Named Entity Recognition in Dutch Colonial Archives: A Replication and Extension of Baseline Experiments

Paul-Antoine Fruchtenreich
ENSAE 2024/2025
paul-antoine.fruchtenreich@ensae.fr

Abstract

Colonial archives, by their very nature, are shaped by the power structures of the eras in which they were created. As a result, they often exclude or obscure the presence of marginalized individuals, including enslaved people, through systemic omissions in indexing and cataloguing. One particularly revealing case is that of the Dutch East India Company (VOC) testaments, where non-European voices are frequently minimized or omitted entirely. In this study, we build upon prior work in historical Named Entity Recognition (NER) by applying and refining modern NLP tools — specifically the Dutch language model BERTje — to this challenging domain. Through our replication and extension of existing experimental setups, we assess how well such models can identify both named and unnamed entities in VOC archival texts. Our findings shed light on the technical limitations of current methods and suggest pathways for more equitable archival practices through machine learning.

1 Introduction

Archives are often perceived as neutral repositories of historical memory, yet this perception has been increasingly challenged by scholars in critical archival theory. Rather than being impartial, archives reflect the priorities, hierarchies, and exclusions of the societies that produced them. These silences—gaps in the historical record—are particularly pronounced in colonial contexts, where indigenous peoples, enslaved individuals, and women were systematically marginalized in both documentation and classification. The mechanisms of these exclusions are not accidental; they are embedded in archival workflows, indexing conventions, and the epistemologies of the colonial bureaucracies that created them.

A striking example of this dynamic can be found in the records of the Dutch East India Company (VOC), which operated as a powerful colonial and commercial entity from the 17th to 18th centuries. Among its administrative records are thousands of notarial testaments produced by VOC officials and settlers living in Asia. While these documents occasionally mention enslaved individuals, non-European women, or local actors, such references are often indirect and fragmentary. In many cases, they are entirely absent from indices and finding aids created during the 19th-century cataloguing efforts—efforts that prioritized the names and roles of European men. The legacy of these archival practices continues to shape research access today, effectively silencing the presence of marginalized groups.

Recent advances in Natural Language Processing (NLP), particularly in the area of Named Entity Recognition (NER), offer new tools for redressing these archival imbalances. By applying machine learning techniques to digitized colonial records, it is now possible to identify and recover entity references that were overlooked or suppressed in earlier indexing systems. This study seeks to explore

Preprint. Under review.

the capabilities and limitations of such approaches by focusing on the annotated VOC testament corpus. We replicate and extend the work of Luthra et al. [2022], evaluating the performance of the Dutch BERT-based model BERTje on this task. Our objective is not only to test the technical feasibility of fine-tuning language models on historical texts, but also to contribute to broader conversations about decolonial methods in digital humanities and archival science.

2 Related Work

NER has long been a cornerstone of NLP research, with benchmarks like CoNLL-2003 and OntoNotes providing standardized datasets for training and evaluation. However, these corpora typically consist of contemporary, well-structured text such as newswire articles or legal documents. When such models are applied to historical sources, their performance drops significantly due to the differences in language usage, spelling, syntax, and document quality. Historical texts often contain inconsistencies introduced by archaic grammar, non-standard spelling, and degradation through time—factors compounded by errors introduced during OCR or handwritten text recognition (HTR).

Over the past decade, historical NER has emerged as a specialized research area concerned with adapting modern NLP tools to the idiosyncrasies of archival material. Surveys such as those by Ehrmann et al. [2021] and Piotrowski [2012] highlight the importance of domain-specific fine-tuning, custom annotation typologies, and character-level modeling. However, the scarcity of annotated corpora for historical languages—and especially for under-resourced contexts such as colonial archives—continues to be a bottleneck for research progress.

Within Dutch-language historical NLP, the work of Hendriks et al. [2020] represented an early attempt to apply off-the-shelf models like spaCy and BERTje to VOC testaments. While promising, these efforts were hampered by low recall rates and insufficient sensitivity to unnamed or ambiguously referenced entities. Building on this, Luthra et al. [2022] introduced a more comprehensive annotated dataset, incorporating not only named entities but also unnamed roles and attributes such as legal status and gender. Their work represents a significant methodological advance, providing both a richer typology and a more nuanced understanding of what constitutes an "entity" in a colonial archival context. Our project extends this research by applying similar methods while experimenting with model fine-tuning, tokenizer alignment, and evaluation strategies aimed at better capturing marginalized references.

3 Dataset

The dataset used in this study is drawn from annotated VOC testament texts digitized as part of the "IJsberg Zichtbaar Maken" project led by the Dutch National Archives. The documents, originally handwritten in Dutch between the 17th and 18th centuries, were transcribed using the Transkribus HTR system and then manually annotated by a team of historians and digital humanists using the BRAT tool. The annotation typology is deliberately inclusive, capturing not only named entities (e.g., "Jan de Vries") but also unnamed references (e.g., "een slaaf"), and annotating them with attributes such as gender, legal status, and role in the testament (e.g., beneficiary, testator).

In total, the corpus contains 2199 unique annotated pages, supplemented by 307 duplicate pages used to calculate inter-annotator agreement. These annotations amount to over 68,000 labeled spans, covering the four core entity types—Person, Place, Organization, and Proper Name—and their associated attributes. The dataset is split into training (70%), validation (10%), and test (10%) sets. Each annotation is encoded in IOB format, making it suitable for token classification tasks using modern NLP frameworks.

4 Methodology

Our experimental design closely follows the baseline setup provided in the original repository by Luthra et al. [2022], while implementing improvements in tokenizer alignment and training stability. We fine-tune the BERTje model, a pre-trained Dutch transformer from HuggingFace, on the VOC dataset. Tokenization is handled using the associated 'bert-base-dutch-cased' tokenizer, with special care taken to preserve prefix spaces and maintain alignment between tokens and IOB labels. Because

subword tokenization can fragment entity mentions, we use the "first subtoken" strategy to ensure consistent label propagation across wordpieces.

The preprocessing pipeline includes data parsing into HuggingFace 'DatasetDict' format, statistical analysis of label distributions, and dynamic padding via a custom data collator. For training, we use the HuggingFace 'Trainer' API with a batch size of 32, learning rate of 5×10^{-5}, linear warmup, and early stopping. The BERTje model is fine-tuned for 5 epochs. Additionally, we experiment with a lightweight transformer trained for 30 epochs as a comparative baseline.

Evaluation is performed using both strict and fuzzy metrics. Strict evaluation requires exact boundary matches between predicted and gold spans, while fuzzy evaluation allows for partial overlap. We report precision, recall, and F1-scores for each core entity type as well as for attribute labels such as gender and legal status. All metrics are micro-averaged over the test set and broken down by entity type. This approach allows us to diagnose which categories the model struggles with and whether performance degrades for unnamed or marginalized references.

5 Results

🔍 seqeval classification report:

	precision	recall	f1-score	support
Organization	0.3755	0.3909	0.3831	243
Organization,B-Place	0.4062	0.4194	0.4127	31
Organization,I-Place	0.4545	0.2500	0.3226	20
Person	0.7629	0.8329	0.7964	2017
Person,B-Place	0.3478	0.6154	0.4444	13
Person,I-Place	0.0000	0.0000	0.0000	10
Place	0.5291	0.6279	0.5743	594
micro avg	0.6739	0.7425	0.7065	2928
macro avg	0.4109	0.4481	0.4191	2928
weighted avg	0.6730	0.7425	0.7055	2928

F1-score (micro): 0.7065

Figure 1: Classification report for BERTje

The evaluation of the BERTje model on the annotated VOC corpus reveals distinct performance patterns depending on the type of entity and attribute involved, which can be noticed on Figure 1. Under strict evaluation criteria, the model achieves an F1 score above 0.70 for standard named entities such as *Person* and *Place*, indicating that BERTje, when fine-tuned, can successfully generalize to structured entities even in noisy, historical text. The high precision observed across entity types reflects the model's conservative prediction strategy, which tends to favor making fewer, more confident predictions—likely a consequence of the abundance of 'O' labels (non-entities) in the training data.

An analysis of the confusion matrix confirms that the model most frequently mislabels true entity tokens as non-entities, indicating a recall deficiency. This is particularly problematic in the context of uncovering marginalized figures, where under-detection is itself a continuation of historical silencing. Performance on unnamed persons was significantly worse than for named ones, illustrating the limitations of token-level classifiers in recognizing context-dependent entities.

We also observed that evaluation scores for 'Role' and 'Gender' attributes tended to correlate strongly with the lexical regularity of the annotated trigger words. For example, testaments often include fixed expressions such as "zijn wettige vrouw" (his lawful wife), which improve detection rates for female gendered persons when such phrases are encountered.

6 Discussion

The results of our experiments illustrate a dual reality. On the one hand, they show that pre-trained language models like BERTje—when properly fine-tuned—are capable of learning to identify many

named entities in colonial texts, even in the presence of HTR noise, archaic spelling, and inconsistent grammar. On the other hand, the persistent difficulty in detecting unnamed, marginalized, or ambiguously referenced individuals reflects deep methodological limitations that go beyond model architecture.

One key issue is the reliance on token classification models that operate locally, without global document context. Many references to enslaved individuals, for instance, are embedded in expressions that span multiple sentences or depend on co-text. Current models trained with token-level supervision often lack the discourse-level understanding required to resolve these references. This limitation calls for experimentation with document-level or span-based models that can incorporate broader semantic dependencies.

Another concern is the annotation sparsity of certain categories. While over 30,000 entity mentions exist in the dataset, labels for some combinations of attributes (e.g., enslaved women beneficiaries) are extremely rare. This introduces class imbalance issues, which conventional loss functions do not handle well. Techniques such as focal loss, class reweighting, or data augmentation might help, but each introduces its own trade-offs and complexities.

7 Conclusion

This study presents a detailed replication and extension of an NER pipeline for colonial Dutch archival texts, applying the BERTje transformer to a high-quality annotated dataset of VOC testaments. Our experiments confirm that modern NLP models can be trained to detect named entities with reasonable success, even under noisy historical conditions. However, the most historically significant cases — those involving marginalized, unnamed, or enslaved individuals — remain the most challenging to identify automatically.

References

Maud Ehrmann, Matteo Romanello, Simon Clematide, and Phillip Benjamin Ströbel. Named entity recognition for historical texts: A survey. *Frontiers in Digital Humanities*, 8:660417, 2021.

Barry Hendriks, Paul Groth, and Marieke van Erp. Recognising and linking entities in old dutch text: A case study on voc notary records. *Proceedings of the International Conference Collect and Connect: Archives and Collections in a Digital Age*, 2020.

Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. Unsilencing colonial archives via automated entity recognition. *arXiv preprint arXiv:2210.02194*, 2022.

Michael Piotrowski. Natural language processing for historical texts. In *Synthesis Lectures on Human Language Technologies*, volume 5, pages 1–157. Morgan & Claypool Publishers, 2012.