

Data Curation, Annotation Quality, and Error Patterns in Old English Automatic Lemmatisation

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

Abstract

Old English (c. 450–1150 CE), preserved in approximately three million words, is a historical language that poses distinctive problems for natural language processing: pervasive spelling variation, rich inflectional morphology, and heterogeneous annotated sources. This paper provides an analysis of the data curation pipeline, annotation quality, and linguistically motivated error patterns for Old English lemmatisation. We describe the extraction, harmonisation, and validation of training data from three historical sources, including a parsed corpus, a scholarly dictionary, and a historical dictionary, each with different formats, POS conventions, and coverage. Quantitative analysis shows that 11.3% of word types receive inconsistent lemma assignments, identity transformations account for 28.3% of training data, and the *ge-* prefix affects 11.7% of annotated instances. Error analysis organised by morphological category shows that strong-verb ablaut accounts for approximately 25% of lemmatisation errors, spelling variation for 20%, and prefix handling for 15%. We contextualise these findings within the broader landscape of low-resource NLP and formulate practical guidelines for building lemmatisation systems for historical low-resource languages, including recommendations on annotation consistency, POS completeness, and source harmonisation.

1 Introduction

Applying modern Natural Language Processing (NLP) to historical languages is hampered by limited annotated data, spelling variation, and the absence of standardised orthography (Piotrowski, 2012). While much attention has gone to neural architectures for lemmatisation, POS tagging, and parsing, far less has addressed data preparation, annotation quality, and systematic error analysis—factors that often determine the ceiling of model performance more directly than architectural choices.

This paper focuses on these under-explored aspects using Old English (c. 450–1150 CE) as a case study. Old English is the earliest documented form of English, preserved in approximately three million words digitised as the Dictionary of Old English Corpus (DOEC; Healey, 2012). It is a paradigmatic low-resource language: generalised spelling variation across scribal traditions, rich morphology, a small annotated corpus relative to the complexity of the language, and multiple digital resources with barely compatible annotation conventions.

Character-level Transformer models achieve the highest lemmatisation accuracy reported to date for Old English: the best model of Martín Arista (2026) reaches approximately 68% lemma accuracy on development data and 67.58% on the held-out test set (the error analysis in Section 5 uses the test figure). This is well below accuracies for better-resourced historical languages such as Latin (~85–90%; Passarotti, 2004) and Middle Dutch (~80–94%; Kestemont et al., 2017). These figures come from different datasets, annotation schemes, and evaluation settings, and are indicative rather than directly comparable; nevertheless, the size of the gap motivates looking beyond model architecture to the training data themselves.

Lemmatisation—the mapping of inflected forms to their dictionary headwords—is fundamental to corpus linguistics and the digital humanities: without it, scholars must manually normalise thousands of variant spellings and inflections before any study. For Old English it enables concordance generation, frequency analysis, and dictionary linking, among other tasks.

2 Old English as a Low-Resource Language

Old English is a West Germanic language attested across roughly seven centuries of scribal activity. Four properties make it both relevant and challeng-

ing for NLP: inflectional morphology, spelling variation, character variants, and limited annotated resources.

Old English marks case, gender, number, and tense through complex, largely explicit paradigms. A single lemma may surface in dozens of forms: the verb *bēon* ‘to be’ has over 15, including *eom*, *eart*, *is*, *sind*, *bēo*, *wæs*, and *wæron*. Nouns inflect for five cases and two numbers, and strong verbs exhibit ablaut (systematic vowel alternation) that changes the stem rather than attaching a suffix.

Orthographically, Old English lacked established spelling conventions. The same word may appear in different forms across or even within a manuscript: ‘king’ may be spelt *cyning*, *kyn- ing*, or *cyng*. It also employs characters absent from Latin script—<þ> (thorn), <ð> (eth), and <æ> (ash): thorn and eth are used interchangeably for dental fricatives, and ash alternates with <a> or <e> across traditions.

Regarding the availability of annotated resources, several digital resources support the computational analysis of Old English, although they differ widely in scope, format, and annotation depth. 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) together provide a POS-tagged and parsed corpus of approximately 1.5 million words, known collectively as the YCOE. The Dictionary of Old English (DOE; Healey, 2018) offers comprehensive digitised dictionary entries for the letters A–I. The Bosworth-Toller dictionary (Crist and Tichý, 2014) is a fully digitised historical Anglo-Saxon dictionary covering the entire alphabet. ParCorOEv3 (Martín Arista et al., 2023) provides a parallel corpus of Old English with Universal Dependencies annotation. The verbal lemmas from the YCOE prose have been semi-automatically gathered by means of morphological generation (Metola Rodríguez, 2015; García Fernández, 2018; Tío Sáenz, 2019), and the lemmas from poetical texts have been compiled by Hamdoun Bghiyel (2025). Finally, the VARIOE morphological dictionary (Cichosz et al., 2021) provides manually compiled lemmas for the YCOE. These resources, while individually valuable, are not mutually compatible: they adopt different lemma inventories, spelling conventions, POS tagsets, and treatments of derivational morphology, which give rise to dire incompatibilities when attempting to combine them for NLP train-

ing.

Lemmatisation of historical languages ranges from rule-based morphological analysers (Lindén and Pirinen, 2009) to dictionary lookup (Passarotti, 2004). Neural approaches have advanced quickly: Manjavacas et al. (2019) showed that sequence-to-sequence models outperform rule-based systems for non-standard varieties, and Kestemont et al. (2017) applied character-level convolutions with word embeddings for Middle Dutch, treating lemmatisation as classification over a fixed lemma set. Bergmanis and Goldwater (2018) introduced context-sensitive lemmatisation with bidirectional LSTMs and attention, handling unseen words effectively, and Kondratyuk and Straka (2019) presented UDPipe 2.0 using contextualised embeddings for multi-task morphological analysis. For Old English, Martín Arista and Metola Rodríguez (2026) showed that character-level processing with custom embeddings outperforms multilingual models for POS tagging, and Martín Arista (2026) found the Transformer to be the best architecture for lemmatisation and established a direct relationship between POS-annotation completeness and lemmatisation accuracy, a finding we return to in Section 4.3.

Low-resource NLP has been widely discussed: Joshi et al. (2020) documented the global inequality in resource availability, Magueresse et al. (2020) reviewed mitigation strategies, Hedderich et al. (2021) surveyed data augmentation, and Anastasopoulos and Neubig (2019) showed that character-level noise injection benefits morphological inflection—particularly relevant for historical languages with pervasive spelling variation. For historical texts specifically, Piotrowski (2012) surveys NLP challenges (tokenisation, normalisation, orthographic variation), while Universal Dependencies (Nivre et al., 2020) has standardised morphosyntactic annotation cross-linguistically.

Annotation quality has been studied mainly for contemporary languages. Plank et al. (2014) found inter-annotator disagreement in POS annotation to be largely systematic and concentrated in debatable cases, so part of what is scored as model error may reflect legitimate alternative analyses; Reiss et al. (2020), by contrast, found substantial proportions of genuinely incorrect labels in benchmarks such as CoNLL-2003. For historical languages such analysis is sparse: Cichosz et al. (2021) built a morphological dictionary for the YCOE, one of the few efforts addressing Old English annotation consistency at scale. We contribute the first quantitative

Language	Ann. tokens	Spell. var.	Lem. acc.
Modern English	>100 M	Low	>95%
Latin	~11 M	Moderate	85–90%
Middle Dutch	130–575 K	High	80–94%
Old English	~84 K	Very high	~68%
Old High German	<50 K	Very high	—

Table 1: Old English in the landscape of historical-language lemmatisation. Accuracy figures are taken from the cited studies and were obtained on different datasets, annotation schemes, and evaluation settings; they provide orientation rather than a controlled comparison. The Old English figure is the development-set accuracy of [Martín Arista \(2026\)](#).

analysis of annotation inconsistency in Old English and its impact on neural lemmatisation.

Table 1 situates Old English within the wider context of historical language NLP. Annotated tokens refers to the size of the training split available for lemmatisation experiments. The combination of generalised spelling variation, rich morphology, and a small annotated corpus places Old English among the most challenging 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% ([Cichosz et al., 2021](#)), which places an empirical ceiling on achievable accuracy.

3 Data Curation Pipeline

Building training data for Old English lemmatisation requires integrating material from sources that were compiled for different purposes, at different times, and following different conventions. We describe the three sources and the harmonisation steps applied. The overall pipeline proceeds in four stages: (i) extraction of raw word–lemma–POS triples from each source; (ii) format validation and column-order correction; (iii) POS tag harmonisation to UPOS; and (iv) quality checks including duplicate detection, encoding normalisation, and consistency analysis.

3.1 Sources

The YCOE contains approximately 1.5 million words of Old English prose and poetry, annotated with phrase-structure trees and POS tags. We extract word–lemma–POS triples from these trees. The corpus is distributed as two subcorpora with different internal formats: Prose (~85,500 entries) with columns ordered as LEMMA–WORD–POS, using uppercase UPOS-style tags (e.g.,

Source	Original	UPOS
YCOE Prose	NOUN	NOUN
YCOE Poetry	Noun	NOUN
DOE	noun	NOUN
BT	noun	NOUN
YCOE Prose	CCONJ	CCONJ
DOE	conjunction	CCONJ
BT	foreign word	X

Table 2: POS tag mapping examples across sources.

NOUN, VERB); and Poetry (~16,200 entries) with columns ordered as WORD–LEMMA–POS, using title-case category names (e.g., Noun, Verb). Failing to detect the reversed column order in the prose subcorpus would silently swap words and lemmas, producing a training set in which the model learns the inverse transformation. This kind of silent format mismatch is a recurrent hazard in historical NLP, where datasets are often manually curated and lack machine-readable metadata.

The DOE ([Healey, 2018](#)) provides comprehensive lemmatised entries for the letters A–I, each listing inflectional forms alongside the headword and its lexical category. We extracted approximately 65,700 inflectional forms, expanding each headword entry into individual word–lemma pairs. For instance, *cyning* ‘king’ yields training pairs from forms *cyning*, *cyninges*, *cyninge*, *cyningas*, *cyninga*, and *cyningum*, all mapped to the lemma *cyning*. Lexical categories follow dictionary conventions (e.g., noun, verb, adjective) rather than the UPOS tagset.

The Bosworth-Toller dictionary ([Crist and Tichý, 2014](#)) is a historical Anglo-Saxon dictionary covering the full alphabet. We extracted approximately 16,000 inflectional forms with lemmas and POS information, providing coverage for lexical items in the letters L–Y not yet published by the DOE.

3.2 Harmonisation

Integrating these three sources requires addressing five interrelated problems. Regarding POS tag mapping, Table 2 illustrates the diversity of POS conventions across sources. All tags were mapped to the Universal Dependencies tagset (UPOS; [Nivre et al., 2020](#)), with empty or unresolvable values assigned the tag X (unknown).

We adopt UPOS rather than a finer-grained, Old English-specific tagset for three reasons. First, UPOS is the most detailed level at which the three sources—the YCOE’s extended morphosyntactic tags and the dictionaries’ traditional lexical-

category labels—can be aligned reliably; any finer inventory would have to be imputed for much of the dictionary data, introducing noise. Second, with only $\sim 84\text{K}$ instances a fine-grained tagset would fragment already sparse evidence per tag. Third, UPOS ensures interoperability with the Universal Dependencies ecosystem (Nivre et al., 2020) and with resources such as ParCorOEv3 (Martín Arista et al., 2023). The cost is that morphological features (case, number, tense, mood) become unavailable to the model; encoding them as UD FEATS in CoNLL-U is a natural extension we leave for future work.

A unified character vocabulary of 98 symbols was constructed, including Old English-specific characters ($\langle p \rangle$, $\langle \delta \rangle$, $\langle \text{æ} \rangle$), macronised vowels ($\langle \bar{a} \rangle$, $\langle \bar{e} \rangle$, $\langle \bar{i} \rangle$, $\langle \bar{o} \rangle$, $\langle \bar{u} \rangle$, $\langle \bar{y} \rangle$), and standard Latin script; out-of-vocabulary characters map to an unknown symbol at inference time.

The *ge-* prefix, one of the most productive in Old English (past participles, perfective verbs, and some nouns and adjectives), is treated inconsistently across sources: some lemmas include it, some exclude it, and some use a parenthetical notation (*ge*)*lyfan* for optionality. In the corpus data, 12,275 instances (11.7% of the total) involve *ge-*, and 221 lemmas use the (*ge*) notation. We normalised these 221 parenthetical cases to canonical forms (selecting the prefixed or non-prefixed variant by predominant corpus usage), but residual inconsistencies remain: unmarked lemmas that are prefixed in one occurrence and unprefixed in another (e.g., *gelȳfde* as *gelȳfan* vs. *lȳfan*) cannot be conflated automatically without risking the merger of genuinely distinct *ge-/non-ge-* pairs, and near-even splits make the canonical choice partly arbitrary. These cases surface as the prefix-inconsistency category of Section 4.1.

Derivational morphology poses similar problems: whether *unriht* ‘injustice’ is a single entry or *un-* + *riht* varies across dictionaries, producing incompatible inventories. Dictionary headwords also follow editorial conventions (e.g., macrons for length: *cȳning* vs. corpus *cyning*). We normalised case but preserved macrons, as they encode phonological information relevant to morphological class.

Quality checks included duplicate detection: exact duplicates of complete word–lemma–POS triples were removed before splitting, so no identical triple spans partitions. Entries where the same word form receives different lemma and/or POS values were deliberately retained, as they repre-

Statistic	Value
Total annotated instances	105,063
Training set	84,050
Development set	10,506
Test set	10,507
Unique word types	90,752
Unique lemmas	37,627
Character vocabulary	98 symbols
Identity transf. (% of train)	23,748 (28.3%)
Unknown POS (X) (% of total)	4,605 (4.4%)
<i>ge-</i> prefix (% of total)	12,275 (11.7%)
Inconsistent types (% of types)	10,256 (11.3%)

Table 3: Dataset statistics after harmonisation (corpus-only condition). The reference base for each percentage is indicated per row.

sent either genuine ambiguity or the inconsistencies analysed in Section 4. Deduplication scripts and logs are in the project repository (Section 5).

After harmonisation, the corpus-only training set comprises 84,050 instances, 10,506 development instances, and 10,507 test instances (80/10/10 random split, seed 42), covering 90,752 unique word types and 37,627 unique lemmas. Table 3 provides the detailed statistics.

4 Annotation Quality Analysis

Annotation inconsistency is a recurrent problem in historical corpora, where editorial decisions about lemma assignment are often subjective and where different annotators may follow different conventions. In this section, we quantify the types and prevalence of inconsistencies in the Old English data.

4.1 Types of inconsistency

Inconsistency was operationalised at the level of the word type: after Unicode normalisation and case folding, with macrons preserved and no further spelling normalisation, a type was counted as inconsistent if two or more distinct lemma strings are assigned to its tokens anywhere in the harmonised corpus data. POS tags were not used to partition types at this stage, so forms whose different lemmas correspond to different POS (homographs) are included in the count; the manual categorisation below separates such genuine ambiguity from true inconsistency.

By this criterion, we identified 10,256 word types (11.3% of the vocabulary) assigned different lemmas across occurrences. A random sample of 200 inconsistent types was manually categorised by a single expert annotator following written guide-

lines with prototypical examples (in the project repository; see Section 5); as a single annotator was involved, the proportions are informed estimates rather than adjudicated figures (see Limitations). Four main categories emerge.

Genuine ambiguity (homographs). Some word forms correspond to multiple lemmas depending on context. For example, *þā* functions as a demonstrative pronoun (lemma: *sē*) or as an adverb (lemma: *þā*); *mæg* can be a verb ‘can’ (lemma: *magan*) or a noun ‘kinsman’ (lemma: *mæg*). These account for approximately 35% of inconsistent types and represent true linguistic ambiguity rather than annotation error.

Variant lemma forms. Different annotators assign different but related lemma forms: *hālgan* may be mapped to *hālig* or *hālgā*, which are spelling variants of the same lexeme. This category accounts for approximately 25% of cases.

Prefix inconsistency (*ge-*). The same verb may be lemmatised with or without the *ge-* prefix in different occurrences: *gelyfde* mapped to *gelyfan* in one instance and to *lyfan* in another. This accounts for approximately 20% of inconsistencies.

Annotation errors. A residual category of clear errors, e.g. a wrong POS tag or a phonologically unrelated lemma. This accounts for approximately 20% of cases.

With $n = 200$, the 95% confidence interval for a 35% proportion is about ± 6.6 points, enough for the coarse ranking but not fine-grained comparisons.

4.2 Identity transformations

A notable feature of the data is the high proportion of identity transformations, where the inflected form equals the lemma (e.g., *cynning* $\rightarrow$ *cynning*): 23,748 instances (28.3% of training data), including uninflected forms, nominative singular nouns, and verb infinitives. Though trivially easy, they teach the model the default ‘copy’ behaviour for invariable words and anchor the distinction between inflected and uninflected forms.

4.3 Impact of POS completeness

The corpus contains 4,605 instances (4.4% of the total) with unknown POS tags (X). Prior experiments on the same dataset (Martín Arista, 2026) demonstrated that the proportion of unknown POS tags in the training data has a direct and measurable impact

Property	YCOE	DOE	BT
Spelling conv.	Corpus	Editorial	Historical
POS format	UPOS	Lex. cat.	Lex. cat.
Macrons	Rare	Systematic	Variable
<i>ge-</i>	Variable	Systematic	Variable
Coverage	A–Z	A–I	A–Z

Table 4: Key discrepancies across the three data sources.

on joint lemma–POS model performance: each percentage point of unknown POS correlates with approximately 0.2% lemma accuracy degradation. When dictionary sources lacking POS annotations are added (raising the unknown POS proportion to 52.9%), lemma accuracy degrades by 13–23 absolute percentage points (Martín Arista, 2026). This finding advises never to train historical-language lemmatisers on data lacking POS annotations.

4.4 Inter-source consistency

The three data sources differ not only in format but in the lemma inventories they presuppose. Table 4 summarises key discrepancies.

The discrepancies span five dimensions. The YCOE records lemmas close to manuscript forms, the DOE uses normalised editorial headwords, and Bosworth-Toller follows nineteenth-century practice, so one lexeme may enter under three spellings. Only the YCOE maps transparently to UPOS; both dictionaries use lexical-category labels, and BT often lacks usable categories. Macrons are systematic in the DOE, rare in the YCOE, and variable in BT, so a naive merge treats *cynning* and *cȳning* as distinct. *ge-* handling is systematic only in the DOE. Coverage also differs (the DOE is restricted to A–I), so harmonisation problems are unevenly distributed across the lexicon.

Dictionary augmentation therefore introduces not just more data but a distributional shift: dictionaries document paradigmatic variation (all inflected forms), whereas the corpus reflects syntagmatic variation (forms attested in running text). We hypothesise that this mismatch, together with POS incompleteness, underlies the finding that adding dictionary data does not improve—and may degrade—accuracy (Martín Arista, 2026, see Section 6.1 and Table 8): the added entries shift the input distribution while raising the unknown-POS proportion from 4.4% to 52.9%. Disentangling the two factors would require ablations we leave for future work.

Category	%	Typical example
Strong verb ablaut	25	<i>sang</i> → * <i>singan</i>
Spelling variation	20	<i>forðon</i> → * <i>forðon</i>
Prefix handling	15	<i>gelyfde</i> → * <i>gelyfan</i>
Rare words	15	Hapax legomena
Annotation incon.	10	Gold standard conflicts
Homographs	10	Context-dependent
Other	5	Miscellaneous

Table 5: Distribution of lemmatisation errors by morphological category, based on a single-annotator categorisation of 500 sampled errors. * marks incorrect model predictions. Gold lemmas: *singan*, *forþam*, *lyfan*, respectively.

5 Morphological Error Analysis

We analysed the errors of the best character-level Transformer lemmatiser of Martín Arista (2026), which achieves 67.58% lemma accuracy on the corpus-only test set of Section 4 (the test-set counterpart of the ~68% development figure), examining only the lemma prediction component. The lemmatiser is a character-level encoder-decoder Transformer that maps the inflected form’s character sequence, plus its UPOS tag as a special input token, to the lemma’s character sequence; it was trained on the 84,050-instance corpus-only set of Section 4. Full specifications, hyperparameters, ablations, and test-set predictions are in the project repository at <https://huggingface.co/Nerthus-Project>, allowing independent assessment.

5.1 Error distribution by category

Table 5 presents the error distribution across morphological categories, from a random sample of 500 errors categorised by a single expert annotator following written guidelines (see the repository); as in Section 4.1, the proportions are informed estimates. With $n = 500$, the 95% confidence interval for a 25% proportion is about ± 3.8 points, enough for the coarse ranking discussed here.

5.2 Strong-verb ablaut

Strong verbs are the single largest source of error. Old English has seven strong-verb classes, each with a distinct ablaut pattern (vowel alternation across tense forms); Class III, for instance, alternates i–a–u–u (*drincan* ‘to drink’: *drincþ*, *dranc*, *druncon*, *gedruncon*). The model must learn seven such patterns while also handling spelling variation within each. Table 6 shows the four principal parts per class: the vowel of a past-tense form maps back

Cl.	Lemma	Inf.	Past	Past pl.	Ptcp.
I	<i>rīdan</i>	ī	ā	i	i
II	<i>bēodan</i>	ēo	ēa	u	o
III	<i>drincan</i>	i	a	u	u
IV	<i>beran</i>	e	æ	æ	o
V	<i>sprecan</i>	e	æ	æ	e
VI	<i>faran</i>	a	ō	ō	a
VII	<i>hātan</i>	ā	ē	ē	ā

Table 6: Ablaut patterns of the seven Old English strong-verb classes. Each column shows the stressed vowel in the relevant principal part.

to a different infinitive vowel, and this transformation is class-dependent and not recoverable from a single form without knowing the class.

Irregular verbs compound the difficulty. Suppletive verbs (*bēon/wesan* ‘to be’, *gān* ‘to go’) have stems so different that character-level matching fails—no cue links *eom* to *bēon*—while preterite-present verbs (*cunnan* ‘to know’, *magan* ‘to be able’) add further irregularity, as their present forms follow strong-verb past patterns.

5.3 Spelling variation

Spelling variation accounts for 20% of errors and affects all POS categories. Three subtypes stand out:

- Thorn/eth interchange: <þ> and <ð> are used interchangeably across (and within) manuscripts. A model trained on *forþam* may fail on *forðam*.
- Vowel variants: *æ* alternates with *a* or *e*; *y* alternates with *i* (e.g., *cyning/kyning*).
- Consonant doubling: inconsistent gemination (*mann/man*; *wolcenn/wolcen*).

Character-level data augmentation—simulating these alternations during training—has been shown to improve accuracy by approximately 2 percentage points (Martín Arista, 2026), but cannot fully resolve variation patterns not represented in the augmentation rules.

5.4 Prefix handling

The *ge-* prefix is the most frequent source of prefix-related errors. Because *ge-* was highly productive in Old English and its inclusion in the lemma varies across lexicographic traditions, the model faces systematic uncertainty. In 15% of errors, the model correctly identifies the stem but incorrectly includes or excludes the prefix.

POS	Approx. acc.	Main difficulty
ADP	~85%	Few errors
CCONJ	~83%	Few errors
DET	~80%	Paradigm size
ADV	~75%	Spelling variation
NOUN	~68%	Declension complexity
ADJ	~65%	Weak/strong inflection
VERB	~55%	Ablaut, <i>ge-</i> prefix
PROPN	~45%	Idiosyncratic forms

Table 7: Approximate accuracy by POS category. Estimates based on error analysis of a stratified sample.

Other productive prefixes causing errors include *un-* (negation), *for-* (intensification/destruction), *be-* (transitivisation), and *ā-* (completion). Unlike *ge-*, these prefixes typically change the meaning and should be retained in the lemma, but the model sometimes strips them by overgeneralising the *ge-* removal pattern.

5.5 Rare words and hapax legomena

Old English has many low-frequency words; many test items appear only once or twice, giving minimal training signal. Proper nouns (personal and place names) are especially affected, with idiosyncratic spellings and inconsistent lemmatisation. The model falls back on character-level heuristics that succeed only when a form closely resembles a more frequent training example.

5.6 Error analysis by POS category

Table 7 shows accuracy broken down by POS category. Verbs have the lowest accuracy, driven by ablaut patterns and the *ge-* prefix. Proper nouns are the next most difficult category. Function words (adpositions, conjunctions, determiners) achieve the highest accuracy because they are frequent and have few inflected forms.

6 Discussion: Lessons for Low-Resource Languages

The challenges documented above are not unique to Old English. Many low-resource languages—historical and contemporary—share limited annotated data, rich morphology, spelling variation or the absence of a standard orthography, and heterogeneous sources with incompatible conventions. We now draw lessons from the case study.

6.1 Data quality and data quantity

The experiments of Martín Arista (2026) show that augmenting training data with dictionary entries—

Condition	Train	Unk. POS	Test acc.
Corpus only	84,050	4.4%	67.58%
+ DOE + BT	~165,750	52.9%	44.6–54.6%

Table 8: Corpus-only vs. dictionary-augmented training. Results from Martín Arista (2026); the augmented condition is 13–23 pp below corpus-only across configurations, with per-configuration figures documented in the project repository.

though it roughly doubles the dataset—does not improve and may substantially degrade performance (Table 8). The critical factor is not data volume but quality, consistency, and relevance to the target distribution: corpus data (attested forms in running text) provides a richer signal than decontextualised dictionary paradigms.

6.2 Annotation consistency

If 11.3% of word types are inconsistently lemmatised, no model can exceed ~89% accuracy even with perfect learning, and the effective ceiling is lower because inconsistencies concentrate in high-frequency ambiguous forms. Low-resource projects should run inter-annotator agreement studies and document explicit annotation guidelines before building NLP tools.

6.3 POS annotations

The degradation from missing POS tags (4.4% to 52.9% unknown POS, costing 13–23 points; Martín Arista, 2026) shows POS information is not optional: joint lemma–POS models learn conditioned patterns (e.g., *-e* maps differently for nouns vs. adjectives), and incomplete POS disrupts this. Projects should prioritise POS annotation alongside lemmatisation.

6.4 Format standardisation

The sources’ heterogeneity—differing column orders, POS conventions, character inventories, and lemma spellings—required preprocessing that consumed a disproportionate share of effort. Digitisation projects should adopt standard formats from the outset (e.g., CoNLL-U) and document conventions explicitly.

6.5 Morphology-specific augmentation

Character-level augmentation mimicking attested variation (thorn/eth interchange, vowel alternation, consonant gemination) yields meaningful gains

(~2 points) at zero annotation cost. For other languages with known orthographic variation, philologically grounded augmentation rules are a cost-effective strategy.

6.6 Character-level processing

Our experiments indicate character-level processing is both necessary and sufficient here: pre-trained word embeddings and subword (BPE) tokenisation both reduced accuracy (Martín Arista and Metola Rodríguez, 2026). A plausible reading is that Old English lemmatisation is primarily a formal task in which distributional-semantic information adds noise; whether this extends to other morphologically rich or orthographically variable languages is an empirical question.

6.7 Cross-linguistic applicability

The conclusions may generalise. Historical Germanic languages (Old High German, Old Saxon, Old Norse) share rich inflection, orthographic instability, and scarce resources; contemporary languages without codified writing face analogous variation; and agglutinative languages (Finnish, Turkish, Quechua) present comparable morphological complexity, so our finding that character-level models beat word- and subword-level approaches may transfer, though we have not tested this. The key question is whether lemmatisation in a given language is primarily morphological (character-level pattern recognition) or semantic (distributional disambiguation); where morphology dominates, we expect—subject to verification—character-level Transformers with morphology-specific augmentation to be a strong baseline.

7 Conclusion

This paper has analysed the data-side challenges of building an NLP lemmatisation pipeline for Old English, a paradigmatic low-resource historical language. Three conclusions stand out. First, annotation inconsistencies are pervasive: the 11.3% of word types with inconsistent lemma assignments set a hard accuracy ceiling and explain many apparent model errors, and the four categories we identify (genuine ambiguity, variant lemma forms, prefix inconsistency, annotation errors) each need different fixes. Second, data provenance matters as much as volume: combining the three sources without harmonisation introduces distributional shift and POS incompleteness that degrade performance, with the jump in unknown POS from 4.4% to

52.9% costing 13–23 points. Third, errors concentrate in well-defined categories—strong-verb ablaut, spelling variation, and prefix handling account for 60%—and are unevenly distributed (function words ~85% vs. verbs ~55%), so targeted strategies could yield real gains.

These implications extend beyond Old English: for any low-resource language, the quality, consistency, and completeness of annotated data set a ceiling no architecture can exceed. Future work should pursue formal inter-annotator agreement studies for Old English, measure the impact of correcting inconsistencies on downstream accuracy, develop verb-class-aware lemmatisation, and test whether these guidelines transfer to languages such as Old High German, Old Norse, and Old Irish.

Limitations

This study has several limitations. First, it is a single-language case study; the cross-linguistic claims of Section 6 remain hypotheses to be tested. Second, the categorisations of inconsistencies (Section 4.1) and errors (Section 5) rest on manual samples of 200 and 500 items, with 95% confidence intervals of roughly ± 7 and ± 4 points for mid-range proportions, and were carried out by a single expert annotator, so the proportions are informed estimates rather than adjudicated figures; a formal multi-annotator study is future work. Third, the error analysis presupposes the model of the companion paper (Martín Arista, 2026), whose full specifications are in the project repository at <https://huggingface.co/Nerthus-Project>. Fourth, the cross-language accuracies in Table 1 come from studies with different corpora and protocols and are orientation only. Fifth, harmonisation is lossy: mapping to UPOS discards morphological features, and (*ge*)-normalisation resolves only explicitly marked cases. Finally, the random 80/10/10 split does not control for lexeme overlap between training and test, so accuracy on genuinely unseen lexemes may be lower than reported.

Acknowledgements

This work was supported by grant PID2023149762NB-I00 (Agencia Estatal de Investigación) and grant PRXX24/00108 (Ministerio de Ciencia, Innovación y Universidades), which are gratefully acknowledged.

References

- Antonios Anastasopoulos and Graham Neubig. 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)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.
- Toms Bergmanis and Sharon Goldwater. 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)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.
- Anna Cichosz, Piotr Pęzik, Marcin Grabski, Magdalena Adamczyk, Paulina Rybińska, and Alicja Ostrowska. 2021. The VARIOE online morphological dictionary for YCOE.
- Sean Crist and Ondřej Tichý. 2014. [Bosworth-toller’s anglo-saxon dictionary online](#).
- Laura García Fernández. 2018. *The Lemmatisation of Old English Preterite-Present, Contracted, Anomalous and Class VII Strong Verbs*. Ph.D. thesis, University of La Rioja.
- Yosra Hamdoun Bghiyel. 2025. *The Lemmatisation of The York-Toronto-Helsinki Parsed Corpus of Old English Poetry with a Relational Database of Old English Dictionaries*. Ph.D. thesis, University of La Rioja.
- Antonette diPaolo Healey. 2012. The dictionary of old english web corpus. With John Price Wilkin and Xin Xiang.
- Antonette diPaolo Healey. 2018. The dictionary of old english: A to i.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- Mike Kestemont, Guy de Pauw, Renske van Nie, and Walter Daelemans. 2017. Lemmatization for variation-rich languages using deep learning. *Digital Scholarship in the Humanities*, 32(4):797–815.
- Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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)*, pages 2779–2795. Association for Computational Linguistics.
- Krister Lindén and Tommi Pirinen. 2009. Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*, pages 89–95, Odense, Denmark. Northern European Association for Language Technology (NEALT).
- Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.
- Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1493–1503. Association for Computational Linguistics.
- Javier Martín Arista. 2026. Neural architectures for old english lemmatisation: A comparative study of corpus-based and dictionary-augmented approaches. *Procesamiento del Lenguaje Natural*. Forthcoming.
- Javier Martín Arista, Sara Domínguez Barragán, Luisa Fidalgo Allo, Laura García Fernández, Yosra Hamdoun Bghiyel, Miguel Lacalle Palacios, Raúl Mateo Mendaza, Carmen Novo Urraca, Ana Elvira Ojanguren López, Esaúl Ruíz Narbona, Roberto Torre Alonso, and Raquel Veja Escarza. 2023. ParCorOE3. an open access annotated parallel corpus old english–english.
- Javier Martín Arista and Darío Metola Rodríguez. 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.
- Darío Metola Rodríguez. 2015. *Lemmatisation of Old English Strong Verbs on a Relational Database*. Ph.D. thesis, University of La Rioja.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043. European Language Resources Association.
- Marco Passarotti. 2004. Development and perspectives of the latin morphological analyser LEMLAT. *Linguistica Computazionale*, 20–21:397–414.

- Susan Pintzuk and Leendert Plug. 2001. The york-helsinki parsed corpus of old english poetry.
- Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*, volume 5 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511. Association for Computational Linguistics.
- Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying incorrect labels in the CoNLL-2003 corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226. Association for Computational Linguistics.
- Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. The york-toronto-helsinki parsed corpus of old english prose.
- Marta Tío Sáenz. 2019. *The Lemmatisation of Old English Weak Verbs on a Relational Database*. Ph.D. thesis, University of La Rioja.