiency and transcription quality.

Together, the diversity of strategies highlights a turning point in the field: while cutting-edge LLMs set a new standard for quality, there is an important niche for optimized and accessible solutions that can facilitate digitization for institutions with limited resources.

As a drawback aspect of the winning approach, the OCRTIST team's results are dependent on Gemini 2.5 Pro, a "black box" model. The exact architecture, training data, and version they used may not be publicly available or could change without notice. Another researcher cannot independently replicate their setup from scratch. The scientific contribution, therefore, is more about demonstrating the application of a powerful tool rather than creating a new one. Nevertheless, the primary goal of a shared task is to find the absolute best solution to a specific problem. By including Gemini, the PastReader 2025 task successfully identified the current, undisputed state-of-the-art for this transcription problem.

8 Conclusions and future work

The PastReader 2025 shared task has highlighted the remarkable progress in the automatic transcription of historical press, underscoring the transformative role of multimodal language models. The success of the LLM-based OCRTIST system establishes a new benchmark, proving that end-to-end approaches can outperform traditional OCR systems and complex modular architectures in producing high-fidelity texts from digitized documents. This breakthrough is fundamental for the preservation and accessibility of newspaper heritage, as it reduces the need for costly human intervention and accelerates mass digitization processes. Future work should focus on optimizing these powerful models for greater efficiency and exploring their application to an even wider range of historical documents, like those with strong content fragmentation (historic press, for example), thus ensuring that our written past is fully accessible for future generations.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, and the scholarship (FPI-PRE2022-105603) from the Ministry of Science, Innovation and Universities of the Spanish Government.

We gratefully acknowledge the Spanish research infrastructure CLARIAH-ES for their essential support in making this lab possible by bringing together specialists from diverse disciplines.

References

Borenstein, N., P. Rust, D. Elliott, and I. Augenstein. 2023. Phd: Pixel-based language modeling of historical documents.

Courty, B., V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, L. O. N. de Araújo, JPW, and MinervaBooks. 2024. mlco2/codecarbon: v2.4.1, May.

Fleischhacker, D., W. Goederle, and R. Kern. 2024. Improving ocr quality in 19th century historical documents using a combined machine learning based approach.

Fleischhacker, D., R. Kern, and W. Göderle. 2025. Enhancing ocr in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries*, 26(3).

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org.*

Greif, G., N. Griesshaber, and R. Greif. 2025. Multimodal llms for ocr, ocr post-correction, and named entity recognition in historical documents.

Macicior-Mitxelena, J. 2025. Transcribing history with tesseract: A gresel1 participation in pastreader 2025. In *Proceedings of the IberLEF 2025 - PastReader Shared Task*. IberLEF 2025, Zaragoza, Spain.

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

Narbona, A. and S. Ros. 2025. Beyond traditional ocr: Exploring the efficiency of llms in document processing. In *Proceedings of the IberLEF 2025 - PastReader Shared Task*. IberLEF 2025, Zaragoza, Spain.

Smith, R. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Torterolo-Orta, Y. A. and M. Míguez-Lamanuzzi. 2025. Fine-tuning a compact multimodal model on consumer-grade hardware at pastreader 2025. In *Proceedings of the IberLEF 2025 - PastReader Shared Task*. IberLEF 2025, Zaragoza, Spain.