Towards Comparable Historical NER: Building a Shared Evaluation Corpus for 18th-Century Historical Texts

Lu Liu[1], Andreas Vlachidis[1], Adam Crymble[1], Deborah Lee[1], and Marco Humbel[1,2]

[1] Department of Information Studies, University College London (UCL), London, United Kingdom
[2] The National Archives, London, United Kingdom

Abstract

Named Entity Recognition (NER) is increasingly applied to historical text analysis. However, differences in evaluation materials, metrics, and annotation guidelines across existing NER projects make it difficult to systematically compare different approaches to historical NER. This study addresses this issue by constructing an evaluation corpus through the normalization of four annotated datasets from the long 18th century. We evaluate the performance of the Edinburgh Geoparser, spaCy and BERT-based tool on this corpus using five evaluation modes. Results show that even under the most lenient criteria, the highest F1-score remains below 70%, highlighting the challenges of applying existing NER systems to historical texts. Through detailed error analysis, we identify common challenges such as spelling and formatting issues. These findings demonstrate the limitations of NER tools in historical documents. We argue that future work should involve collaboration with historians to ensure that evaluation corpus align with real user needs.

Keywords: named entity recognition, evaluation corpus, historical documents, digital humanities, natural language processing

1 Introduction

With the development of mass digitization in cultural heritage, an increasing number of historical documents have become available in digital formats. In this context, Named Entity Recognition (NER) can be considered a crucial step for extracting structured and meaningful information from digitized historical texts. NER is a subtask of Natural Language Processing (NLP) that involves identifying and classifying entities of common interest from texts, such as persons, places, organizations, etc. It has been widely used in contemporary projects and serves as the basis for many text-mining applications, such as semantic annotation, question answering, ontology population and opinion mining [14]. In recent years, applications in the historical domain have become one of the main developments of NER tasks. Some studies shown that person and place names are the most frequent search elements in various digital libraries [3; 5]. Extracting these key entities by using NER not only helps historians retrieve relevant texts but also benefits downstream tasks such as entity linking and relation extraction. In other words, high-quality NER can greatly support the exploration and study of historical materials.

NER is often regarded as a solved problem in some major contemporary NER evaluation forums with reported success ratios over 95% [6], these results are largely based on contemporary texts, such as newspaper articles, journal articles, blog posts, and other sources [14]. NER applied to historical texts can face more challenges than applied to contemporary well-edited journalism materials. For instance, historical documents are often heterogeneous and may contain OCR errors

Lu Liu, Andreas Vlachidis, Adam Crymble, Deborah Lee, and Marco Humbel. "Towards Comparable Historical NER: Building a Shared Evaluation Corpus for 18th-Century Historical Texts." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 968–982. https://doi.org/10.63744/dwCJ80qwvAtr.

and spelling variations that have evolved over time, all of which can influence the performance of NER [8]. Moreover, historians' requirements and expectations for named entities and NER may differ from those defined in mainstream NER tasks, and this is rarely discussed in existing research. Given these challenges and continuous advancements in NER technologies, it is crucial to investigate the performance of existing NER systems on historical texts and explore ways to adapt them to achieve satisfactory results.

However, it is difficult to systematically compare the shortcomings and advantages of different NER approaches applied to historical materials due to the varying settings adopted in most historical NER projects [12]. More specifically, existing historical NER projects often use different definitions and annotation boundaries for named entities. The measurements they use to report the performance of NER systems are also varying. These variations make it difficult to give comparisons of NER methods across different projects. As a result, it remains difficult to make absolute claims about the performance of NER on historical documents. The incomparability of historical NER approaches makes it difficult to build upon existing work to advance the technologies. To address this issue, we propose the development of a shared evaluation historical corpus to enable fair and systematic comparison of NER systems and drive progress in historical NER, which motivates the study.

This study uses datasets from the long 18th century[1] (1688-1815) as a case study to create an evaluation corpus and establish evaluation standards for assessing the capabilities of NER applied to materials from this period. In the process, we introduce normalised annotations to evaluate the NER performance of three tools representing different approaches: the Edinburgh Geoparser,[2] spaCy,[3] and a BERT-based NER model (`bert-base-cased`),[4] covering methods from rule-based systems to deep learning models. The creation of the corpus makes it possible to employ a set of evaluation measurements to compare the strengths and weaknesses of different NER tools on historical text from this period. Meanwhile, the experience gained from applying NER to these materials could also be transferred to the work on other historical periods.

2 Related Work

2.1 Evaluation Development

Annotated datasets are commonly used to evaluate the performance of systems and models. With the development of NER, various evaluation tasks and datasets in the NLP field have gradually emerged to assess NER performance.

In 1995, the term 'Named Entity' was first introduced at the 6th Message Understanding Conference (MUC-6) [10]. The task mainly focused on identifying the names of people, organizations, and geographic locations in texts drawn from the 1993 and 1994 *Wall Street Journal*. In MUC events, system outputs were compared against manually created gold standard annotations, and each extracted object was classified into categories such as **Correct** (exact match), **Partial** (partial match), **Incorrect** (does not match), **Missed** (items present in the gold standard but not identified), and **Spurious** (items incorrectly predicted by the system) [4]. The evaluation framework then calculated performance scores using Precision, Recall and F-measure accordingly. These scores have been widely adapted ever since as the standard way for measuring the NER performance.

Following MUC, system development competitions for languages other than English emerged, such as IREX (Information Retrieval and Extraction Exercise) and CoNLL (Conference on Com-

[1] Typically from around 1688 (the Glorious Revolution) to 1815 (the end of the Napoleonic Wars).

[2] https://www.ltg.ed.ac.uk/software/geoparser/

[3] https://spacy.io/

[4] Implemented using the `bert-base-NER` model (https://huggingface.co/dslim/bert-base-NER) from Hugging Face. This model is a pre-trained BERT-base model fine-tuned on the CoNLL-2003 English dataset, and we use it with default inference settings without additional fine-tuning.

putational Natural Language Learning), which included not only English but also Japanese and German. The CoNLL-2003 shared task focused on four types of entities: Persons, Locations, Organizations, and Miscellaneous. It included English texts from Reuters Corpus and German texts extracted from the newspaper *Frankfurter Rundshau* [19]. IREX targeted the identification of eight entity types (Persons, Locations, Organizations, Aircrafts, Date, Time, Money and Percent) in a Japanese newspaper corpus [13]. In contrast to MUC, they did not include partial matches but only entities that exactly matched the corresponding entities in the data file were counted as correct and contributed to the scores [15]. The CoNLL-2003 dataset remains widely used today, with many NER tools employing it for training or testing models to demonstrate their effectiveness.

The exact match criterion used in CoNLL is not always applicable to certain domains. For example, in biomedical tasks, researchers often value useful information provided by partial matches. To address this, SemEval-2013 (Semantic Evaluation) introduced an extended evaluation regime [20], which included the following four criteria: (1) **Strict**, which requires exact matches of both entity boundaries and types; (2) **Exact**, which allows type misclassification; (3) **Partial**, allowing boundary overlaps and incorrect entity types; and (4) **Type**, which requires correct classification of entity types, regardless of boundaries.

Unlike traditional NER evaluations that rely on straightforward Precision and Recall metrics, the ACE (Automatic Content Extraction) program employed a more complex scoring system. Rather than merely measuring whether entities were correctly identified, it recognizes all mentions of entities and assigns them weighted values [7]. The ACE evaluation calculates scores using a formula in which an entity's final score is the product of the entity's inherent value and the sum of the weighted values of all its mentions [1].

In recent years, NER has begun to be applied to the field of history, and historical NER evaluation conferences have also emerged. The HIPE (Identifying Historical People, Places and other Entities) shared task provided an opportunity to evaluate the NER on historical newspapers in French, German, and English [9]. HIPE used both **Strict mode** where correct matches must get both entity type and boundary exactly matched and **Fuzzy mode** where partially matched boundary is allowed.

2.2 System Comparability

With the rise of historical NER, it is important to understand the capabilities of NER systems applied to historical materials. However, most research on historical NER has been conducted only in the context of individual projects, which makes it difficult to compare results. Challenges in system comparability arise from differences in materials, measurements, and annotation guidelines.

Historical texts tend to be highly heterogeneous, including various document types and languages. This heterogeneity leads to great variations between collections, while domain shifts inevitably affect NER performance, involving factors such as transcription quality, multilingualism, and formatting. With these variables, NER methods applied to different materials are difficult to compare. For example, Rodriquez et al. compared the performance of four NER systems on Holocaust survivor testimonies and letters of soldiers [18]. Stanford CRF performed best on the testimonies with a 54% F-score, whereas OpenCalais was the most accurate for the letters with a 36% score. This demonstrates not only performance loss across document types, but also the incomparability of NER applied to different materials.

Although most NER projects adopt standard metrics such as Precision, Recall and F-score, the modes of which entities contribute to the final scores vary. For instance, in CoNLL, only exact matches are considered correct, whereas MUC also includes partial matches in its scoring. As a result, it is difficult to compare NER systems across projects when the evaluation modes differ, and unified measurements are required to enable meaningful comparisons.

In different historical corpora, there is no standardised guideline to aid annotators in making

consistent decisions about challenging interpretations that affect the scope and boundaries of annotations. For example, should honorifics like 'King' and 'Queen' be considered person names? How should entities such as 'London Bridge' be annotated? Should 'Queen's garden' be tagged as a person or a place? The different choices make NER faces challenges of different degrees. Determining entity span is another difficulty. For instance, should honorifics following person names such as 'Sir Hans Sloane' be included in person annotation? Identifying the correct span for place names can be even more complex, especially for compound place names, such as 'cliffs of a high mountain of slego in Ireland'. How to identify references of individual objects and the boundaries between objects and their descriptions consistently become a challenge for researchers.

Ortolja-Baird et al. discussed this issue for encoding Sloane's data [17]. They consulted several domain experts, and one view suggested that the head noun of the phrase is the best candidate for identifying the object (e.g., 'corall' in 'Red corall growing on a rock with shells'), while another argued that descriptive details like color and size are meaningful and should be included (e.g., 'Red corall'). Ortolja-Baird noted these interpretive differences depend on data use. Although debates on defining entity boundaries remain, there is a consensus that emphasizes considering context and historical sensitivity. Overall, annotation guidelines vary with background and requirements. Consequently, NER systems assessed in these projects are challenged distinct criteria, which makes it difficult to achieve accurate and systematic comparisons of NER approaches.

3 Method

To address the comparability challenges of historical NER, this paper develops a shared evaluation corpus based on four annotated datasets, including Mary Hamilton Papers[5], Old Bailey Online[6], Sloane's Catalogue[7] and HIPE2020.[8] These historically significant and heterogeneous datasets help reveal the problems faced by NER systems when applied to long 18th-century documents. Moreover, they are already annotated with person and place names using TEI or IOB tags, which helps reduce the cost of corpus construction.

3.1 Datasets

Mary Hamilton (1756-1816), one of the most well-connected women in 18th-century British high society, maintained good relationships with royalty and aristocracy [2]. Her letters provide good opportunities to explore the cultural and social life of the time. Approximately 1,600 digitized letters are available online in XML with TEI tags. After excluding those without relevant annotations, 1,589 were selected for evaluation.

Sir Hans Sloane (1660–1753) was an early modern physician and naturalist, and he left a vast collection that became foundational to British cultural heritage [16]. The Enlightenment Architectures project digitized and TEI-encoded several of Sloane's manuscript catalogue, and Volume I and Volume V were chosen for evaluation.

Old Bailey Online is a searchable archive of criminal trials held at the Old Bailey criminal court from 1674 to 1913 [11]. Given the corpus size and annotation density, 500 files from trials between 1674 and 1747 were selected.

HIPE2020 is a shared task aimed at evaluating NER performance on historical newspapers [9]. It includes newspapers from 1790 to 1960. According to the research scope, 19 newspapers dated between 1790 and 1840 were selected.

[5] https://www.maryhamiltonpapers.alc.manchester.ac.uk/
[6] https://www.oldbaileyonline.org/
[7] https://sloanelab.org/
[8] https://impresso.github.io/CLEF-HIPE-2020/

3.2 Annotations

Since all selected data are already annotated for person and place names, manual annotation is unnecessary. However, as discussed previously, the lack of uniform annotation guidelines means annotators from different projects may interpret person and place names inconsistently. For instance, Mary Hamilton's letters and HIPE2020 primarily focus on locations with geopolitical borders, such as cities, countries, and Parishes. Specifically, the Mary Hamilton Papers include English locations ranging from towns like 'Taxal' and 'Sandleford' to broader regions such as 'Richmond' and 'Northamptonshire', while HIPE2020 favours modern political entities. In contrast, Sloane Catalogue contains place names referring to natural features such as forests, mountains, and rivers. These places tend to be more detailed, e.g., 'chalk pitts near Greenwich in Kent.' Old Bailey place names mostly refer to urban buildings and specific sites, such as streets, pubs, and landmarks like 'White-cross-street' and 'High-gate'.

3.3 Normalization

As discussed, the selected historical datasets originate from different projects, each with its own annotation priorities and focus. As a result, the scope and span of annotated entities vary. For example, different corpora give different answers on whether 'the Duke of Buckingham' should be tagged as a person name, or whether titles such as 'Mr.' in 'Mr. Bradbourn' should be included. Our goal is to normalize these annotations to ensure consistency across the evaluation corpus. This process involves standardizing the span and scope of existing annotations by establishing a set of design principles that ensure a unified standard. By doing so, we aim to build a comprehensive and heterogeneous evaluation corpus that serves as a fair benchmark for testing and comparing different NER systems.

Based on the instances of entities in the historical corpora, the scope and span of person and place entities across the different projects are classified and defined in the Table 1 and Table 2:

Person	
Classification	**Definition**
Names	Proper names of individuals e.g., *Ann Petty*
Honorifics	Words or expressions implying or expressing high status, politeness, or respect. Including common titles (Mr., Ms. Etc.), professional titles (President, Dr. etc.), nobility titles (King, Duke etc.), religious titles (Bishop, Minister etc.) and military titles (Captain etc.).
Names with honorifics	Honorifics followed by person names e.g., *Dr. Campbell*
Conjoined entities	Entities that are combined or linked together, typically with a conjunction such as 'and', 'or' or ','. e.g., *Ann Slow* alias *Ebram*, *James* and *William Greffis*
Nominal mentions	The use of a noun or noun phrase to mention a person entity without using their proper name. Generally, occupation, status and demonym are used to be the nominal mentions. e.g., Cittizen of London, two young men, Prisoner

Table 1: Span and Scope of Person Entities

The normalization should ensure that person and place annotations in the evaluation corpus, derived from multiple corpora, follow a consistent scope. The normalization rules should be based on the following principles:

a) Annotations should have a clear, consistent, and generalizable scope that applies across the included corpora and remains valid if new datasets are added in the future.

b) The evaluation corpus should provide a level playing field for assessing various NER systems, helping to reveal their limitations when applied to long 18th-century texts.

To ensure the rules are both universal and effective in exposing the limitations of historical NER, we reviewed annotation guidelines from mainstream NER datasets to get their annotation

Place	
Classification	**Definition**
Administrative Locations	Geographic areas that are defined and managed by a government or administrative body e.g., convulsed all *Europe*, from *Virginia*, in *Kent*
Natural Features	Naturally occurring physical formations or areas on the earth's surface e.g., famed *Mount Vesuvius*, *River Thames*
Facilities	Functional, primarily man-made structures, including buildings and similar facilities designed for human habitation e.g., apartments at *Louvre*
Thoroughfares	Names of streets, squares, roads, highways, addresses, etc. e.g., *Oxford Road*
Possessive and Descriptive	Constructs that show ownership, possession, or a relationship between things or proper noun is used to describe an extensive entity. e.g., *Fairford* church, *London*'s park In this example, 'Fairford church' is not a proper noun, while 'Fariford' is the descriptive noun of the church (a church in Fairford).
Nested and Embedded Entities	Location entities are embedded in other person or location named entities. e.g., Bay of *Naples*, duke of *Buckingham* In this example, 'Naples' is a location, and 'Naples' is nested in another location entity 'Bay of Naples'
Locative Designators	Terms used to provide specific information about the geographical context of a place. e.g., *River* Thames, Cripplegate *Parish*, *Parish of* Paddington *Parish of..*, *District of..*, *City of..* are categorized in locative designators rather than locative specifiers (e.g. on the coast of California), because they specifies the nature of the entities as being a parish, a district and a city.
Locative Specifiers	Terms used to provide additional details that specify or clarify the geographical position of a place. e.g., *the capital of* Corsica, *coast of* Spain
Locative Modifiers	Terms that add context or specify features of a place name. e.g., *North* London
Nominal mentions	The use of a noun or noun phrase to mention a location entity without using their proper name. e.g., the *Turnpike road*, the *Park*
Compound toponyms	Structures that combine multiple geographical units e.g., *Globe Tavern* in *Fleet-street*, *Sun Tavern* behind the *Royal Exchange*

Table 2: Span and Scope of Place Entities

selections regarding the span and scope of person and place entities. By comparing selections across our historical datasets and four widely used benchmark corpora (MUC, ACE, CoNLL-2003, and OntoNotes), we applied a majority voting approach to derive normalization rules and adjust the existing annotations accordingly. That is, for each category in the Table 1 and Table 2, we adopt the annotation choice used by the majority of the corpora, ensuring resulting rules are generalizable. The detailed majority voting results are provided in Table 5 and Table 6 in Appendix A.

According to the review, all NER corpora classify names, including surnames, given names, and middle names, as person entities, while most exclude conjoined entities and nominal mentions. Among the generic datasets, only ACE and OntoNotes include honorifics, although OntoNotes limits annotation to noble and royal titles, excluding common titles such as 'Mr.' or 'Dr.' In contrast, honorifics appear more prominently in historical corpora. In historical Europe, they were frequently used to refer to nobility or royalty, often without mentioning specific personal names. For example, 'Duke of Buckingham' in the Old Bailey refers to John Smith, and 'the Queen' in Mary Hamilton's letters refers to Queen Charlotte. Excluding such terms could result in losing significant information about these individuals. Therefore, honorifics are included in the normalized evaluation corpus to evaluate NER performance on this feature.

For the majority voting results of place entities, most corpora agree that administrative locations, natural features, and thoroughfares are subtypes of place entities. Regarding facilities, only CoNLL explicitly classifies them as places among the generic corpora. However, facilities are included in the normalized evaluation corpus for two reasons: (1) all historical corpora treat them as places; and (2) although ACE and OntoNotes do not classify them as locations, they define them

as a separate entity type 'Facility,' which indicates their relevance and importance. For annotation boundaries of place entities, both generic and historical corpora give lower-than-average votes to categories such as 'Possessive and Descriptive', 'Nested and Embedded Entities', 'Locative Modifiers', and 'Nominal Mentions'. Therefore, these will be excluded or revised in the normalized corpus. In cases of compound toponyms, historical corpora show partial agreement. Considering their treatment in generic corpora, we decided to normalize such cases into separate location units and include them. Finally, due to ongoing disagreement about locative designators and specifiers in historical corpora, we refer to generic guidelines and choose to retain locative designators while excluding specifiers.

In summary, based on the majority voting results discussed above, the existing annotations were normalized to ensure consistency across the evaluation corpus. Table 3 presents the size of each dataset after normalization, including the number of entities annotated as persons and places. The normalized evaluation corpus supports a fair comparison of historical NER system performance. A detailed description of the normalization procedures is provided in Appendix B.

Dataset	Person entities	Place entities
Old Bailey	110291	18549
Sloane's Catalogue	1964	2324
Mary Hamilton	33536	6208
HIPE2020	69	63
Overall	**145860**	**27144**

Table 3: Overview of the Normalized corpora

4 NER Performance Evaluation

4.1 Tools Selection

Using the normalized corpus, we evaluated the performance of three out-of-the-box NER tools in recognizing persons and places in 18th-century historical texts. We selected Edinburgh Geoparser, spaCy, and a BERT-based model as they are representative of different NER approaches.

In addition, we conducted preliminary experiments with modern Large Language Models (LLMs).[9] However, they were ultimately excluded from the main evaluation for two reasons. First, the output of the LLM is often unstable, making reproducible comparisons impossible. Second, current LLMs cannot reliably provide position indices for each recoginzed entity, which prevents accurate alignment with the gold annotations. For example, in the sentence "Duke of Buckingham lived in Buckingham", the LLM may identify 'Buckingham' as a place entity without specifying which one it refers to. For these reasons, LLMs were not included in the current evaluation.

4.2 Measuring Performance

As discussed in Section 2.1, the performance of typical NER tasks is usually evaluated in terms of Precision, Recall and F1 measures. However, the definition of what counts a correct entity can vary across different tasks. CoNLL adopts a strict evaluation scheme where only entities with exact matches are considered correct, whereas in MUC events, partially correct entities are assigned a 50% weight and contribute to the scores.

[9] We used GPT-4 here.

To illustrate these differences more intuitively, consider the example of 'Duke of Buckingham.' When comparing the recognition results of an NER system with the annotations in the evaluation corpus, several types of outcomes may occur:

- **Complete correctness** occurs when both the entity type and boundary perfectly match the evaluation corpus, such as correctly identifying 'Duke of Buckingham' as a Person entity.
- **Boundary errors** arise when the entity type is correct but the boundary is incomplete or excessive, such as tagging 'Duke' as a Person but missing 'of Buckingham.'
- **Type errors** involve incorrect classification regardless of boundary accuracy, for instance, misclassifying 'Duke of Buckingham' as a Place rather than a Person, even though 'Buckingham' is matched.
- **Missed entities** occur when the system fails to detect entities present in the corpus, omitting 'Duke of Buckingham' entirely.
- **Spurious entities** are false positives, where the system incorrectly identifies non-entities as entities, such as tagging 'Represent' as a Person.

These categories form the foundation for calculating Precision, Recall, and F-score in Section 4.3, providing a comprehensive assessment of system performance across different types of recognition errors.

Given these types of recognition outcomes, we designed five evaluation modes based on the MUC classification of entity results and SemEval's multiple evaluation models to more accurately score system performance under different criteria. These five evaluation modes reflect varying levels of stringency in determining which entities are eligible to contribute to the final score.

Strict evaluation represents the most rigorous approach, where only entities with perfect type classification and exact boundary matches receive full weight, while all other cases are treated as incorrect with no credit. **Type evaluation** relaxes boundary requirements by giving partial credit (50% weight) to entities with correct classification but overlapping spans, while maintaining strict penalties to errors in type classification, as well as missed and spurious entities.

Partial evaluation introduces the most nuanced scoring system. In addition to fully correct entities and those with correct type but overlapping boundaries, entities with incorrect type but identical or overlapping spans are also considered. These categories are weighted at 100%, 50%, and 25% respectively, recognising that boundary detection remains valuable even when classification fails. When considering partial matches, it is essential to recognize the significance of instances where the NER system assigns incorrect types (misclassifications) but achieves partial or complete boundary matches. These cases still provide valuable insights into the system's recognition capabilities. For example, in the phrase 'Queen's garden', the evaluation corpus may annotate it as a place entity, whereas the NER system might recognize only 'Queen' as a person entity. Our goal is not to attempt to demonstrate the failures of NER systems, but rather to analyse the types of errors and their causes. To achieve this, as shown in equation below, instead of completely discarding misclassified matches, we assign a 25% weight to entities with incorrect classifications that have exact or partial boundary matches. In Equation 1, *COR* denotes the number of completely correct entities, *PAR_type* refers to entities with correct type but partially matching boundaries, and *PAR_weak* represents entities with incorrect type but overlapping or matching boundaries.

$$
\begin{aligned}
Precision &= \frac{COR + 0.5 \times PAR\_type + 0.25 \times PAR\_weak}{related\ annotations\ produced\ by\ NER\ system} \\
Recall &= \frac{COR + 0.5 \times PAR\_type + 0.25 \times PAR\_weak}{related\ entities\ in\ evaluation\ corpus} \\
F_1 &= \frac{2 \times Precision \times Recall}{Precision + Recall}
\end{aligned} \tag{1}
$$

Lenient evaluation mirrors Type mode but awards full weight to entities with both correct type and either exact or overlapping boundaries, emphasising boundary detection over perfect precision. **Ultra-lenient evaluation** builds on the Partial mode framework but assigns full 100% weight to all three categories, providing maximum credit for any form of entity recognition regardless of classification accuracy.

4.3 Results and Discussions

We evaluated the performance of three NER tools: Edinburgh Geoparser, spaCy and BERT by using the normalized evaluation corpus. Table 4 presents the evaluation results of these NER tools across the five evaluation modes. For detailed implementation of the data preprocessing and evaluating procedures, please refer to the GitHub repository[10] linked in Appendix B.

	Strict	Type	Partial	Lenient	Ultra-Lenient
Edinburgh Geoparser					
Person	0.502	0.577	0.584	0.652	0.680
Place	0.166	0.317	0.340	0.410	0.504
Macro-F1	0.334	0.447	0.462	0.531	0.592
Micro-F1	0.454	0.532	0.541	0.610	0.650
Spacy					
Person	0.440	0.482	0.483	0.524	0.530
Place	0.086	0.188	0.209	0.254	0.338
Macro-F1	0.263	0.335	0.346	0.389	0.434
Micro-F1	0.386	0.432	0.437	0.478	0.498
BERT					
Person	0.513	0.622	0.626	0.731	0.748
Place	0.208	0.405	0.408	0.497	0.508
Macro-F1	0.360	0.513	0.517	0.614	0.628
Micro-F1	0.470	0.575	0.580	0.681	0.696

Table 4: Scores of NER systems

The results show that even in the ultra lenient evaluation mode, the highest score remains below 70%, which indicates that NER faces significant challenges when applied to historical documents. In addition to the inherent limitations of the approaches, features specific to historical documents, such as multilingual content, noise introduced during digitisation, and language variation over time, making it difficult for NER systems trained on contemporary data to perform effectively [8]. Overall, the BERT-based model demonstrates the best performance across all evaluation modes, which can be attributed to its context-aware architecture and large-scale pretraining, which provides robustness against the complex historical texts. The Edinburgh Geoparser performs moderately well. Its reliance on rule-based methods and gazetteer limit its general applicability to more diverse or ambiguous contexts. In contrast, spaCy shows the lowest scores. This can be attributed to its

[10] `https://github.com/SepNmoon/NER-evaluation-historical-corpus`

heavy reliance on models trained on contemporary, clean data, which makes it difficult to adapt to historical documents. Consistently, across all modes, all NER tools perform better on person entities compared to place entities. Person names typically follow more recognisable patterns, and the presence of titles and honorifics aids identification. As shown in Table 1 and Table 2, the scope and span of place entities are more complex and diverse, which makes NER tools harder to identify consistently. Moreover, compared to person names, the identification of place names is easier affected by OCR errors.

Next, we analysed the types and causes of errors made by spaCy, BERT, and the Edinburgh Geoparser on the historical evaluation corpus. Four broad categories of errors were identified. **Error 1** represents missed entities where systems fail to detect entities present in the gold annotations. These are commonly caused by formatting issues such as extra spaces, transcription errors, abbreviations, the use of hyphens, and the omission of honorifics or small-scale locations. **Error 2** involves spurious entities where non-entities are incorrectly identified as entities. This is mainly due to formatting issues and case sensitivity in spaCy and the Geoparser, while BERT specifically struggles with sub-word splitting errors. **Error 3** is about type misclassification where entities are detected but assigned incorrect categories, commonly due to nested entities, ambiguity and formatting issues. **Error 4** covers boundary mismatches where entity types are correct but spans are incomplete or excessive, typically resulting from entity merging errors, definite articles, nested entities, and various formatting issues.

As frequently noted above, spelling and formatting issues are the most common sources of errors. spaCy and Geoparser struggle to identify entities correctly when their forms are disrupted by elements such as extra spaces introduced during transcription, hyphens, or apostrophes, which break the expected entity patterns. An illustrative example is shown in Figure 1. The figure consists of two elements: the left one represents the actual annotation in the evaluation corpus, while the right one displays the result identified by the NER system.

```
[[316256, 316271, 'Dr. - Lavater .', 'pers'], ['O']]
[[317396, 317413, 'Dr. Green= =voelt', 'pers'], ['O']]
```

Figure 1: Entities with Spelling Issues

In comparison, BERT is less affected by entities with formatting and spelling issues, although hyphens and abbreviations still impact its performance. Errors introduced during transcription also present a significant challenge for NER systems. In HIPE dataset, for instance, many occurrences of ‘s’ were transcribed as ‘f’ incorrectly, which disrupts the word spelling and reduces the accuracy of recognition. Additionally, the presence of historical orthographic features, such as the long s, also disrupt recognition (as shown in Figure 2).

```
[2581, 2592, 'Miſs Burney', 'pers'], ['O']
```

Figure 2: Entities with Long S

Furthermore, the recognition capabilities of NER tools are highly dependent on capitalized initials, especially in the case of Geoparser. While this may work well in well-edited modern texts, it introduces more errors when applied to complex historical texts. For example, Geoparser frequently misclassifies capitalized non-entities as entities, or fails to recognize entities that appear in lowercase. Figure 3 shows an extreme example where the Geoparser incorrectly merged all adjacent capitalized words into a single entity.

For person names, a significant source of score loss stems from the inconsistent recognition of honorifics. Both BERT and spaCy rarely identify honorifics as part of person entities. The Edinburgh Geoparser holds a slight advantage in this regard, as it can recognize some common honorifics, such as ‘Mr.’ This limitation also affects nested entity recognition. For example, in

```
[948, 964, 'Charles Goodhand', 'pers'], [937, 1108, 'John Parry Charles Goodhand Thomas
Whit John Preston Thomas Wilkinson Samuel Brettridge William Finch John Peacock William
Griffin John Becket Thomas Tatlock Thomas Pattle', 'person']
```

Figure 3: Over-merging Capitalized Words

cases where a place name is nested within a person entity, such as 'Prince of Wales,' spaCy only recognizes 'Wales' as a place entity. Another instance is the place entity 'Davis's Straits,' where spaCy incorrectly tags 'Davis' as a person entity.

For place names, the evaluation corpus includes many small-scale locations extracted from the Old Bailey dataset, such as streets, pubs, and churches. Both spaCy and the Edinburgh Geoparser perform poorly on these types of entities, while BERT demonstrates stronger recognition capabilities. However, BERT's performance declines when processing entities that lack explicit place designating terms. For example, as shown in Figure 4, 'Prince's Head' is the name of a pub and does not contain common locative indicators, BERT fails to correctly classify it as a place entity. In this figure, the 'O' on the right side means the entity was not recognized by the NER system.

```
[9647, 9661, "Prince 's Head", 'place'], ['O']
```

Figure 4: Entities without Explicit Indicators

Additionally, for BERT, most errors are concentrated in the splitting of sub-words. Due to its tokenization mechanism, BERT may split an entity into multiple sub-word tokens, and some sub-words fail to merge after prediction, which results in a large number of meaningless sub-word tokens.

Besides the aforementioned errors, NER tools also make errors in identifying the boundaries of entities with definite articles, locative specifiers and locative designators. Although this issue appears less severe—since they often capture the core element of the entity—it still impacts overall accuracy. For example, historians may prefer recognizing the full phrase 'North of France' rather than just 'France.'

Overall, the errors made by the three NER tools are concentrated in several specific areas, particularly issues related to entity formatting and spelling, including transcription errors and inherent spelling variations in the texts. The performance of spaCy and Geoparser in recognizing small-scale places is poor, whereas BERT performs significantly better. However, some place entities are still missed. Additionally, there are also issues such as the failure to recognize honorifics, nested entities and locative designators. These findings reveal the limitations of NER when applied to historical texts and provide valuable insights to guide the future development of historical NER.

5 Conclusion and Future Work

In this study, we address the issue of incomparability of different NER systems on historical documents by developing a shared evaluation corpus. This corpus was constructed from four historically significant datasets with annotations normalized to ensure consistency across sources. Using this corpus, we evaluated the performance of three representative NER tools in recognizing person and place entities. The testing results provide an initial understanding of the limitations of existing NER tools and reveal the challenges inherent in applying NER to historical texts.

However, this study also has some limitations. Firstly, we only focus on the recognition of person and place names, whereas historical research may involve more entity types. In addition, the annotations in the corpus were normalized through majority voting across historical and generic datasets; whether these annotation choices truly align with historians' expectations remains uncertain. The normalization rules should be adjusted to better reflect historians' actual research

needs.

In future work, we plan to communicate with historians through interviews and questionnaires to better understand their requirements. Based on the results, we aim to develop a gold standard corpus that aligns with their needs and expectations, which can be used to comprehensively evaluate the performance of NER applied to historical documents.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. We are also grateful for the creators of the datasets used in this study, including the Old Bailey Online, Mary Hamilton Papers, Sloane's Catalogue and HIPE2020. These resources made this research possible.

References

[1] ACE08. "Automatic Content Extraction 2008 Evaluation Plan (ACE08) ". 2008. URL: https://my.eng.utah.edu/~cs6961/papers/ACE-2008-description.pdf.

[2] Barker, Hannah, Coulombeau, Sophie, Denison, David, Oudesluijs, Tino, Ulph, Cassandra, Wallis, Christine, and Yáñez-Bouza, Nuria. "Unlocking the Mary Hamilton Papers". 2023. URL: https://www.maryhamiltonpapers.alc.manchester.ac.uk/.

[3] Bates, Marcia J. "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions ". In: *College & Research Libraries 57*, no. 6 (1996), pp. 514–523. DOI: 10.5860/crl_57_06_514.

[4] Chinchor, Nancy and Sundheim, Beth. "MUC-5 Evaluation Metrics ". In: *Proceedings of the 5th conference on Message understanding*. 1993, pp. 69–78. DOI: 10.3115/1072017.1072026.

[5] Chiron, Guillaume, Doucet, Antoine, Coustaty, Mickael, Visani, Muriel, and Moreux, Jean-Philippe. "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information ". In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Toronto, ON, Canada, 2017. DOI: 10.1109/JCDL.2017.7991582.

[6] Cunningham, Hamish. "Information extraction, automatic ". In: *Encyclopedia of Language & Linguistics*. Elsevier, 2006, pp. 665–677.

[7] Doddington, George, Mitchell, Alexis, Przybocki, Mark, Ramshaw, Lance, Strassel, Stephanie, and Weischedel, Ralph. "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation ". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal, 2004, pp. 837–840.

[8] Ehrmann, Maud, Hamdi, Ahmed, Pontes, Elvys Linhares, Romanello, Matteo, and Doucet, Antoine. "Named Entity Recognition and Classification in Historical Documents: A Survey ". In: *ACM Computing Surveys 56*, no. 2 (2023), pp. 1–47. DOI: 10.1145/3604931.

[9] Ehrmann, Maud, Romanello, Matteo, Flückiger, Alex, and Clematide, Simon. "Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers ". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 2020, pp. 288–310. DOI: 10.1007/978-3-030-58219-7_21.

[10] Grishman, Ralph and Sundheim, Beth. "Design of the MUC-6 evaluation ". In: *MUC6 '95: Proceedings of the 6th conference on Message understanding*. 1995, pp. 1–11. DOI: 10.3115/1072399.1072401.

[11] Hitchcock, Tim, Shoemaker, Robert, Emsley, Clive, Howard, Sharon, and McLaughlin, Jamie. "The Old Bailey Proceedings Online". 2023. URL: www.oldbaileyonline.org.

[12] Humbel, Marco, Julianne Nyhan, Andreas Vlachidis, Sloan, Kim, and Ortolja-Baird, Alexandra. “Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future ”. In: *Journal of Documentation 77*, no. 6 (2021), pp. 1223–1247. DOI: 10.1108/JD-02-2021-0032.

[13] “IREX. IREX Named Entity Task Homepage”. URL: https://nlp.cs.nyu.edu/irex/index-e.html.

[14] Marrero, Mónica, Urbano, Julián, Sanchez-Cuadrado, Sonia, Morato, Jorge, and Gómez-Berbís, Juan Miguel. “Named Entity Recognition: Fallacies, challenges and opportunities ”. In: *Computer Standards & Interfaces 35*, no. 5 (2012), pp. 482–489. DOI: 10.1016/j.csi.2012.09.004.

[15] Nadeau, David and Sekine, Satoshi. “A survey of named entity recognition and classification ”. In: *Lingvisticæ Investigationes. International Journal of Linguistics and Language Resources 30*, no. 1 (2007), pp. 3–26. DOI: 10.1075/li.30.1.03nad.

[16] Nyhan, Julianne, Flinn, Andrew, Vlachidis, Andreas, Carine, Mark, Hill, JD, and Jansari, Sushma. “Welcome to the Sloane Lab - Sloane Lab”. 2023. URL: https://sloanelab.org/.

[17] Ortolja-Baird, Alexandra, Pickering, Victoria, Nyhan, Julianne, Sloan, Kim, and Fleming, Martha. “Digital Humanities in the Memory Institution: The Challenges of Encoding Sir Hans Sloane’s Early Modern Catalogues of His Collections ”. In: *Open Library of Humanities 5*, no. 1 (2019). DOI: 10.16995/olh.409.

[18] Rodriguez, Kepa J., Bryant, Mike, Blanke, Tobias, and Luszczynska, Magdalena. “Comparison of Named Entity Recognition Tools for Raw OCR Text ”. In: *Proceedings of KONVENS 2012 (LThist 2012 workshop)*. 2012, pp. 410–414. DOI: 10.13140/2.1.2850.3045.

[19] Sang, Erik F. Tjong Kim and Meulder, Fien De. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition ”. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147.

[20] Segura-Bedmar, Isabel, Martínez, Paloma, and Herrero-Zazo, María. “SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013) ”. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*. Atlanta, Georgia, USA, 2013, pp. 341–350.

A Majority Voting Results

Table 5 summarizes the annotation choices for person entities in both generic and historical NER datasets respectively. A ‘Yes’ indicates that the dataset includes this type of person entity, whereas a ‘No’ indicates it does not. ‘Partial yes’ indicates that the entity type exists in the corpus, but only some of its occurrences are annotated, revealing inconsistencies in the labelling process.

Table 6 presents the annotation choices regarding place entities in both generic and historical corpora. Empty cells indicate cases where information was not available in the official documentation.

	Generic Corpora				Historical Corpora			
	MUC7	ACE	CoNLL-2003	OntoNotes	Old Bailey	Sloane's Catalogue	Mary Hamilton	HIPE2020
Names	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Honorifics	No	Yes	No	Partial yes	Yes	Yes	Yes	No
Honorifics with names	No	No	No	Partial yes	Yes	Yes	Yes	Yes
Conjoined Entities	No	No	No	No	Yes	No	No	No
Nominal Mentions	No	Yes	No	No	No	No	No	No

Table 5: Person Annotations in Generic vs. Historical Corpora

	Generic Corpora				Historical Corpora			
	MUC7	ACE	CoNLL-2003	OntoNotes	Old Bailey	Sloane's Catalogue	Mary Hamilton	HIPE2020
Administrative Locations	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Natural Features	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Facilities	No	No	Yes	No	Yes	Yes	Yes	Yes
Thoroughfares	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Possessive and Descriptive	No		No	No	No	No	No	No
Nested and Embedded Entities	No	Yes	No	No	No	No	No	No
Locative Designators	Yes	Yes	Yes	Yes	Partial yes	Yes	Partial yes	Partial yes
Locative Specifiers	No	Yes	No	No	No	Yes	Partial yes	No
Locative Modifier	No	Yes	No	No	No		No	No
Nominal mentions	No	Yes	No	No	No	No	No	Yes
Compound toponyms	No	Yes	No	No	Yes	Yes	No	No

Table 6: Place Annotations in Generic vs. Historical Corpora

B Implementation Details of Normalization

For person entities, the normalization process involved two main steps: adding missing honorifics in HIPE2020 and splitting conjoined entities in the Old Bailey corpus. Since HIPE2020 is relatively small, we manually reviewed the dataset and add annotations of honorifics. Conjoined entities, such as 'Susan Banster, alias Green', 'John Clarke, otherwise Maiden', 'Robert, William, and Margaret Dine' or 'Hester the Wife of Peter Sayre', were mainly present in the Old Bailey dataset. We automatically identified these cases using regular expressions by extracting patterns containing keywords such as 'alias', 'and', 'otherwise' or 'wife', and then manually split each instance into separate annotations to ensure that every name was treated as a distinct entity.

For place entities, normalization was required across all datasets based on the results of majority voting. We adopted a hybrid approach combining automatic extraction with manual correction. The HIPE2020 was first manually reviewed to remove nominal mentions and add locative designators. In addition, both Old Bailey and Mary Hamilton datasets required to include locative designators. We used regular expressions to extract capitalized word followed by 'of' immediately preceding a `<placeName>` tag, such as 'City of', 'Parish of', and 'District of'. We then expanded the annotation boundaries to incorporate these designators into the corresponding place names (City of

<placeName>London</placeName> → <placeName>City of London</placeName>). By contrast, locative specifiers were removed by extracting place names that contained a lowercase token followed by 'of', such as 'sea coast of Sussex'. Finally, compound toponyms were normalized by splitting multi-token place entities that contained lowercase elements, for example, 'Sun Tavern behind the Royal Exchange'. Since many church names (e.g., 'St Martin in the Fields') follow this pattern but should be treated as a single entity, we excluded names beginning with 'St.' from this operation.

All code used for preprocessing and evaluation is available at: https://github.com/SepNmoon/NER-evaluation-historical-corpus