A Comparative Study of Static and Contextual Embeddings for Analyzing Semantic Changes in Medieval Latin Charters

Yifan Liu[1]*, Gelila Tilahun[2], Xinxiang Gao[1], Qianfeng Wen[1], Michael Gervers[3]
[1] Department of Computer Science, University of Toronto
[2] DEEDS Project, University of Toronto
[3] Department of Historical and Cultural Studies, University of Toronto Scarborough
yifanliu.liu@mail.utoronto.ca

Abstract

The Norman Conquest of 1066 C.E. brought profound transformations to England's administrative, societal, and linguistic practices. The DEEDS (Documents of Early England Data Set) database offers a unique opportunity to explore these changes by examining shifts in word meanings within a vast collection of Medieval Latin charters. While computational linguistics typically relies on vector representations of words like static and contextual embeddings to analyze semantic changes, existing embeddings for scarce and historical Medieval Latin are limited and may not be well-suited for this task. This paper presents the first computational analysis of semantic change pre- and post-Norman Conquest and the first systematic comparison of static and contextual embeddings in a scarce historical data set. Our findings confirm that, consistent with existing studies, contextual embeddings outperform static word embeddings in capturing semantic change within a scarce historical corpus.

1 Introduction

The Norman Conquest of 1066 is a pivotal event in English history, marked by the introduction of new administrative and cultural practices by the Normans. This transformation is evident in the Medieval Latin charters — official documents recording legal agreements, grants, rights, and privileges — preserved in the DEEDS (Documents of Early England Data Set) corpus (Gervers et al., 2018). One implication of these transformations is the shift in language usage and word meanings within the Medieval Latin charters, illustrated by the following examples: *comes* generally meant "official" in Anglo-Saxon charters, but in Norman documents, it consistently appeared as a title meaning "earl" or "count"; *proprius* ("one's own") was used by the Anglo-Saxons to indicate signing a document "with one's own hand," whereas the Normans used it to refer to property ownership. Investigating these changes in word meanings before and after the Norman Conquest — a process known as lexical semantic change (LSC) — provides insights into the cultural and societal transformations while also posing challenging research questions on how to systematically model this change.

In the field of computational linguistics, various methods have been proposed for modeling lexical semantics and thereby for studying semantic changes. In earlier years, static word embedding approaches, where each word was mapped to a fixed vector representation based on its co-occurrence patterns with other words within a corpus (Mikolov et al., 2013; Bojanowski et al., 2017), were dominant and proven effective in LSC studies (Kim et al., 2014; Hamilton et al., 2016). In more recent years, contextual representations, which provide different vectors for the different contexts in which a word appears (Devlin et al., 2019; Peters et al., 2018), have achieved state-of-the-art performance in LSC studies, likely due to their ability to handle phenomena like polysemy and homonymy more effectively than static representations (Martinc et al., 2019; Giulianelli, 2019; Kutuzov et al., 2022).

Despite the successes of contextual embeddings in LSC research, they are typically trained on large corpora (Davies, 2010; Michel et al., 2011) and require significantly more training data than static embeddings due to their more complex architectures and larger parameter sizes (Bommasani et al., 2021). This poses a challenge for studies involving smaller data sets such as the DEEDS Medieval Latin corpus, which contains only 17k charters and 3M tokens — considerably smaller than the billion-token corpora typically used to train contextual embeddings (Davies, 2010; Michel et al., 2011). Meanwhile, the Medieval Latin charters contain a rich and expansive vocabulary, including local dialects and borrowings from other languages (e.g.,

*Corresponding author

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 182–192
January 20, 2025. ©2025 Association for Computational Linguistics

the Anglo-Saxon manuscripts include an extensive amount of Old English). These factors collectively raise concerns about the adaptability and relative performance of existing embedding methods in this scarce and heteroglossic data set.

Therefore, this paper aims to address the research gap in Medieval Latin charters with the following contributions:

- We present the **first LSC study** on **Medieval Latin charters** from England to understand the semantic change induced by the Norman conquest. These English Latin charters are exclusively a collection of legal documents pertaining to property rights whose topic and genre are quite different from other medieval Latin corpora described in section 2.3.
- We provide a **systematic comparison** between static embeddings and contextual embeddings in modeling semantic change within Medieval Latin charters, which offers insights into the adaptability of these models within the context of a scarce and heteroglossic corpus.

The rest of this paper is organized as follows. Section 2 summarizes the previous literature on static and contextual embeddings. Section 3 provides a detailed introduction to the DEEDS data set. Section 4 outlines the training process for the different embedding methods on this corpus.[1] Sections 5 and 6 present the experiments, results, and discussions related to evaluating these embedding methods in capturing semantic change.

2 Related Work

The standard computational approach for lexical semantic change (LSC) analysis involves separately training embeddings for different periods within a corpus (Gulordava and Baroni, 2011), and then measuring the distance between the representations of a given word across these periods. In this section, we review the current approaches of semantic change analysis using static and contextual embeddings and their applications to Medieval Latin corpora.

2.1 Static Word Embeddings

Early methods for word embeddings relied on co-occurrence count-based techniques (Deerwester et al., 1990; Turney and Pantel, 2010). With the rise of deep neural networks, prediction-based models became more popular. These include the **Continuous Bag-of-Words** model (Mikolov et al., 2013), which encodes contextual information by predicting target words from their surrounding context; the **Continuous Skip-gram** model (Mikolov et al., 2013), which predicts surrounding words based on the target word; and the **Subword model** (Bojanowski et al., 2017), which improves these approaches by learning context vectors through subword tokenization.

The integration of these prediction-based embeddings into LSC studies began with Kim et al. (2014). Building on this, Hamilton et al. (2016) showed that neural-based diachronic embeddings outperform traditional count-based methods. Subsequent research further enhanced these techniques by incorporating subword models to improve representation quality, particularly for low-resource and morphologically rich languages (Xu et al., 2019; Xu and Zhang, 2021).

In LSC, aligning embedding spaces across periods is important for meaningful semantic change analysis. One effective strategy is weight initialization, where word embeddings share initial training weights across periods. Kim et al. (2014) introduced **incremental initialization**, initializing each year's weights with the previous year's vectors. For scarce corpora, Montariol and Allauzen (2019) proposed **internal initialization**, which trains a base model on the entire corpus before fine-tuning for each period, and **backward external initialization**, which starts with pre-trained embeddings for the last period and trains in reverse. These strategies align embeddings across periods and address data scarcity, making them suitable for Medieval Latin charters.

2.2 Contextual Embeddings

Unlike static word embeddings, which provide a single fixed vector for each word, contextual embeddings generate unique representations for each word usage based on its context. **BERT** (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is a leading example of such models. Early studies, including Hu et al. (2019), Giulianelli (2019), and Martinc et al. (2019), applied contextual embeddings to lexical semantic change (LSC). For instance, Martinc et al. (2019) fine-tuned a pre-trained BERT model on another corpus and aggregated embeddings to represent all

[1] Corpus and codes available at: `https://anonymous.4open.science/r/historical-text-embedding-C328/README.md`

instances of a word within a time-slice subcorpus. Contextual embeddings have since demonstrated strong performance in LSC tasks across languages such as English, German, and others (Kanjirangat et al., 2020; Rodina et al., 2021; Montariol and Allauzen, 2021; Kurtyigit et al., 2021; Kutuzov et al., 2022).

However, most contextual representations are trained on large, modern corpora, leaving historical corpora underexplored. Addressing this gap, Qiu and Xu (2022) introduced histBERT, a BERT model adapted to historical American English (COHA), which outperformed the standard BERT in detecting semantic changes in historical texts. Another approach is training BERT models from scratch for historical data. Manjavacas Arevalo and Fonteyn (2021) developed MacBERTh, trained on historical English from 1450–1900, showing better results than adaptation-based methods. Similarly, Beck and Köllner (2023) extended this approach to German with GHisBERT, trained on texts dating back to 750 C.E. These methods not only align contextual embeddings with historical data but also provide valuable insights for developing embeddings suited to Medieval Latin, a scarce and historical language.

2.3 Towards Medieval Latin Embeddings

Training word embeddings for Medieval Latin presents unique challenges due to a limited size of training corpora when compared to contemporary and modern languages. Several efforts have been made to construct Medieval Latin corpora to improve embedding training. Notable examples include the Dictionary of Medieval Latin from British Sources (Latham et al., 1975), which documents the Latin vocabulary used in Britain from 540 to 1600 C.E; Index Thomisticus, a digital corpus of Thomas Aquinas's 13th-century works (Busa, 1973); the Polish Medieval Latin Lexicon (Plezia and Weyssenhoff-Brożkowa, 1992), covering the 10th to mid-15th centuries; and the Frankfurt Latin Lexicon (Mehler et al., 2020), spanning the 6th to 9th centuries. These efforts have facilitated the development of high-quality static Latin word embeddings using CBOW, Skip-gram, and subword models. However, the topics and genres on which they focus differ from the DEEDS corpus in that DEEDS corpus is a collection of legal charters which primarily focuses on the rights of ownership and transfer of properties within Anglo-Saxon and Norman periods, which are critical sources for understanding impacts of the Norman conquest.

Contextual embeddings are believed to require even larger corpora, making their training on Medieval Latin languages more challenging than static embeddings. Although no contextual embeddings have been directly trained on Medieval Latin, some works have focused on Latin more broadly: Devlin et al. (2019) introduced Multilingual BERT, trained on the Wikipedia corpus for over 100 languages, including Latin; Bamman and Burns (2020) trained a BERT model specifically for Latin on a vast corpus of 600M tokens spanning from 200 B.C.E. to the present; Luis A. Vasquez trained a Latin BERT model on the Classical Language Toolkit (CLTK) corpus.[2]

The historical language change of Latin has long attracted scholarly interest, and with the development of Latin corpora and word embeddings, researchers can now understand these changes computationally. For example, Sprugnoli et al. (2020) analyzed Latin language change between the Classical and Medieval/Christian eras and evaluated different Latin embeddings on this task; Ribary and McGillivray (2020) detected semantic split in words with general and legal meanings by building Latin word embeddings from a 6th-century Roman law sourcebook; and SemEval 2020 (Schlechtweg et al., 2020) included a task to calculate semantic change between the pre-Christian and Christian eras, using carefully annotated data from the LatinISE corpus (McGillivray and Kilgarriff, 2013).

However, significant research gaps still remain in the analysis of semantic change in Medieval Latin. First, there has been no computational evaluation of semantic change in the context of the Norman Conquest, a period marked by substantial administrative, cultural, and linguistic shifts (Gervers et al., 2018). Second, although contextual embeddings have proven more powerful than static embeddings in large contemporary corpora, there is a lack of contextual embeddings specifically trained on scarce and historical Medieval Latin corpora, so a systematic comparison between these approaches is still needed.

3 Data

For our analysis, we used Medieval Latin charters from DEEDS (Documents of Early England Data

[2] https://huggingface.co/LuisAVasquez/simple-latin-bert-uncased

Set).[3] The DEEDS database contains transcripts of over 70K Latin charters from the 7th to the 14th century. Of these, 40K pertain to England, and 17k are dated. They are official documents issued by kings and commoners and deal with the transfer of property and property rights.

In this study, we focused on the 17k dated charters, as the dates were essential for splitting the corpus for semantic change analysis. We split the corpus into three sets: the Anglo-Saxon period (from 589 to 1066 CE), referred to as **ANG** in later sections; the Norman period (from 1066 to 1153 CE), referred to as **NOR**; the later post-conquest period up to 1272 CE (also called Plantagenet period), referred to as **PLA**. Table 1 provides a summary of the corpus data.

	ANG	**NOR**	**PLA**
Time Span	589-1065	1066-1153	1154-1272
# of Charters	1432	4050	12926
# of Tokens	0.49M	0.76M	2.80M

Table 1: Overview of the Medieval Latin corpus

The main focus of this paper is the semantic change induced by the Norman conquest (i.e., the transition from **ANG** to **NOR** periods, referred to as *AN* in the later section). For comparison, we also examine the transitions from **NOR** to **PLA**, referred to as *NP*.

4 Models

4.1 Static Word Embeddings

We used the Continuous Skip-gram model with subword information (Mikolov et al., 2013; Bojanowski et al., 2017), as implemented in the FastText module in the Gensim library (Řehůřek and Sojka, 2010), to generate static word embeddings for each period. We adopted the incremental initialization from Kim et al. (2014) as well as internal and backward external initialization from Montariol and Allauzen (2019). Due to resource constraints, we only tuned the embedding sizes (100 and 300) and the number of training epochs (10, 30, and 50) for each period and reported the best results.[4] All other hyperparameters were kept at their default settings in the FastText module.

Incremental Initialization: The embeddings from the previous period were used to initialize the embeddings for the subsequent period (incrementally). We refer to this model as `Incremental` in later sections.

Internal Initialization: We trained a base model on the full corpus for 50 epochs, which was then used to initialize the embeddings for the first period, with subsequent period embeddings being updated incrementally. We refer to this model as `Internal` in later sections.

Backward External Initialization: We utilized pre-trained Latin word embeddings from Grave et al. (2019) on Common Crawl and Wikipedia as the base model. Then, we incrementally updated each period's embeddings from the most recent to the oldest, a reverse updating process that might be beneficial to our corpora, which have lower volumes in the older periods (Montariol and Allauzen, 2019). We refer to this model as `External` in later sections.

4.2 Contexual Embeddings

BERT Trained from Scratch: We pre-trained a BERT model from scratch on the full Medieval Latin charters corpus using the hyperparameters recommended by Manjavacas and Fonteyn (2022) in historical English and Beck and Köllner (2023) in historical German. The model consists of 12 hidden layers, each with 768-dimensional embeddings, and 12 attention heads, with a vocabulary size of 32,000 tokens. Training was conducted over 10 epochs with a batch size of 8 using the masked language modeling (MLM) task, where 10% of the tokens were randomly masked. We refer to this model as `MLatin-BERT` in later sections.

BERT Adapted from Pre-trained Models: For comparison, we continued training two Latin BERT models on the Medieval Latin charters corpus: the first, Latin-BERT by Bamman and Burns (2020) [5], which was trained on a diverse range of Latin corpora with 600M tokens spanning from 200 B.C.E. to the present, and the second, `simple-latin-bert-uncased` by Luis A. Vasquez [6], which was trained using corpora from the Classical Language Toolkit (CLTK). Both models were configured with standard BERT hyperparameters with a hidden size of 768 and 12 layers. They were further trained from their last check-

[3] `https://deeds.library.utoronto.ca/content/about-deeds`

[4] See Appendix A for details

[5] `https://github.com/dbamman/latin-bert`

[6] `https://huggingface.co/LuisAVasquez/simple-latin-bert-uncased`

points on the Medieval Latin corpus for an additional 4 epochs, as recommended by the original BERT paper (Devlin et al., 2019). We refer to these models as `Ada-BERT-Bam` and `Ada-BERT-Vas`, respectively, in later sections.

Tokenizer: We pre-trained a tokenizer for all described models, which accounts for the diverse word forms in the Medieval Latin charters. The tokenizer was trained with the same hyperparameter settings outlined by Beck and Köllner (2023) using the HuggingFace `BertWordPieceTokenizer` module with a vocabulary of 32000 and a maximum sequence length of 512.

Extract Word Embeddings: To enable direct comparison between contextual and static embeddings in the semantic change analysis, we followed the method described by Martinc et al. (2019) to extract word embeddings from contextual embeddings for each time period (discussed in Section 3), as detailed in Algorithm 1.

Algorithm 1 Extract and average word embeddings from contextual embeddings for a time period

```
Input: Medieval Latin texts for a given time period, $\mathcal{C} = \{S_1, S_2, \ldots, S_n\}$, where $S_i$ is a sentence. Contexual embeddings $\mathcal{E} = \{\mathbf{E}_{S_1}, \mathbf{E}_{S_2}, \ldots, \mathbf{E}_{S_n}\}$, where $\mathbf{E}_{S_i} \in \mathbb{R}^{L \times d}$ is the embedding matrix for sentence $S_i$.
Output: Word embeddings $\mathbf{W} \in \mathbb{R}^{M \times d}$, where $M$ is the number of distinct words in $\mathcal{C}$.
Initialize word embedding matrix $\mathbf{W}$
for each distinct word $w_j \in \mathcal{C}$ do
    Initialize embedding sets $\mathcal{W}_j = \{\}$
end for
for each sentence $S_i \in \mathcal{C}$ do
    $\mathbf{S}_i \leftarrow \frac{1}{4} \sum_{l=L-3}^{L} \mathbf{E}_i^{(l)}$ {Compute sentence embedding using last four layers}
    for each word $w_j \in S_i$ do
        Identify the word pieces $\mathbf{P}_j$ corresponding to word $w_j$ using offset mappings.
        Compute word embedding: $\mathbf{w}_j^{(S_i)} \leftarrow \frac{1}{|\mathbf{P}_j|} \sum_{p \in \mathbf{P}_j} \mathbf{S}_i^{(p)}$ {Compute word embedding for $w_j$ in sentence context $S_i$}
        Store $\mathbf{w}_j^{(S_i)}$ in set $\mathcal{W}_j$
    end for
end for
for each word $w_j$ in vocabulary do
    $\bar{\mathbf{w}}_j \leftarrow \frac{1}{|\mathcal{W}_j|} \sum_{\mathbf{w}_j^{(S_i)} \in \mathcal{W}_j} \mathbf{w}_j^{(S_i)}$ {Compute average embedding}
    Store $\bar{\mathbf{w}}_j$ in $\mathbf{W}$
end for
return $\mathbf{W}$
```

5 Lexical Similarity Analysis

5.1 Similarity Measures

To evaluate the applicability of different embedding models in analyzing semantic change within the Medieval Latin charters, we conducted a semantic similarity analysis across various periods following the approach of Beck and Köllner (2023). Specifically, for a given word w occurring in two periods t_1 and t_2, we computed the cosine similarity between their embeddings $\mathbf{w}_{t_1}$ and $\mathbf{w}_{t_2}$ using the following formula:

$$\text{Cos}(\mathbf{w}_{t_1}, \mathbf{w}_{t_2}) = \frac{\mathbf{w}_{t_1} \cdot \mathbf{w}_{t_2}}{\|\mathbf{w}_{t_1}\| \|\mathbf{w}_{t_2}\|} \tag{1}$$

A lower cosine similarity score between periods suggests a potential semantic shift in the word's meaning (Kim et al., 2014; Giulianelli, 2019).

In our analysis, we divided the data into three periods, as outlined in Section 3, and therefore, for each word, we computed two cosine similarity measures: COS$_{AN}$, representing the transition from **ANG** to **NOR** and COS$_{NP}$, representing the transition from **NOR** to **PLA**. We will refer to the above labels in later sections.

5.2 Data Set Labeling

To quantitatively assess the performance of different embedding methods, we applied the following labeling procedure to the data set. We selected commonly occurring words with a relative frequency exceeding five occurrences per 100,000 in all periods, resulting in 662 words in total. Three Latin specialists with domain knowledge were asked to make a binary decision on whether the meaning of each word had changed from the Anglo-Saxon to the Norman period (marked as 1) or remained unchanged (marked as 0), which were then used as **semantic change labels** for subsequent studies. For each period, the labelers made their decisions on a word by reviewing 10 sample sentences containing the word. If all three labelers agreed on a label, the word was classified as either *changed* (for positive cases, 41 words) or *unchanged* (for negative cases, 297 words)[7]. Examples of *changed* words include *finis*, which shifted from meaning

[7] The list of changed and unchanged words can be found at: `https://anonymous.4open.science/r/historical-text-embedding-C328/README.md`

		Static			Contextual		
		Incremental	Internal	External	MLatin-BERT	Ada-BERT-Bam	Ada-BERT-Vas
AN	δ_μ	0.054*	-0.004	0.002	0.047*	0.037*	0.055*
	ρ	-0.169*	0.018	-0.120	-0.481*	-0.395*	-0.360*
NP	δ_μ	0.011	-0.015	-0.003	0.009	0.006	0.012
	ρ	-0.003	0.055	-0.072	-0.135*	-0.126*	-0.141*

Table 2: Quantitative results of static and contextual embeddings in semantic for the *AN* and *NP* periods. Two metrics are reported: δ_μ indicates the difference in mean cosine similarity between the *unchanged* and *changed* word groups, and ρ represents the correlation between semantic change labels and cosine similarity measures for each target word across two periods. An asterisk (*) denotes statistically significant results (t-test, $p < 0.01$).

"end" or "completion" in Anglo-Saxon times to "fine" as a payment in a final agreement in Norman, and *honorifice*, which originally meant "honorable" or "honorably" in the context of a king's duties, but in Norman documents referred specifically to the manner in which land was held by a feudal lord. Examples of *unchanged* words include pronouns (e.g., *meus*, "my"), numbers (e.g., *centum*, "hundred"), greetings (e.g., *salute*, "hello"), and prepositions (e.g., *post*, "after"; *usque*, "until"). In cases where no consensus was reached, the words were excluded from both categories. Our analysis focused solely on the 338 target words that were clearly categorized as either *changed* or *unchanged*.

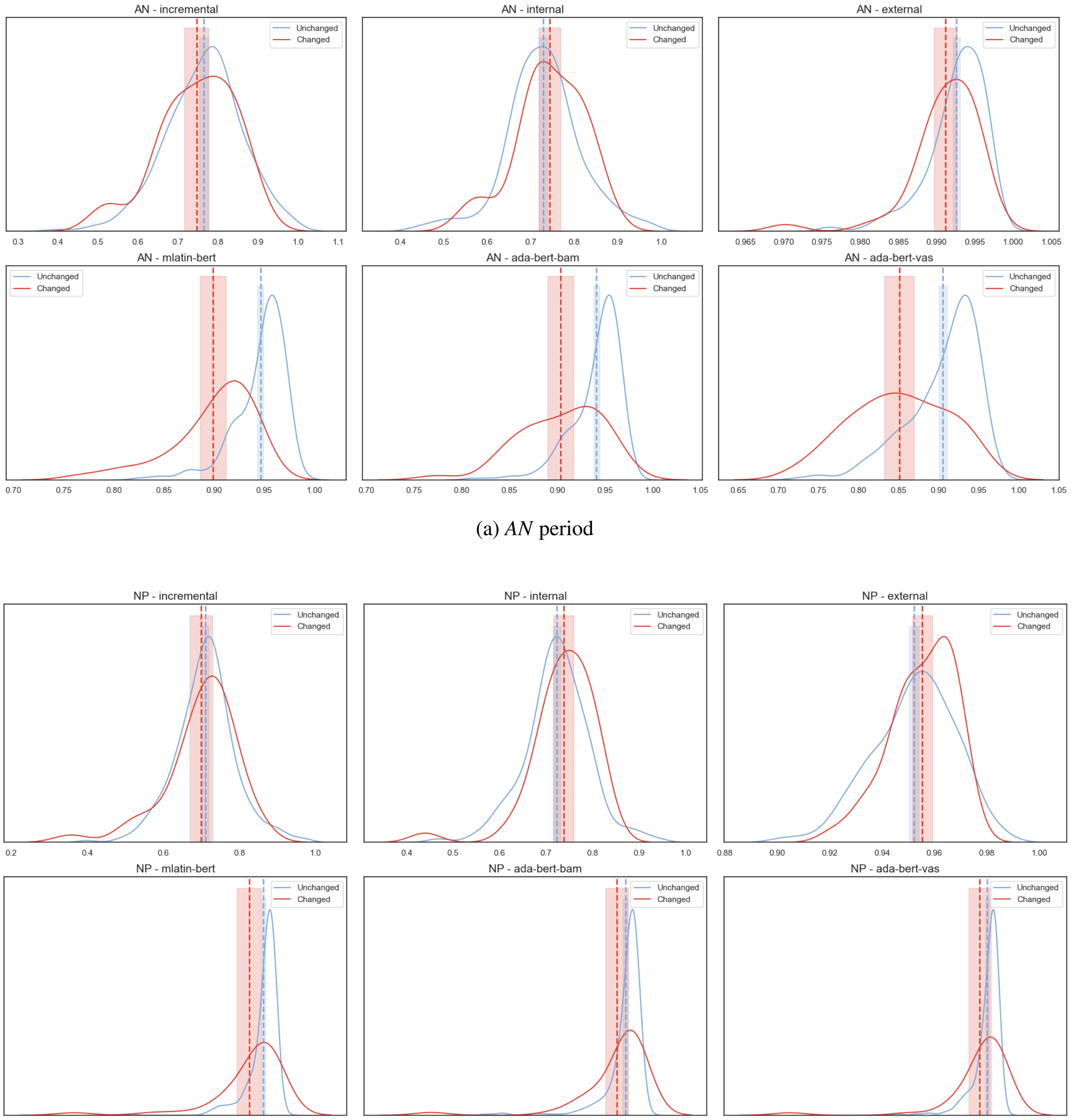
6 Results

6.1 Semantic Change in *AN* Period

Given our primary focus on the semantic changes induced by the Norman Conquest, we first present the results of COS_{AN} (i.e., the cosine similarity between the embeddings from the Anglo-Saxon and Norman periods for a given word) across different embedding models (as discussed in Section 4). The *AN* section of Table 2 reported two performance metrics: the difference in the averages of the COS_{AN} between *unchanged* and *changed* words (as discussed in Section 4), δ_μ, where a larger difference indicates a better ability to distinguish between the two groups; the Pearson correlation, ρ, between the binary change labels and COS_{AN} for all target words, with values ranging from -1 (strong negative correlation, the most desirable outcome) to 1 (strong positive correlation, the least desirable outcome). All contextual embeddings demonstrated statistically significant δ_μ values. The correlation coefficient further highlighted the better performance of contextual embeddings in semantic change analysis, with MLatin-BERT achieving the strongest negative correlation ($\rho = -0.481$) and outperforming models adapted from pre-trained Latin BERT. Among the static embedding methods, Incremental and External showed fair results, with the correct direction of δ_μ and a moderate negative correlation between true semantic change labels and cosine similarity, although the correlation was much weaker than that of the contextual models. In contrast, Internal produced results opposite to those expected.

Figure 1a displays a more detailed distributions, mean values, and 95% confidence intervals of COS_{AN} for both the *changed* and *unchanged* word groups. Contextual embeddings consistently showed an obvious difference between the distributions of *changed* and *unchanged* words, with *changed* words centering around much lower cosine similarity scores. Notably, Ada-BERT-Vas produced lower similarity for both word groups compared to MLatin-BERT and Ada-BERT-Bam. The results for static embeddings reveal several concerns: while Incremental identified the correct difference in mean values (with the mean cosine similarity being smaller for the *changed* word group), it did not show a significant difference in the distribution shapes between the two word groups. The External model exhibited a difference in distribution, but the absolute difference in mean cosine similarity was marginal (only around 0.002). The Internal approach produced completely opposite to the expected results.

Overall, these results suggest that contextual embeddings are more effective at capturing semantic changes and distinguishing *changed* words from *unchanged* words, even in a scarce and historical language setting, which demonstrates the adaptabil-



(a) *AN* period

(b) *NP* period

Figure 1: Distribution of cosine similarity for *changed* and *unchanged* words across different embedding models – *AN* period (top) and *NP* period (bottom). The dashed lines represent the mean cosine similarity for *changed* and *unchanged* words across the two periods and for each model. The shaded areas represent the 95% confidence intervals.

ity of contextual embeddings to smaller data sets beyond what has been shown in existing literature. Additionally, we found that both static and contextual models trained from scratch (`Incremental` and `MLatin-BERT`) performed better than those adapted from pre-trained embeddings, likely due to the lack of high-quality base representations for Medieval Latin texts.

6.2 Comparison Across Periods

For comparison, we also report the distributions, δ_μ between the *unchanged* and *changed* groups of COS_{NP} (i.e., the cosine similarity between the embeddings from the Norman and Plantagenet periods for a given word), and the correlation ρ between semantic change labels and COS_{NP}. We expect the *AN* period to have a smaller mean value across all words, a larger mean difference between *changed* and *unchanged* words, and a more negative correlation between COS_{NP} and semantic change labels than for *NP* period, based on the assumption that the semantic change from the Anglo-Saxon period to the Norman period is more significant than from Norman to Plantagenet (often seen as a continuation of Norman ruling) due to the profound linguistic, cultural, and sociological shifts triggered by the Norman Conquest (Clanchy, 2012).

The results from Figure 1b indicate that all contextual embeddings find higher distribution center values for both *changed* and *unchanged* words during the *NP* period than *AN* period. Additionally, the *NP* section of Table 2 reveals that LL contextual embeddings identify significantly larger δ_μ and more negative ρ during the *AN* periods. These results suggest that contextual embeddings effectively differentiate periods of dramatic semantic change from relatively stable periods. Among the static embeddings, although the `Incremental` and `External` approaches correctly demonstrate smaller δ_μ and weaker ρ in the *NP* period compared to the *AN* period, they fail to capture the difference in absolute mean cosine similarity, as both models display lower mean cosine similarity across all word groups in the *NP* period than in the *AN* period.

7 Conclusion

This paper represents the first effort to explore semantic changes in the Medieval Latin charters as a result of the Norman Conquest, and the first to systematically implement and compare static and contextual word embeddings in the context of the scarce and historical corpus. Our evaluation on the DEEDS Medieval Latin charters corpus with manually labeled semantic changes demonstrates that contextual embeddings outperform static word embeddings, even on a scarce and complex historical data set. This finding is consistent with results from large contemporary data sets and confirms the adaptability of contextual embeddings to smaller data sets beyond what has been shown in existing literature. Furthermore, consistent with previous work on building contextual embeddings for historical corpora (Manjavacas Arevalo and Fonteyn, 2021; Beck and Köllner, 2023), training from scratch yields better performance in capturing the correlation between semantic change labels and similarity measures.

Limitations

This research opens new avenues for historical linguistics by providing a framework to explore semantic change in Medieval Latin charters and understand the social, cultural, and political impacts of the Norman Conquest. One could utilize the semantic change analysis framework discussed in this paper as a knowledge discovery process to learn previously unrealized shifts in word meaning.

However, this study also faces certain limitations. As an initial exploration of diachronic embeddings in Medieval Latin charters, we lack a gold standard data set for semantic change detection and were only able to construct binary semantic change labels due to resource constraints. Future work could involve collaboration with more Medieval Latin scholars to develop a continuous semantic change index ranging from zero to one, which could allow for more informative and rigorous quantitative evaluations of our models and establish a benchmark for subsequent research in this field. Additionally, this study has primarily used cosine similarity between word embeddings from different periods as the metric for modeling semantic change, which may not be the most appropriate measure. Future research could explore alternative distance-based metrics, such as Average Pairwise Distance (APD) and Inverted Cosine Similarity over Prototypes (PRT), as suggested in previous studies (Giulianelli et al., 2020; Kutuzov et al., 2022).

References

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Christin Beck and Marisa Köllner. 2023. Ghisbert–training bert from scratch for lexical semantic investigations across historical German language stages. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 33–45.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Roberto Busa. 1973. *Sancti Thomae Aquinatis Operum omnium indices et concordantiae*. Fromman-Holzboog.

Michael T Clanchy. 2012. *From memory to written record: England 1066-1307*. John Wiley & Sons.

Mark Davies. 2010. The corpus of historical American English: 400 million words, 1810-2009.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Michael Gervers, Gelila Tilahun, Shima Khoshraftar, and Roderick A Mitchell. 2018. The dating of undated medieval charters. *ARCHIVES: The Journal of the British Records Association*, 53(137):1–33.

Mario Giulianelli. 2019. Lexical semantic change analysis with contextualised word representations. *Unpublished master's thesis, University of Amsterdam, Amsterdam*.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. *arXiv preprint arXiv:2004.14118*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3899–3908.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. Sst-bert at semeval-2020 task 1: Semantic shift tracing by clustering in bert-based embedding spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. *arXiv preprint arXiv:2106.03111*.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized language models for semantic change detection: lessons learned. *arXiv preprint arXiv:2209.00154*.

Ronald Edward Latham et al. 1975. Dictionary of medieval latin from british sources. *(No Title)*.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, (Digital humanities in languages).

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India. NLP Association of India (NLPAI).

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New methods in historical corpus linguistics*, 1(3):247–257.

Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, and Wahed Hemati. 2020. The frankfurt latin lexicon: From morphological expansion and word embeddings to semiographs. *arXiv preprint arXiv:2005.10790*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Syrielle Montariol and Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. *arXiv preprint arXiv:1909.01863*.

Syrielle Montariol and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marian Plezia and Krystyna Weyssenhoff-Brożkowa. 1992. *Słownik łaciny średniowiecznej w Polsce: Lexicon mediae et infimae latinitatis Polonorum*, volume 1. Zakład Narodowy im. Ossolińskich.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis. *arXiv preprint arXiv:2202.03612*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Marton Ribary and Barbara McGillivray. 2020. A corpus approach to roman law based on justinian's digest. In *Informatics*, volume 7, page 44. MDPI.

Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. 2021. Elmo and bert in semantic change detection for russian. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*, pages 175–186. Springer.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. Building and comparing lemma embeddings for Latin. Classical Latin versus Thomas Aquinas. *IJCoL. Italian Journal of Computational Linguistics*, 6(6-1):29–45.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Yang Xu, Jiasheng Zhang, and David Reitter. 2019. Treat the word as a whole or look inside? subword embeddings model language change and typology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 136–145, Florence, Italy. Association for Computational Linguistics.

Yang Xu and Zheng-sheng Zhang. 2021. Historical changes in semantic weights of sub-word units. *Computational approaches to semantic change*, pages 169–187.

A Hyperparameter Experiments for Static Embeddings

This section details the hyperparameter selection for static embeddings. Figure 2 illustrates how the evaluation metrics in *AN* period, δ_μ and ρ (see detailed definitions and significance in Section 6.1), vary across different hyperparameter settings, specifically the number of training epochs (10, 30, and 50) and the embedding size (100 and 300).

For the `Incremental` approach, the best hyperparameters were found when the embedding size was set to 100 and the number of training epochs was 50. A clear trend emerges where an embedding size of 100 outperforms a size of 300. Additionally, with a embedding size of 100, increasing the number of training epochs leads to better results, whereas with a embedding size of 300, fewer training epochs yield better outcomes.

In the `External` approach, the optimal hyperparameters were identified when the embedding size was 100 and the training epochs were set to 10. There is a trend indicating that smaller embedding sizes and fewer training epochs produce better results for this approach.

For the `Internal` approach, the best performance was observed when the embedding size was 300 and the number of training epochs was 10. However, the results do not exhibit a consistent trend across different hyperparameter settings and embedding sizes.

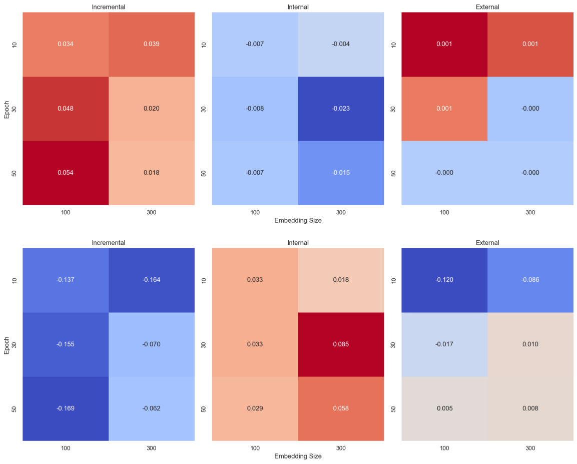


Figure 2: Heatmaps showing the evaluation metrics varying across different hyperparameter settings, with δ_μ (top) and ρ (bottom).

B Effect of Model Size on Contextual Embeddings

In this section, we examine how the size of a BERT model trained from scratch affects performance during the *AN* period. In addition to the `MLatin-BERT` model, we trained two smaller models: a small BERT model (4 attention heads, 4 hidden layers, and an embedding size of 256) and a medium BERT model (8 attention heads, 8 hidden layers, and an embedding size of 512), both of which are smaller than `MLatin-BERT`.[8] As shown in Table 3, there is a clear trend where larger model sizes result in better performance, evidenced by the greater differences in mean cosine similarity and stronger correlations between the semantic change labels and cosine similarity for larger models. These findings are consistent with established scaling laws (Kaplan et al., 2020).

	`Small`	`Medium`	`MLatin-BERT`
δ_μ	0.012	0.028	0.047
ρ	-0.250	-0.327	-0.481

Table 3: Evaluation metrics (δ_μ and ρ) across different model sizes: Small, Medium, and Large (`MLatin-BERT`).

C Effect of Adaption on Contextual Embeddings

In this section, we examine how adapting a pre-trained BERT model to Medieval Latin charters affects performance. We replicate the study for the *AN* period using Latin-BERT (Bamman and Burns, 2020). Table 4 shows that domain adaptation of the pre-trained Latin BERT model to Medieval Latin charters enhances its ability to identify semantic change, as evidenced by the greater difference in mean cosine similarity and the stronger correlation between the semantic change labels and cosine similarity observed in the `Ada-BERT-Bam` model.

	`Latin-BERT-Bam`	`Ada-BERT-Bam`
δ_μ	0.020	0.037
ρ	-0.326	-0.395

Table 4: Evaluation metrics (δ_μ and ρ) for Bamman and Burns (2020)'s Latin BERT (`Latin-BERT-Bam`) and the adapted version (`Ada-BERT-Bam`).

[8] Future work could explore larger BERT models, which we did not pursue due to resource constraints.