Cost-Aware Pre-Annotation Strategies for Nested NER in Historical Latin Notarial Deeds

Charlene Ellul, Vanessa Buhagiar, Claudia Borg and Charlie Abela
Department of Artificial Intelligence, University of Malta, Malta
{charlene.ellul, vanessa.buhagiar, claudia.borg, charlie.abela}@um.edu.mt

Abstract

Manual annotation for Named Entity Recognition in historical documents remains expensive and time-consuming, particularly for complex nested entity structures in domain-specific texts such as Latin notarial deeds. Active learning frameworks like the Humanities Entity Recognizer (HER) reduce annotation requirements by iteratively selecting informative samples for expert annotation, nevertheless, existing sentence-based sampling strategies create unpredictable annotation costs when sentence lengths vary extensively. We extend the HER to support nested entities through composite BIO label encoding and introduce token-budgeted sample selection to address annotation cost variability. Under token-budgeting, each annotation iteration targets a fixed token budget rather than a fixed sentence count, while Active Curriculum Learning (ACL) ensures diverse sentence length representation in initial samples. Experiments on seventeenth-century Latin notarial deeds from Malta's Notarial Registers Archive demonstrate that token-budgeted sampling achieves comparable macro-span F1 to sentence-based sampling while exhibiting more stable learning trajectories across iterations. Additional experiments examining entity-level performance revealed systematic variation by semantic granularity, with higher-level categorical entities achieving stronger recognition than role-based middle-level entities, which depend on discourse context. Our results demonstrate that controlling sample selection at the token level rather than sentence level provides more predictable annotation planning for active learning in historical document corpora with heavy-tailed sentence length distributions.

Keywords: Latin notarial documents, active learning, nested entities, historical documents

1. Introduction

The Notarial Registers Archive (NRA)[1] in Valletta, Malta preserves an extensive collection of more than 24,000 bound volumes, fragments, and loose folios that span seven centuries of history of the Maltese islands, with the earliest records dating to the fifteenth century. Managed by the Notarial Archives Foundation (NAF)[2], this growing collection encompasses a wide range of legal documents, including marriage contracts, property transactions, and documents relating to commerce and maritime trade. Despite being a fundamental source for the study of Maltese history, this collection is limited in the meaningful access it can provide to the researcher, as the only systematic finding aid is the *Repertorio*, which is an index, typically preserved in the form of a bound manuscript that lists the party names and folio references, offering no insight into the semantic content of the deeds themselves. Another barrier to accessing this collection is that of scribal variability, and it is an issue that is shared by other archival collections in Malta and beyond. The case of the NRA adds a layer of complexity owing to the incorporation of Italian, Sicilian, and Maltese linguistic elements within the main textual corpus that is written in Latin and in later centuries in Italian.

Accurate interpretation of these records requires years of study to build familiarity with the different script types and hands in which the deeds were written, and hinges on a confident understanding of the large compendium of abbreviations and non-standardised historical orthography scribes used in their work. These competencies are increasingly rare, and without them the textual content of the archive remains effectively inaccessible. As a result, a vast documentary heritage that is meticulously preserved risks becoming a silent archive: its knowledge locked within pages that few can decipher.

Recent work in archival and library studies has increasingly argued that archival catalogues and descriptive metadata should not be treated merely as static finding aids, but as structured data capable of supporting large-scale computational analysis and reuse. The adoption of Linked Data and ontology-based models enables archival descriptions to be transformed into interconnected semantic representations that support querying, aggregation, and integration across collections and institutions (Lapôtre, 2017; Bensmann et al., 2017).

This conceptual shift reframes archives as dynamic knowledge infrastructures rather than passive repositories, in which descriptive metadata and document-level content form part of a unified data system (Koch et al., 2023). This underpins initiatives such as the NotaryPedia (Ellul et al., 2019) project that leverages this perspective of a dynamic knowledge infrastructure and seeks to inte-

[1] https://nationalarchives.gov.mt/en/notarial-registers-archive-valletta/
[2] https://nafmalta.org/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 407–417.
11 May 2026.

grate archival standards, digitised notarial sources, and language technologies for knowledge extraction into a unified knowledge management platform. Within NotaryPedia, these Linked Data principles are operationalised by enriching archival catalogue descriptions with Named Entity Recognition (NER)-derived entities extracted from the deeds themselves, enabling queryable knowledge structures over deed content.

A major obstacle in developing effective NER models for historical archives is the cost of transcribed datasets and the manual annotations carried out on them. As highlighted in recent surveys (Ehrmann et al., 2023; Keraghel et al., 2024), historical NER typically depends on transcriptions that are often produced or validated with palaeographic expertise, and on entity annotation performed by domain experts, such as historians. Scribal and orthographic variation, abbreviations, domain-specific language, and the need for interpretive judgement make large-scale gold-standard annotation impractical.

In this work, we focus on annotation cost as a planning problem rather than as a direct measurement of annotation time. For historical corpora with heavy-tailed sentence-length distributions, selecting a fixed number of sentences per iteration leads to highly unpredictable annotation workloads. We therefore investigate whether controlling sample selection at the token level can stabilise annotation effort proxies while maintaining learning effectiveness in active learning–based pre-annotation for nested NER in seventeenth-century Latin notarial deeds from the NRA. Preliminary palaeographic assessment of the corpus found that the notarial deeds from the seventeenth century were more consistently legible and less heavily abbreviated than those of the preceding century, thus making this subset suitable for an initial controlled study.

We compare three pre-annotation strategies reflecting common practice and documented limitations: (i) sentence-based sampling as implemented in HER (Erdmann et al. (2019)), (ii) token-budgeted sampling combined with ACL to improve early-stage coverage across sentence lengths and address cost variability noted in historical NER, and (iii) a staggered strategy that incrementally introduces semantic complexity across nested entity levels. The goal is not only to improve model performance but also to develop annotation workflows that are realistic and sustainable for archival institutions.

2. Background And Related Work

2.1. Annotation Cost As A Central Constraint In Historical NER

In a comprehensive survey, Ehrmann et al. (2023) identifies lack of resources as one of the four core challenges of historical NER, alongside document diversity, noisy input, and linguistic variation. Settles et al. (2008) emphasise that annotation effort is not constant per sentence. It varies with sentence length and the complexity of language used. These are properties that are especially pronounced in administrative and legal records, such as notarial deeds.

2.2. Active Learning for NER

Pre-annotation reframes annotation as a human-machine collaboration, in which models generate candidate annotations that are then validated and corrected by language experts. This approach reduces cognitive load and accelerates annotation while preserving expertise.

Pre-annotation can be integrated with active learning, allowing models to select the most informative samples for annotation. A key contribution in this area is the Humanities Entity Recognizer (HER, Erdmann et al., 2019), explicitly designed for humanities data. HER combines sampling seeds to train CRF models iteratively, using only a subset of the manually annotated data.

Active learning is a well-established methodology in NLP. Settles (2009) describes core strategies, including uncertainty sampling, query-by-committee, and density-weighted methods. After being developed and most extensively studied for classification, active learning query strategies were later adapted to structured prediction, including sequence labelling tasks such as NER. Shen et al. (2017) show that standard uncertainty measures for NER operate at the sequence level and that common least-confidence formulations can disproportionately select longer sentences because uncertainty is aggregated across tokens. They propose length-normalised scoring to mitigate this bias. This characteristic is particularly pronounced in historical texts, where long and complex sentences can dominate active learning selection.

2.3. NER in Latin and Historical Documentary Texts

Recent work on NER for historical documentary texts has focused on medieval charters. Chastang et al. (2021) utilise a CRF-based model to handle the shift from single names to complex multi-component denominations and the prevalence of nested entities within legal Latin char-

ters using Burgundian diplomatic documents from the 9th to the 14th century. To show that formulaic legal discourse can support strong cross-regional and cross-chronological generalisation, Aguilar and Stutzmann (2021) train spACy, Flair, and a custom BiLSTM-CRF model on French medieval charters. Integrating NER into Handwritten Text Recognition (HTR) pipelines has also become a priority to reduce cascading errors in noisy archival data. At manuscript level, Boroş et al. (2020) demonstrated that combined HTR+NER approaches showed superior performance over sequential pipelines HTR→NER approaches on multilingual corpora of Latin, German, and Czech charters. Torres Aguilar (2022) uses multi-lingual medieval charters including Latin, with stacked embeddings and BERT-based models to address complex multilingual writing practices, effectively modelling code-switching and bilingual sequences without a performance drop compared to monolingual counterparts. Novotny et al. (2023) presented a bootstrapping pipeline for late medieval charters in Czech, Latin, and German, demonstrating the scalability of these techniques for large-scale historical datasets. Related workflow-oriented research is the research on the late medieval Bolognese Memoriali series (1265–1452) Loss et al. (2025), where they developed a tailored tagging system for facilitating direct metadata extraction from handwritten notarial records into structured databases. Our study shifts attention from medieval charter corpora to seventeenth-century Latin notarial deeds and focuses specifically on nested NER within a cost-aware pre-annotation and active-learning workflow.

2.4. Nested Named Entity Recognition

Nested NER has been recognised as a distinct problem in NLP for over a decade. Finkel and Manning (2009) introduced one of the earliest models for nested entity recognition using constituency parsing, while Lu and Roth (2015) introduce *mention hypergraphs* for joint mention extraction and classification, designed to capture overlapping entity spans. More recent neural approaches achieve strong results on standard nested NER benchmarks, including merge-and-label architectures (Fisher and Vlachos (2019)) and sequence-based learning/decoding methods for extracting inner entities (Shibuya and Hovy (2020)). However, these approaches are typically evaluated on contemporary datasets with abundant annotations and relatively stable languages.

The lack of work explicitly addressing active learning for nested NER in historical contexts represents a significant gap in the literature. Addressing this gap requires rethinking pre-annotation not merely as a speed-up mechanism, but as a structural annotation methodology aligned with semantic granularity and annotation effort.

3. Archival Context And Corpus

The corpus used in this study is selected from the NRA whose main collection consists of registri, which are the official copies of the notarial acts. Buttigieg and Abela (2020) describe a notary's workflow beginning with the *bastardellum* (draft copy) that functioned as a notebook in which the notary recorded the essential details of a deed, such as the date, type of contract and the names of the contracting parties. These notes were eventually expanded into their full legal form in the *minutarium* (original), a volume containing the original version of the act which also carried the highest significance in legal terms. A faithful copy of the original deed was transcribed chronologically in the *registrum* (register). These registers are one of the richest sources for understanding social, economic, and legal life in Malta and the central Mediterranean.

In this work, we use a sample of 102 deeds from a single bound register: the Register of Notary Giovanni Battista Micallef, 1633 – 1634, Volume 1 (R352/1). This register contains notarial deeds primarily written in Latin, although instances occur in which deeds begin in Latin and subsequently continue in Italian/Sicilian. In addition, where no suitable Latin or Italian/Sicilian equivalent exists for a Maltese vernacular term, such words are occasionally rendered in a Latinised form (e.g., *dublectum*[3], skirt). The texts, therefore, exhibit lexical and syntactic influence from Latin and Italian, alongside extensive use of abbreviations typical of notarial practice. The deeds vary in content and length, and also display considerable variation in sentence length, with some clauses spanning several hundred tokens. Owing to the absence of terminal punctuation, which is typical of the Latin language in general, paragraph boundaries are also treated as sentence boundaries.

Figure 1 shows an analysis of sentence length in the corpus, revealing a strongly right-skewed distribution. The majority of sentences contain fewer than 30-40 tokens, while a long tail of very long sentences extends beyond 100 tokens and, in some cases, exceeds 300 tokens. Heavy-tailed distributions of this kind are common across linguistic data. Bentz et al. (2014) have shown that languages with richer inflectional morphology tend to exhibit heavier-tailed patterns in their word-frequency distributions, suggesting that morphological complexity can influence the shape of such distributions.

The corpus is accompanied by expert-curated reference material, including gazetteers derived from

[3]NRA, R494/1/SUB/3, Notary Giacomo Zabbara, f.3r, 11 July 1495

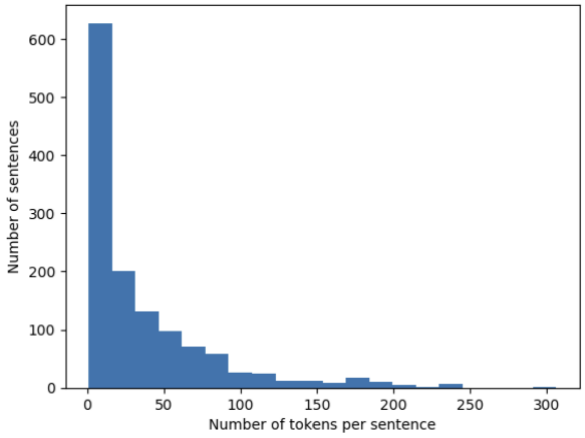

Figure 1: Token distribution per sentence in fullCorpus.

established publications (Wettinger, 2000; Fiorini, 1996).

The transcriptions used in this study were produced as part of the broader NotaryPedia workflow and were normalised for the present experiments by removing inline markup and converting them into a token-per-line format, resulting in 48,197 tokens. [4]

These corpus characteristics motivate the pre-annotation and active learning strategies described in the next section.

4. Annotation Scheme and Entity Hierarchy

4.1. Motivation for Nested Entity Annotation

In our work, nested entities have two purposes. They capture both functional and ontological hierarchies present in notarial deeds. For person entities (PER), nesting reflects both role-based distinctions, such as whether a person is a party or a witness (PER-P/PER-W), and structural components, namely first names and last names (PERFN and PERLN). Similarly, keyword entities can represent part-whole entities, such as a geographical location and these can be embedded within a phrase. For instance, in expressions such as *commrij commendae de Rocchabruna* (the Commander of the Commandery of Roccabruna), the full phrase is annotated as (KEY), while the embedded place name *Rocchabruna* is simultaneously annotated with the geographical entity (GEO), yielding the composite label (explained further in Sec 4.3) `KEY-I+GEOB-B`. Although these nested entities may share token boundaries, they represent distinct semantic layers that are essential for the accurate interpretation of notarial records.

4.2. Entity Levels

Entities are organised into three conceptual layers that reflect different semantic roles and levels of detail. Higher-level entities are the core semantic categories. Middle-level entities are the role-based distinctions, capturing the legal role within a deed. Parties are annotated only when explicitly listed in the marginal notes. Witnesses are usually mentioned at the end of a deed. Lower-level entities are the structural or ontological components, capturing the fine-grained semantic details of a higher-level entity. A definition of all entity labels per level is found in Table 1.

Level	Label	Description
Higher	`DATE`	Temporal expressions
Higher	`DEEDT`	Deed type, typically assigned by the notary, often recorded in the margin
Higher	`GEO`	Geographic entities including place names, territorial references and buildings
Higher	`KEY`	Interesting keyphrases
Higher	`PER`	Person entities
Higher	`INS`	Institution
Middle	`PERP`	Party of a deed
Middle	`PERW`	Witness of a deed
Lower	`DATEAN`	*a nativitate* date format
Lower	`DATEANIN`	*a nativitate* and indiction date format
Lower	`DATEAI`	*ab incarnatione* date format
Lower	`DATEAIIN`	*ab incarnatione* and indiction date format
Lower	`DATEIN`	Indiction date format
Lower	`GEOB`	Buildings
Lower	`GEOPC`	Geographic location where the deed was drawn
Lower	`PERFN`	Person first name
Lower	`PERLN`	Person last name
Lower	`PERNN`	Person nickname

Table 1: Entity labels grouped by semantic level in the annotation scheme.

4.3. Encoding Nested Entities

HER does not inherently support nested entities. To preserve the original architecture while extending its functionality, nested entities were linearised using composite BIO labels (Straková et al., 2019)

[4] The original transcriptions preserves additional inline annotations for palaeographic and documentary phenomena, such as marginal notes, abbreviations, and layout cues. These annotations are not used in the present NER experiments, but were retained because the corpus is intended for reuse in future NotaryPedia modules, including automatic transcription and richer information extraction.

that encode hierarchical structure by concatenating entity tags with a '+' delimiter (e.g. `GEO-B+GEOPC-B` or `PER-B+PERP-B+PERFN-B`). This strategy enables nested entities to be predicted within a single tagging layer. Annotation was conducted in a staggered manner, progressively introducing lower-level entities after higher-level entities had been established.

5. Pre-Annotation Approach

We used the HER (Erdmann et al., 2019) as the baseline active learning framework and describe the extensions required to support nested entity annotation in seventeenth-century Latin notarial deeds. HER was developed for humanities and historical collections, where annotated data are scarce and pre-annotation is intended to support expert correction rather than replace it. Accordingly, our aim is not to compare state-of-the-art NER architectures, but to examine how alternative pre-annotation and sample-selection strategies influence learning behaviour and annotation planning in a low-resource historical setting. We therefore adopt HER's Conditional Random Field (CRF)-based sequence labelling models, which are computationally lightweight and suitable for iterative active learning workflows.

This study does not measure annotation time directly. Instead, annotation effort is approximated using two observable values: (i) the number of annotated tokens per iteration and (ii) the distribution of nested entity depths within each annotation batch. Token count captures gross workload size, while nesting depth reflects semantic and structural complexity. These measures allow us to compare annotation strategies under controlled and reproducible conditions, while acknowledging that actual annotation time may vary with document variation and annotator expertise.

5.1. HER-Based Pre-Annotation Baseline

While Erdmann et al. (2019) report learning curves at fixed token milestones for experimental consistency, our baseline follows the HER repository workflow and uses fixed sentence batches (200 sentences per iteration) to mirror typical tool usage. The individual deeds, distributed across multiple text files, are pre-processed to remove inline annotations and to resolve hyphenated line breaks. The deeds are then concatenated and converted into a one-token-per-line format of the form `<label><tab><token>`, with all labels initialised to `O` (Outside). This constitutes the full unlabelled corpus used throughout the experiments.

The workflow begins with selecting 200 randomly sampled sentences, which are manually annotated as the seed corpus. The remaining sentences constitute the unannotated corpus. Gazetteers are used to perform rule-based pre-annotation, thereby expanding the manual annotation of the seed corpus.

The seed corpus is used to train a CRF tagger. The trained model can be used to pre-annotate the candidate entities of the unannotated corpus. If the performance of the trained CRF model is deemed unsatisfactory, the gazetteers are updated with the newly annotated entities, and a ranking mechanism is applied on the unannotated corpus based on the Pre-Tag DeLex (PTDL) (Erdmann et al. (2019)) strategy to rank them according to how useful they would be to annotate next.

PTDL prioritises sentences likely to contain entities by targeting informative out-of-vocabulary (OOV) tokens. It begins with pre-tagging the unannotated corpus using gazetteer matches extracted from the annotated seed. HER splits it into UNE (sentences containing a pre-tagged entity) and UnoNE (all remaining sentences). It then trains feature-based linear-chain CRF models (CRFsuite; Okazaki (2007)) using only delexicalized (context-based) features and ranks sentences by summing the weighted frequency scores of unique OOV tokens, normalised by sentence length. In our implementation, we ran PTDL for four active-learning rounds. In each round we manually annotated the top-ranked 200 additional sentences, retrained the models, and re-ranked the remaining pool of unannotated corpus. Model training and feature selection are described in Section 5.3.

5.2. Gazetteers Construction

In this study, the initial gazetteers were manually compiled by an expert annotator using established scholarly reference works. These include Place-Names of the Maltese Islands, ca. 1300–1800 (Wettinger, 2000) and the indexes of the Documentary Sources of Maltese History series (Fiorini, 1996). These publications cover personal first names, last names, place names, buildings, institutional entities, and keywords. After each annotation iteration, the gazetteers are expanded using newly annotated data, with additional gazetteers also created for nested entity types.

5.3. Model Training and Feature Selection

For each iteration, a CRF model was trained using three-fold cross-validation on the seed corpus, with two-thirds of the data used for training and one-third for testing in each fold. We trained multiple CRF models using predefined feature-set templates, including word-shape features, character n-grams, contextual token windows (e.g., previous

word/bigram), and gazetteer indicators, and evaluated each configuration by macro span-F1 averaged across folds. The best-performing feature configuration was then used to train the final model on the full seed set for that iteration. Experiments were run on a CPU-only Linux environment (Ubuntu 24.04 LTS, 2 vCPUs, 8 GB RAM).

6. Methodology

6.1. Token-Budgeted Seed Selection Approach

A key contribution of this work is the replacement of sentence-based seeds with token-budgeted seeds. Because the sentence length distribution is skewed, selecting a fixed number of sentences can result in a disproportionate share of annotation effort being allocated to a small number of very long sentences. By introducing a constraint whereby each annotation batch contains approximately 2,000 tokens, this imbalance in annotation cost is substantially reduced. We set 2,000 tokens as a practical compromise: large enough for meaningful retraining, small enough for frequent iterations and to avoid domination by very long sentences.

6.2. Active Curriculum Learning

To address representation imbalance in the initial seed, an ACL strategy is introduced for the first 2,000-token batch. Under this strategy, instead of selecting a random set of sentences from the unlabelled corpus, sentences are first grouped into bins based on token length (e.g. 1–20, 21–50, 51–100, 101–200, >200 tokens). This curriculum-based constraint is applied only in the first iteration to accelerate model learning by exposing the CRF to a representative range of sentence lengths early, following the general principles of ACL described by Jafarpour et al. (2021). From the second iteration onward, the token-budgeted condition followed the same HER/PTDL ranking procedure as the sentence-based baseline, except that sentences were accumulated until the batch reached approximately 2,000 tokens rather than a fixed count of 200 sentences.

6.3. Staggered Training By Entity Level Approach

The HER baseline using the sentence-based approach was also tested without using the nested entity composite label encoding. CRF models were trained separately on Higher-Level, Middle-Level, and Lower-Level entities (refer to Table 1). This staggered training strategy reflects the hierarchical structure of the annotation scheme and the varying semantic complexity associated with each entity group. It also mirrors realistic annotation workflows in archival contexts, where coarse-grained categories are typically established before finer distinctions are introduced.

6.4. Evaluation Metrics

Evaluation focuses on metrics that are meaningful for pre-annotation quality. Token-level accuracy reflects overall tagging correctness but is heavily influenced by non-entity tokens because of class imbalance. Macro-span F1 is computed over exact entity spans, which better captures boundary detection and nested entity correctness. All results are reported using span-level precision, recall, and macro-averaged F1.

7. Results

This section reports the results of three pre-annotation experiments designed to assess (i) the effectiveness of sentence-based pre-annotation as implemented in HER, (ii) the impact of token-budgeted pre-annotation combined with ACL, and (iii) the effect of staggered annotation by entity level.

7.1. Sentence-Based Pre-Annotation (HER Baseline)

We use HER's standard sentence-based active-learning loop as the baseline, adding 200 manually annotated sentences per iteration. Note that this baseline mirrors the HER toolkit's practical annotation loop (fixed-size sentence batches) rather than the token-based batch schedule used in Erdmann et al. (2019) to standardise learning-curve evaluation. The best baseline performance resulted in a macro span-F1 of 0.6675 with feature configuration including (`wordShape`, `charNgrams`, `prevWord`, `prevBiWord`, `prevWordShape`, `contextPosition`, `gazetteers`). As shown in Table 2, macro span-F1 fluctuates between iterations, but improves overall relative to the initial seed, suggesting that additional entity annotations generally increase entity coverage. A decrease in macro-span F1 is observed in the final iteration, despite additional annotation effort. This highlights the limitations of sentence-level sampling under highly skewed sentence-length distributions and motivates the token-budgeted approach introduced next. While sentence-based pre-annotation improves overall entity coverage, annotation costs remain unpredictable due to extreme variation in sentence length.

Iteration	Seed Sentences	Macro span-F1
1	200	0.6288
2	400	0.6090
3	600	0.6675
4	800	0.6412

Table 2: Results of the sentence-based pre-annotation using HER.

7.2. Token-Budgeted Pre-Annotation With Active Curriculum Learning

To address manual annotation cost imbalance, we introduced a token-budgeted condition in which annotation batches contained approximately 2,000 tokens rather than a fixed number of sentences. In the first iteration, an ACL strategy was applied by sampling sentences proportionally across sentence-length bins to ensure representative exposure to both short and long sentences. Figure 2 shows the macro-averaged span-level F1 across iterations for sentence-based pre-annotation and token-budgeted pre-annotation with ACL. Each point in Figure 2 is obtained by three-fold cross-validation on the labelled seed available at that iteration. The figure, therefore, reflects learning progress under increasing labelled data rather than performance on a single fixed held-out test set. Token-based sampling exhibits a smoother learning trajectory across iterations compared to sentence-based sampling. The two conditions are compared as alternative annotation-budget regimes rather than as matched batch sizes: the sentence-based baseline adds 200 sentences per round, which in our runs corresponded to 6,076–13,780 newly annotated tokens, whereas the token-budgeted condition remained close to the intended budget of approximately 2,000 tokens per round.

As shown in Table 3, token-budgeted sampling maintains a consistent annotation workload and reaches macro span-F1 values that are comparable to the sentence-based baseline, while exhibiting a smoother learning trajectory across iterations. Performance peaks at iteration 5 (macro span-F1 = 0.6377) with the feature set (`wordShape`, `charN-grams`, `prevWord`, `prevBiWord`, `gazetteers`). The curriculum-constrained first iteration accelerates early gains relative to random sentence selection, particularly for entities occurring in longer and syntactically complex sentences. Overall, these results demonstrate that controlling annotation cost at the token level leads to smoother learning behaviour while preserving, and in some cases improving, entity coverage.

Full per-entity precision, recall, and F1 scores for the best-performing iteration of each experiment are provided in Appendix 10.

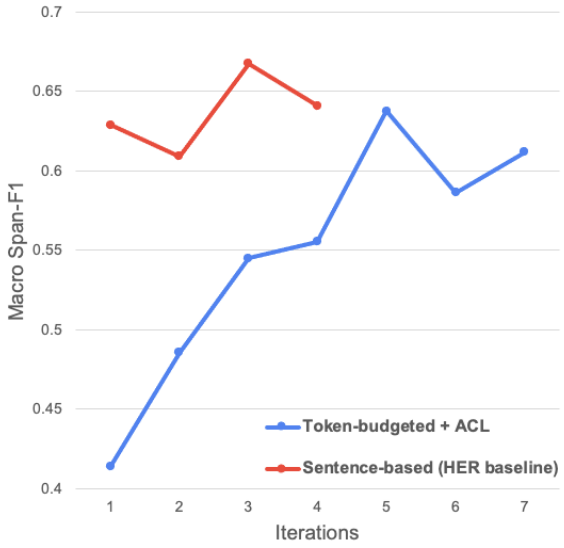

Figure 2: Macro span-F1 across iterations for sentence-based HER and token-budgeted HER with ACL-initialised seed selection. Each point is computed by three-fold cross-validation on the labelled seed available at that iteration.

Iteration	Seed Tokens	Macro span-F1
1	1,996	0.4137
2	4,047	0.4856
3	6,084	0.5449
4	8,231	0.5555
5	10,254	0.6377
6	12,274	0.5864
7	14,365	0.6119

Table 3: Results of token-budgeted pre-annotation with Active Curriculum Learning.

7.3. Annotation Complexity Analysis

To investigate whether token-budgeted sampling implicitly reduces annotation complexity by favouring simpler samples, we analyse the distribution of nested entity depths across annotation strategies. We count the number of annotated entity spans involving single entities, two nested entities, and three nested entities within each annotation batch.

Table 4 reports the average number of annotations by nesting depth across runs for sentence-based sampling and token-budgeted sampling with ACL. While sentence-based sampling results in substantially larger and more variable annotation batches in terms of token count, the relative distribution of nested entity depths is comparable across strategies. These findings suggest that controlling annotation batch size at the token level improves predictability of annotation workload without reducing semantic complexity.

Strategy	Single	2-Nested	3-Nested	Tokens / Iter.
Sentence-based	870 (8.46%)	685.8 (6.67%)	152.8 (1.49%)	10,280.5
Token-based	197.3 (9.78%)	181.1 (8.97%)	35.6 (1.76%)	2,018

Table 4: Average extracted entities and annotation cost per iteration.

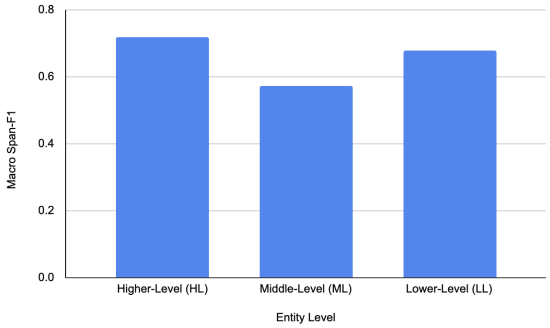

Figure 3: Macro span-F1 by Entity Level (Staggered Annotation)

7.4. Staggered Annotation By Entity Level

To analyse the impact of semantic granularity and nesting complexity, additional experiments were conducted in which models were trained separately on Higher-Level, Middle-Level, and Lower-Level entities. As shown in figure 3, higher-level entities achieve the highest and most stable performance, benefiting from higher frequency and clearer context. Middle-level entities show moderate performance, reflecting the difficulty of distinguishing legal roles such as parties and witnesses. Lower-level entities show some performance relative to middle-level entities, but remain challenging due to their fine-grained semantics, frequent abbreviations, and dependence on nested structures. Performance decreases with increasing semantic granularity, highlighting the challenges posed by fine-grained nested entities in historical notarial documents.

7.5. Entity-Level Error Analysis

The staggered annotation experiments provide the clearest view of the classification of labels by the model. We perform an entity-level error analysis using aggregated confusion matrices across iterations and summarise the distribution of error types in Figure 4. The figure shows that the dominant error mode at all levels is missed entity tokens, indicating that the remaining errors are primarily recall-driven. At the Higher-level, the most frequently misclassified labels are KEY and PER, which are most often missed as O. 10.2% are labels that have their BIO boundary flipped. Errors in rarer categories (e.g., INS and DEEDT) remain consistent with data sparsity. Middle-level entities, PERP

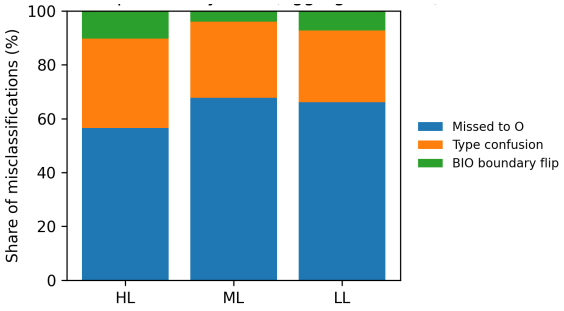

Figure 4: Error composition by level aggregated across iterations

and PERW, consistently obtain the lowest F1 scores and are erroneously classified as O. These entities encode legal roles that are not lexically marked but inferred from their position in the deeds and their formulaic context, making them difficult to capture using local sequence features alone. At the Lower-level, fine-grained components are again predominantly missed (e.g., PERFN−B→O = 522, PERLN−B→O = 336, GEOB−I→O = 201), suggesting that these classes require more annotations and are sensitive to boundary variation. In addition, Lower-level exhibits clear semantic-adjacency confusions, most notably between closely related date subtypes (DATEANIN−I→DATEIN−I = 105 and DATEIN−I→DATEANIN−I = 49). Overall, the staggered setting confirms that increased semantic granularity amplifies recall limitations and subtype ambiguity, with ML role labels (PERP/PERW) emerging as the most consistently misclassified and most likely to benefit from additional targeted annotation and document-level cues.

8. Conclusion

This paper investigated cost-aware pre-annotation strategies for nested NER in seventeenth-century Latin notarial deeds from the Maltese NRA. Extending the HER to support nested entities, we compared sentence-based pre-annotation as a baseline, token-budgeted pre-annotation with ACL, and staggered annotation by entity level. The results show that sentence-based sampling leads to unstable learning behaviour under highly skewed sentence-length distributions, whereas token-budgeted sampling provides more predictable annotation cost and more stable gains in span-level F1.

Staggered annotation experiments reveal differences across semantic levels: Higher-level entities are recognised with the highest accuracy, middle-level role-based entities remain challenging due to their dependence on discourse structure, and lower-level structural entities benefit from regular

patterns and gazetteer support despite nesting. An entity-level error analysis indicates that remaining errors are driven primarily by semantic granularity and document-level dependencies rather than boundary detection failures. Overall, the findings demonstrate that cost-aware pre-annotation can produce high-quality silver-standard annotations and support more efficient annotation planning, laying foundations for scalable knowledge extraction workflows for complex historical archival corpora.

Future work will extend token-budgeted sampling with additional iterations by adding unannotated deeds from a different notary and a different decade of the seventeenth century, testing whether gains in macro span-F1 persist under temporal and scribal shift. We will furthermore track dataset provenance during ranking to quantify which sources are sampled across iterations and how this affects learning stability and label performance. In addition, future work will also investigate encoder-based transformer models for Latin, such as LatinBERT (Bamman and Burns, 2020), to evaluate whether contextual representations can improve nested entity recognition beyond the lightweight CRF-based pre-annotation setting adopted here.

9. Acknowledgments

This research is contributing to the NotaryPedia project, a collaboration between the Notarial Archives Foundation and the Department of Artificial Intelligence at the University of Malta, funded by the Ministry for the National Heritage, the Arts and Local Government. The project aims to bridge traditional archival standards with AI-assisted content analysis, demonstrating how archival collections can be transformed into dynamic, interconnected knowledge infrastructures.

10. Bibliographical References

Sergio Torres Aguilar and Dominique Stutzmann. 2021. Named Entity Recognition for French medieval charters. In *NLP4DH*.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology.

Felix Bensmann, Benjamin Zapilko, and Philipp Mayr. 2017. Interlinking Large-scale Library Data with Authority Records. *Frontiers in Digital Humanities*, Volume 4 - 2017.

Christian Bentz, Douwe Kiela, Felix Hill, and Paula Buttery. 2014. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10:175 – 211.

Emanuela Boroş, Verónica Romero, Martin Maarand, Kateřina Zenklová, Jitka Křečková, Enrique Vidal, Dominique Stutzmann, and Christopher Kermorvant. 2020. A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 79–84.

Emanuel Buttigieg and Joan Abela. 2020. NAV: a survey of the past, present, and future of the Notarial Archives of Valletta, Malta. *Nuovi Annali della Scuola Speciale per Archivisti e Bibliotecari*, 34:5–26.

Pierre Chastang, Xavier Tannier, and Sergio Aguilar. 2021. A Named Entity Recognition Model for Medieval Latin Charters. *Digital Humanities Quarterly*, 15.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Charlene Ellul, Joel Azzopardi, and Charlie Abela. 2019. NotaryPedia: A Knowledge Graph of Historical Notarial Manuscripts. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, pages 626–645, Cham. Springer International Publishing.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics.

S. Fiorini. 1996. *Documentary Sources of Maltese History Series*. University of Malta Press, Malta.

Joseph Fisher and Andreas Vlachos. 2019. Merge and Label: A Novel Neural Network Architecture for Nested NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5840–5850, Florence, Italy. Association for Computational Linguistics.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active Curriculum Learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, Online. Association for Computational Linguistics.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study.

Inês Koch, Carla Teixeira Lopes, and Cristina Ribeiro. 2023. Moving from ISAD(G) to a CIDOC CRM-based Linked Data Model in the Portuguese Archives. *J. Comput. Cult. Herit.*, 16(4).

Raphaëlle Lapôtre. 2017. Library Metadata on the web: the example of data.bnf.fr. *JLIS.it*, 8(3):58–70.

Edward Loss, Fabiana Guernaccini, and Manuel Carassai. 2025. From Manuscript to Metadata: experiments on Handwritten Text Recognition, Tagging and Importation for the Memoriali series (1265-1452). *JLIS.it*, 16(2):59–85.

Wei Lu and Dan Roth. 2015. Joint Mention Extraction and Classification with Mention Hypergraphs. In *Conference on Empirical Methods in Natural Language Processing*.

Vit Novotny, Kristina Luger, Michal Štefánik, Tereza Vrabcova, and Ales Horak. 2023. People and Places of Historical Europe: Bootstrapping Annotation Pipeline and a New Corpus of Named Entities in Late Medieval Texts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14104–14113, Toronto, Canada. Association for Computational Linguistics.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Burr Settles. 2009. Active Learning Literature Survey. Technical Report TR1648, University of Wisconsin–Madison Department of Computer Sciences.

Burr Settles, Mark W. Craven, and Lewis A. Friedland. 2008. Active Learning with Real Annotation Costs.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Takashi Shibuya and Eduard Hovy. 2020. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Sergio Octavio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters using Stacked Embeddings and BERT-based Models. In *Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA, 2022)*.

G. Wettinger. 2000. *Place-Names of the Maltese Islands, ca. 1300–1800*. Publishers Enterprises Group, Malta.

Appendix

Per-Entity Evaluation Metrics for Main Pre-Annotation Strategies

The following tables report pooled per-entity precision, recall, and F1 scores across the three cross-validation folds for the two main pre-annotation strategies evaluated in this work: the sentence-based HER baseline and the token-budgeted approach with Active Curriculum Learning. These detailed metrics are included for completeness and reproducibility. Results for the HL, ML, and LL staggered models are summarised in Figure 3 in the main text, as these experiments serve as auxiliary analysis rather than primary comparisons.

Entity	Precision	Recall	F1	Support
DATE	0.825	0.786	0.805	84
DATEAN	0.000	0.000	0.000	1
DATEANIN	0.778	0.636	0.700	11
DATEIN	0.779	0.767	0.773	60
DEEDT	0.811	0.577	0.674	52
GEO	0.837	0.788	0.812	421
GEOB	0.274	0.103	0.150	29
GEOPC	0.655	0.776	0.710	49
INS	0.357	0.435	0.392	23
KEY	0.827	0.767	0.796	1676
PER	0.874	0.815	0.843	1224
PERFN	0.923	0.870	0.896	1123
PERLN	0.956	0.880	0.916	785
PERP	0.552	0.408	0.469	103
PERW	0.474	0.394	0.430	137

Table 5: Pooled precision, recall, and F1 across the three folds for the best sentence-based HER baseline iteration (Iteration 3 with macro span-F1 = 0.6675).

Entity	Precision	Recall	F1	Support
DATE	0.781	0.735	0.758	34
DATEAN	0.000	0.000	0.000	1
DATEANIN	0.500	0.625	0.556	8
DATEIN	0.500	0.625	0.555	16
DEEDT	1.000	0.400	0.571	5
GEO	0.864	0.827	0.845	185
GEOB	0.889	0.533	0.667	15
GEOPC	0.586	0.629	0.607	27
INS	0.667	0.222	0.333	9
KEY	0.833	0.682	0.750	547
PER	0.830	0.729	0.776	483
PERFN	0.896	0.798	0.844	445
PERLN	0.957	0.805	0.874	328
PERP	0.000	0.000	0.000	15
PERW	0.547	0.460	0.500	76

Table 6: Pooled precision, recall, and F1 across the three folds for the best token-budgeted + ACL iteration (Iteration 5 with macro span-F1 = 0.6377).