Neural Architectures for Old English Lemmatisation: A Comparative Study of Corpus-Based and Dictionary-Augmented Approaches

Arquitecturas neuronales para la lematización del inglés antiguo: comparativa de enfoques basados en corpus y aumentados con diccionarios

Javier Martín Arista
Universidad de La Rioja
javier.martin@unirioja.es

Abstract: Lemmatisation is an essential step in the computational analysis of historical corpora, such as Old English, which presents problems of spelling variation, widespread inflectional morphology, and limited annotated resources. This paper conducts a detailed comparison of neural architectures for Old English lemmatisation: encoder-decoder (BiLSTM with attention) and Transformer under three experimental conditions (corpus-only, dictionary augmentation without POS tags, with POS tags) and advanced techniques (character-level augmentation, learning cycles with restarts). The Transformer achieves 66.65% development accuracy and our improved model 67.58% test accuracy. Dictionary augmentation without POS degrades performance (≈53% unknown POS); with POS it recovers but does not exceed corpus-only training. Pre-trained embeddings (Word2Vec) and BPE tokenization reduce accuracy (up to −12.67% and −12.92%), showing that lemmatisation is a morphological task where semantic signals introduce noise.
Keywords: lemmatisation, Old English, neural networks, Transformer, historical linguistics, low-resource NLP, data augmentation.

Resumen: La lematización es un paso esencial del análisis computacional de corpus históricos, como el del inglés antiguo, que presenta problemas de variación ortográfica, morfología flexiva generalizada y recursos anotados limitados. Este artículo lleva a cabo una comparación detallada de arquitecturas neuronales para la lematización del inglés antiguo: encoder-decoder (BiLSTM con atención) y Transformer en tres condiciones experimentales (solo corpus, aumento con diccionario sin etiquetas POS, con etiquetas POS) y técnicas avanzadas (aumento a nivel de carácter y ciclos de aprendizaje con reinicios). El Transformer alcanza 66,65% de precisión en desarrollo y nuestro modelo mejorado 67,58% en test. El aumento con diccionario sin POS empeora el rendimiento (≈53% de POS desconocidos); con POS se recupera, pero no supera el entrenamiento solo con corpus. Los embeddings preentrenados (Word2Vec) y la tokenización BPE reducen la precisión (hasta −12,67% y −12,92%), lo que evidencia que la lematización es una tarea morfológica donde las señales semánticas introducen ruido.
Palabras clave: lematización, inglés antiguo, redes neuronales, Transformer, lingüística histórica, PLN con pocos recursos, aumento de datos.

1. Introduction

Applying modern natural language processing (NLP) tools to historical languages can be problematic for reasons of spelling variation, limited annotated resources, and the absence of native-speaker intuitions (Piotrowski, 2012). Lexical and morphological tools trained on contemporary languages often fail when applied to earlier stages of the same language. Amongst historical languages, Old English (c. 450–1150 CE) is a relatively well documented language if compared with other old Germanic languages but is clearly an under-resourced language if compared with Present-Day English. The *Dictionary of Old English Corpus* (DOEC, Healey 2012), which has gathered all the written records of the language, contains approximately 3 million words of text.

Along with the scarcity of linguistic data, the computational processing of Old English poses some difficulties: the existence of generalised inflectional morphology, the lack of standardised orthography, the generalisation of character variants, and the limited availability of annotated resources.

Lemmatisation, the mapping of inflected word forms to their dictionary headwords, is fundamental to corpus linguistics and digital humanities research. Without automated lemmatisation, scholars must manually normalise thousands of variant spellings and inflections before conducting frequency analysis or building concordances. For Old English, lemmatisation allows for concordance generation, frequency analysis, dictionary linking, and downstream NLP. A reliable lemmatiser would accelerate research across philology, literary studies, and historical linguistics.

We approach lemmatisation as a joint task with part-of-speech (POS) tagging, following recent work showing that multi-task learning improves both objectives (Kondratyuk and Straka, 2019). This is particularly relevant for Old English, where the same surface form may correspond to different lemmas depending on grammatical function. For instance, *þā* can function as a demonstrative pronoun (lemma: *sē*) or as an adverb (lemma: *þā*). Joint lemma-POS prediction guarantees that the model disambiguates such cases using grammatical information learnt from the training data.

This said, the study addresses four research questions: (i) How do encoder-decoder (Seq2Seq) and Transformer architectures compare for Old English lemmatisation? (ii) Can dictionary data augmentation improve performance in low-resource settings? (iii) What is the impact of POS tag availability in augmented training data? (iv) Can advanced training techniques (data augmentation, cyclical learning rates, label smoothing) further improve performance?

2. Related Work

The lemmatisation of historical languages has traditionally relied on rule-based morphological analysers (Linden and Pirinen, 2009) or dictionary lookup systems. For Latin, the LEMLAT morphological analyser achieves high accuracy through comprehensive morphological rules (Passarotti, 2004). For Middle English, *The Penn-Helsinki Parsed Corpus* provides manually annotated lemmas that pave the way for dictionary-based approaches (Kroch and Taylor, 2000). For Old English, the lemmas from poetical texts have been semi-automatically gathered by means of morphological generation (Hamdoun Bghiyel, 2025); while the lemmas of *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* have been manually compiled by Cichosz et al. (2021). In spite of the quality of their philological data, these lemmatised inventories of Old English are hardly compatible with one another due to spelling conventions and, above all, for reasons related to the treatment of derivational morphology, which results in substantially different lemma list depending, for instance, on the overall consequences of the adoption of the strong or the weak forms of affixes.

Neural approaches to historical language lemmatisation have shown promising results. Manjavacas et al. (2019) demonstrated that neural sequence-to-sequence models outperform rule-based systems for non-standard historical language varieties. Of particular relevance is Kestemont et al. (2017), who developed a lemmatiser for Middle Dutch using temporal (1D) convolutions at the character level combined with word2vec embeddings for contextual

disambiguation. Although it is effective for Middle Dutch (achieving 90–94% overall accuracy), this architecture has fundamental limitations for Old English. First, it cannot predict lemmas not seen during training (approximately one half of the Old English corpus are hapax legomena). Second, their reliance on word embeddings for disambiguation assumes sufficient corpus size to learn meaningful distributional representations, but Middle Dutch corpora are substantially larger (130K–575K training tokens) than available Old English resources.

We adopt a sequence-to-sequence generation approach that can produce novel lemmas character-by-character, thus addressing these limitations and guaranteeing generalisation to unseen vocabulary. Modern lemmatisation systems predominantly use neural sequence-to-sequence models that treat lemmatisation as a character-level transduction task. Bergmanis and Goldwater (2018) introduced context-sensitive neural lemmatisation using bidirectional LSTMs with attention and demonstrate that character-level models can handle unseen words effectively. Transformer-based approaches have recently achieved state-of-the-art results for morphological tasks. Kondratyuk and Straka (2019) presented UDPipe 2.0, which uses contextualised embeddings for multi-task morphological analysis. However, pre-trained multilingual models may not capture the specific characteristics of historical languages with limited training data, which motivates task-specific approaches like ours.

Data augmentation has proven effective for low-resource NLP tasks. Sennrich et al. (2016) showed that back-translation improves neural machine translation. Wei and Zou (2019) introduced Easy Data Augmentation (EDA) techniques including synonym replacement, random insertion, swap, and deletion. For morphological tasks, character-level augmentation is particularly relevant. Anastasopoulos and Neubig (2019) demonstrated that morphological inflection systems benefit from character-level noise injection that mimics spelling variation. This insight is especially applicable to Old English, where spelling variation is pervasive.

Several resources support the computational analysis of Old English. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al., 2003) and *The York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug 2001) provide a POS-tagged and parsed corpus of Old English containing approximately 1.5 million words. The combination of both resources is known as *YCOE*. *The Dictionary of Old English* (Healey, 2018) offers comprehensive digitised dictionary entries. The Bosworth-Toller dictionary is a fully digitised historical Anglo-Saxon dictionary (Crist and Tichý, 2014). ParCorOEv3 provides a parallel corpus of Old English with Universal Dependencies annotation (Martín Arista et al., 2023). Recent computational work includes Martín Arista and Metola Rodríguez (2026), who have demonstrated that character-level processing with custom word embeddings outperforms multilingual models for Old English POS tagging.

3. Data and Resources

3.1 Data Sources

Our primary training data comes from a POS-tagged and parsed corpus of Old English, the *YCOE*. We extracted approximately 65,700 inflectional forms from *The Dictionary of Old English* (DOE, Healey 2018) with lemmas and lexical categories. Each entry follows the structure: LEMMA INFLECTIONAL FORMS (comma-separated) LEXICAL CATEGORY. For example, *cyning* 'king' lists *cyning, cyninges, cyninge, cyningas, cyninga, cyningum* as noun; *findan* 'to find' lists *findan, findeþ, findaþ, findende, fand, funde, fundon, funden* as verb. Each inflectional form was expanded to a separate training example. We also extracted approximately 16,000 inflectional forms from the Bosworth-Toller dictionary with lemmas and POS tags, which provided additional coverage of lexical items not published by the DOE yet (letters L-Y). For instance, *wyn* 'pleasure' lists *wyn, wynn, wynne, wynna, wynnum* as noun; *lufian* 'to love' lists *lufian, lufie, lufast, lufaþ, lufiaþ, lufode, lufodest, lufodon, lufod, gelufod* as verb. Each inflectional form gives rise to a separate training example.

3.2. Preprocessing and Harmonisation

Preprocessing and validation were applied consistently across all data sources. Poetry and prose subcorpora used different column orders and POS tag conventions (e.g. Title Case vs lowercase category names); dictionary entries used lexical category names rather than standard tags. We validated field order and format for each subcorpus and mapped all POS labels to the Universal Dependencies (UPOS) tagset so that a single training pipeline could be used. All were harmonised to standard UPOS tags (NOUN, VERB, ADJ, ADV, ADP, CCONJ, PRON, NUM, etc.), with empty or missing values mapped to X. The character vocabulary comprises 98 symbols, including Old English-specific characters (<æ, þ, ð, p>), macronised vowels (<ā, ē, ī, ō, ū, ȳ>), and standard Latin script; unknown characters at inference time are mapped to a dedicated symbol.

3.3. Experimental Conditions

To illustrate the representation used by the model, the inflected form *cyninges* 'king' (gen. sg.) is encoded as the character sequence ⟨c, y, n, i, n, g, e, s⟩, framed by beginning- and end-of-sequence markers. The expected output is the character sequence <c, y, n, i, n, g> corresponding to the lemma *cyning*, together with the POS tag NOUN. Likewise, the verb form *findende* 'finding' is encoded as <f, i, n, d, e, n, d, e> and the target output is <f, i, n, d, a, n> (lemma *findan*) with the tag VERB. The form *wæron* is encoded as <w, æ, r, o, n> and mapped to <b, ē, o, n> with the tag VERB. This character-level representation enables the model to generate previously unseen lemmas character by character rather than selecting from a closed inventory of labels, which is essential given the large proportion of low-frequency and hapax forms in Old English.

We define three experimental conditions to isolate the effects of data augmentation and POS tag availability: (i) Corpus-Only (84,050) (ii) Corpus+Dictionary, no POS (164,908) + DOE + BT (no POS); (iii) Corpus+Dictionary with POS (146,798) + DOE + BT with POS. The main difference between conditions (ii) and (iii) is POS tag availability: without POS tags, 52.9% of dictionary data receives the unknown tag X, compared to only 5.1% with POS tags. This distinction is central for understanding our results.

Final dataset statistics after preprocessing are shown in Table 1. The Corpus-Only condition comprises 84,050 training instances, 10,506 development, and 10,507 test, with approximately 90,753 unique words. Dictionary-augmented conditions yield larger training sets but with differing unknown-POS proportions that significantly affect learning.

Condition	Train	Dev	Test
Corpus-Only	84,050	10,506	10,507
Corpus+Dict (no POS)	164,908	20,614	20,615
Corpus+Dict (with POS)	146,798	18,349	18,351

Table 1: Dataset statistics.

4. Methodology

4.1. Task Formulation and Input Representation

We formulate lemmatisation as a character-level sequence-to-sequence task with auxiliary POS classification. The input is an inflected word form as a character sequence (e.g. *cyninges*); the output is the lemma as a character sequence (e.g. *cyning*) plus a POS tag (e.g. NOUN). We use joint prediction with a shared encoder. This context-free approach processes words in isolation, suitable for dictionary-style lemmatisation and applications where sentence context is unavailable. We use a shared character vocabulary of 98 symbols, including Old English special characters (<æ>, <þ>, <ð>, <p>) and standard Latin script. The encoder consumes the character sequence of the inflected form; the decoder generates the lemma character-by-character, with an end-of-sequence token signalling completion; and a separate classification head predicts the POS tag from the encoder representation.

4.2. Context-Free Design

A deliberate constraint of our setup is that lemmatisation is performed in a context-free manner: each inflected form is processed in

isolation, without access to surrounding words or sentence-level information. This decision has clear consequences. It limits the ability of the model to disambiguate homographs whose lemma depends on syntactic context (e.g. *þā* as demonstrative pronoun vs. adverb), and is the main architectural restriction of the present study. We adopted it for three reasons: (i) it makes the system directly applicable to dictionary-style processing of word lists, concordances, and editorial workflows where sentence context is unavailable or unreliable; (ii) it keeps the architecture simple, reproducible, and easy to deploy; and (iii) preliminary context-aware experiments, reported in Section 5, produced substantially lower development accuracy (16–18%) than the context-free model (67.58%).

4.3. Seq2Seq Architecture

Our Seq2Seq model uses 64-dimensional character embeddings, a 2-layer bidirectional LSTM encoder (512 hidden units per direction), Bahdanau additive attention over encoder states, a 2-layer LSTM decoder (512 hidden units), and a POS classifier implemented as a linear layer on the mean-pooled encoder output. The model is trained with a dropout rate of 0.3, a learning rate of 1e-3, and a batch size of 128. The model comprises approximately 8 million parameters.

4.4. Transformer Architecture

Our Transformer model uses a model dimension of 512, 8 attention heads, 4 encoder and 4 decoder layers, a feed-forward dimension of 1024, dropout of 0.1 (baseline) or 0.15 (improved model), and sinusoidal positional encoding. The baseline is trained with a learning rate of 5e-5 and a batch size of 64. The total parameters are approximately 21.5 million. No pre-trained foundation or multilingual model is used as a starting point. Both the Seq2Seq and the Transformer models are trained from scratch on the Old English data described in Section 3. The Transformer follows the standard encoder–decoder design of Vaswani et al. (2017): the encoder produces contextual character representations and the decoder generates the lemma character sequence autoregressively. Sinusoidal positional encodings are added to the character embeddings so that the model has access to position information without learnt positional parameters. No subword tokeniser, contextualised embedding, or pre-trained checkpoint is used in the main models; the only experiments involving pre-trained or subword representations are the ablations reported in Table 5 (Word2Vec and BPE).

4.5. Training Techniques

For our improved model, we applied several techniques. Firstly, we applied character-level data augmentation: we implemented four operations designed for Old English spelling variation, including (i) character substitution (þ↔ð, æ↔a/e, i↔y), (ii) deletion (random character removal for words > 2 characters), (iii) swap (adjacent character swap to mimic scribal errors), and (iv) duplication, with probability 0.2 per word and re-applied each epoch to create fresh variations. Secondly, we applied cyclical learning rate with warm restarts (Loshchilov and Hutter, 2017), involving cosine annealing with range [1e-5, 1.5e-4], 4 cycles, and 5% warmup steps. This schedule allows the model to escape local minima and explore multiple regions of the loss landscape. Thirdly, we used label smoothing with factor 0.15 to prevent overconfident predictions. Fourthly, we relied on reduced POS loss weight: from 0.5 to 0.2 to prioritise lemmatisation over POS tagging. The improved model was trained for 80 epochs with batch size 24 (effective 72 with gradient accumulation).

In the corpus-only condition, approximately 23,700 training instances are identity transformations (inflected form identical to lemma), which the model learns to reproduce and which help stabilise training for invariable words and uninflected forms.

Full specifications will be documented in the repository https://huggingface.co/Nerthus-Project.

4.6. Evaluation Protocol

We report results separately for the development and test sets. During model exploration and selection — i.e. the comparison of architectures, the three data conditions, and the ablations involving pre-trained embeddings and BPE — all configurations are compared on the development

set only. The test set is held out and used exclusively for the final improved model selected on the development set. This protocol is intended to avoid indirect tuning on the test data and explains the structure of the result tables: Table 2 reports development-set figures across all baseline conditions, whereas Table 3 reports test-set figures only for the final configuration that was selected on development data. Intermediate rows in Table 3 show development figures only, since these models were never selected for final reporting and were not evaluated on the test split. Lemma accuracy is computed as exact match between the predicted character sequence and the gold lemma after normalisation; POS accuracy is computed as exact match on the predicted UPOS tag.

5. Results

Table 2 shows baseline results across experimental conditions. The Transformer consistently outperforms Seq2Seq. Under corpus-only training, the Transformer achieves 66.65% lemma accuracy versus 53.15% for Seq2Seq. This is a 25.4% relative improvement. Adding dictionary data without POS tags causes severe degradation: Seq2Seq drops to 40.86%, Transformer to 58.03%. With POS tags added to dictionary data, Seq2Seq recovers to 52.07% and Transformer to 50.26%, but neither exceeds corpus-only performance.

Dataset	Model	Lemma Acc	POS Acc
Corpus-Only	Seq2Seq	53.15%	80.27%
Corpus-Only	Transformer	66.65%	81.21%
Dict (no POS)	Seq2Seq	40.86%	69.38%
Dict (no POS)	Transformer	58.03%	68.48%
Dict (POS)	Seq2Seq	52.07%	84.42%
Dict (POS)	Transformer	50.26%	79.44%

Table 2: Baseline model results (development set).

Table 3 shows the progression with advanced techniques. Notice that intermediate rows are reported on the development set only; the test set was evaluated exclusively for the final selected configuration (see 4.6). Our best model achieves 68.68% development accuracy and 67.58% test accuracy (POS accuracy 82.75%), representing a +2.03% absolute improvement over the Transformer baseline and a +27% relative improvement over Seq2Seq. To our knowledge, this is the highest accuracy reported so far for character-level neural lemmatisation of Old English on the YCOE. Previous work on Old English lemmatisation has used different lemma inventories, different segments of the available corpora, and different evaluation protocols, which makes direct, like-for-like comparison difficult. Therefore, our figures do not constitute a benchmark but rather a reference point for future work that adopts the same training/development/test split.

Model	Dev Lemma	Test Lemma	Test POS
Transformer baseline	66.65%	—	—
+ Pre-training	66.98%	—	—
+ Data Aug + Cyclical LR	68.68%	67.58%	82.75%

Table 3: Advanced training results.

Table 4 quantifies the correlation between unknown POS proportion and accuracy. Each percentage point of unknown POS correlates with approximately 0.2% lemma accuracy degradation. With 52.9% unknown POS in the dictionary (no POS) condition, average lemma accuracy drops to 49.45%.

Condition	Unknown POS	Avg Lemma Acc	Avg POS Acc
Corpus-Only	4.4%	59.90%	80.74%
Dict no POS	52.9%	49.45%	68.93%
Dict (POS)	5.1%	51.17%	81.93%

Table 4: Impact of unknown POS tags.

We tested pre-trained Word2Vec embeddings (trained on the DOEC, 150 dimensions, 106K vocabulary) and Byte Pair Encoding (BPE) with 500 merges. These experiments followed Kestemont et al. (2017), who reported improvements from word embeddings for Middle Dutch. Contrary to expectations, word embeddings reduced test accuracy: character-only baseline 67.58%, Word2Vec context fusion 63.53% (−4.05%), Word2Vec + character concatenation 49.37% (−12.67%). Most strikingly, words with embeddings (47.24% accuracy) performed worse than out-of-vocabulary words without embeddings (51.41%). BPE subword tokenisation achieved only 54.66% test accuracy, a −12.92% decrease from the character-level baseline (Table 5). The BPE vocabulary grew to 14,687 subword units versus 98 characters in the character-level model. It must be borne in mind that all figures are test-set accuracies; the character-only baseline corresponds to the final improved model in Table 3. These results establish that for Old English lemmatisation, character-level processing is both necessary and sufficient.

Configuration	Test Accuracy	Change
Character-only baseline	67.58%	—
Word2Vec context fusion	63.53%	−4.05%
Word2Vec + character concat	49.37%	−12.67%
BPE subword (500 merges)	54.66%	−12.92%

Table 5: Word embedding and BPE experiments.

The improved model exhibited characteristic cyclical learning-rate behaviour: four distinct training cycles were visible in the learning curves, with performance peaking at cycle boundaries (epochs 22, 41, 58, 74). The best development result (68.68%) was obtained at epoch 74. Warm restarts enabled the optimiser to escape local minima and explore multiple regions of the loss landscape, which contributed to the final gain over the baseline.

To test whether sentence-level context could improve lemmatisation, we ran additional experiments in which the model received surrounding words as context (following the idea of context-sensitive lemmatisation). Under two configurations (small and larger context encoders), development accuracy remained around 16–18%, far below the 67.58% achieved by the context-free character-level Transformer.

This result supports our design choice: for Old English, lemmatisation behaves as a morphological pattern-matching task at the word form, and contextual semantic information does not compensate for the added complexity and potential noise. We therefore did not pursue context-aware architectures further and report only the context-free results in the main tables.

6. Discussion, Error Analysis and Limitations

The first question that arises from the previous sections is why dictionary augmentation underperforms. Our results challenge the intuition that more training data always improves performance. We suggest three possible factors. Firstly, isolated vs. contextual learning: dictionary entries provide inflection–lemma pairs in isolation, whereas corpus data embeds words in various grammatical contexts. Models learn morphological patterns better from diverse contexts: seeing *cyninges* in genitive constructions teaches more than listing *cyninges* as a genitive form of *cyning*. Secondly, distribution mismatch: dictionary compilation follows lexicographical conventions that differ from corpus usage. Rare inflections may be overrepresented in dictionaries (documenting the full paradigm) but underrepresented in actual texts. Thirdly, the central role of POS tags: without POS tags, 53% of dictionary data provides incomplete signal. The model cannot learn POS-conditioned lemmatisation patterns. This seems to advise never training historical text lemmatisers on data lacking POS annotations.

These observations correspond to our specific setting — Old English, the YCOE training split, and the DOE / Bosworth-Toller dictionaries with the proportions of unknown POS reported in Table 4 — and should not be read as a universal law for historical lemmatisation. Whether

dictionary augmentation without POS would also harm performance in larger corpora, in languages with more standardised orthography, or with different lemma conventions, is an open empirical question. The safer reading of our results is therefore methodological: the absence of POS information in augmentation data should be treated as a risk factor that needs to be tested in each new setting, rather than assumed to be harmless.

The second question addressed in this discussion is character-level data augmentation. Data augmentation proved particularly effective for Old English because it simulates scribal variation (relatively unpredictable spellings), improves generalisation (the model learns that *cyning* and *kyning* map to the same lemma), and achieves meaningful gain (+2% accuracy) with no additional annotation effort.

The next question is why Transformers excel at lemmatisation. Transformers outperform Seq2Seq for character-level lemmatisation because self-attention may capture relationships between any characters, all character positions may attend to each other simultaneously, and layer normalisation and residual connections may lead to training deeper networks with better gradient flow. The complementary question is why word embeddings and BPE fail. Word2Vec captures semantic similarity (e.g. *cyning* ≈ *ealdormann* in distribution), but lemmatisation is a morphological task. The assignment *cyninges* → *cyning* requires recognising the genitive suffix *-es*, rather than knowing that *cyning* 'king' resembles *ealdormann* 'earl'. Semantic information is irrelevant and often noise. Character-level features already capture morphology optimally; word embeddings add conflicting signals that interfere with learnt character patterns. Furthermore, 51% of test words were out-of-vocabulary in the Word2Vec vocabulary, which gives inconsistent learning signals across the dataset. BPE fails for converging reasons. Old English lacks standardised spelling, that is to say, the same morpheme may be spelt differently across texts (e.g. *-es, -æs, -as* for the genitive). BPE learns surface patterns from the training data, not underlying morphological structure. The resulting vocabulary of 14,687 subword units fragments the input without capturing generalisable units, and character-level models retain fine-grained control to learn arbitrary transformations. For morphologically complex languages with high spelling variation, character-level processing remains optimal. This finding contrasts with Kestemont et al. (2017)'s positive results for Middle Dutch, where much larger corpora and a classification-based setup made word embeddings useful for contextual disambiguation among known lemmas.

The 67.58% test accuracy should be interpreted in the context of other historical and morphologically rich languages. Reported lemmatisation accuracies for Latin and for Middle Dutch with ample data often reach the mid-80s or low 90s (e.g. Passarotti, 2004; Kestemont et al., 2017), in part because of more standardised orthography, larger annotated resources, or classification over a closed lemma set. Old English combines extreme spelling variation with rich inflectional morphology and a relatively small annotated corpus. It is therefore among the most difficult settings for automatic lemmatisation. Studies of annotation consistency suggest that human agreement on Old English lemmatisation is itself in the range of roughly 90–95%, so the achievable ceiling is limited not only by model capacity but by the consistency of the gold standard and the ambiguity of certain forms.

The model correctly lemmatises many frequent patterns: *cyninges* → *cyning* (NOUN), *cwæð* → *cweðan* (VERB), *gōdan* → *gōd* (ADJ), *þurh* → *þurh* (ADP). Common errors, which confirm that morphological complexity and spelling variation remain the primary sources of difficulty, include strong verb ablaut (*sang* predicted as *singan*), lemma variants (*halgan* → *halig* vs *halga*), and spelling variants (*forðon* vs *forðam*). In quantitative terms, strong verb ablaut represents ~25%, spelling variants amount to ~20%; prefix handling (like *ge-* inconsistency) and rare words reach ~15% each; annotation inconsistencies and homographs represent ~10%.

The best model can be run at hundreds of words per second on standard hardware and supports batch processing for large texts. Output is produced in a simple word-lemma-POS format suitable for concordancing, frequency lists, and linking to dictionary entries. For ambiguous or uncertain cases, beam search can return several lemma candidates with scores, which allows

downstream tools or annotators to choose or validate. These characteristics make the system usable in practice for digitised Old English corpora and editorial projects and the context-free design keeps deployment simple and avoids the need for sentence-segmented or parsed input.

The limitations of the research reported in this article include context-free processing (no sentence-level disambiguation), single-language scope, evaluation on held-out corpus data, and lower accuracy on proper nouns. Performance on entirely new texts may differ. Evaluation is performed on a held-out split from the same corpus (YCOE); we have not evaluated on independently collected texts or other editions, where spelling and morphological variation might differ. The gold data themselves contain a non-negligible proportion of inconsistent or debatable lemma assignments, which caps the maximum accuracy that can be attributed to the model alone. Proper nouns and very low-frequency types remain particularly challenging, and the model has not been tuned or evaluated for named-entity-aware lemmatisation. Finally, the benchmark status of the reported figures is also limited by the absence of a direct comparison with earlier Old English lemmatisers.

Our findings have implications for historical NLP: (i) data quality matters more than quantity: increasing training data through dictionary augmentation is not a substitute for high-quality corpus annotation; (ii) task-specific solutions can outperform pre-trained multilingual models for historical languages; (iii) preprocessing and format validation of corpus sources are essential. Future work may explore pre-trained language models on the full DOEC, transfer learning from related Germanic languages, dictionary-enhanced inference as post-processing, and extension to Middle and Early Modern English.

7. Conclusion

This article has conducted a comprehensive study of neural architectures for Old English lemmatisation. The Transformer with character-level data augmentation and cyclical learning rates achieves 67.58% test accuracy. Six main conclusions can be drawn. First, the Transformer outperforms Seq2Seq by 25% relative improvement (66.65% vs 53.15% on corpus-only data), thus establishing the superiority of attention-based architectures for this task. Second, dictionary augmentation without POS tags degrades performance catastrophically (52.9% unknown POS causes 13–23% accuracy drop) which suggests, at least in our setting, that POS-less dictionary augmentation should be avoided or, when unavoidable, carefully validated. Third, even with POS tags, dictionary augmentation does not exceed corpus-only performance, as contextual corpus data provides richer learning signals than isolated dictionary entries. Fourth, character-level augmentation provides +2.03% improvement by simulating scribal spelling variation. Fifth, pre-trained word embeddings reduce accuracy by up to 12.67%, and BPE subword tokenisation by 12.92%; lemmatisation is a purely morphological task where semantic and subword abstractions introduce noise. Sixth, data quality matters more than quantity. The results of this task involving historical language lemmatisation with rich morphology, generalised spelling variation and limited resources advise corpus-only training with Transformer architecture, character-level augmentation, complete POS annotations, and avoidance of word embeddings and subword tokenisation.

Acknowledgements

This work was supported by grant PID2023149762NB-100 (Agencia Estatal de Investigación) and grant PRXX24/00108 (Ministerio de Ciencia, Innovación y Universidades), which are gratefully acknowledged.

References

Anastasopoulos, A., and Neubig, G. (2019). Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 984–996).

Bergmanis, T., and Goldwater, S. (2018). Context sensitive neural lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1391–1400).

Cichosz, A., Pęzik, P. Grabski, M., Adamczyk, M., Rybińska, P. and Ostrowska, A. (2021). *The VARIOE online morphological dictionary for YCOE*; University of Łódź.

Crist, S., and Tichý, O. (2014). *Bosworth-Toller's Anglo-Saxon Dictionary online.* Charles University, Faculty of Arts. Retrieved from https://bosworthtoller.com.

Hamdoun Bghiyel, Y. (2025). *The Lemmatisation of The York-Toronto-Helsinki Parsed Corpus of Old English Poetry with a Relational Database of Old English Dictionaries*. Ph.D. Dissertation, University of La Rioja.

Healey, A. diPaolo. (2012). *The Dictionary of Old English Web Corpus*. With John Price Wilkin and Xin Xiang. Dictionary of Old English Project, University of Toronto.

Healey, A. diPaolo. (2018). *The Dictionary of Old English: A to I*. Dictionary of Old English Project, University of Toronto.

Kestemont, M., de Pauw, G., van Nie, R. and Daelemans, W. (2017). Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities* 32:4, pp. 797-815.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. EMNLP.

Kroch, A. and Taylor, A. (2000). *Penn-Helsinki Parsed Corpus of Middle English*, *second edition.*

Lindén, K., and Pirinen, T. (2009). Weighted finite-state morphological analysis of Finnish compounding with HFST-LEXC. In *Proceedings of the 17th Nordic Conference on Computational Linguistics* (NODALIDA 2009) (pp. 89–95).

Loshchilov, I., and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Toulon, France.

Manjavacas, E., Kádár Á. and Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. NAACL.

Martín Arista, J. and Metola Rodríguez, D. (2026). Outperforming multilingual models: Character-level processing and custom word embeddings for Old English. *International Journal of Humanities and Arts Computing,* 20 (1), 18–35.

Martín Arista, J. (ed.), Domínguez Barragán, S., Fidalgo Allo, L., García Fernández, L., Hamdoun Bghiyel, Y., Lacalle Palacios, M., Mateo Mendaza, R., Novo Urraca, C., Ojanguren López, A. E., Ruíz Narbona, E., Torre Alonso, R. and Vea Escarza, R. (2023). *ParCorOEv3. An open access annotated parallel corpus Old English-English*. Nerthus Project, Universidad de La Rioja.

Passarotti, M. (2004). Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica Computazionale* 20-21, 397-414.

Pintzuk, S. and Plug, L. (2001). *The York-Helsinki Parsed Corpus of Old English Poetry*. University of York.

Piotrowski, M. (2012). *Natural language processing for historical texts*. Synthesis Lectures on Human Language Technologies 5:2.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 86–96).

Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. University of York.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5998–6008). Curran Associates, Inc. Long Beach, California.

Wei, J., and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6382–6388).