# How to age BERT Well: Continuous Training for Historical Language Adaptation

**Anika Harju** **Rob van der Goot**
IT University of Copenhagen, Denmark
{aniha, robv}@itu.dk

## Abstract

As the application of computational tools increases to digitalize historical archives, automatic annotation challenges persist due to distinct linguistic and morphological features of historical languages like Old English (OE). Existing tools struggle with the historical language varieties due to insufficient training. Previous research has focused on adapting pretrained language models to new languages or domains but has rarely explored the modeling of language variety across time. Hence, we investigate the effectiveness of continuous language model training for adapting language models to OE on domain-specific data. We compare the continuous training of an English model (EN) and a multilingual model, and use POS tagging for downstream evaluation. Results show that continuous pre-training substantially improves performance. More concretely, EN BERT initially outperformed mBERT with an accuracy of 83% during the language modeling phase. However, on the POS tagging task, mBERT surpassed EN BERT, achieving an accuracy of 94%, which suggests effective performance to the historical language varieties.[1]

## 1 Introduction

Applying Natural Language Processing (NLP) techniques to historical archives is a complex undertaking exacerbated by data scarcity (Biagetti et al., 2024). The limited availability of historical training data has impeded the advancement of NLP applications in archives such as OE due to the labor-intensive task required for manual annotation, leaving this domain relatively underexplored (Wunderlich et al., 2015b). Efforts to reduce the cost and human labor in sequence labeling tasks, such as POS tagging through semi-automation, have fallen short of capturing the full complexity of morphosyntactic alignment, highlighting the need for manually annotated corpora to obtain meaningful insights in NLP tasks involving historical archives (Moon and Baldridge, 2007).

Despite the capabilities of automated techniques in handling different levels of linguistic annotation (Bollmann, 2013; Hardmeier, 2016; Hämäläinen et al., 2021), manual annotation, though tedious, is an effective method to handle the complexities of varying dialects and the intricate linguistic phenomena of historical language (Beck et al., 2020). Furthermore, orthographic inconsistencies in historical archives pose significant challenges to corpus-based analytical linguistic techniques, including automated tagging, which can sometimes diminish the effectiveness and reliability of the analytical outcome (Baron and Rayson, 2008). One approach to overcome this issue is to normalize the OE data to modern English, thereby enhancing the accuracy of POS tagging (Bollmann, 2019), with manual normalization shown to improve performance across the nuanced historical linguistic features and spelling variations of ancient text (Moon and Baldridge, 2007; Scheible et al., 2011). However, normalization models require annotated training data, which is not available for all varieties of historical languages.

In this paper, we focus on re-training a discriminative language model (i.e. BERT) on OE, a West Germanic language related to Old Frisian and Old Saxon (Yang and Eisenstein, 2016), and demonstrate the refinement of historical archives with the ISWOC corpus (ISWOC, 2014), Complete Corpus of Anglo-Saxon Poetry (Hidley and Macrae-Gibson, 2014), and the Plaintext Wikipedia dump 2018 (Rosa, 2018). Our paper focuses on OE, an earlier stage (mid-fifth century), of the language with unique morphological patterns and features (Baker, 2012). We use POS tagging as a downstream evaluation, to evaluate the effectiveness of the re-training procedure. An example of sentences annotated with POS tags can be seen in Figure 1.

[1] Code and language model will be made public upon acceptance.



*Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, pages 258–267
January 20, 2025. 

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ac | hi | wunedon | on | clænnysse | oð | heora | lifes | ænde | mid | mycclum | geleafan |
| but | they | lived | in | purity | until | their | lives | end | with | great | faith |
| C- | Pp | V- | R- | N | R- | Ps | Nb | Nb | R- | Py | N |

Figure 1: Annotated example from the dataset, including a literal translation. First row: original OE data, Second row: literal English translation, last row: POS tags

| Text | # words |
|---|---|
| Unlabeled OE | |
| Wikipedia | 311,793 |
| Anglo-Saxon Poetry | 1,810,636 |
| ISWOC OE corpus | |
| Orosius | 1,728 |
| Ælfric's Lives of Saints | 3,137 |
| Apollonius of Tyre | 5,541 |
| Anglo-Saxon Chronicles | 5,939 |
| West-Saxon Gospels | 13,061 |
| Total | 29,406 |

Table 1: OE data

Our contributions can be summarized as follows:

- Adaptation of English BERT-Base-Uncased and Multilingual Bert-Base-Uncased models to OE through language modeling to enhance the generalization of the unique linguistic structures inherent in the OE language.
- A downstream evaluation of POS tagging tasks assessed the effectiveness of the BERT models on the historical archives.
- In-depth analysis and interpretation of the performance metrics, providing insights into the capabilities of the BERT models.

## 2 Old English

Historical OE is a West-Germanic language connected to Old Frisian and Old Saxon within the Ingvaeonic language used in England following the settlement of the Angles, Saxons, Jutes, and Frisians from Britain (Brigada Villa and Giarda, 2024). During the mid-fifth century, English-speaking settlers known as the Anglo-Saxons established themselves in Britain until the Norman Conquest. OE was inflected across various POS to denote first, second, and third person, singular and plural forms, and for mood, indicative, subjunctive, and imperative. (Fischer et al., 2017) The OE alphabet (Figure 2) consists of 24 letters (Wunderlich et al., 2015a).

| Split | unannotated | annotated |
|---|---|---|
| Train | 2,039,393 | 1,000 |
| Dev | 41,772 | 615 |
| Test | 41,264 | 615 |
| Ælfric's Lives of Saints (Out-of-domain) | – | 200 |
| Orosius (Out-of-domain) | – | 111 |

Table 2: Dataset splits

As time progressed, OE evolved into four dialects - Northumbrian, spoken north of the river Humber; Mercian, spoken in the Midlands; Kentish, spoken in Kent; and West Saxon, spoken in the southwest (Baker, 2012; Yang and Eisenstein, 2016). These dialects played a critical role in shaping the development of the English language. American regional dialects also have origins in OE dialects, with Standard Modern English primarily influenced by the Mercian dialect (Baker, 2012).

a æ b c d ð e f ȝ/g h i l m n o p r s/ſ t þ u ƿ/w x y

Figure 2: The OE alphabet

### 2.1 Data

For the language modeling step, we collected an unlabelled OE corpus (Table 1) using the Complete Corpus of Anglo-Saxon Poetry, which includes nine collections of unlabeled OE historical archives (Hidley and Macrae-Gibson, 2014) with the Plaintext Wikipedia dump 2018 (Rosa, 2018), comprising over two million words combined. We excluded fully capitalized texts to prevent potential misrepresentation of the data during pre-training. The ISWOC corpus (Table 1), which includes 2,541 human-annotated sentences in the West Saxon OE dialect, was utilized for supervised learning during our experiment, combining a total of 2,230 sentences for the training and development split and 311 combined sentences from Ælfric's Lives of Saints and Orosius (smallest files) for an out-of-domain dataset to assess and compare the learning capability of the BERT models (Table 2). The monolingual OE corpus contains morphosyntactic annotation at the sentence segmentation level, list-

ing POS, grammatical features, and lemma form for each token (ISWOC, 2014).

## 3 Related Work

Research on historical text processing has spanned various low-resourced languages, with efforts dedicated to refining NLP methodologies for better handling ancient and historical data. Previous studies have concentrated on overcoming the unique challenges posed by historical archives, such as developing tools and techniques to improve POS tagging accuracy. The preliminary efforts have paved the way for more effective NLP applications in historical linguistics, offering new opportunities for studying and preserving invaluable linguistic resources (Rayson et al., 2007; Scheible et al., 2011). Prior work included (Rögnvaldsson and Helgadottir, 2008) study on morphosyntactic tagging for Old Norse texts. Sanchez-Marco et al. (2011) also adapted methods for Old Spanish by enhancing dictionaries with word variants and retraining taggers with limited annotated data, demonstrating some applied NLP techniques. Sukhareva and Chiarcos (2014) mapped annotations from English to ancient Germanic languages highlighting the potential to advance our understanding of ancient texts (Yang and Eisenstein, 2016). The Qiu and Xu (2022) study concluded that incorporating historical data during training improved the capacity of BERT for diachronic semantic analysis.

In our experiments, we rely on the domain-specific pre-training technique with unlabeled data using masked language modeling (MLM) to enable BERT and mBERT to learn general language patterns from the unannotated OE archives for the digitization of important works like the OE Beowulf (Brodeur, 1959) poem to preserve historical records (Gururangan et al., 2020).

## 4 Method

**Language Modeling** In the first stage of the experiment, the BERT models underwent training to predict masked tokens (Figure 3) using an unlabeled OE corpus following the original procedure proposed by Devlin et al. (2019). The unsupervised learning process enabled the models to learn underlying patterns from the raw OE data without the constraints of pre-existing labels (Berg-Kirkpatrick et al., 2010). The goal of the pre-training phase was to provide the models with a foundational understanding of OE language patterns and morphological features for downstream evaluation supervised tasks. We compared batch sizes of 8, 16, and 32 (best performance with 16 batch_size) with a set peak learning rate of 1e-3 (Appendix A) during the language modeling phase to optimize the learning ability of the model to generalize the intricate linguistic patterns of the archives.

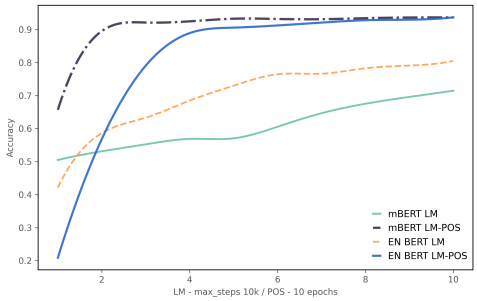


Figure 3: Learning curves on OE test data

| Model | POS | LM-POS |
|---|---|---|
| EN BERT | 86.37 | 92.16 |
| mBERT | 88.79 | 93.70 |

Table 3: Accuracy scores

We train both the English trained bert-base-uncased, and the multilingual bert-base-multilingual-uncased to evaluate the effect of multilingual training. Upon inspection of the vocabulary of the tokenizers, we find that the special characters used in OE (Section 2) are present in both tokenizers. However, for the English model, they are often only used as separate characters, whereas for the multilingual model, they are only used for subwords from other languages (e.g. Danish, Icelandic), so the tokenizers are likely not trained on much OE data.

**POS Tagging** The second stage of the experiment involved fine-tuning the BERT models for POS tagging on a manually annotated OE corpus (Table 1) containing 29,406 tokens (ISWOC, 2014). The supervised learning process also involved fine-tuning the BERT models on batch sizes of 8, 16, and 32 (best performance with 8 batch_size) with a set learning rate of 2e-5 (Appendix A) for a controlled evaluation of the learning capacity of the model across tasks.

## 5 Results

**Language Modeling** Before the unsupervised task, EN BERT and mBERT demonstrated closely

| Metric | A- | C- | Df | Du | F- | G- | I- | N- | Nb | Ne | Pd | Pi | Pp | Ps | Px | Py | R- | V- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EN BERT | | | | | | | | | | | | | | | | | | |
| Recall | 0.52 | 0.96 | 0.76 | 0 | 0 | 0.88 | 0 | 0 | 0.93 | 0.72 | 0.91 | 0.57 | 0.98 | 0.98 | 0.15 | 0.85 | 0.97 | 0.95 |
| Precision | 0.49 | 0.96 | 0.83 | 0 | 0 | 0.82 | 0 | 0 | 0.85 | 0.90 | 0.98 | 0.50 | 0.97 | 0.90 | 0.50 | 0.83 | 0.92 | 0.95 |
| F1 Score | 0.50 | 0.96 | 0.79 | 0 | 0 | 0.85 | 0 | 0 | 0.89 | 0.80 | 0.95 | 0.53 | 0.97 | 0.95 | 0.23 | 0.84 | 0.95 | 0.95 |
| mBERT | | | | | | | | | | | | | | | | | | |
| Recall | 0.66 | 0.96 | 0.82 | 0.07 | 0 | 0.83 | 0 | 0.30 | 0.92 | 0.82 | 0.94 | 0.86 | 0.98 | 0.94 | 0.35 | 0.94 | 0.97 | 0.95 |
| Precision | 0.59 | 0.97 | 0.80 | 1.00 | 0 | 0.81 | 0 | 1.00 | 0.89 | 0.78 | 0.96 | 0.43 | 0.98 | 0.94 | 0.64 | 0.93 | 0.94 | 0.98 |
| F1 Score | 0.62 | 0.97 | 0.81 | 0.13 | 0 | 0.82 | 0 | 0.46 | 0.91 | 0.80 | 0.95 | 0.57 | 0.98 | 0.94 | 0.45 | 0.94 | 0.95 | 0.96 |

Table 4: Performance scores on OE test data

comparable performance on the OE archives (Table 3). During language modeling, EN BERT exhibited stable accuracy across various configurations, with minor deviations suggesting consistent learning and effective convergence on the linguistic structures within OE (Figure 3). The stability underscored the capacity of EN BERT to adapt to historical linguistic patterns during the unsupervised phase, forming a robust basis for subsequent tasks. Although mBERT started with a higher accuracy, the model was quickly outperformed by EN BERT when training on more data, suggesting differing adaptation capabilities (Figure 3).

**POS Tagging** Results from the downstream POS tagging task revealed that, despite lower performance in the language modeling phase, mBERT outperformed EN BERT in the fine-tuning stage, demonstrating better generalization across linguistic features in OE. In the POS tagging task, a reversal in model performance patterns emerged compared to the language modeling task. mBERT achieved higher accuracy, ultimately reaching optimal performance (Appendix B). EN BERT, in contrast, which exhibited progressively improving accuracy and a stable learning trajectory during language modeling, achieved lower performance in the supervised POS tagging task (Appendix C). The shifted learning trend suggested that, although EN BERT adapted effectively to historical linguistic patterns in the unsupervised language modeling phase, mBERT proved more adaptable to generalize the unique linguistic historical OE archives (Figure 3). mBERT also outperformed EN BERT on the out-of-domain data, demonstrating its ability to handle diverse linguistic variations. (Appendix D & E). Based on the results (Table 3, 4 & 5), we hypothesize that mBERT outperformed EN BERT in the downstream POS tagging task due to its multilingual training (both for language modeling and the tokenizer), which allowed the model to generalize the unique linguistic features of the OE archives to achieve optimal results.

| | Out-of-domain | |
|---|---|---|
| Model | POS | LM-POS |
| EN BERT | 71.96 | 76.71 |
| mBERT | 77.87 | 84.13 |

Table 5: Out-of-domain accuracy scores

## 6 Analysis

**Performance** The personal pronouns (Pp) label attained the highest F1 scores, with mBERT recording 0.98, closely followed by EN BERT achieving 0.97 (Table 4) on the unique OE POS labels. A breakdown of the findings revealed that the EN and ML models demonstrated similar trends in capturing the same distribution of three of the 18 POS categories - proper noun (Ne), demonstrative pronoun (Pd), and preposition (R-). mBERT outperformed EN BERT across most of the 18 categories, particularly for personal pronouns (Pp), common nouns (Nb), quantifiers (Py), conjunctions (C-), and verbs (V-). mBERT demonstrated lower performance on the interrogative adverb (DU) and infinitive marker (N-) labels, while EN BERT did not identify the labels. Both models failed to recognize the foreign word (F-) and interjection (I-) labels during the downstream task (Figure 4).

**Tagging Discrepancies** In some instances, although the BERT models indicated a high confidence level in predicting the POS label for some tokens, the predictions were incorrect, while in a few cases, lower confidence levels aligned with correct classifications (Appendix F).

**Misclassifications** Tagging discrepancies observed throughout the corpus showed the predicted frequency for adjectives (A-) indicated an over-

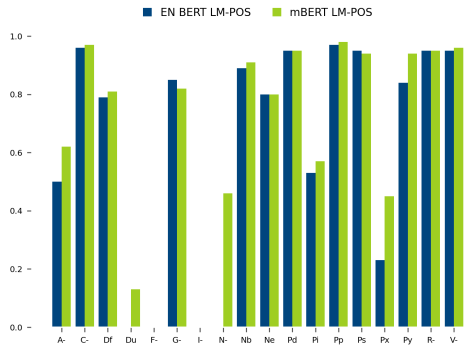


Figure 4: F1 results on the OE test data

prediction, manual inspection revealed that this is mainly due to contextual ambiguities in the Ælfric's Lives of Saints archive. Other notable discrepancies included challenges predicting the conjunctions (C-) label, misclassifications for subjunctions (G-) and pronouns (Pp), and underpredictions for adverbs (Df) and possessive pronouns (Ps) (Appendix G).

**Low Predictions** Underrepresentation of labels in the West Saxon Gospels, particularly for foreign words (F-) and interjections (I-), recorded zero predictions in a few instances despite having actual labels, indicating the challenges of the models to recognize less common POS categories (Appendix H). The EN BERT model also failed to make any predictions for interrogative adverbs (Du) despite 53 representations of the label throughout the biblical archive (Appendix G).

**Contextual Errors** The POS interjection (I-) label demonstrated a 100% error rate due to the nuanced characteristics of the label to exhibit considerable variability in context and form, which likely obstructed the tagging process. Similarly, the interrogative adverb (DU) also exhibited 100% error, with its syntactic complexity reflecting morphological challenges (Appendix I).

## 7 Conclusion and Future Work

In this paper, we introduced, to the best of our knowledge, the first historical language model specifically developed for OE. We demonstrated that retraining on limited data can lead to substantial improvements in performance, as evidenced by state-of-the-art scores in part-of-speech (POS) tagging (Eiselen and Gaustad, 2023). The pretraining of the BERT models on raw historical OE archives enhanced the POS tagging performance. The fine-tuning of the BERT models on a manually annotated OE corpus allowed the models to refine predictions to achieve high accuracy (Figure 3). The findings underscored the value of combining unsupervised and supervised training techniques to enhance POS tagging for historical languages. Nevertheless, our analysis highlighted that employing NLP techniques on historical OE archives is a difficult task. Future research should address the misclassification errors while developing strategies to enhance the generalization of the unique grammatical structures inherent in OE, including testing different models to optimize performance.

## 8 Acknowledgments

We want to express our gratitude to Professor Kristin Bech from the Faculty of Humanities, Department of Literature, Area Studies and European Languages at the University of Oslo, Norway, for her invaluable guidance in distinguishing and categorizing the OE dialect of the ISWOC corpus, which was essential to this paper.

## References

Peter S Baker. 2012. *Introduction to old English*. John Wiley & Sons.

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590.

Erica Biagetti, Martina Giuliani, Silvia Zampetta, Silvia Luraghi, and Chiara Zanchi. 2024. Combining neo-structuralist and cognitive approaches to semantics to build wordnets for ancient languages: Challenges and perspectives. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024*, pages 151–161, Torino, Italia. ELRA and ICCL.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the*

*7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria. Association for Computational Linguistics.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Luca Brigada Villa and Martina Giarda. 2024. From YCOE to UD: Rule-based root identification in Old English. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 22–29, Torino, Italia. ELRA and ICCL.

Arthur Gilchrist Brodeur. 1959. *The art of Beowulf*. Univ of California Press.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct South African languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53, Dubrovnik, Croatia. Association for Computational Linguistics.

Olga Fischer, Hendrik De Smet, and Wim van der Wurff. 2017. *A brief history of English syntax*. Cambridge University Press.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of historical old literary finnish texts in modern orthography. *arXiv preprint arXiv:2107.03266*.

Christian Hardmeier. 2016. A neural model for part-of-speech tagging in historical texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 922–931, Osaka, Japan. The COLING 2016 Organizing Committee.

Greg Hidley and O.d. Macrae-Gibson. 2014. Complete corpus of anglo-saxon poetry. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

ISWOC. 2014. The ISWOC treebank.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 390–399, Prague, Czech Republic. Association for Computational Linguistics.

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

Paul Rayson, Dawn E Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of the Corpus Linguistics conference: CL2007*.

Rudolf Rosa. 2018. Plaintext wikipedia dump 2018. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Eirikur Rögnvaldsson and Sigrun Helgadottir. 2008. Morphological tagging of old norse texts and its use in studying syntactic variation and change. In *Proceedings of the LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008). ELRA, Paris*.

Cristina Sanchez-Marco, Gemma Boleda, and Lluis Padro. 2011. Extending the tool, or how to annotate historical language varieties. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, OR, USA. Association for Computational Linguistics.

Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf'pos-tagger on early modern german text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 19–23.

Maria Sukhareva and Christian Chiarcos. 2014. Diachronic proximity vs. data sparsity in cross-lingual parser projection. a case study on germanic. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Martin Wunderlich, Alexander Fraser, and Paul Sander Langeslag. 2015a. God wat Þæt ic eom god-an exploratory investigation into word sense disambiguation in old english. In *GSCL*, pages 39–48.

Martin Wunderlich, Alexander M. Fraser, and Paul Sander Langeslag. 2015b. God wat Þaet ic eom god - an exploratory investigation into word sense disambiguation in old english. In *German Society for Computational Linguistics*.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328, San Diego, California. Association for Computational Linguistics.

## A Hyperparameters

| Model | dropout rate | learning rate | weight decay | batch size | steps / epochs | optimizer |
|---|---|---|---|---|---|---|
| **LM** | 0.1 | **1e-3** | 0.01 | **8**, 16, 32 | 10k | adamw |
| **POS** | 0.1 | **2e-5** | 0.1 | **8**, 16, 32 | 10 | adamw |

Table 6: Training hyperparameters, best in bold.

## B mBERT Metrics

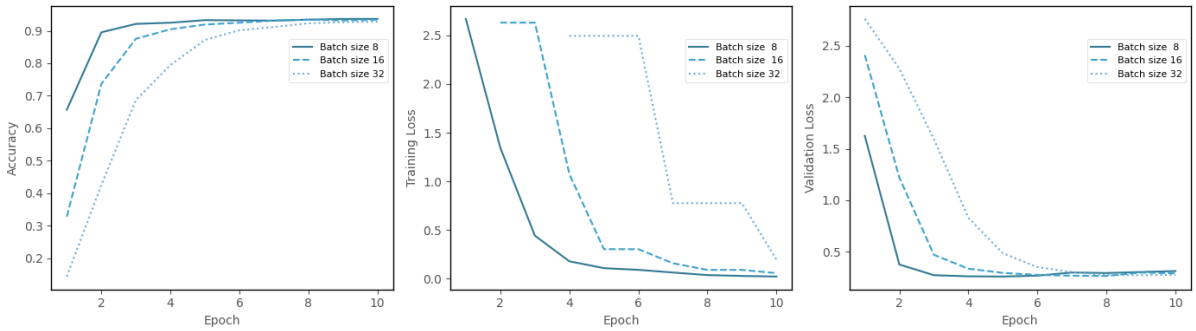


Figure 5: mBERT accuracy and loss metrics across different batch sizes

## C EN BERT Metrics

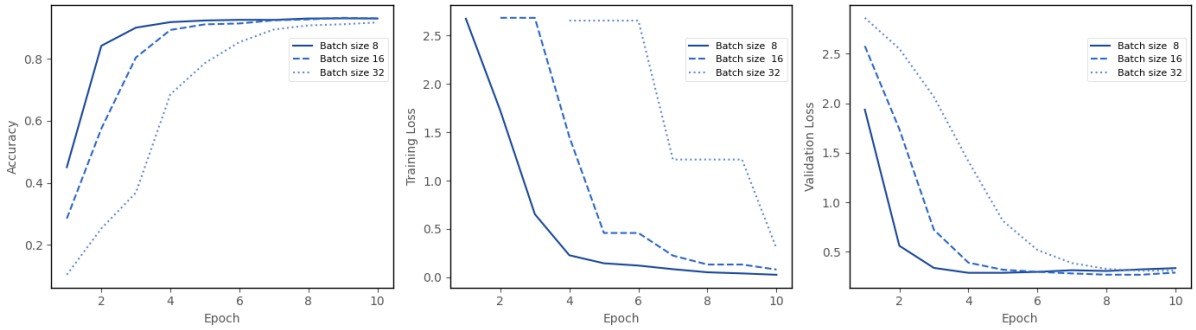


Figure 6: EN BERT accuracy and loss metrics across different batch sizes

## D mBERT Out-of-domain Metrics

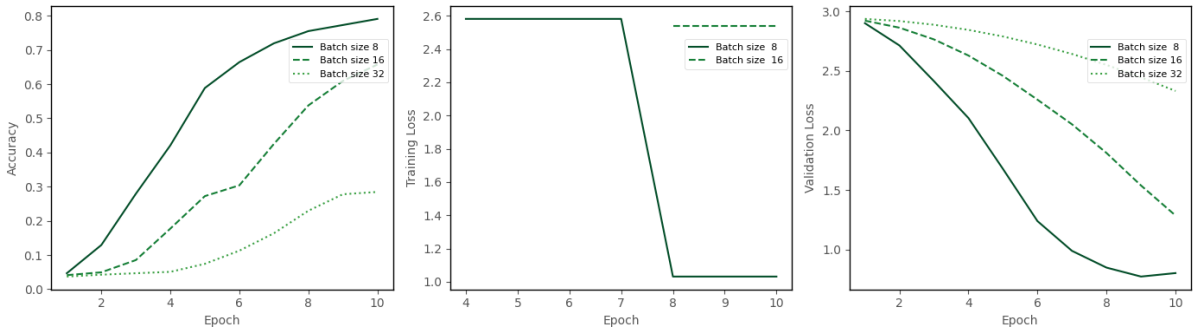


Figure 7: mBERT Out-of-domain accuracy and loss metrics across different batch sizes

## E EN BERT Out-of-domain Metrics

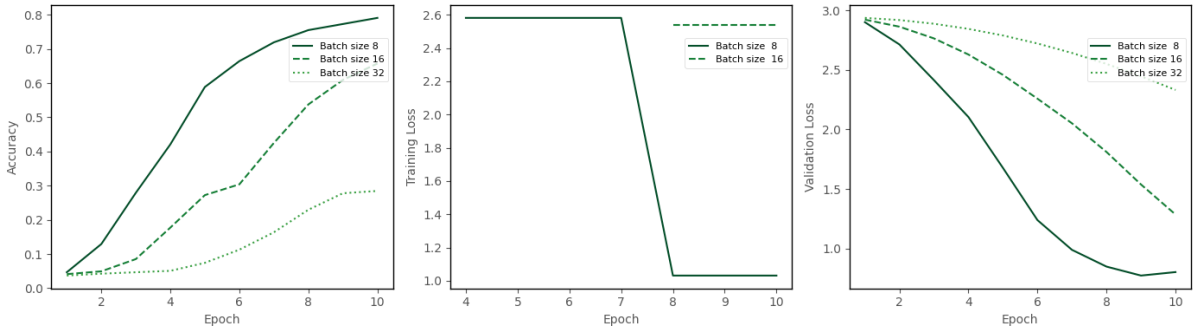


Figure 8: EN BERT Out-of-domain accuracy and loss metrics across different batch sizes

## F Tagging Errors

| POS | Summary | Actual FQ | Bert FQ | mBERT FQ | Error |
|---|---|---|---|---|---|
| A | adjective | 168 | 174 | 195 | misclassification |
| Du | interrogative adverb | 53 | 0 | 0 | no prediction |
| F- | foreign word | 12 | 0 | 0 | no prediction |
| G- | subjunction | 111 | 128 | 131 | misclassification |
| I- | interjection | 10 | 0 | 0 | no prediction |
| Nb | common noun | 264 | 311 | 368 | misclassification |

Table 7: Most frequent POS tagging errors

## G Lowest Predicted Frequency

| POS | Actual FQ | Bert FQ | mBERT FQ |
|---|---|---|---|
| A- | 331 | 184 | 189 |
| C- | 1141 | 382 | 383 |
| Df | 1076 | 379 | 365 |
| Du | 53 | 0 | 0 |
| F- | 12 | 0 | 0 |
| G- | 528 | 287 | 284 |
| I- | 10 | 0 | 0 |
| N- | 10 | 6 | 7 |
| Nb | 1830 | 1011 | 969 |
| Ne | 341 | 182 | 176 |
| Pd | 765 | 356 | 354 |
| Pi | 57 | 21 | 22 |
| Pp | 1836 | 993 | 1002 |
| Ps | 326 | 192 | 194 |
| Px | 40 | 8 | 9 |
| Py | 412 | 221 | 244 |
| R- | 895 | 508 | 503 |
| V- | 2835 | 1524 | 1553 |

Table 8: West-Saxon Gospels

## H OE POS Tags

| POS | Summary |
|---|---|
| A- | adjective |
| C- | conjunction |
| Df | adverb |
| Du | interrogative adverb |
| F- | foreign word |
| G- | subjunction |
| I- | interjection |
| N- | infinitive marker |
| Nb | common noun |
| Ne | proper noun |
| Pd | demonstrative pronoun |
| Pi | interrogative pronoun |
| Pp | personal pronoun |
| Ps | possessive pronoun |
| Px | indefinite pronoun |
| Py | quantifier |
| R- | preposition |
| V- | verb |

Table 9: A list of the POS labels in the ISWOC Corpus

# I Actual Frequency vs. Predicted Frequency

| POS | Actual FQ | Bert FQ | mBERT FQ |
|---|---|---|---|
| A- | 175 | 170 | 171 |
| C- | 587 | 330 | 330 |
| Df | 518 | 409 | 410 |
| F- | 6 | 3 | 4 |
| G- | 136 | 126 | 126 |
| N- | 2 | 2 | 2 |
| Nb | 1042 | 1042 | 1041 |
| Ne | 630 | 629 | 629 |
| Pd | 478 | 450 | 450 |
| Pi | 1 | 1 | 0 |
| Pp | 274 | 273 | 272 |
| Ps | 76 | 76 | 77 |
| Px | 12 | 11 | 12 |
| Py | 280 | 278 | 277 |
| R- | 653 | 647 | 647 |
| V- | 858 | 850 | 849 |

(a) Anglo-Saxon Chronicles

| POS | Actual FQ | Bert FQ | mBERT FQ |
|---|---|---|---|
| A- | 78 | 65 | 77 |
| C- | 121 | 85 | 85 |
| Df | 164 | 131 | 129 |
| Du | 2 | 0 | 0 |
| G- | 78 | 85 | 85 |
| Nb | 264 | 311 | 268 |
| Ne | 97 | 75 | 89 |
| Pd | 151 | 134 | 134 |
| Pi | 3 | 2 | 1 |
| Pp | 132 | 122 | 119 |
| Ps | 29 | 29 | 31 |
| Px | 11 | 6 | 7 |
| Py | 125 | 119 | 127 |
| R- | 174 | 161 | 161 |
| V- | 272 | 265 | 277 |

(b) Orosius

| POS | Actual FQ | Bert FQ | mBERT FQ |
|---|---|---|---|
| A- | 237 | 225 | 226 |
| C- | 317 | 222 | 221 |
| Df | 531 | 418 | 420 |
| Du | 9 | 0 | 4 |
| F- | 16 | 14 | 14 |
| G- | 267 | 261 | 257 |
| I- | 9 | 0 | 0 |
| N- | 3 | 3 | 3 |
| Nb | 852 | 838 | 838 |
| Ne | 171 | 140 | 140 |
| Pd | 434 | 400 | 401 |
| Pi | 27 | 22 | 20 |
| Pp | 645 | 606 | 606 |
| Ps | 166 | 162 | 162 |
| Px | 13 | 14 | 13 |
| Py | 124 | 123 | 123 |
| R- | 412 | 380 | 380 |
| V- | 1102 | 1064 | 1064 |

(c) Apollonius of Tyre

| POS | Actual FQ | Bert FQ | mBERT FQ |
|---|---|---|---|
| A- | 168 | 174 | 195 |
| C- | 211 | 147 | 144 |
| Df | 245 | 193 | 193 |
| Du | 7 | 0 | 0 |
| F- | 20 | 0 | 4 |
| G- | 111 | 128 | 131 |
| I- | 7 | 0 | 0 |
| N- | 1 | 0 | 1 |
| Nb | 575 | 589 | 529 |
| Ne | 169 | 149 | 145 |
| Pd | 256 | 242 | 239 |
| Pi | 3 | 2 | 3 |
| Pp | 220 | 232 | 215 |
| Ps | 85 | 55 | 84 |
| Px | 3 | 2 | 4 |
| Py | 92 | 80 | 73 |
| R- | 303 | 298 | 297 |
| V- | 562 | 547 | 581 |

(d) Ælfric's Lives of Saints

Table 10: Actual and predicted POS frequencies