Historical Ink: 19th Century Latin American Spanish Newspaper Corpus with LLM OCR Correction

Laura Manrique-Gómez[1] **Tony Montes**[2] **Arturo Rodríguez-Herrera**[3] **Rubén Manrique**[2]
[1] History and Geography Department, Universidad de los Andes, Bogotá D.C.
[2] Systems and Computing Engineering Department, Universidad de los Andes, Bogotá D.C.
[3] Civil and Environmental Engineering Department, Rice University, Houston TX

{l.manriqueg, t.montes, rf.manrique}@uniandes.edu.co
da.rodriguezh@rice.edu

Abstract

This paper presents two significant contributions: First, it introduces a novel dataset of 19th-century Latin American newspaper texts, addressing a critical gap in specialized corpora for historical and linguistic analysis in this region. Second, it develops a flexible framework that utilizes a Large Language Model for OCR error correction and linguistic surface form detection in digitized corpora. This semi-automated framework is adaptable to various contexts and datasets and is applied to the newly created dataset.

1 Introduction

The computational processing of historical newspaper texts is crucial due to the valuable information these texts contain about political, economic, and cultural history. Over the past three decades, Digital Humanities has driven extensive digitization efforts, resulting in numerous curated digital collections (Berry and Fagerjord, 2017; Dobson, 2019). However, converting these images into machine-readable texts remains challenging, particularly in achieving accurate transcription. A primary challenge is the accuracy of OCR technology, especially with the extremely diverse newspaper layouts, materially degraded documents, and non-standardized fonts typical of historical texts. Traditional OCR methods often produce errors that complicate subsequent analysis.

To address these challenges, we employed GPT-4o-mini (OpenAI, 2024), a Large Language Model (LLM), within a pipeline for OCR error correction. While the LLM is capable of fixing OCR-related errors that traditional systems often miss (Langlais, 2024), our pipeline also detects and classifies potential hallucinations to avoid further issues and streamline the process. Additionally, it contributes by identifying surface forms—specific word occurrences—within the dataset.

1.1 Related Work

The "Chronicling America" initiative marks a significant advancement in the digitization of historical newspaper materials (Humanities). Another major effort, is the "Atlas - Oceanic Exchanges" collection, which traces global information networks in 19th-century newspaper materials (Exchanges). Similarly, “Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines” (Cordell and Smith) explores the culture of reprinting in the U.S. before the Civil War, while the European “Project Impresso: Media Monitoring the Past” (SNSF and FNR, 2023) addresses the OCR challenges specific to English and Germanic languages.

Despite these advancements, historical newspapers are scarcely digitized in the Global South (LeBlanc, 2024). Consequently, a gap remains in specialized corpora for 19th-century Latin American newspapers, limiting the study of the region’s unique historical and linguistic features. Our research addresses this gap by introducing a new dataset of Latin American newspaper texts in old Spanish. This dataset was post-processed with LLM models for addressing OCR errors and distinguishing them from historical linguistic surface forms[1].

ICDAR post-OCR correction competitions in 2017 and 2019 (Chiron et al., 2017; Rigaud et al., 2019) presented interesting solutions to error detection and correction in 10 European languages, such as Clova AI model based on multi-lingual BERT. Similarly, Nguyen et al. (2020) achieved comparable results by initializing embeddings with popular static embeddings such as GloVe (Pennington et al., 2014). In another approach, Veninga (2024) examined the fine-tuning of ByT5, a character-

[1]The dataset is available at https://huggingface.co/datasets/Flaglab/latam-xix in its three versions: "original", "cleaned", and "corrected"

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 132–139
November 16, 2024.

level LLM, emphasizing the importance of preprocessing and context length optimization. This results aligns with earlier studies on character-level models, such as Amrhein and Clematide (2018), which demonstrated the potential of character-based sequence-to-sequence models in improving OCR correction.

The application of LLMs for post-OCR correction has gained traction, especially in improving the accuracy of digitized historical texts. Early work by Nguyen et al. (2021) laid the foundation by categorizing post-OCR correction methods, highlighting the challenges associated with isolated-word and context-dependent approaches. As discussed by Thomas et al. (2024), the introduction of Transformers' architecture leads to state-of-the-art performance in various text correction tasks and also presents a new baseline for post-OCR correction.

Langlais (2024) builds on this foundation by addressing the persistent issue of OCR quality in cultural heritage texts. They propose that LLMs can significantly enhance correction accuracy through context-aware processing, although challenges like hallucinations and language switching remain. More recent work by Thomas et al. (2024) demonstrates the superiority of a prompt-based approach using Llama 2 over traditional models like BART (Soper et al., 2021a), reducing character error rates (CER) by over 54%. These findings are consistent with those of Soper et al. (2021b), who reported comparable improvements using fine-tuned BART models. These studies highlight the evolution from traditional correction methods to LLM-based approaches. Nevertheless, further studies are needed to test correction methods in historical documents containing linguistic and regional variants.

2 Sourcing

The dataset was initially compiled from Colombian digital newspaper archives. The primary focus was on publications that included cartoons or illustrations, which were intended for subsequent multimodal modeling. This review also extended to the physical collections on-site, as only approximately 50% of the physical collection had been digitized. Through this process, 64 newspaper titles were identified, representing 7% of the total 1,655 publications in the collections. This first iteration resulted in a dataset consisting of 4,032 pages of scanned pages of newspapers, primarily from

Figure 1: El Oso, Peru. An example of a scanned newspaper image. The corresponding OCR-extracted text and the corrected version can be found in Appendix A, for reference.

Nueva Granada—a former country encompassing Colombia, Panama, Venezuela, and Ecuador—.

A second iteration completed the revision of 3,038 digitized newspapers of 58 digital collections across Mexico, Argentina, Colombia, Peru, Chile, Panama, Venezuela, Uruguay, Bolivia, Cuba, and Ecuador as shown in Table C1 . Some countries, such as Bolivia, Cuba, and Venezuela have very limited or no web collections, resulting in their underrepresentation or absence from the final dataset. Additionally, some newspapers were printed in Europe due to lower costs; in some cases, printing outsourcing was utilized. The final dataset comprises 197 newspaper titles and 23,522 pages of scanned images, primarily from Mexico City (Mexico is the only country that has digitized its entire collection), but also includes publications from other Latin American cities, such as Buenos Aires, Lima, Bogota, and Santiago de Chile. An example of a newspaper image can be observed in Figure 1.

Originally, the Latin American 19-century newspaper dataset consists of scanned images. These images were processed using a layout model, followed by an OCR service. The layout model was specifically trained using data from annotated newspapers available in Roboflow OCR (2022); Alpha (2023); RSCOE (2023); GrabadosXIX (2023). These datasets were merged into a single dataset (CD) consisting of 1368 images of newspapers annotated for binary layout classification: images and texts. The CD dataset includes 10% of images from our newspapers dataset, labeled by hand, and it was enriched with data augmentation for shear and rotation. These techniques help to increase the model's performance in images with scanning errors.

The CD dataset was used to train an image

recognition model from Azure Cognitive Services[2], which can extract the images in the newspaper page and extract the text through the OCR. The model's performance scored MAP@75 of 87.0%, resulting in a collection of annotations and coordinates for both text and images. These coordinates were used to crop the original image, and then process it with the OCR model. Once the OCR results were obtained, we merged the processed text with the images, creating a dataset that contains the newspaper images and their associated text. From a sample of 2,500 transcribed texts, each containing 1,000 characters, manual supervision revealed that 8.5% were unreadable. The remaining texts contained multiple transcription errors, primarily due to the artisanal printing techniques and the grammatical and lexical variations of the era. These errors significantly impacted readability, introducing bias when using the texts as input for NLP-LLM models.

3 Processing

The dataset includes samples of newspapers that were either handwritten or produced using early carving machines. Over time, these machines would wear out, leading to text features that were easily confused with backward accent marks, unwanted punctuations, or misplaced characters between words. Such misreadings disrupted the continuity of the text without adding any semantic meaning.

Detecting these errors automatically poses a challenge due to the linguistic shifts between modern and 19th-century Spanish. OCR models trained on such historical texts are lacking, especially considering the semantic and orthographic changes over time. For instance, what might appear as an OCR error could instead be a historical surface form of a word; for example, the conjunction "y" (and) was often written as "I".

Additionally, some texts were completely unintelligible for OCR, and challenging for humans to interpret, due to the fonts used in certain newspapers. The varied layouts of these newspapers also resulted in texts filled with scores or numbers, or in some cases, samples containing only chapter titles or numbering (e.g., "III IV V"), which added noise to the dataset. A general overview of the pipeline from the source until the final post-processed, is observed in Figure 2.

[2]Model available through Azure cloud services at https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/

3.1 Cleaning and filtering

Some of the most common cleaning steps for text data include removing duplicates and noisy data, which are particularly crucial for subsequent analysis. In this case, 3.08% of rows were removed due to duplicates or empty texts. Additionally, 1.74% of rows were filtered out where over 50% of the characters were non-alphabetic, as these rows are more likely to be noise than useful content. Rows with four or fewer tokens were also removed, accounting for 0.61% of the data; this was achieved by training a new tokenizer with a vocabulary size of 52,000, derived from the BETO (Spanish BERT) pre-trained tokenizer (Cañete et al., 2020).

3.2 Post-OCR LLM Correction

As previously discussed, LLMs have established a baseline for correcting OCR errors in historical texts (Thomas et al., 2024; Langlais, 2024). Detecting and fixing OCR errors from newspapers is challenging because these errors are often subtle and numerous. This problem is especially pronounced with 19th-century newspapers, where the quality of the paper and the outdated printing methods contribute to a high frequency of errors. These errors create significant noise and complicate the text correction process (Lopresti, 2008).

In this paper, we use a technique for detecting OCR errors and correcting them using GPT-4o-mini and taking advantage of the fact that LLMs were trained mostly in modern language. This way, manually checked rules can classify corrections between errors, word surface forms, or none of both (hallucinations). These rules, explained in the following section, were revised and selected by a field expert who served as well as an evaluator for these corrections testing their precision for this case.

We employed a *diff* algorithm to detect the differences between the original and corrected texts. This approach allowed us to fully leverage the LLM's ability to correct the text while ensuring a reliable and structured output. The *diff* algorithm identifies added, removed, and changed parts between the two texts, similar to the functionality seen in GitHub's blame feature. By doing so, we can specify the exact changes made during the correction process, enabling us to classify these alterations effectively.

This method proved more effective than instructing the LLM to return corrections in a specific

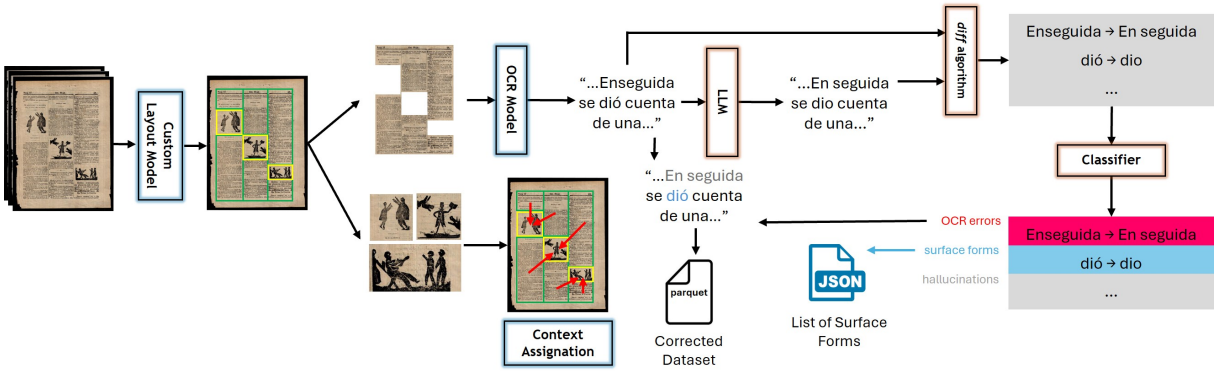

Figure 2: Overview of the full methodology pipeline. The blue components correspond to the Layout+OCR stage to get to digitized text, and the orange components correspond to the Post-OCR LLM Correction stage. The two outputs of the pipeline are the LatamXIX **Corrected Dataset** and the **List of Surface Forms**. The *Custom Layout Model* also extracts the images of the newspaper which are then assigned to the related texts (context). The final version of the text has the OCR errors corrected but not the surface forms, as they are part of the language.

format, such as JSON, as the *diff* algorithm produced shorter, more consistent, and less variable outputs. Additionally, this differentiation allows us to ignore any additions or deletions that result from LLM hallucinations, focusing instead on meaningful changes. An example of the original text, the corrected version, and the detected differences can be found in Appendix A, as well as the parameters chosen for this step.

3.3 Corrections Classification

Once the corrections are detected and isolated through the *diff* algorithm, the last step is to classify them. Still, first, it is important to state the main differences between the possible labels for each correction:

- **Surface form:** In linguistics, the term *surface form* (or word form) denotes the specific appearance of a word in a given context, contrasting with its lexical form, which pertains to its meaning (Sarveswaran et al., 2019). During the 19th century in Latin America, certain words were documented with variant spellings reflecting language shifts over time. It's important to note that changes in surface forms do not necessarily alter the semantic content of the word, but rather represent orthographic modifications.
- **OCR error:** An OCR error, on the other hand, refers to every possible misread text from the real newspaper text. The OCR errors must be corrected but must be carefully separated newspaper linguistic "errors" that contribute to the linguistics of the time.
- **Hallucinations:** If none of the above is the case, the correction is an LLM hallucination or a translation to modern Spanish, which would be wrong, so these corrections must be omitted.

To enhance classification rule analysis, corrections were noted along with their frequency across the dataset to assess relevance. All corrections were converted to lowercase for effective grouping. Many corrections were reviewed and consolidated into a set of linguistic rules for categorization. This framework can be used to identify and analyze similar changes and classification rules in other languages and contexts. This paper presents a validated set of standardized rules and exceptions for classifying corrections in the LatamXIX dataset.

3.3.1 Accent changes

Corrections involving only accent changes (addition or removal) between the original and corrected texts refer mostly to **surface forms**, given the differences between 19th-century Spanish accent rules and modern ones (Montgomery, 1966). This includes varied accent expressions for the same word, such as "antes" sometimes written as "ántes". Surface forms pose problems for NLP tasks because, in Spanish, words without accents can have different meanings, such as "acepto" (present) and "aceptó" (past). Thus, for some NLP tasks, focusing on surface forms without accent changes may be preferable, which is another outcome presented in this paper.

Feature	Value
Size	$\sim 128MB$
Rows	64,077
Words	$\sim 22M$
Tokens	$\sim 28.7M$
Newspapers	197
Years Range	1806 - 1899
Total Corrections	830,951
Surface Forms	37,492
Non-Accent Surface Forms	7,466
% of OCR Error Corrections	12.33%
% of Hallucinations Detected	77.96%

Table 1: Final Historical Ink: LatamXIX LLM Post-OCR corrected dataset

3.3.2 Specific changes

A set of letter-to-letter changes was extracted to represent key **surface words** and common **OCR errors**. For surface words, common changes include "y" for "i" or "g" for "j", e.g., "mui" for "muy" and "jeneral" for "general"; in fact, the connector "y" used to be written as "i" in most of the early 19th-century texts (Bouzouita and Gutiérrez, 2015). Common OCR errors include accent misreading or number confusion, such as "ó" read as "6" or "i" as "1". Appendix B shows a list of surface form changes.

3.3.3 Other letter-to-letter changes

When the number of letters in the original and corrected texts matches, changes generally refer to **OCR errors**, e.g., "la" misread as "In" or "señor" as "sefor".

3.3.4 Remaining changes

Corrections not fitting the preceding categories are challenging to classify as OCR errors or hallucinations, particularly with multiword corrections. A text similarity ratio was computed based on positional character matches between the original and corrected texts. This ratio, combined with the number of words in the corrected text and correction frequency, helped categorize corrections. For instance, "ascripeión" to "suscripción" had a ratio of 0.76, while "que" to "como" had a ratio of 0.0, effectively distinguishing most cases.

4 Results

Following the outlined steps, we produced the LatamXIX dataset, as shown in Table 1 and detailed in Appendix C, alongside a flexible LLM OCR correction framework. This framework allows for easy interchange between datasets or LLMs, facilitating further research. We also compiled a list of 19th-century Latin American Spanish surface forms from newspapers and developed a general framework for detecting these forms in diverse contexts.

Old Spanish surface forms are particularly useful for semantic change detection, capturing meaning variations of specific words and aiding comparisons of their historical evolution across different periods and Spanish-speaking regions.

In terms of LLM post-OCR corrections, the system generated 830,951 corrections. However, a notable 78% of these were classified as hallucinations, indicating the model's tendency to generate incorrect or fabricated content when uncertain. Only 12% addressed actual OCR errors, reflecting the core objective of the framework. This gap highlights a key limitation of current LLM models in historical OCR correction, where distinguishing between genuine errors and hallucinations remains a challenge, especially in specialized datasets.

Moreover, due to Azure OpenAI's API content policy for the chosen LLM (GPT-4o-mini), 2,899 rows (4.52%) were excluded from processing because they contained content flagged as harmful, violent, or sexual. This limitation underscores the challenges content moderation policies pose when applying LLMs to historical texts. The percentage of flagged content provides insight into the prevalence of such material in 19th-century Spanish, offering valuable perspectives for comparative analysis with modern Spanish [3].

5 Future Work

While the OCR correction using LLMs has progressed towards a more automated pipeline, a substantial portion of rule definition within the presented framework still requires manual professional input. To advance this process, future work should aim to enhance the automation of the rule-defining procedures. By reducing the reliance on human expertise, we can improve both the efficiency and accuracy of the OCR correction framework.

[3]The dataset, surface forms, and processing steps are available in https://github.com/historicalink/LatamXIX

6 Limitations

A significant limitation of this work is the reliance on manual evaluations for assessing OCR accuracy, as most evaluations and rule definitions were performed by experts. This manual process introduces subjectivity and limits scalability. The absence of a comprehensive automated evaluation method prevents more consistent accuracy assessments and restricts the ability to refine the framework based on objective metrics like Character Error Rate (CER).

7 Acknowledgements

We would like to thank the two anonymous reviewers from the EMNLP NLP4DH conference for their helpful feedback and suggestions.

References

Alpha. 2023. Newspaperbox Dataset.

Chantal Amrhein and Simon Clematide. 2018. Supervised OCR error detection and correction using statistical and neural machine translation methods. *Journal for Language Technology and Computational Linguistics*, 33(1):49–76.

David M. Berry and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and critique in a Digital age*. Polity Press.

Miriam Bouzouita and Mara Fuertes Gutiérrez. 2015. Spanish studies: Language and linguistics. *The Year's Work in Modern Language Studies*, 75:171–185.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017. Icdar2017 competition on post-OCR text correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1423–1428.

Ryan Cordell and David Smith. The viral texts project. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines (2022)*.

James E. Dobson. 2019. *Critical Digital Humanities: The search for a methodology*. University of Illinois Press.

Oceanic Exchanges. The atlas. *Mapping the Histories and Metadata of Digitised Newspapers Collections Around the World. (2021)*.

GrabadosXIX. 2023. Grabados_Sample Dataset.

National Endowment for the Humanities. Chronicling america: Library of congress. *News about Chronicling America RSS*.

Pierre-Carl Langlais. 2024. Post-OCR-correction: 1 billion words dataset of automated OCR correction by llm. *Hugging Face*.

Zoe LeBlanc. 2024. More than keywords. *The American Historical Review*, 129(1):164–168.

Daniel Lopresti. 2008. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, page 9–16. Association for Computing Machinery.

Thomas Montgomery. 1966. On the development of spanish y from "et". *Romance Notes*, 8(1):137–142.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-OCR processing approaches. *ACM Comput. Surv.*, 54(6).

Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 2020. Neural machine translation with BERT for post-OCR error detection and correction. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 333–336, New York, NY, USA. Association for Computing Machinery.

OCR. 2022. OCR_project dataset.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Christophe Rigaud, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2019. ICDAR 2019 competition on post-OCR text correction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1588–1593.

RSCOE. 2023. Newspaper Dataset.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden, Germany. Association for Computational Linguistics.

SNSF and FNR. 2023. Impresso - Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021a. BART for post-correction of OCR newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Elizabeth Soper, Stanley Fujimoto, and Yen-Yun Yu. 2021b. BART for post-correction of OCR newspaper text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 284–290, Online. Association for Computational Linguistics.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

M.E.B. Veninga. 2024. LLMs for OCR post-correction.

A LLM Correction

A.1 Prompt

Below is the prompt used to request the LLM to correct the historical text extracted by the OCR model. This prompt remained unchanged for the correction of the entire dataset and was generated through manual trial and error, ensuring it was concise enough to accommodate the potential length of the text.

```
Dado el texto del siglo XIX
entre ```, retorna únicamente
el texto corrigiendo los errores
ortográficos sin cambiar la
gramática. No corrijas la
ortografía de nombres:
```

{text}
```
```

Equivalent to the following prompt in English:

```
Given the 19th-century text
between ```, return only the text
with spelling errors corrected
without changing the grammar.
Do not correct the spelling of
names:
```

{text}
```
```

A.2 Example

The LLM response was successful for most of the texts except for some cases where Azure's Content Policy was triggered due to text content, and for very long texts where the model started to hallucinate the whole text. An example of an original text, its retrieved LLM correction, and all the changes detected by the diff algorithm is the following (surface forms and OCR errors) is:

- **Original:** La publicacion del Oso se harà dos veces cada se mana, y constará de un pliego en cuarto ; ofreciendo à mas sus redactores, dar los gravados oportunos, siempre que loexija el asuntode que trate. Redactado por un Num. 8. TEMA del Periodico. POLITICA MILITAR. OCTAVA SESION. Abierta la sesion á las dore y un minuto de la noche , 25 de Febrero de 1845 , con asistencia de todos los Señores Representantes, se leyó y aprobó la acta de la Asamblea anterior , ménos en lo tocante à la torre del Convento de Santo Domingo, punto que quedó para ventilarse en mejor ocasion. En seguida se dió cuenta de una nota del Ejecutivo , referente à que urjía la necesidad de organizar un Ejército ; pues decia el Excmo. Decano: - "Un poder sin bayonetas vale tanto como un cero puesto á la izquierda."

- **Corrected:** La publicación del Oso se hará dos veces cada semana, y constará de un pliego en cuarto; ofreciendo además sus redactores, dar los grabados oportunos, siempre que lo exija el asunto de que trate. Redactado por un Num. 8. TEMA del Periódico. POLÍTICA MILITAR. OCTAVA SESIÓN. Abierta la sesión a las dos y un minuto de la noche, 25 de Febrero de 1845, con asistencia de todos los Señores Representantes, se leyó y aprobó la acta de la Asamblea anterior, menos en lo tocante a la torre del Convento de Santo Domingo, punto que quedó para ventilarse en mejor ocasión. Enseguida se dio cuenta de una nota del Ejecutivo, referente a que urgía la necesidad de organizar un Ejército; pues decía el Excmo. Decano: - "Un poder sin bayonetas vale tanto como un cero puesto a la izquierda."

B Specific Surface Form Changes

For the surface form extraction from the texts and its differentiation from OCR errors and LLM hallu-

cinations, a set of surface form changes was constructed for 19th-century Latin American Spanish. The complete set of known changes with an example for each case is presented in Table B1.

Change	Example
á ↔ a	hara → hará
é ↔ e	fué → fue
í ↔ i	decia → decía
ó ↔ o	ocasion → ocasión
ú ↔ u	ningun → ningún
i ↔ y	mui → muy
j ↔ g	jente → gente
v ↔ b	gravado → grabado
s ↔ x	espiró → expiró
j ↔ x	méjico → méxico
c ↔ s	faces → fases
s ↔ z	dies → diez
z → c	doze → doce
q → c	quatro → cuatro
n → ñ	senor → señor
ni → ñ	senior → señor
k → qu	nikel → níquel
k → c	kiosko → quiosco
ou → u	boulevar → bulevar
s → bs	suscriciones → subscripciones
c → pc	suscriciones → subscripciones
s → ns	trasportar → transportar
t → pt	setiembre → septiembre
rt → r	libertar → liberar
r ↔ rr	vireinato → virreinato
...lo → lo ...	cambiólo → lo cambió
...se → se ...	acercóse → se acercó

Table B1: Set of Surface Form change rules to extract them from the LatamXIX dataset

C Dataset Overview

A more specific overview of the dataset is described in Figure C1 and Table C1.

Country	Presence (%)
Mexico	49.59%
Argentina	21.23%
Colombia	12.53%
Peru	8.43%
Chile	6.39%
Panama	0.83%
Venezuela	0.52%
Uruguay	0.17%
France	0.16%
Ecuador	0.09%
Spain	0.06%

Table C1: LatamXIX dataset presence distribution grouped by country

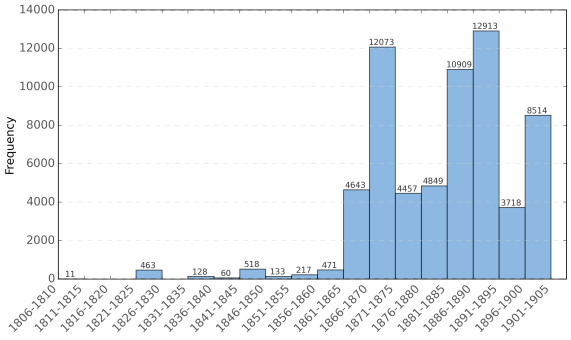

Figure C1: LatamXIX dataset decade distribution