A Dual-View Analysis of Multiple Languages in Colonial Newspapers

Zhan Su[1], Xiaoya Chen[2], Fengran Mo[3], Ida L. Vos[4],
Prayag Tiwari [5], Yazhou Zhang[6], Qian Zheng[2]*, Natália da Silva Perez [1,4]*,
[1] University of Copenhagen, Denmark [2] Zhengzhou University of Light Industry, China
[3] University of Montreal, Canada [4] Erasmus University Rotterdam, Netherlands
[5] Halmstad University, Sweden [6] Tianjin University, China
zhan.su@di.ku.dk, zhengqian@zzuli.edu.cn, dasilvaperez@eshcc.eur.nl

Abstract

Historical newspapers from the colonial period offer valuable evidence of how racializing language evolved over time. However, there are challenges in studying this type of historical data: 1) Data scarcity: acquiring large, annotated historical datasets is difficult, hindering the possibility of analyzing racialization comprehensively; 2) Digitized materials frequently contain Optical Character Recognition (OCR) errors and other types of noise that complicate text extraction and computational analysis; 3) Colonial newspapers are often multilingual and written in archaic prose, hindering the effectiveness of NLP tools developed for modern, single language texts. This paper addresses these challenges by conducting a dual-view, jointly studying multilingual event extraction and temporal semantic shift tasks. Specifically, we introduce a contextual question answering (CQA) and a visual question answering (VQA) derived from eighteenth- and nineteenth-century colonial newspapers. Content-wise, we focus on how enslaved people were described by enslavers as well as how they articulated their own condition through QA pairs of newspapers written in Dutch, English-French, and Spanish. Our results show that LLMs are still limited for low-resource VQA tasks. For temporal semantic change, we train temporal word embedding with a compass. The study concludes that racialization is a fluid process of linguistic recalibration where the decline of slavery merely shifted the language of control onto new categories of labor and identity.

1 Introduction

The historical study of racialization examines how racial identities are formed and maintained over time. This topic is central to research in history (Silva Perez, 2025), sociology, and sociolinguistics (Alim, 2016; Costa and Sokolovska, 2025; Grieve et al., 2025; Meng et al., 2026). Within this field, historical colonial newspapers are especially valuable. They served as public forums where ideas of racial hierarchy and resistance were openly expressed (da Silva Perez and Borenstein, 2025; Borenstein et al., 2023a; Newman, 2022; Mäkinen, 2022; Yu et al., 2025b). Historians traditionally work closely with the texts they study. Nevertheless, NLP tools can automate parts of the analysis and speed up the research process. This allows historians to extract evidence from large corpora more efficiently and focus more on interpretation (Borenstein et al., 2023a).

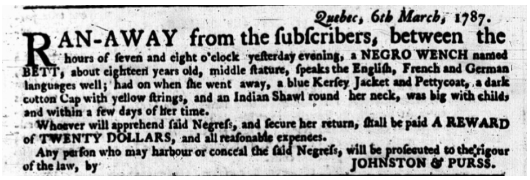

Quebec, 6th March, 1787.
RAN-AWAY from the ſubſcribers, between the hours of ſeven and eight o'clock yeſterday evening, a NEGRO WENCH named BETT, about eighteen years old, middle ſtature, ſpeaks the Engliſh, French and German languages well; had on when ſhe went away, a blue Kerſey Jacket and Pettycoat, a dark cotton Cap with yellow ſtrings, and an Indian Shawl round her neck, was big with child, and within a few days of her time.
Whoever will apprehend ſaid Negreſs, and ſecure her return, ſhall be paid A REWARD of TWENTY DOLLARS, and all reaſonable expences.
Any perſon who may harbour or conceal the ſaid Negreſs, will be proſecuted to the rigour of the law, by JOHNSTON & PURSS.

Figure 1: A screenshot from the historical colonial newspapers. The resulting texts often contain unusual spellings, layout noise, and many OCR errors.

However, leveraging these archives on NLP tasks remains a profound challenge for researchers (Hill and Hengchen, 2019; Rigaud et al., 2019; Yu et al., 2025a). There are three main challenges. First, building high-quality datasets still depends heavily on manual work, which is slow and costly. This makes it hard to draw general conclusions about racialized language (Borenstein et al., 2023a; Huang et al., 2026). Second, and more importantly, the digitization of these fragile, centuries-old documents creates serious data quality problems. The resulting texts often contain unusual spellings, layout noise, and many OCR errors (Fig. 1). Finally, colonial newspapers are often multilingual and written in archaic languages, hindering the effectiveness of standard NLP tools, which are usually trained on clean, modern text (Hill and Hengchen, 2019; Todorov and Colavizza, 2022; Mo et al., 2026b).

These challenges are especially acute for less-

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 20559–20573

studied tasks (Borenstein et al., 2023a,b). In this work, we argue that event extraction (EE) and semantic shift offer two complementary perspectives on racialization in colonial newspapers. **On the one hand**, EE captures the explicit, event-level manifestations of racialization, such as acts of enslavement, labor control, punishment, or resistance, as they are described in colonial texts (Sprugnoli and Tonelli, 2019; Lai et al., 2021; Dagdelen et al., 2024). Prior work shows that extracting events from historical advertisements can reveal how racialized groups were placed within colonial power structures. However, performance remains limited because of OCR errors and the use of archaic language (Borenstein et al., 2023b). **On the other hand**, the semantic shift task captures the implicit, long-term evolution of racialized meaning, revealing how terms associated with race, labor, and identity gradually changed as colonial systems transformed (Borenstein et al., 2023b). In summary, EE focuses on what happens in specific historical moments while semantic shift reveals how language itself adapts to sustain racial hierarchies over time.

By studying event extraction and semantic shift together, we link specific racialized events to broader changes in language. This dual view helps connect historical actions with evolving meanings that still influence modern racial bias. We address the shortcomings of previous work and make the following contributions: (1) To address the dataset scarcity and OCR noisy challenges, we use LLMs to study multilingual event extraction in Dutch, English-French, and Spanish by constructing contextual question answering (CQA) and visual question answering (VQA), centered on historical racialization processes in colonial newspapers from 18-19 centuries. (2) To study how racializing language evolved, we study the semantic shift in historical newspapers from the Dutch in the colonial period between 1816 and 1882, since the abolition of slavery in the Dutch colonies in 1863 serves as a useful case study for examining temporal shifts. We train the temporal word embeddings with a compass (TWEC) framework in each year from 1816 to 1882 to study the semantic change at the word level.

We find that current LLMs perform well on CQA tasks when provided with clean, modern text. However, when the context consists of images of original colonial newspapers, even recent state-of-the-art multimodal models perform poorly, highlighting persistent challenges in processing archaic language and historical print. Beyond model performance, our semantic shift analysis shows that racialization is a changing process in language. As formal slavery declined, words of control shifted to new kinds of labor and identity. The code is available on the link. [1]

2 Related Work

2.1 LLM for historical research

The use of LLMs in historical research has changed digital humanities. Research now goes beyond keyword search and topic modeling and moves toward methods that focus on meaning, generation, and active reasoning (Grossmann et al., 2023; Piper and Wu, 2025; Cherukuri et al., 2025; Zhang et al., 2024b; Su et al., 2024; Simons et al., 2025; Su et al., 2024; Hauser et al., 2024; Karch et al., 2025; Zhang and Colavizza, 2025; Levchenko, 2025; Humphries et al., 2024; Zhang et al., 2026; Mo et al., 2025, 2026a; Zhang et al., 2025). Levchenko (2025) demonstrates that multimodal LLMs achieved a character error rate (CER) as low as 1.8% on 18th and 19th-century English manuscripts, surpassing state-of-the-art tools. Beyond direct transcription, LLMs are increasingly used to post-correct noisy OCR outputs. Tudor et al. (2025) explored this in the context of "prompting the past," using LLMs to clean historical text transcripts. Zhang et al. (2024a) explore the potential of OpenAI's GPT models on the post-OCR correction task containing English texts published between 1559 and 1928 on versification and pronunciation. Unlike prior studies focusing on using LLM to OCR issues, we study the historical research by studying the multilingual event extraction and tracking the semantic evolution of racialization.

2.2 Event Extraction

Event Extraction (EE) is a task of organizing natural texts into structured events (Hogenboom et al., 2011; Xiang and Wang, 2019; Chen et al., 2024; Ma et al., 2024). Usually, EE is decomposed into smaller, less complex tasks (Lin et al., 2020). such as detecting the existence of an event (Weng and Lee, 2011; Nguyen and Grishman, 2018; Sims et al., 2019), identifying its participants (Du et al., 2021; Li et al., 2020), and extracting the attributes associated with the event (Li et al., 2020). Some research work has shown the benefit of framing the

[1] https://github.com/shuishen112/historical_nlp

EE as QA tasks (Li et al., 2020). The work most closely related to ours is (Borenstein et al., 2023a), which studies event extraction by building multilingual QA pairs from newspaper advertisements in the early modern colonial period. In contrast, their work focuses only on advertisements. We use LLMs to generate data from entire newspaper pages and construct both contextual QA and visual QA tasks.

2.3 Temporal Semantic Change

Historical corpora have been widely used to study social phenomena such as language change (Kutuzov et al., 2018; Borenstein et al., 2023b; Qi et al., 2026). For example, Garg et al. (2018) use word embeddings to measure changes in stereotypes and attitudes toward women and ethnic minorities in the United States during the twentieth and twenty-first centuries. Levis Sullam et al. (2022) analyze a large and diverse text collection, including books, newspapers, songs, and essays, to track how Jews are linked to different semantic topics and how these links change over time.

Borenstein et al. (2023b) also use word embeddings to study how bias in historical newspapers persists and changes. However, training separate word embeddings for each year makes it hard to compare meanings across time. In this paper, we use the TWEC framework to study word-level semantic change in a shared vector space.

3 Dataset Collection

We constructed our multilingual dataset from three digitized colonial newspapers available in open access:

- *De Curaçaosche Courant*, held by Delpher.nl, was published in Willemstad, Curaçao, and is mostly in Dutch with small portions in other languages: 3249 newspaper issues, approximately 13000 pages, spanning 1816-1882.
- *Quebec Gazette*, held by the Bibliothèque et Archives Nationales du Quebéc, was published in Quebéc City, and is bilingual in French-English: 51 newspaper issues, approximately 200 pages, spanning 1765-1807.
- *Revista Economica*, held by the Biblioteca Virtual de Prensa Histórica of the Ministry of Culture of Spain, was published in Havana and is in Spanish: 50 newspaper issues, approximately 400 pages, spanning 1878-1882.

We focused on newspapers that contained texts related to slavery, enslaved labour, or racialized populations (fugitive ads, sale notices, offers of enslaved workers for hire, etc), were published in different languages beyond English, in colonies governed by different empires, and were made available in open access by the respective repositories.

4 Event Extraction from Colonial Newspapers

As discussed in the introduction, we study event extraction from colonial newspapers by framing it as a question answering task and using LLMs to extract event arguments. We introduce both CQA and VQA benchmarks based on newspapers from the eighteenth and nineteenth centuries in Dutch, French, and Spanish.

4.1 Contextual Question Answering Tasks

We formalize the contextual question answering task as a tuple (C, q, a). Here, C denotes the context, representing a specific historical text segment such as a fugitive advertisement or a sales notice, which serves as the unstructured source of linguistic evidence. q represents a query designed to probe specific dimensions of racialization, ranging from physical descriptors to markers of resistance, effectively acting as a sociological prompt. The objective of the model is to map the input pair (C, q) to an answer a, which extracts the relevant semantic span or generates a synthesized response grounded in the text. The dataset preprocessing setting is presented in Appendix A.7.2. We evaluate the current strong open LLM (e.g., Qwen-72B (Team et al., 2024)) and closed-formed LLM (e.g., GPT-4o (Hurst et al., 2024)). The details are presented in Appendix A.7.3.

4.1.1 Data Quality and Human Verification

In our benchmark, we use Open LLM Gemini to extract the datasets from PDF files, which inevitably causes hallucination issues and fabricates details extrinsic to the source text. To guarantee the reliability of our benchmark and mitigate the risk of hallucination inherent in synthetic data generation, we use a rigorous *human-in-the-loop* quality assurance protocol. Specifically, we subjected the generated CQA triples to a manual verification process. Human annotators reviewed the samples with a specific criterion: the answer a must be strictly derivable solely from the evidence provided within

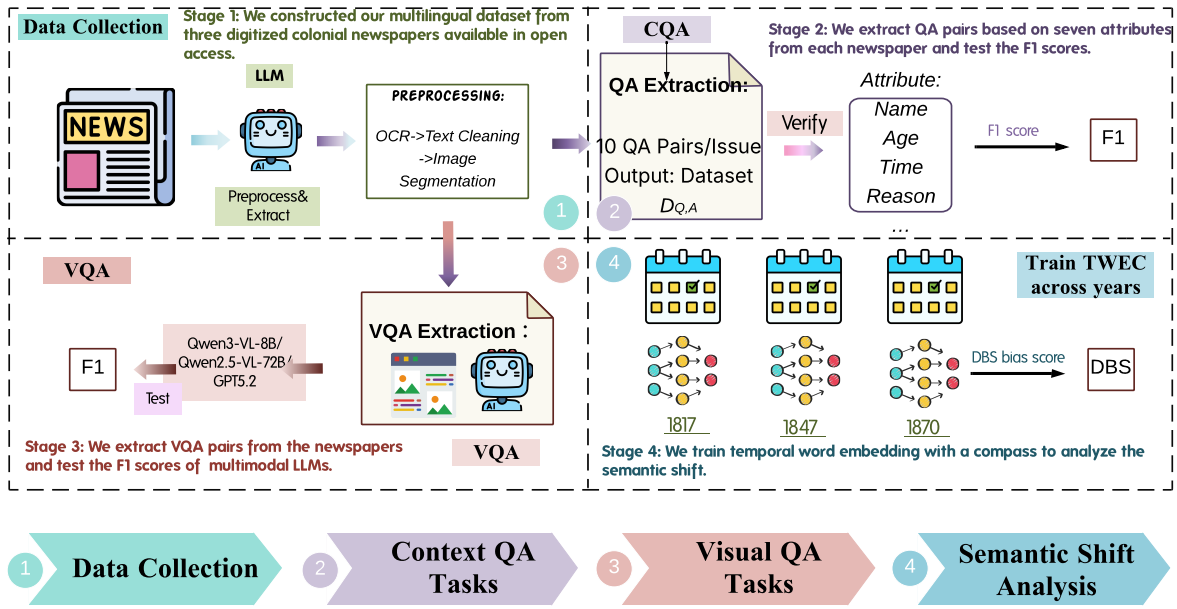


Figure 2: We conduct data collection using LLM (stage 1) and then study on multilingual event extraction (CQA tasks in stage 2 and VQA tasks in stage 3) and semantic shift tasks (stage 4).

the context (C), without recourse to external information. This validation step effectively filtered out ungrounded responses and hallucinated facts, ensuring that our benchmark evaluates a model's capability to reason over specific context from the newspaper.

4.1.2 Results

Table 1 reports the performance of different models on multilingual question answering tasks in Dutch, French, and Spanish. The scores are averaged across seven question types. Overall, larger models perform better than smaller ones in all languages. The 1.5B models show limited performance, especially on reasoning-related categories. Performance improves substantially when scaling to 72B parameters. Across all three languages, GPT-4o achieves the highest average score. It consistently outperforms open-weight models and Claude 3.5. This indicates strong multilingual and cross-task generalization. For Dutch, performance increases steadily with model size. GPT-4o obtains the best average score, followed by Claude 3.5 and Qwen2-72B. Smaller models lag behind, particularly on Reason and Position questions. For French, all models perform better than in Dutch. GPT-4o again achieves the highest average score. Large open models, such as Qwen2-72B, perform competitively but remain below proprietary models. For Spanish, overall performance is the highest among the three languages. GPT-4o achieves the best result, while Qwen2-72B and Claude 3.5 show similar strong performance. Small models again perform significantly worse.

In summary, the results show a clear effect of model scale and model structure. Large models perform better across languages and question types, with GPT-4o consistently achieving the strongest results.

4.2 Visual Question Answering Tasks

In addition to CQA tasks, we test whether LLMs can answer questions when the input is a newspaper image. In this setting, the model uses the image as context and produces an answer directly. To support this, we introduce a visual question answering (VQA) benchmark that captures the multimodal nature of colonial newspapers and reduces reliance on noisy text digitization. By asking multimodal LLMs to answer questions from raw document images, we test whether they can connect language with visual text and recover meaning without using error-prone OCR systems. The statistics of the dataset are presented in Table 9. We evaluate the current state-of-the-art multimodal LLMs, such as Qwen2.5-VL-72B and closed-form LLM GPT5.2.

4.2.1 Results

As shown in Table 3, we present the results on visual question answering tasks across Dutch, En-

Model	Name	Age	Time	Reason	Position	Reward	Other	AVG
Dutch								
Qwen2-1.5B	0.616	0.416	0.497	0.372	0.322	0.550	0.472	0.464
Llama 3-8B	0.601	0.392	0.487	0.353	0.311	0.502	0.452	0.443
Qwen2-72B	0.745	0.812	0.672	0.512	0.683	0.612	0.632	0.667
Claude 3.5	0.800	0.812	0.788	0.572	0.721	0.700	0.673	0.724
GPT-4o	**0.819**	**0.836**	**0.792**	**0.580**	**0.741**	**0.716**	**0.685**	**0.738**
English-French								
Qwen2-1.5B	0.723	0.752	0.654	0.512	0.519	0.831	0.590	0.654
Llama 3-8B	0.712	0.723	0.632	0.502	0.520	0.801	0.581	0.639
Qwen2-72B	0.821	0.912	0.872	0.621	0.517	0.877	0.623	0.749
Claude 3.5	0.824	0.934	**0.901**	0.601	0.506	**0.887**	0.643	0.756
GPT-4o	**0.831**	**0.935**	0.899	**0.642**	**0.527**	**0.887**	**0.653**	**0.768**
Spanish								
Qwen2-1.5B	0.653	0.624	0.570	0.373	0.478	0.677	0.596	0.567
Llama 3-8B	0.642	0.614	0.564	0.321	0.473	0.664	0.573	0.550
Qwen2-72B	0.890	0.921	0.900	0.710	0.701	0.831	0.712	0.809
Claude 3.5	0.900	0.927	**0.922**	0.670	0.682	0.832	0.702	0.805
GPT-4o	**0.901**	**0.931**	0.921	**0.715**	**0.707**	**0.857**	**0.728**	**0.823**

Table 1: Evaluation results on contextual question answering tasks are reported using F1 scores. The best and second-best results are highlighted in bold and underline, respectively. Dataset details are provided in Table 8.

glish–French, and Spanish newspapers. Scores are averaged over all available question types. Overall performance is low for all models, especially for Dutch and English–French. Compared with open-sourced language models, small multimodal models such as Qwen2.5-VL-7B show limited ability to answer questions from historical newspaper images. Models such as Qwen3-VL-8B perform better but still struggle in several categories. GPT-5.2 achieves the highest average score in all three language settings. It consistently outperforms open multimodal models across most question types. Large open models such as Qwen2.5-VL-72B and LLaVA-OV-72B perform competitively but remain behind GPT-5.2. Specifically, Spanish yields much higher scores than Dutch and English–French. This trend is consistent across all models and suggests that visual and textual cues are easier to extract in Spanish newspapers. In summary, VQA on historical newspapers is challenging. Larger models perform much better, with GPT-5.2 achieving the best results across all languages.

To better understand the underlying causes of these failures, we provide a detailed analysis of the LLM's performance on VQA tasks. We sample 28 images to analyze VQA failure cases. All samples are drawn from a one-page newspaper format and exhibit complex layouts. We manually annotate two potential degradation factors: font degradation and image corruption.

Among the 28 examples, 7 cases involve font degradation and 12 cases involve image destruction. The corresponding F1 scores are as follows: layout complexity (0.163), font degradation (0.133), and image destruction (0.192). The results suggest that font degradation may have a stronger negative impact than image destruction in this setting.

VQA failure	Num	F1-score
Layer-out Complexity	28	0.163
Font Degradation	7	0.133
Image destruction	12	0.192

Table 2: VQA failure modes for historical newspaper in Dutch. We select 28 samples in our VQA pairs and manually annotate the font degradation and image destruction.

Model	Name	Age	Time	Reason	Feature	Other	AVG
Dutch							
Qwen2.5-VL-7B	0.083	0.142	0.093	-	0.060	0.168	0.109
Qwen3-VL-8B	0.142	0.312	0.135	-	0.151	0.200	0.188
LLaVA-OV-72B	0.143	0.374	0.143	-	0.142	0.210	0.202
Qwen2.5-VL-72B	0.153	0.478	0.163	-	0.152	0.203	0.230
GPT-5.2	**0.240**	**0.676**	**0.273**	-	**0.429**	**0.283**	**0.380**
English-French							
Qwen2.5-VL-7B	0.102	0.013	0.145	0.092	0.132	0.204	0.115
Qwen3-VL-8B	0.212	0.043	0.232	0.113	0.232	0.298	0.188
LLaVA-OV-72B	0.223	0.032	0.210	0.133	0.252	0.332	0.197
Qwen2.5-VL-72B	0.234	0.043	0.243	0.132	0.263	0.324	0.207
GPT-5.2	**0.284**	**0.083**	**0.481**	**0.209**	**0.275**	**0.384**	**0.286**
Spanish							
Qwen2.5-VL-7B	0.342	-	0.463	0.362	0.132	0.323	0.324
Qwen3-VL-8B	0.645	-	0.693	0.501	0.241	0.392	0.494
LLaVA-OV-72B	0.682	-	0.732	0.492	0.242	0.403	0.510
Qwen2.5-VL-72B	0.672	-	0.743	**0.521**	0.282	0.452	0.534
GPT-5.2	**0.710**	-	**0.776**	0.500	**0.306**	**0.471**	**0.553**

Table 3: Evaluation on Visual Question Answer tasks, "-" indicates that there is no dataset extracted from the newspaper. The best and second-best results are highlighted in bold and underline, respectively.

5 Semantic Shift analysis

To capture the semantic shift in the colonial newspaper, we go beyond static event extraction and perform a diachronic analysis. The period from the eighteenth to the nineteenth century includes major social and political changes. For example, the abolition of slavery in the Dutch colonies in 1863 provides a key case for studying language change over time in low-resource colonial archives.

5.1 Measures

To mathematically rigorously quantify the semantic shift in racializing language, we introduce the *diachronic bias shift* (DBS) metric, which draws upon the Word Embedding Association Test (WEAT) framework (Caliskan et al., 2017). By projecting target identity terms (e.g., "freedman") onto an axis defined by opposing attribute sets, DBS explicitly tracks the transition of the racialized subject from an object of capital to an object of social anxiety.

Specifically, the DBS metric measures the temporal evolution of the semantic distance between a specific target group and opposing attribute concepts. We define two distinct attribute sets $\{\mathcal{A}, \mathcal{B}\}$ representing opposing semantic framings. For example, we set archaic vocabulary of coerced servitude ($\mathcal{A}_{\text{servitude}}$) from the emerging modern framework of free market contractuality ($\mathcal{B}_{\text{market}}$).

- $\mathcal{A}_{servitude} = \{slave, toil, forced, master\}$
- $\mathcal{B}_{market} = \{wage, contract, strike, pay\}$

Let $\mathbf{w}_t \in \mathbb{R}^d$ be the aligned vector representation of a target term w at time step t. The bias score $Bias_t(w)$ is defined as the difference between the mean cosine similarity of the target word to $\mathcal{A}$ and $\mathcal{B}$.

$$\text{Bias}_t(w) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \cos(\mathbf{w}_t, \mathbf{a}_t) - \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \cos(\mathbf{w}_t, \mathbf{b}_t) \quad (1)$$

where $\mathbf{a}_t$ and $\mathbf{b}_t$ denote the vector representations of the attribute words at time t. The sign of $\text{Bias}_t(w)$ reveals the dominant semantic framing in the corpus for a given era. To construct temporally aligned vector spaces for DBS analysis, we employ the Temporal Word Embeddings with a Compass (TWEC) framework (Di Carlo et al., 2019) and generate a temporally aligned vector space for each annual data slice from 1816 to 1882 (Sec. 5.2).

5.2 Temporal Word Embeddings with a Compass (TWEC)

TWEC operates in two stages: (1) training independent word embedding models for each temporal slice, and (2) aligning these embeddings into a shared coordinate system using stable "compass" words.

Let $\mathcal{D} = \{D_{t_1}, D_{t_2}, \ldots, D_{t_n}\}$ denote our diachronic corpus partitioned into n temporal slices, where D_{t_i} represents the document collection from time period t_i. For each slice D_{t_i}, we train an independent Word2Vec model M_{t_i} using the Skip-Gram architecture with negative sampling (Algorithm 1).

Algorithm 1 TWEC Training Framework

Require: Diachronic corpus $\mathcal{D} = \{D_{t_1}, \ldots, D_{t_n}\}$, embedding dimension d, window size w, minimum frequency f_{min}
Ensure: Aligned embeddings in shared coordinate system: $\{\mathbf{W}^{(t_1)}, \ldots, \mathbf{W}^{(t_n)}\}$

Stage 1: Train Independent Slice Models
for each time slice $t_i \in \{t_1, \ldots, t_n\}$ **do**
  $D_{t_i}^{\text{clean}} \leftarrow \text{Tokenize}(D_{t_i})$
  $\mathcal{V}_{t_i} \leftarrow \{w : \text{count}(w, D_{t_i}) \geq f_{min}\}$
  $\mathbf{W}_0^{(t_i)} \leftarrow \text{SkipGram}(D_{t_i}^{\text{clean}}, \mathcal{V}_{t_i})$
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{(t_i, M_{t_i}, \mathbf{W}_0^{(t_i)})\}$
end for
Stage 2: Procrustes Alignment
Select reference time: $t_{\text{ref}} \leftarrow t_1$
$\mathbf{W}^{(t_{\text{ref}})} \leftarrow \mathbf{W}_0^{(t_{\text{ref}})}$ {Reference unchanged}
for each slice $t_i \in \{t_2, \ldots, t_n\}$ **do**
  Extract compass vectors from reference:
    $\mathbf{C}_{\text{ref}} \leftarrow [\mathbf{w}_1^{(t_{\text{ref}})}, \ldots, \mathbf{w}_{|\mathcal{C}|}^{(t_{\text{ref}})}] \in \mathbb{R}^{|\mathcal{C}| \times d}$
  Extract compass vectors from current slice:
    $\mathbf{C}_{t_i} \leftarrow [\mathbf{w}_1^{(t_i)}, \ldots, \mathbf{w}_{|\mathcal{C}|}^{(t_i)}] \in \mathbb{R}^{|\mathcal{C}| \times d}$
  Compute alignment matrix via Procrustes (Algorithm 2):
    $\mathbf{Q}_{t_i} \leftarrow \text{Procrustes}(\mathbf{C}_{t_i}, \mathbf{C}_{\text{ref}})$
  Align all embeddings: $\mathbf{W}^{(t_i)} \leftarrow \mathbf{W}_0^{(t_i)} \cdot \mathbf{Q}_{t_i}$
end for
return $\{\mathbf{W}^{(t_1)}, \ldots, \mathbf{W}^{(t_n)}\}$

In the second stage, we align the embeddings in stage 1 into a shared coordinate system using stable "compass" words (refer to Algo 2).

5.3 Results

Figure 4 illustrates the temporal evolution of bias scores for target terms. The commodification set $\mathcal{A}_{prop}$ is $\{sale, price, buy, owner\}$ and the criminalization set $\mathcal{B}_{threat}$ is $\{riot, dangerous, vagrant, steal\}$ (refer to the Appendix A.4). We observe significant shifts in the semantic framing of several terms across the 19th century. (1) **Slave terms** (*slaaf*) exhibited strong commodification bias (Bias $> +0.3$) in 1817–1840, reflecting economic objectification. Post-1847, we observe a gradual shift toward neutral framing (Bias ≈ 0), corresponding with abolitionist discourse. (2) **Free persons** (*vrijman*) maintained consistently positive commodification bias throughout the period, though decreasing from $+0.42$ (1817) to $+0.28$ (1870), suggesting evolving conceptions of free labor. (3) **Indigenous terms** (*inboorling*): Demonstrated oscillation between framing categories, with peak criminalization bias (-0.31) in 1847 during labor shortage periods, reverting to near-neutral by 1870. (4) **Workers** (*arbeider*): Showed increasing commodification bias over time ($+0.15$ to $+0.37$), reflecting the rise of wage labor discourse post-emancipation.

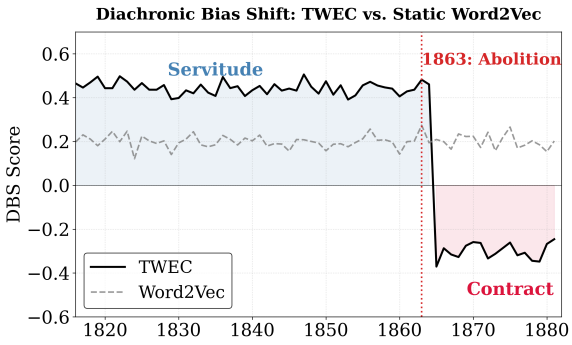
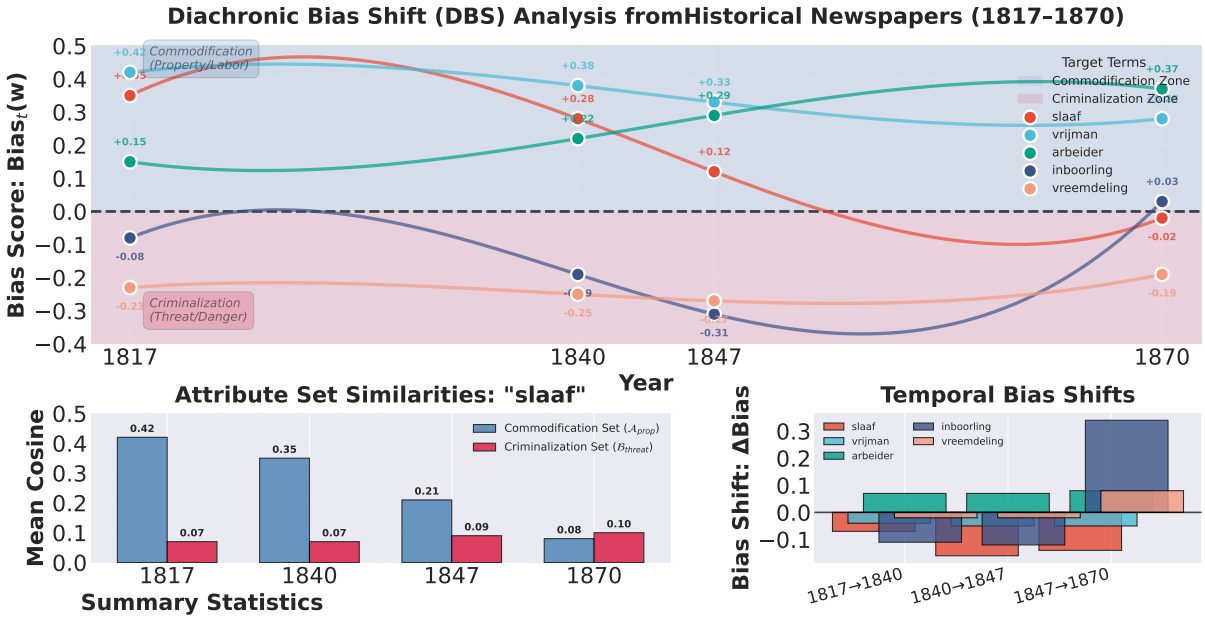


Figure 3: DBS score for "labor". The trajectory reveals a semantic phase transition. The vertical dashed line marks the 1863 Abolition of Slavery, where the discursive framing shifts from "Servitude" to "Contract".

The results reveal that racialization is a fluid process of linguistic recalibration, where the decline of formal enslavement did not eliminate bias but instead transferred commodification and criminalization frameworks onto "workers" to sustain colonial power structures.

5.4 Discussion

To ensure robust results, we conduct several analyses. First, we evaluate the alignment quality of TWEC (Sec.5.4.1). Next, we compare TWEC with a static word2vec model (Sec.5.4.2).



Target Term	Initial Bias (1817)	Final Bias (1870)	Change (Δ)	Dominant Framing
slaaf	+0.350	-0.020	-0.370	Commodified
vrijman	+0.420	+0.280	-0.140	Commodified
arbeider	+0.150	+0.370	+0.220	Commodified
inboorling	-0.080	+0.030	+0.110	Criminalized
vreemdeling	-0.230	-0.190	+0.040	Criminalized

Figure 4: Diachronic Bias Shift trajectories for target demographic terms in Dutch newspapers. Positive values indicate commodification bias (association with property/labor), while negative values indicate criminalization bias (association with threat/danger) (In this case, $\mathcal{A}_{prop} = \{sale, price, buy, owner\}$, $\mathcal{B}_{threat} = \{riot, dangerous, vagrant, steal\}$). Shaded regions denote semantic framing zones.

5.4.1 Alignment Quality

We test the alignment error during 1817-1840, 1817-1847, 1817-1870. Table 4 shows the Procrustes alignment quality for each temporal slice relative to the 1817 reference. The alignment error increases monotonically with the temporal gap. This trend is accompanied by a gradual reduction in the number of usable compass words, suggesting that diminishing anchor vocabulary may partially explain the degradation in alignment quality. Despite this, the overall error remains relatively low, demonstrating the robustness of the proposed alignment method.

Metric	1817→1840	1817→1847	1817→1870	Mean
Alignment error	0.087	0.112	0.134	0.111
Compass words used	487	473	456	472

Table 4: Procrustes Alignment Quality (Mean Euclidean Distance)

5.4.2 TWEC vs Word2vec

To test the quality of TWEC vs static Word2vec, we compare the TWEC and static word embedding training. As shown in Figure 3, the term "arbeid" (Labor in Dutch) exhibits a distinct structural break, transitioning sharply from the positive domain associated with Servitude ($\mathcal{A}_{\text{servitude}}$) to the negative domain associated with Contract ($\mathcal{B}_{\text{market}}$) immediately following the 1863 threshold. This "crossover" trajectory quantitatively confirms that the abolition of slavery precipitated a rapid discursive realignment, effectively decoupling labor from the lexicon of coerced bondage and integrating it into the framework of free market economics. In contrast, Word2Vec serves as a stability check. This demonstrates the advantages of TWEC in temporal semantic analysis.

5.4.3 Statistical Significance

We assess the statistical significance of observed bias shifts using bootstrap resampling ($N = 1000$ iterations). Table 5 presents p-values for temporal comparisons.

The term 'slaaf' exhibited highly significant semantic shifts across all periods ($p < 0.001$ for 1817→1840 and 1847→1870), providing strong evidence for diachronic change in the seman-

Target Term	1817→1840	1840→1847	1847→1870
slaaf	$< 0.001^{***}$	0.023^{*}	$< 0.001^{***}$
vrijman	0.089	0.156	0.041^{*}
arbeider	0.012^{*}	0.003^{**}	$< 0.001^{***}$
inboorling	0.034^{*}	$< 0.001^{***}$	0.067
vreemdeling	0.234	0.412	0.378

Table 5: Statistical significance of bias shifts (p-values indicate the probability of observing the measured bias shift by chance alone. Values below 0.05 represent statistically significant temporal semantic change.). $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

tic framing of enslaved persons. In contrast, 'vreemdeling' showed no significant shifts (all p > 0.05), suggesting stable criminalization framing throughout the corpus period.

6 Conclusion

In this paper, we study event extraction and semantic shift in colonial newspapers. We find that LLMs perform well on contextual question answering tasks. However, their accuracy drops sharply when the context is images of original newspapers. This shows that current multimodal models still struggle with archaic language, even though they perform well on modern text. Our semantic shift analysis also shows that racialization changed after the end of formal slavery. Bias moved from enslaved people to "workers" through the new language of control and criminalization.

7 Limitations

We study racializing language in several languages from the eighteenth and nineteenth centuries. One limitation of our work is that we focus only on Dutch, English, French, and Spanish newspapers. Many other colonial languages and regions are not included. In future work, we plan to use more historical newspapers and study a wider range of languages. In addition, our analysis depends on large language models. These models may reflect biases from their modern training data, which can influence the results. Finally, our semantic shift analysis is limited by data size. Some years contain fewer documents, which may reduce the reliability of the observed language changes. Future work could use larger and more balanced datasets to address this issue.

8 Ethical Considerations

Studying texts about the history of slavery raises important ethical concerns (Rickford, 2016). The enslaved people described in the newspapers used in this study lived centuries ago, so issues of personal privacy and data protection do not directly apply. However, the newspapers contain many examples of racist and demeaning language, and this language can be harmful and distressing, even when presented in a historical context. Preserving this linguistic characteristic was necessary for us to assess how accurately LLMs process and analyze these historical documents. Nonetheless, this requires appropriate interpretation and careful handling.

Acknowledgments

Research for this article was funded by a Sapere Aude grant from the Independent Research Fund Denmark under Grant ID 10.46540/2063-00035B.

References

H. Samy Alim. 2016. Introducing raciolinguistics: Racing language and languaging race in hyper- racial times. In *Raciolinguistics*, pages 1–30.

Nadav Borenstein, Natália da Silva Perez, and Isabelle Augenstein. 2023a. Multilingual event extraction from historical newspaper adverts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10304–10325.

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natália da Silva Perez, and Isabelle Augenstein. 2023b. Measuring intersectional biases in historical documents. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2711–2730.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Komala Subramanyam Cherukuri, Pranav Abishai Moses, Aisa Sakata, Jiangping Chen, and Haihua Chen. 2025. Large language models for oral history understanding with text classification and sentiment analysis. *arXiv preprint arXiv:2508.06729*.

James Costa and Zorana Sokolovska. 2025. Historical foundations: Some threads for integrating and interrogating historiography in critical sociolinguistics. *Critical sociolinguistics: dialogues, dissonances, developments*.

Natália da Silva Perez and Nadav Borenstein. 2025. Annotating "privacy" to train semi-supervised event extraction models for historical newspapers. *Current Research in Digital History*, 8.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6326–6334.

Xinya Du, Alexander M Rush, and Claire Cardie. 2021. Grit: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main Volume*, pages 634–644.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, and 1 others. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and 1 others. 2024. Large language models' expert-level global history knowledge benchmark (hist-llm). *Advances in Neural Information Processing Systems*, 37:32336–32369.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty ocr on historical text analysis: Eighteenth century collections online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2026. A survey on large language models with multilingualism: Recent advances and new frontiers. *Artificial Intelligence Review*.

Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2024. Unlocking the archives: large language models achieve state-of-the-art performance on the transcription of handwritten historical documents. *Available at SSRN*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tristan Karch, Jakhongir Saydaliev, Isabella Di Lenardo, and Frederic Kaplan. 2025. Llm agents for interactive exploration of historical cadastre data: framework and application to venice. *Computational Humanities Research*, 1:e11.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400.

Maria Levchenko. 2025. Evaluating llms for historical document ocr: A methodological framework for digital humanities. *arXiv preprint arXiv:2510.06743*.

Simon Levis Sullam, Giorgia Minello, Rocco Tripodi, and Massimo Warglien. 2022. Representation of jews and anti-jewish bias in 19th century french public discourse: Distant and close reading. *Frontiers in big Data*, 4:723043.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th annual*

meeting of the association for computational linguistics, pages 7999–8009.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2024. Star: boosting low-resource information extraction by structure-to-text data generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18751–18759.

Susanna Mäkinen. 2022. Stability and variation in the genre of runaway slave notices in american newspapers 1704–1865. *Ennen ja nyt: Historian tietosanomat*, 22(3):83–87.

Chuan Meng, Litu Ou, Sean MacAvaney, and Jeff Dalton. 2026. Revisiting text ranking in deep research. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, and 1 others. 2025. Uniconv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6936–6949.

Fengran Mo, Zhan Su, Yuchen Hui, Jinghan Zhang, Jia Ao Sun, Zheyuan Liu, Chao Zhang, Tetsuya Sakai, and Jian-Yun Nie. 2026a. Opendecoder: Open large language model decoding to incorporate document quality in rag. *arXiv preprint arXiv:2601.09028*.

Fengran Mo, Jinghan Zhang, Yuchen Hui, Jia Ao Sun, Zhichao Xu, Zhan Su, and Jian-Yun Nie. 2026b. Convmix: A mixed-criteria data augmentation framework for conversational dense retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 15555–15563.

Simon P Newman. 2022. *Freedom seekers: Escaping from slavery in restoration London*. University of London Press.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Andrew Piper and Sophie Wu. 2025. Evaluating large language models for narrative topic labeling. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 281–291.

Rui Qi, Fengran Mo, Yufeng Chen, Xue Zhang, Shuo Wang, Hongliang Li, Jinan Xu, Meng Jiang, Jian-Yun Nie, and Kaiyu Huang. 2026. Language-coupled reinforcement learning for multilingual retrieval-augmented generation. *Preprint*, arXiv:2601.14896.

John R Rickford. 2016. *Raciolinguistics: How language shapes our ideas about race*. Oxford University Press.

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. IEEE.

Natália da Silva Perez. 2025. Racialised language in colonial newspaper advertisements during the eighteenth and nineteenth centuries. In *The Routledge Handbook of Information History*, pages 95–109.

Arno Simons, Michael Zichert, and Adrian Wüthrich. 2025. Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *arXiv preprint arXiv:2506.12242*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3623–3634.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265.

Zhan Su, Fengran Mo, Prayag Tiwari, Benyou Wang, Jian-Yun Nie, and Jakob Grue Simonsen. 2024. Mixture of latent experts using tensor products. *arXiv preprint arXiv:2405.16671*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Konstantin Todorov and Giovanni Colavizza. 2022. An assessment of the impact of ocr noise on language models. *arXiv preprint arXiv:2202.00470*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Crina Tudor, Beáta Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering llms. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 401–408.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Yongan Yu, Xianda Du, Qingchen Hu, Jiahao Liang, Jingwei Ni, Dan Qiang, Kaiyu Huang, Grant McKenzie, Renee Sieber, and Fengran Mo. 2025a. Weatherarchive-bench: Benchmarking retrieval-augmented reasoning for historical weather archives. *arXiv preprint arXiv:2510.05336*.

Yongan Yu, Qingchen Hu, Xianda Du, Jiayin Wang, Fengran Mo, and Renée Sieber. 2025b. Wximpact-bench: A disruptive weather impact understanding benchmark for evaluating large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4016–4035.

James Zhang, Wouter Haverals, Mary Naydan, and Brian W Kernighan. 2024a. Post-ocr correction with openai's gpt models on challenging english prosody texts. In *Proceedings of the ACM Symposium on Document Engineering 2024*, pages 1–4.

Jinghan Zhang, Fengran Mo, Tharindu Cyril Weerasooriya, Ruimin Dai, Xiaoyan Han, Yanjie Fu, Dakuo Wang, and Kunpeng Liu. 2026. Starpo: Stability-augmented reinforcement policy optimization. *Preprint*, arXiv:2604.08905.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model. *arXiv preprint arXiv:2508.18090*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

A Appendix

A.1 Algorithm

We present the alignment algorithm used in our Algo 1.

Algorithm 2 Orthogonal Procrustes Alignment

Require: Source matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$, target matrix $\mathbf{B} \in \mathbb{R}^{m \times d}$

Ensure: Optimal orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$

1: Center matrices:

2: $\boldsymbol{\mu}_A \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mathbf{a}_i, \quad \boldsymbol{\mu}_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mathbf{b}_i$

3: $\mathbf{A}^c \leftarrow \mathbf{A} - \mathbf{1}_m \boldsymbol{\mu}_A^T, \quad \mathbf{B}^c \leftarrow \mathbf{B} - \mathbf{1}_m \boldsymbol{\mu}_B^T$

4: Compute cross-covariance: $\mathbf{M} \leftarrow (\mathbf{B}^c)^T \mathbf{A}^c$

5: Singular Value Decomposition: $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$

6: Optimal rotation: $\mathbf{Q} \leftarrow \mathbf{U}\mathbf{V}^T$

7: **return** $\mathbf{Q}$

A.2 TWEC Model Configuration

Table 6 presents the hyperparameters used for TWEC training. For Word2vec, we use the gensim [2] to train the word2vec seperately for each year. All the training is conducted on an A100 80G GPU.

Table 6: TWEC Hyperparameters

Parameter	Value
Embedding dimension (d)	100
Context window (w)	5
Minimum word frequency ($f_{\min}$)	3
Skip-gram negative samples	5
Training epochs (T)	15
Learning rate	0.025
Compass vocabulary size ($\lvert\mathcal{C}\rvert$)	500
CV stability threshold ($\tau_{\text{stability}}$)	0.5

A.3 Computational Complexity

The TWEC framework has time complexity $\mathcal{O}(n \cdot T \cdot |V| \cdot d)$ for Stage 1, where n is the number of time slices, T is the number of training epochs, $|V|$ is the vocabulary size, and d is the embedding dimension. Stage 2 has a complexity $\mathcal{O}((n-1)\cdot d^3)$ due to SVD computation. DBS computation for each target word requires $\mathcal{O}(n \cdot (|\mathcal{A}| + |\mathcal{B}|) \cdot d)$ operations.

A.4 DBS attribute sets

Table 7 presents the attribute sets we used in the computation of DBS scores. In this paper, we restrict the attribute set to a small subset of words. Exploring methods for constructing a larger and more comprehensive attribute set is left for future work.

Attribute	Wordlist
$\mathcal{A}_{servitude}$	$\{slave, toil, forced, master\}$
$\mathcal{B}_{market}$	$\{wage, contract, strike, pay\}$
$\mathcal{A}_{prop}$	$\{sale, price, buy, owner\}$
$\mathcal{B}_{threat}$	$\{riot, dangerous, vagrant, steal\}\}$
$\mathcal{A}_{nature}$	$\{water, flow, stream, bank\}$
$\mathcal{B}_{politics}$	$\{vote, law, war, rights\}$

Table 7: Attribute sets used for DBS evaluation.

[2] https://pypi.org/project/gensim/

A.5 Prompt of Data Generation

We present the prompt template we used to generate the CQA and VQA pairs from the newspaper.

Context-based Question Answer Pairs Prompt Template

User Prompt:
You are a historian specializing in 19th-century history. I will provide you with an original runaway slave advertisement (Context) in newspapers in different languages. Based on this text, please extract key information and convert it into ten context-based question answer (CQA) pairs from every newspaper.
Task requirements:

1. Identify key attributes mentioned in the advertisement (e.g., name, age, time, reason, position, reward, etc.).
2. For each attribute, generate a question in its original language.
3. First prioritize slavery-related content, including runaway enslaved persons, reward notices, sales, ownership, and physical descriptions. Only if insufficient, use other colonial advertisements or notices.
4. Extract the corresponding text span from the original advertisement as the answer.
5. Answers must be short and factual. Do not infer, summarize, or use external knowledge.
6. The output format must be a Python tuple list in the form: `(Context, Question, Answer)`.

Visual-based Question Answer Pairs Prompt Template

User Prompt:
You are a historian specializing in 19th-century history. I will provide you with an original runaway slave advertisement (Context) in newspapers in different languages. Based on this text, please analyze the given newspaper page image and extract some high-quality Visual–Question–Answer (VQA) pairs.
Task requirements:

1. Identify key attributes mentioned in the advertisement (e.g., name, age, time, reason, position, reward, etc.).
2. For each attribute, generate a question in its original language.
3. First prioritize slavery-related content, including runaway enslaved persons, reward notices, sales, ownership, and physical descriptions. Only if insufficient, use other colonial advertisements or notices.
4. Extract the corresponding text span from the original advertisement as the answer.
5. Answers must be short and factual. Do not infer, summarize, or use external knowledge.
6. Page_Num must be an integer (starting from 1).
7. The output format must be a Python tuple list in the form: `(Page_Num, Context, Question, Answer)`.

A.6 Prompt of Data Evaluation

Data Evaluation Prompt Template

User Prompt:
You are a historian specializing in 19th-century history. Answer the question using only information explicitly stated in the provided context.
Task requirements:

1. The answer must be in the same language as the question.
2. Do not translate between languages.
3. If the answer appears verbatim in the context, copy it exactly.
4. If multiple possible spans exist, choose the shortest correct span.
5. Do not explain, paraphrase, or add extra words.
6. The answer should be as short and factual as possible.

A.7 Dataset setting

A.7.1 Data Preprocessing for VQA

Building upon the filtered corpus, we construct a visual question answering (VQA) dataset to incorporate visual and layout information. For this purpose, newspaper pages are randomly sampled, and VQA pairs are generated by aligning questions and answers with the corresponding full-page newspaper images. Unlike the text-only QA setting, the contextual input in the VQA dataset consists of the entire newspaper page containing the original textual evidence, rather than a localized text excerpt. This design aims to better reflect realistic historical newspaper reading scenarios. Table 9 presents the dataset statistics for VQA pairs in each language.

A.7.2 Dataset Processing for CQA

Our dataset is derived from digitized *De Curaçaosche Courant* newspapers published between 1816 and 1882. *Quebec Gazette* newspapers published between 1765 and 1807. *Revista Economica* published between 1878 and 1882. We employ Gemini 2.5 to automatically extract question–answer (QA) pairs from the newspaper text. In the initial sampling stage, we adopt a year-wise strategy: for each year in the target period, one newspaper is randomly selected, from which ten QA pairs are extracted, and the resulting QA pairs predominantly concern slavery and closely related topics.

We observe a substantial decline in slavery-related content in newspapers published after 1863, the year in which slavery was officially abolished in the Netherlands. As a consequence, newspapers from 1864 to 1882 yield very few relevant QA instances. To maintain topical coherence and data consistency, we therefore restrict our dataset to the period 1816–1863, excluding later years from subsequent analysis.

Attribute	Dutch	French	Spanish
Name	605	208	163
Age	158	31	14
Time	163	70	58
Reason	21	4	15
Position	94	76	64
Reward	36	8	3
Other	250	104	169

Table 8: Data statistics for contextual question answer tasks.

Attribute	Dutch	French	Spanish
Name	86	38	32
Age	19	4	-
Time	24	43	20
Reason	-	5	11
Feature	9	4	7
Other	63	38	96

Table 9: Data statistics for visual question answer tasks.

A.7.3 Comparison Models

To evaluate the proposed benchmark across varying scales of compute and accessibility, we selected a diverse set of baseline models categorized into three distinct tiers. First, to test the efficacy of lightweight, locally deployable models, we include Qwen2-1.5B (Team et al., 2024), llama 3-8B (Touvron et al., 2023). Second, to assess the capabilities of high-performance open-weight architectures, we evaluate Qwen2-72B. Finally, we establish a performance upper bound using state-of-the-art proprietary models, including Claude 3.5 (Handa et al., 2025), GPT-4o (Hurst et al., 2024), thereby enabling a critical comparison between closed-source LLMs and accessible alternatives in the context of

historical text analysis.

For VQA tasks, we include these comparison models to provide a fair and representative benchmark across model scale, architecture, and training paradigms in multilingual VQA. Specifically, Qwen2.5-VL-7B and Qwen3-VL-8B represent strong small-to-mid-scale open-source vision–language models, allowing us to assess performance under realistic resource constraints. Qwen2.5-VL-72B and LLaVA-OV-72B serve as large-scale open-source baselines, enabling analysis of how scaling impacts multilingual and reasoning-heavy visual tasks. GPT-5.2 is included as a state-of-the-art proprietary model, providing an upper-bound reference for current multimodal capabilities. Together, this selection spans diverse model sizes, training recipes, and openness levels, ensuring that observed performance differences reflect fundamental model capabilities rather than special design choices, and allowing us to contextualize gains across languages and question types.

A.8 Case Study

To provide qualitative evidence complementary to the quantitative evaluation, we present one correctly answered example for both context-based question answering (CQA) and visual question answering (VQA), selected directly from our dataset.

Contextual QA example

Context:
"de Negerin Markita slavin van den Heer John Harrison"
Question: Wie is de eigenaar van de slavin Markita?
Answer: De Heer John Harrison

In this text-only setting, the model accurately retrieves an explicitly stated factual detail from a short and unambiguous context. The question requires direct fact lookup without multi-sentence reasoning, resulting in an exact match between the predicted answer and the ground truth.

Visual QA example

Context:
[Image:] `1851-08-30-Decura..._p4.jpg`
Question: Wat is de leeftijd van de slaaf François?
Answer: 26 jaren

For the VQA case, the model successfully grounds its answer in the visual content of the document.