Lost in Translation?
Exploring the Shift in Grammatical Gender from Latin to Occitan

Ahan Chatterjee[1,2], Matthias Schöffel[1,2],
Matthias Aßenmacher[2,3], Marinus Wiedner[4], Esteban Garces Arias[2,3]

[1]Bavarian Academy of Sciences (BAdW), Munich [2]LMU Munich
[3]Munich Center for Machine Learning (MCML) [4]University of Freiburg
Correspondence: ahan.chatterjee@badw.de

Abstract

The diachronic evolution from Latin to the Romance languages involved a restructuring of the grammatical gender system from a tripartite configuration (masculine, feminine, neuter) to a bipartite one (masculine, feminine) in most Romance languages. In this work, we introduce an interpretable deep learning framework to investigate this phenomenon at both lexical and contextual levels. First, we show that conventional tokenization strategies are insufficiently robust for this low-resource historical setting, and that our proposed tokenizer improves performance over these baselines. At the lexical level, we evaluate the contribution of morphological features to gender prediction. At the contextual level, we quantify the contributions of different part-of-speech categories to grammatical gender prediction. Together, these analyses characterize the distribution of gender information between the lemma and its sentential context. We make our codebase, datasets, and results publicly available at https://github.com/ahan-2000/Lost-in-Translation-.

1 Introduction

Despite substantial advances in natural language processing (NLP), contemporary research remains concentrated on fewer than two dozen of the nearly 7,000 languages spoken worldwide. The vast majority of historical and regional languages are categorized as low-resource languages, defined by data scarcity, minimal digital presence, and a lack of standardized resources (Singh, 2008). Medieval Occitan, a Romance language historically spoken in southern France, the Val d'Aran, and parts of Piedmont (cf. Figure 1), played an important role in medieval cultural and economic life all over Europe. Despite this, UNESCO currently classifies it as an endangered language (Mothe, 2024). Medieval Occitan presents many of the challenges typical of low-resource languages. In addition to severe data scarcity and the lack of annotated gold-standard

Figure 1: Historical spread of the Occitan language (Poujade et al., 2024).

resources, the language displays substantial instability: it shows extensive orthographic variation, with lexical items attested in multiple spellings both across and within texts (Garces Arias et al., 2023; Schöffel et al., 2025), as well as dialectal fragmentation stemming from the absence of a standardized norm (Zampieri et al., 2020). As a result, existing work consistently describes Occitan as a neglected low-resource Romance language with severe resource limitations (Woller et al., 2021).

As a direct descendant of Vulgar Latin (Pasquini and Serva, 2021), Occitan, too, underwent a transition from a tripartite to a bipartite gender system, as did most Romance languages. As Vulgar Latin evolved, morpho-phonological changes weakened the stable neuter category of Classical Latin, ultimately leading to the collapse and absorption of the neuters from the second declension class predominantly into the masculine gender (Szlovicsák, 2023). However, the specific factors that governed this reassignment in Occitan, whether semantic, phonological, or morphological, remain insufficiently studied, especially for nouns inherited from the third declension class (Marzo and Wiedner, 2025; Polinsky and Van Everbroeck, 2003). This work addresses this gap through a computational study of Medieval Occitan, examining how gram-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 276–296

matical gender information is distributed between morphological features and morpho-syntactic context, and how these two sources contribute to model predictions for nouns descended from the Latin neuters. Methodologically, we propose a general framework for disentangling lemma-internal and contextual signals in grammatical gender prediction for low-resource historical languages. Our study is based on annotated corpora spanning law (*Lo Codi*), medicine (*Albuc*), and poetry (*Croisade*). Using these heterogeneous but sparse resources, we examine how morphological features and morpho-syntactic context jointly contribute to gender prediction for nouns descended from the Latin neuter class, through the following research questions:

RQ1 (Lexical-Level Analysis): To what extent can the grammatical gender of Occitan nouns derived from the Latin neuter class be predicted by word-level features, including their phonological and morphological characteristics?

RQ2 (Contextual Analysis): How is grammatical gender information distributed between morphological features and morpho-syntactic context in Occitan, and how do these sources contribute to model predictions?

2 Related Work

During the medieval period, the collapse of Latin neuter nouns often led to their absorption into the masculine gender in Romance languages, i.e., for neuters from the second declension class ending in *-um* (Klingebiel, 2019; Loporcaro, 2018). Although morpho-phonological cues provide strong signals (e.g., nouns in *-a* are typically feminine, while many others default to masculine), there are important exceptions, such as consonant-final gender-ambiguous nouns such as *mar* ('sea'), as well as Grecisms in *-a*, e.g., *propheta* ('prophet'). These irregularities suggest that accurate gender assignment may require additional information, including stress patterns, Latin etyma, and, of course, sentence context, given that gender is not a morphological but a morpho-syntactic category and that Old Occitan lacks an overt gender system.

One core research question is how effectively grammatical gender can be assigned to a noun solely on the basis of its form, including its lexical, phonological, and morphological characteristics. Early work by Brugmann (1897) emphasized the critical role of both phonological and semantic cues in gender assignment. However, these approaches are largely rule-based and language-specific, limiting their generalizability across diverse languages or linguistic families. Classic typological research, such as Corbett (1991), highlights that noun gender assignment typically involves a combination of morpho-phonological cues and semantic principles (e.g., natural gender for animates; but see the Greek loanwords as mentioned before). In Occitan, purely semantic gender applies in certain contexts, but for inanimate nouns, formal phonological and morphological cues predominate in gender determination and sometimes even supersede semantic criteria, e.g. the Greek loanwords.

Computational studies have attempted to quantify and predict gender from lexical features. Rule-based approaches to gender assignment have been extensively developed for languages such as French, producing long lists of endings and their most probable genders (Lyster, 2006). Nastase and Popescu (2009) analyze the prediction of grammatical gender using orthographic features and report that using a word's orthographic form in a statistical classifier improves gender prediction beyond baseline. These studies confirm that morpho-phonological cues have strong predictive power for gender. However, purely form-based prediction is not enough. Recent work by Williams et al. (2019) took an information-theoretic approach to languages such as German and Czech, measuring how much of gender assignment can be explained by a noun's form, meaning, or inflection class. They found that a combination of features provides the best predictions, highlighting that no single feature (orthography, phonology, semantics) accounts for everything, which is supported by recent experimental evidence (Basirat et al., 2021). Chronologically, the literature progressed from early descriptive grammars and implicit rules to manual rule compilations, then to data-driven classification, and now to neural and interpretable models. For Occitan, however, published computational work is still sparse. While some nouns have inherent grammatical gender, sentence context helps to identify gender assignment. In Occitan, as in related Romance languages like French and Spanish, determiners and adjectives agree in gender with nouns, participles, or pronouns. For example, the presence of the feminine article *la* before a noun signals that the noun is feminine in that context. Thus, the noun *torista* ('visitor') may be ambiguous in isolation, but in the phrase *la torista*, the

article disambiguates it as feminine in this context. While nouns predominantly have fixed grammatical gender, a few remnants, primarily from Latin neuter, exhibit atypical behaviour. Early computational work by Cucerzan and Yarowsky (2003) demonstrated that combining morphological analysis with contextual information significantly improves grammatical gender identification. Using a small annotated lexicon together with contextual cues such as co-occurrence with gendered articles and adjectives, their approach infers the gender of previously unseen words with high accuracy.

3 Data Description

The primary dataset for this study is drawn from three key Medieval Occitan sources. The first, *Lo Codi*, was annotated by Tobias Schmid as part of the ALMA Project (Heidelberg Academy of Sciences and Humanities, Bayerische Akademie der Wissenschaften, Academy of Sciences and Literature Mainz, 2025; Prifti et al., 2023). The second, the *Chanson de la Croisade Albigeoise*, was prepared and revised by Marinus Wiedner. The third source is the DOM Dictionary Project (Bayerische Akademie der Wissenschaften, 2025). The resulting annotated dataset comprises Latin–Occitan pairs, including Latin words, their corresponding Occitan lemmata, and the grammatical gender of each form, with a data distribution of 40.85% of unique lemmas from the *DOM* data source, 46.39% from *Lo Codi*, and 12.76% from *Croisade*. In addition, we use raw Occitan texts to analyze contextual cues (cf. Appendix A.2).

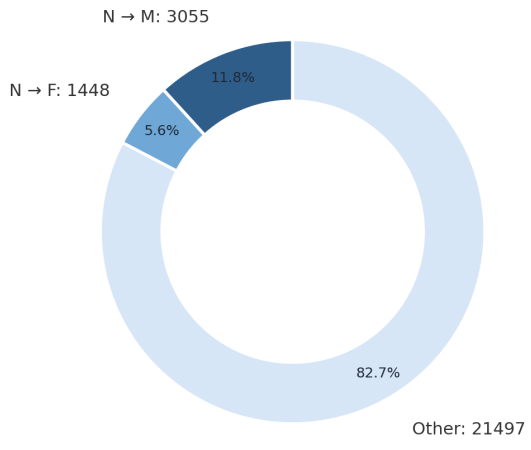

Figure 2: Gender Shift Frequencies across all three investigated corpora.

Our initial analysis confirms the complete absorption of the Latin neuter class into masculine and feminine genders in Occitan. As Figure 2 shows, the dominant shift is from neuter to masculine (3,055 cases), while a smaller but still substantial number of nouns shift to feminine (1,448 cases). A closer look at the orthographic features driving this divergence (Figure 7, Appendix A.1) reveals that specific endings are highly predictive of the outcome. The role of endings in this process is more nuanced than raw frequency counts suggest. While the ending *-um* is overwhelmingly associated with a masculine outcome, it is also, paradoxically, the single most common ending for nouns that become feminine. This is explained by the fact that the overall shift to masculine was far more prevalent, meaning any frequent neuter ending would appear dominant in that category. This finding underscores the importance of moving beyond simple frequency counts to understand the underlying mechanisms. By contrast, other endings, such as *-ia* and *-la*, provide a clearer signal and correlate strongly with feminine outcomes, further supporting the importance of morphological cues in gender (re)assignment.

4 Preliminary Analysis: Model and Tokenization Selection

We run a targeted set of probes to select (i) the embedding family that best captures Medieval Occitan variation in a Latin–Occitan setting and (ii) a tokenization policy that is robust to heavy orthographic noise. Concretely, we evaluate embedding models under three complementary criteria: (P_1) a frozen-encoder linear probe for Occitan gender prediction, (P_2) retrieval of Occitan orthographic variants given a Latin lemma, and (P_3) unsupervised structure in the embedding space via clustering.

4.1 Embedding Model Selection

We conduct a preliminary backbone selection study comparing FastText, mBERT, and ByT5 on three complementary probes: frozen gender prediction, Latin→Occitan variant retrieval, and clustering of Occitan forms. mBERT performs best across all three probes, suggesting that it provides the most reliable representation of lexical and cross-lingual structure for Medieval Occitan. We therefore adopt it as the embedding backbone in all downstream experiments (cf. Appendix B.1 for detailed results).

4.2 Tokenization Policy Selection

Medieval Occitan exhibits frequent spelling variation and sparse type coverage, making segmentation a primary bottleneck. We therefore evaluate tokenization policies via (i) OOV rate and (ii) masked token recovery accuracy on an Occitan masked language modeling (MLM)-style objective.

Example 1	**Example 2**
`de lay del primpcipat`	`En sa cambra secretament`
Hybrid: `de, la, y, del, pri, mp, ci, pat`	*Hybrid:* `En, sa, cambra, s, ec, ret, amen, t`

Figure 3: Examples of hybrid tokenization capturing orthographic and morphological variation in Medieval Occitan. In *primpcipat*, the subword `mp` isolates consonant-cluster variation, helping the model remain robust to spelling differences such as `nc`/`mp`. In *secretament*, the final `t` is segmented separately, reflecting a common Old Occitan alternation where the adverbial `-t` may be elided (e.g., *secretamen* vs. *secretament*). Such fine-grained segmentation supports better generalization across predictable historical variants.

Subword vs. Hybrid Segmentation. We evaluate tokenization policies using (i) **OOV rate**, defined as the proportion of tokens mapped to `[UNK]`, and (ii) **masked token recovery**, defined as top-1 accuracy at masked *subword* positions. In this experiment, a hybrid policy (Occitan-adapted BPE with a word-level fallback) preserves full coverage (zero `[UNK]`) and yields the best masked recovery (25.23%), indicating that explicit fallback coverage is crucial while still benefiting from corpus-adapted subword units (cf. Appendix B.2). Qualitatively, the hybrid tokenizer also produces more interpretable subword boundaries than generic WordPiece segmentation (Fig. 3).

4.3 Domain-Adaptive MLM Fine-Tuning

Given the consistent advantages of mBERT and hybrid tokenization, we apply domain-adaptive MLM fine-tuning for 10 epochs with identical hyperparameters across runs and evaluate on a held-out validation split. Fine-tuning substantially improves fit to the Occitan corpus: standard MLM adaptation reduces validation perplexity from 942.85 to 10.44, while the hybrid-vocabulary variant attains the best validation perplexity (9.52). Since perplexity is tokenization-dependent, we interpret these values as within-configuration diagnostics; taken together with the probing and tokenization results, they motivate our final setup: mBERT with a hybrid tokenizer and domain-adaptive MLM fine-tuning.

Based on the preliminary analyses, we adopt mBERT with hybrid tokenization and MLM adaptation as the backbone for all subsequent experiments. We now address our first research question by investigating grammatical gender prediction from lexical features alone, setting up a contrast with the contextual models introduced next.

5 Methodology

5.1 Lexical Grammatical Gender Prediction

Grammatical gender is a nominal classification system; in Occitan, it is bipartite (masculine and feminine) and is typically realized through noun morphology and agreement. In this section, we examine gender assignment based solely on lexical information, without relying on sentential context.

5.1.1 Feature Representation and Engineering

Data Normalization. We lowercase lemmas and apply Unicode NFKD normalization, stripping combining diacritics for character-level features; original forms are retained for the embeddings.

Task and Imbalance Handling. We predict *Occitan grammatical gender* as a bipartite label $y \in \{\mathrm{M}, \mathrm{F}\}$. Since outcomes are highly skewed (the Latin neuter most frequently maps to Occitan masculine), we use class-weighted training and focal loss; we also perform ablations to quantify the contribution of each feature group.

Morphological and Phonotactic Features. From both Latin and Occitan lemmas, we extract initial word substrings and suffix character n-grams ($1 \leq n \leq 4$), emphasizing word-final cues consistent with Romance gender marking (Table 1). We further encode syllabic shape using (i) vowel-run syllable count $S(w)$ and (ii) VC templates $P(w)$ (Table 2), and include length features $|w_{\text{lat}}|$, $|w_{\text{occ}}|$, their difference, and ratio.

Lang	**Lemma**	**Initial Substrings**	**Suffix**
Latin	*domus*	`do`	`us`
Occitan	*dom*	`do`	`om`

Table 1: Example of initial word substrings/suffix bigram extraction ($n = 2$) for aligned Latin–Occitan lemmas.

Stress as a Coarse Proxy. We include a lightweight stress-position proxy (ultimate/penultimate/antepenultimate), derived from a syllable-weight heuristic: monosyllables

Word (w)	**Language**	**Syllable Count** $S(w)$
festum	Latin	2
festa	Occitan	2
tempus	Latin	2
temps	Occitan	1
Word (w)	**Language**	**VC Pattern** $P(w)$
festum	Latin	CVCCVC
festa	Occitan	CVCCV
tempus	Latin	CVCCVC
temps	Occitan	CVCCC

Table 2: Syllabic structure features: vowel-run syllable count $S(w)$ and VC template $P(w)$.

are stressed on the only syllable; disyllables on the penult; for polysyllables, we stress the penult if it is heavy (long vowel or closed syllable), otherwise the antepenult. We treat this feature as an approximate cue rather than as a definitive phonological annotation.

Embedding Features We use frozen pretrained representations as *feature extractors* and compare them as alternative embedding feature sets rather than concatenating them: FastText (subword n-gram composition), mBERT, and ByT5. For mBERT/ByT5, each lemma is embedded in isolation and represented by mean pooling over subword/byte final-layer states; FastText uses standard word-type vectors. These embeddings are then used directly as input features to the downstream classifier.

5.1.2 Experimental Setup

We evaluate feature sets using lemma-grouped 10-fold cross-validation to prevent leakage across orthographic variants. Let $\mathcal{D} = \{(x_i, y_i, \ell_i)\}_{i=1}^{N}$, where x_i are features, $y_i \in \{\mathrm{M}, \mathrm{F}\}$ is the label, and ℓ_i is a lemma ID; folds are formed over lemmas and scores are averaged across folds. We evaluate a diverse set of classifiers to cover complementary inductive biases, ranging from transparent linear models (Logistic Regression), to non-linear tree ensembles (Random Forest, XGBoost), to sequence-aware neural architectures (FFN, BiLSTM, and attention-based variants). This design allows us to test whether grammatical gender is primarily recoverable from simple lexical cues or whether stronger performance requires models that capture higher-order or sequential interactions in the feature space. Hyperparameters are tuned with Optuna (Bayesian optimization), maximizing validation Macro-F1 within the cross-validation protocol.

Although lexical features are highly informative, they do not fully determine grammatical gender in all cases. For nouns such as *psalmista*, the intended gender may only become clear from sentence-level agreement cues, especially the article (*lo*/*la*). We therefore turn to our second research question, examining whether contextual information improves prediction beyond lemma-internal evidence alone.

5.2 Context-based Grammatical Gender Prediction

In the previous section, we examined the contribution of lexical features to gender prediction in isolation. Here, we study the contribution of *sentence-level context* as a second source. In Occitan, gender is jointly encoded by the noun and its agreeing dependents (articles, adjectives, and other modifiers); we exploit this distributed encoding as a prediction signal when lemma-internal cues are weak.

5.2.1 Dataset & Data Preparation

We use ∼130k tokens of unannotated Occitan texts spanning multiple genres (law, poetry, and medicine). We normalize the corpus by lowercasing, stripping diacritics, and standardizing punctuation. Because parallel Latin sentences are unavailable, we rely on an existing Occitan–Latin lemma lexicon and link each Occitan lemma (cf. Algorithm 1) occurrence to its containing sentence, yielding contextual instances for downstream analysis.

5.2.2 Proposed Methodology

We quantify the contribution of sentential context to Occitan gender prediction using three input settings. Each instance is (X, i, L, G_L, y), where $X = (x_1, \ldots, x_T)$ is an Occitan sentence, i indexes the target noun token $w = x_i$, L is its Latin lemma with Latin gender $G_L \in \{\mathrm{M}, \mathrm{F}, \mathrm{N}\}$, and $y \in \{\mathrm{M}, \mathrm{F}\}$ is the gold Occitan label. A pretrained encoder produces contextual states:

$$H = \mathrm{BERT}_\theta(X) = (h_1, \ldots, h_T), \qquad h_t \in \mathbb{R}^d. \tag{1}$$

All configurations share the same MLP head f_ϕ, with

$$p(y \mid r) = \mathrm{softmax}(f_\phi(r)). \tag{2}$$

(i) Word-only. We form a lexical representation from isolated embeddings and Latin metadata:

$$r_{\text{word}} = [\, e(w);\ e(L);\ \mathrm{onehot}(G_L) \,]. \tag{3}$$

Algorithm 1 Construction of Occitan–Latin Lemma–Gender Dataset

Require: Raw Occitan corpus D
Require: Occitan–Latin lemma lexicon $\mathcal{L}$
Require: Similarity function $\textsc{Sim}(\cdot,\cdot)$ (cf. C.1)
Ensure: Table T of aligned lemmas, contexts and genders

$D_{\text{pos}} \leftarrow \textsc{PoSTag}(D)$ ▷ tag every token in the corpus (cf. E)
$N \leftarrow \{(w, s, \ell_{\text{oc}}) \in D_{\text{pos}} : \text{PoS}(w) = \text{NOUN}\}$ ▷ collect noun tokens with sentence s and lemma ℓ_{oc}
$T \leftarrow \emptyset$
for all $(w, s, \ell_{\text{oc}}) \in N$ **do** ▷ iterate over all noun instances
  if $\exists(\ell_{\text{oc}}, \ell_{\text{la}}, g_{\text{oc}}, g_{\text{la}}) \in \mathcal{L}$ **then**
    $(\hat{\ell}_{\text{oc}}, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}}) \leftarrow (\ell_{\text{oc}}, \ell_{\text{la}}, g_{\text{oc}}, g_{\text{la}})$ ▷ exact lemma match
  else
    Find $(\ell', \ell'_{\text{la}}, g'_{\text{oc}}, g'_{\text{la}}) \in \mathcal{L}$ s.t. $\ell' = \arg\max_{\tilde{\ell} \in \mathcal{L}} \textsc{Sim}(\ell_{\text{oc}}, \tilde{\ell})$ and $\textsc{Sim}(\ell_{\text{oc}}, \ell') \geq \tau$ (τ=0.85).
    if a candidate exists **then**
      $(\hat{\ell}_{\text{oc}}, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}}) \leftarrow (\ell', \ell'_{\text{la}}, g'_{\text{oc}}, g'_{\text{la}})$ ▷ fuzzy lemma match
    else
      continue ▷ skip if no reliable match is found
    end if
  end if
  Append row $(\hat{\ell}_{\text{oc}}, s, \hat{\ell}_{\text{la}}, \hat{g}_{\text{oc}}, \hat{g}_{\text{la}})$ to T ▷ store Occitan lemma, context, Latin lemma, and both genders
end for
return T

(ii) Context-focused. To target the noun within its sentence, we use noun-conditioned attention over H (cf. Architecture in Figure 4):

$$r_{\text{ctx}} = [\,\text{Attn}(h_i, H, H);\ e(L);\ \text{onehot}(G_L)\,]. \tag{4}$$

(iii) Masked-context. To isolate contextual cues, we mask the noun $x_i \leftarrow [\text{MASK}]$, re-encode $H^{\text{mask}} = \text{BERT}_\theta(X^{\text{mask}})$, and use the state at position i:

$$r_{\text{mask}} = [\,h_i^{\text{mask}};\ e(L);\ \text{onehot}(G_L)\,]. \tag{5}$$

The masked-context setting evaluates how much of a noun's gender can be recovered from the surrounding sentence alone. Because Occitan articles, adjectives, and other dependents inflect to agree with the noun, the surrounding sentence jointly encodes the noun's gender; we therefore read masked-context performance as a measure of this distributed encoding rather than as an independent contextual signal. Comparing the word-only, context-focused, and masked-context configurations lets us bound how much predictive signal each source contributes (cf. Algorithm 2).

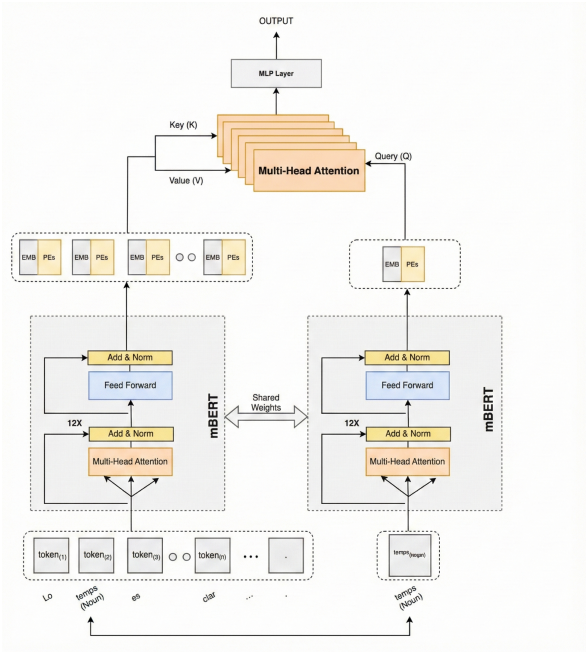

Figure 4: Proposed architecture to assess the impact of contextual cues on nouns' grammatical gender prediction.

All experiments use 3-group k-fold cross-validation, preserving the class distribution of Occitan gender labels across splits. To prevent label leakage, splits are constructed at the lemma level so that orthographic variants of the same lemma do not appear in both training and validation folds. A fixed random seed (13) is used throughout for reproducibility. We further analyze which contextual categories drive predictions by aggregating token-level contributions by PoS tag, e.g., determiners, adjectives, and verbs (cf. Appendix E), yielding tag-wise estimates of their influence on gender prediction.

6 Results and Discussion

6.1 Lemma-level gender prediction

Table 3 reports mean Accuracy and Macro-F1 (10-fold CV) for lemma-level gender prediction across model families and embedding feature sets. Overall, neural sequence models outperform shallow baselines, and attention generally yields further gains. The best Macro-F1 is achieved with a 2× BiLSTM + multi-head self-attention (MHSA) model trained with imbalance-aware objectives (focal loss/class weighting), yielding the strongest results with multilingual encoders (mBERT/ByT5), with pretrained representations providing more informative lexical cues than static embeddings.

FastText (baseline = 0.7734)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.7606	0.0128	1.66%
Meta-features	0.7640	0.0094	1.22%
Occitan n-grams	0.7667	0.0067	0.87%
Syllable counts	0.7667	0.0067	0.87%
VC patterns	0.7710	0.0024	0.31%
Stress patterns	0.7746	-0.0012	-0.16%
mBERT (baseline = 0.8224)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.8092	0.0132	1.61%
Meta-features	0.8168	0.0056	0.68%
Occitan n-grams	0.8169	0.0055	0.67%
Syllable counts	0.8194	0.0030	0.37%
VC patterns	0.8220	0.0004	0.05%
Stress patterns	0.8239	-0.0015	-0.18%
ByT5 (baseline = 0.8106)			
Block	**F1**	**Δ**	**% drop**
Latin n-grams	0.7958	0.0148	1.83%
Meta-features	0.8006	0.0100	1.23%
Occitan n-grams	0.8035	0.0071	0.88%
Syllable counts	0.8087	0.0019	0.23%
VC patterns	0.8091	0.0015	0.19%
Stress patterns	0.8123	-0.0017	-0.21%

Table 4: Feature ablation comparison across FastText, mBERT, and ByT5.

6.2 Feature ablation

To quantify the contribution of each feature group, Table 4 removes one block at a time from the best configuration (per embedding set) and reports the resulting Macro-F1 drop. Latin and Occitan character n-grams, especially suffix cues, are the most influential, producing the largest decreases (1.6–1.8 Macro-F1 points). Length/meta-features are the next strongest contributors (0.7–1.3 points), while VC templates and stress proxies have comparatively smaller effects.

6.3 Feature attributions (SHAP)

SHAP attributions (cf. Figure 5) are broadly consistent with the ablation results: suffix features and length-related meta-features dominate the decision signal across embedding modalities. Stress-related cues occasionally receive non-trivial attribution; however, since stress is derived from a heuristic proxy, we interpret these effects cautiously.

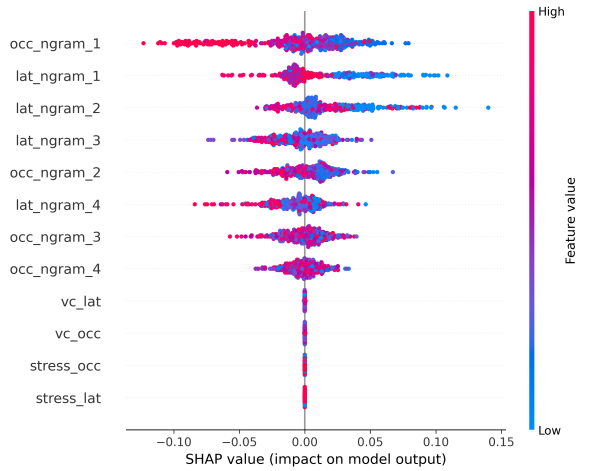

Figure 5: SHAP summary plot for the best-performing lemma-level model.

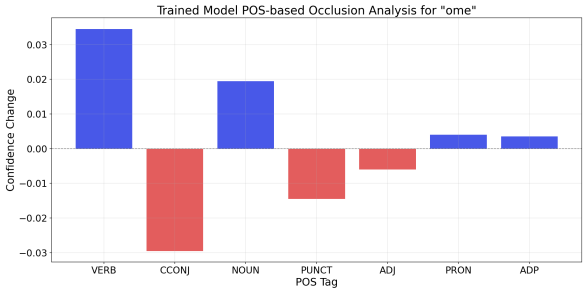

Figure 6: Example in which the lemma-only model misclassifies *ome* as feminine: *"aquill ome qui tenunt uera fe e pois tornunt en heresia deuent auer atrestal pena cum li altre e tant maior quant maior peccat ill fant."* With sentence-level context, the prediction shifts to masculine, with attribution distributed across agreement-bearing tokens such as *aquill*, illustrating how gender information distributed between the lemma and its local context lets the contextual model recover the correct label when the lemma representation alone is insufficient.

6.4 Impact of Contextual Cues

We evaluate contextual induction following Algorithm 2. Table 5 compares three mBERT-based configurations. Adding sentence context yields a substantial gain over the word-only baseline (Macro-F1: 0.665 → 0.929). Masking the noun remains substantially better than word-only (Macro-F1 0.902), but underperforms the unmasked setting, consistent with the noun form carrying most gender signal while context provides additional disambiguation. In cases where the lemma representation alone is insufficient, the contextual model produces a different prediction than the lemma-only model, as illustrated in Figure 6.

Model	Accuracy	Macro F1
ByT5		
Logistic Regression	0.6418 ± 0.0372	0.6139 ± 0.0440
Random Forest	0.7333 ± 0.0265	0.6974 ± 0.0333
XGBoost	0.7141 ± 0.0364	0.7354 ± 0.0492
BiLSTM	0.7550 ± 0.0410	0.7427 ± 0.0397
2×BiLSTM+Attn. (CE, 100 ep)	0.7902 ± 0.0389	0.7867 ± 0.0354
2×BiLSTM+MHSA (CE, LS=0.1, 100 ep)	0.8380 ± 0.0249	0.8106 ± 0.0237
FastText		
Logistic Regression	0.6578 ± 0.0384	0.6319 ± 0.0459
Random Forest	0.7167 ± 0.0392	0.6917 ± 0.0397
XGBoost	0.7152 ± 0.0524	0.7068 ± 0.0528
Feedforward NN (FFN)	0.7266 ± 0.0231	0.7113 ± 0.0283
BiLSTM	0.7763 ± 0.0323	0.7561 ± 0.0293
BiLSTM + Attn. (CE, 100 ep)	0.7480 ± 0.0492	0.7224 ± 0.0536
2×BiLSTM+Attn. (CE+ES, 100 ep)	0.7904 ± 0.0349	0.7512 ± 0.0354
2×BiLSTM+MHSA (CE+LS=0.1, 50 ep)	0.8134 ± 0.0416	0.7734 ± 0.0344
mBERT		
Logistic Regression	0.6749 ± 0.0254	0.6572 ± 0.0273
Random Forest	0.7287 ± 0.0397	0.7097 ± 0.0441
XGBoost	0.7352 ± 0.0402	0.7168 ± 0.0425
BiLSTM	0.7525 ± 0.0355	0.7252 ± 0.0384
BiLSTM+Attn. (CE, 100 ep)	0.7846 ± 0.0272	0.7419 ± 0.0309
2×BiLSTM+Attn. (FL + Class Wts, 50 ep)	0.7840 ± 0.0369	0.7460 ± 0.0341
2×**BiLSTM + MHSA (CE, LS=0.1, 100 ep)**	$\mathbf{0.8327 \pm 0.0365}$	$\mathbf{0.8224 \pm 0.0385}$

Table 3: Mean $\pm$ SD over 10-fold lemma-grouped cross-validation for lemma-level grammatical gender prediction. Best per-embedding rows are bolded. Under the shared 2×BiLSTM+MHSA head, the mBERT advantage over ByT5 is significant at the instance level on pooled out-of-fold predictions (paired bootstrap, Δ Macro-F1 $= +0.0395$, 95% CI $[+0.0250, +0.0543]$, $p < 10^{-6}$; full procedure in Appendix F.1).

Experiment (mBERT)	Acc.	Macro F_1
Word-only Model	0.808 ± 0.154	0.665 ± 0.108
Context model (explicit noun attention)	$\mathbf{0.979 \pm 0.012}$	$\mathbf{0.929 \pm 0.034}$
Context model (noun masked)	0.977 ± 0.008	0.902 ± 0.097

Table 5: Classification performance across the three experimental settings. The word-only baseline is lower than in the lemma-level experiments because, for comparability with the contextual models, it uses only the Latin lemma, Occitan lemma, and Latin gender, without the richer lemma-level feature set introduced earlier.

Δ Statistic	Mean	95% CI
Δ_1^{prob}	0.283	[0.281, 0.285]
Δ_2^{prob}	0.279	[0.277, 0.281]
$\Delta_1^{\log p}$	0.294	[0.285, 0.303]
$\Delta_2^{\log p}$	0.340	[0.334, 0.346]

Table 6: Mean values and 95% confidence intervals for the Δ statistics.

To examine whether context increases confidence in the correct label, we report mean probability and log-probability deltas for the gold class (Table 6). Both Δ_1 (context vs. word-only) and Δ_2 (masked-context vs. word-only) are positive, indicating that contextual cues systematically raise the model's confidence in the ground-truth class.

6.5 Model Explainability

For the context model, we use 8-head attention with the target noun state as query and sentence states as keys/values. Figure 8 in Appendix G illustrates that attention concentrates on the noun token, with the associated article typically receiving the next-highest mass, matching Occitan morpho-syntax where articles (e.g., *lo*/*la*) are strong gender cues. Across heads, attention mass is broadly distributed, with no single head consistently specializing in a particular Part-of-Speech (PoS) category.

To quantify which contextual categories contribute most, we run PoS-conditioned occlusion (cf. Algorithm 3 in Appendix C.3) and aggregate token-level deltas by tag. Table 7 shows that nouns contribute the largest positive delta, followed by determiners and adjectives, consistent with gender information being distributed across the noun and its agreeing dependents.

PoS tag	Mean Δ	Count (n)	Sign-flip p
NOUN	$+0.0026$	39,521	$< 10^{-4}$
DET	$+0.0010$	24,042	$< 10^{-4}$
ADJ	$+0.0003$	29,920	$< 10^{-4}$
CCONJ	-0.0010	27,492	$< 10^{-4}$
ADP	-0.0007	26,577	$< 10^{-4}$
VERB	-0.0003	29,336	$< 10^{-4}$
PUNCT	-0.0001	30,408	0.096
PRON	-0.0002	23,893	0.997

Table 7: PoS-wise mean occlusion deltas with sign-flip permutation test (10,000 permutations, two-sided). NOUN, DET, and ADJ contribute reliably positive contextual evidence; CCONJ, ADP, and VERB contribute reliably negative effects; PUNCT and PRON are not significant. Effect magnitudes are small in absolute terms; we read them as stable but modest contextual cues. Full procedure in Appendix F.2.

7 Conclusion

Gender information in Medieval Occitan is distributed across two sources: lemma-internal morphology and sentence-level context. Suffix morphology carries the strongest single signal; articles, adjectives, and other agreeing dependents provide additional morpho-syntactic cues that may inform gender assignment in context-sensitive interpretation, and when the lemma alone is ambiguous, they can shift a model's prediction. Taken together, these findings support a two-layer view of gender in Medieval Occitan: lexical morphology provides the primary structural encoding, while agreement and contextual patterns, that is, morpho-syntactic cues, reflect its realization in usage. Methodologically, the work highlights that historical orthographic instability makes standard tokenization brittle; hybrid tokenization with domain-adaptive MLM enables models to exploit meaningful subword regularities while remaining robust to spelling variation. More broadly, the proposed lexical-versus-contextual comparisons and attribution analyses offer a useful framework for studying grammatical change in noisy historical corpora, though future work with richer gold annotation and improved morpho-syntactic resources would allow for more fine-grained analyses.

Limitations

It is important to acknowledge several limitations. First, while our corpus is genre-diverse, it remains relatively small and label-imbalanced (approximately 2:1 masculine-to-feminine), which may limit minority-class generalization despite mitigation via focal loss and class-weighted training. Second, key components of the preprocessing pipeline were set heuristically, most notably the fuzzy-matching threshold ($\tau = 0.85$) and the stress-position proxy, and our ablations suggest that the stress feature can introduce mild noise. Third, our PoS-conditioned analyses (cf. Appendix E) rely on automatic tagging, and our evaluation shows $\sim$71% tagging accuracy, implying that PoS-based attribution results may be biased by tagging errors. Finally, the contextual model is less reliable in sentences where the target noun occurs at sentence boundaries or where agreement-bearing cues (cf. Appendix I) are sparse, which motivates future work on boundary-aware modeling and richer syntactic supervision. More broadly, our conclusions pertain to the Latin-to-Occitan neuter collapse and should be tested across additional Medieval Romance varieties. Our experiments quantify how gender information is *distributed* between lexical and contextual sources for synchronic prediction; they do not directly test the diachronic question of what drove the historical reassignment of former Latin neuters, which requires parallel diachronic data and a different experimental design.

Ethics Statement

We affirm that our research adheres to the ACL Ethics Policy. This work uses publicly available datasets and involves no human subjects or personally identifiable information. All data and code, including preprocessing, modeling choices, and evaluation protocols, are released to enable reproducible research and further investigation. Our work is intended exclusively for research purposes, and we encourage careful interpretation of results, particularly in low-resource and historical language settings where annotation uncertainty and data scarcity are common.

Acknowledgments

Esteban Garces Arias sincerely thanks the Mentoring Program of the Faculty of Mathematics, Statistics, and Informatics at LMU Munich and the Munich Center for Machine Learning (MCML) for their ongoing support. Matthias Aßenmacher received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581 - BERD@NFDI.

References

Ali Basirat, Marc Allassonnière-Tang, and Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*, 7(1):20200048.

Bayerische Akademie der Wissenschaften. 2025. Dictionnaire de l'occitan médiéval (dom). https://dom.badw.de/fr/le-projet.html. Accessed: 25 November 2025.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Karl Brugmann. 1897. *The nature and origin of the noun genders in the Indo-European languages: A lecture delivered on the occasion of the sesquicentennial celebration of Princeton University*. C. Scribner's sons.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Silviu Cucerzan and David Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 40–47.

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.

Esteban Garces Arias, Vallari Pai, Matthias Schöffel, Christian Heumann, and Matthias Aßenmacher. 2023. Automatic transcription of handwritten old Occitan language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15416–15439, Singapore. Association for Computational Linguistics.

Heidelberg Academy of Sciences and Humanities, Bayerische Akademie der Wissenschaften, Academy of Sciences and Literature Mainz. 2025. Alma: Knowledge networks of medieval romance-speaking europe.

Kathryn Klingebiel. 2019. Occitan studies: Language and linguistics. *The Year's Work in Modern Language Studies*, 79(1):181–197.

Michele Loporcaro. 2018. *Gender from Latin to Romance: History, geography, typology*, volume 27. Oxford University Press.

Roy Lyster. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies*, 16(1):69–92.

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 and Short Papers)*, pages 1493–1503, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniela Marzo and Marinus Wiedner. 2025. Remarks on grammatical gender in romance. In *Parla, e sie breve e arguto. Festschrift für Maria Selig / Studies in Honor of Maria Selig*, ScriptOralia 147, pages 201–207, Tübingen. Narr.

Josiane Mothe. 2024. Shaping the future of endangered and low-resource languages—our role in the age of llms: A keynote at ecir 2024. In *ACM SIGIR Forum*, volume 58, pages 1–13. ACM New York, NY, USA.

Vivi Nastase and Marius Popescu. 2009. What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1377, Singapore. Association for Computational Linguistics.

Michele Pasquini and Maurizio Serva. 2021. Stability of meanings versus rate of replacement of words: an experimental test. *Journal of Quantitative Linguistics*, 28(2):95–116.

Maria Polinsky and Ezra Van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from latin to french. *Language*, 79(2):356–390.

Clamenca Poujade, Myriam Bras, and Assaf Urieli. 2024. CorpusArièja: Building an annotated corpus with variation in Occitan. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 66–71, Torino, Italia. ELRA and ICCL.

Elton Prifti, Wolfgang Schweickard, Maria Selig, and Sabine Tittel. 2023. Sprachdatenbasierte modellierung von wissensnetzen in der mittelalterlichen romania (alma): Projektskizze. *Zeitschrift für romanische Philologie*, 139(2):301–332.

Matthias Schöffel, Esteban Garces Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025. Unveiling factors for enhanced pos tagging: A study of low-resource medieval romance languages. *arXiv preprint arXiv:2506.17715*.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Béla Szlovicsák. 2023. Preliminary examination of the latin neuter on inscriptions. *Acta Antiqua Academiae Scientiarum Hungaricae*, 62(4):419–434.

Marinus Wiedner. 2025. Cometa: Corpus de l'occitan médiéval comparatif et annoté: Provence et languedoc. Zenodo.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. 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 5734–5739, Hong Kong, China. Association for Computational Linguistics.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do not neglect related languages: The case of low-resource occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

A Data Description

A.1 Gender Shift by Lemma Ending

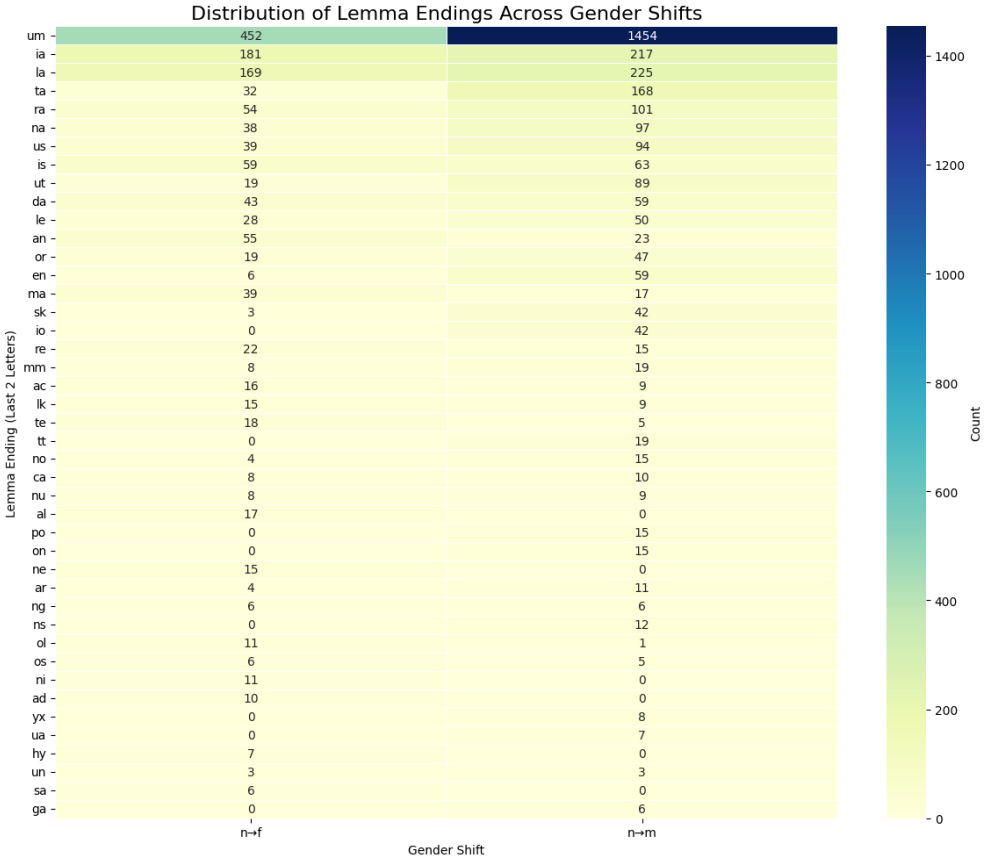

Figure 7: Gender shift frequencies for different lemma endings.

A.2 Lexical Diversity in Raw Occitan Texts

File	Tokens	Types	TTR	MATTR@50	MATTR@100	MATTR@500
Harley_3041.txt	443	227	0.512	0.852	0.776	0.512
Latin_901.txt	1105	492	0.445	0.830	0.739	0.537
Nouvelle_acquisition_française_11180.txt	4312	1324	0.307	0.824	0.735	0.520
Arsenal_6355.txt	11277	2491	0.221	0.849	0.762	0.545
Français_13504.txt	35802	6163	0.172	0.808	0.724	0.536
Roman_de_Flamenca.txt	46724	7852	0.168	0.861	0.787	0.585
Nouvelle_acquisition_française_11151.txt	20268	3151	0.155	0.766	0.661	0.429
Croisade_Albigeoise.txt	81305	12370	0.152	0.822	0.747	0.562
Add__10323.txt	53564	7609	0.142	0.850	0.764	0.552
Add__21218_ohne1-6.txt	31713	4485	0.141	0.805	0.717	0.507
Français_13503.txt	35352	4773	0.135	0.820	0.722	0.493
Français_13509.txt	62489	7867	0.126	0.849	0.763	0.548
Français_2232.txt	34217	4194	0.123	0.799	0.705	0.489
Lays_d-amors.txt	125475	13515	0.108	0.791	0.698	0.493
Français_1049.txt	130220	10425	0.080	0.778	0.680	0.463

Table 8: Lexical diversity metrics for the 15 raw Occitan corpora, including the size-dependent TTR and the more robust MATTR metric, computed with varying window sizes (Data Source: Wiedner (2025)).

B Preliminary Analysis

B.1 Embedding Model Backbone

(1) Frozen-encoder probing: We compare FastText (Bojanowski et al., 2016), mBERT (Devlin et al., 2019), and ByT5 (Xue et al., 2022) as frozen feature extractors with an identical linear classifier for Occitan *grammatical gender* prediction. Each instance is a bilingual pair $(w_{\text{lat}}, w_{\text{occ}})$ with Latin gender g_{lat}; we embed words in isolation, mean-pool subword/byte states, and classify $[e(w_{\text{lat}}); e(w_{\text{occ}}); \text{onehot}(g_{\text{lat}})]$. mBERT performs best on this probe (**Macro F1 = 72.04**), outperforming FastText and slightly exceeding ByT5; we therefore adopt mBERT as our default backbone.

Embedding Model	Macro F1	Accuracy %
FastText	57.51	58.30
mBERT	**72.04**	**73.37**
ByT5	71.24	72.56

Table 9: Comparison of frozen embedding models on the Occitan gender prediction task.

(2) Variant retrieval: To test whether embeddings place one-to-many Latin → Occitan realizations near each other, we cast variant identification as retrieval: given a Latin lemma w_{lat}, rank all candidate Occitan forms w_{occ} in the corpus by cosine similarity between isolated word embeddings (mean-pooled over subword/byte units). We report Recall@3 (and nDCG@3) since each query can have up to three attested variants in this experiment. Again, mBERT performs best on this probe (**Recall@3 = 0.59**), outperforming ByT5 and FastText, indicating that multilingual contextual encoders better cluster orthographic variants than static monolingual embeddings.

Embedding Model	Recall@k	MRR	nDCG
FastText	0.41	0.31	0.33
mBERT	**0.59**	**0.47**	**0.49**
ByT5	0.51	0.40	0.42

Table 10: Retrieval performance for identifying Occitan orthographic variants from a Latin lemma.

(3) Unsupervised structure: We probe intrinsic geometry by applying K-Means to isolated Occitan form embeddings and evaluating cluster agreement with canonical lemma labels. mBERT yields the best cluster separation on this probe (**Silhouette = 0.049**), outperforming ByT5 and FastText; taken together, these findings motivate our use of mBERT in the downstream pipeline.

Embedding Model	Silhouette Score
FastText	0.026
mBERT	**0.049**
ByT5	0.042

Table 11: Clustering performance of different embedding models. Higher scores indicate better-defined and more pure clusters with respect to canonical lemmas.

B.2 Tokenization Policy on BPE

We compare the standard mBERT WordPiece tokenizer against corpus-trained BPE tokenizers (vocabulary sizes 600 and 800) and a hybrid tokenizer that combines Occitan-adapted BPE with a word-level fallback. Tokenizers are evaluated using two criteria: **OOV rate**, defined as the proportion of tokens mapped to `[UNK]`, and **masked token recovery**, defined as top-1 accuracy at masked subword positions.

BPE formulation. BPE iteratively builds a subword vocabulary by merging the most frequent adjacent pair of symbols. At step t,

$$V_{t+1} = V_t \cup \{ab\}, \qquad (a, b) = \arg\max_{(x,y)} f(x, y), \tag{6}$$

where $f(x, y)$ is the corpus frequency of the pair (x, y). Repeating this process for a fixed number of merges yields a vocabulary of reusable subword units.

Summary of findings. Table 12 shows that corpus-trained BPE alone incurs non-zero OOV and very low masked recovery, while the standard mBERT tokenizer preserves full coverage but yields only moderate recovery. The hybrid tokenizer achieves the strongest overall trade-off, retaining zero OOV while substantially improving masked token recovery, which motivates its use in the downstream pipeline.

Tokenization Policy	OOV Rate (%)	Masked Token Recovery (%)
mBERT Tokenizer	0.0	15.78
BPE (vocab=600)	2.63	3.43
BPE (vocab=800)	2.86	4.76
Hybrid (BPE+word-level)	**0.0**	**25.23**

Table 12: OOV rate and masked token recovery for tokenization policies on the Occitan corpus.

C Algorithms

C.1 Algorithm 1: Construction of Occitan–Latin Lemma–Gender Dataset

We define

$$\textsc{Sim}(x, y) = \alpha\, \textsc{CosSim}(x, y) + (1 - \alpha)\, \textsc{LevSim}(x, y),$$

with

$$\textsc{LevSim}(x, y) = 1 - \frac{d_{\text{Lev}}(x, y)}{\max(|x|, |y|)}.$$

Since both $\textsc{CosSim}(x, y)$ and $\textsc{LevSim}(x, y)$ are normalized to $[0, 1]$, $\textsc{Sim}(x, y) \in [0, 1]$. We set $\alpha = 0.3$, i.e.,

$$\textsc{Sim}(x, y) = 0.3\, \textsc{CosSim}(x, y) + 0.7\, \textsc{LevSim}(x, y),$$

and accept a candidate iff

$$\textsc{Sim}(x, y) \geq 0.85.$$

The threshold and the value for α were chosen through qualitative assessment across samples and threshold settings with an Occitan linguistic expert.

C.2 Algorithm 2: Evaluation of Contextual Induction in Grammatical Gender Prediction

Algorithm 2 Evaluation of Contextual Induction in Grammatical Gender Prediction

Require: Dataset D of input instances
Require: Models M_{word} (word-only), M_{ctx} (context), M_{mask} (context with noun masked)
Ensure: Mean delta-probabilities and log-likelihood deltas for contextual induction; classification metrics

for all sample (X, i, W, L, G_L, Y) in D **do**
  $p_{\text{word}} \leftarrow M_{\text{word}}(X, i, W, L, G_L)$
  $p_{\text{ctx}} \leftarrow M_{\text{ctx}}(X, i, W, L, G_L)$
  $p_{\text{mask}} \leftarrow M_{\text{mask}}(X, i, W, L, G_L)$
  ▷ Ground-truth probability under word-only, context, and masked-context settings
  $\Delta_{p1} \leftarrow p_{\text{ctx}} - p_{\text{word}}$ ▷ prob. deltas
  $\Delta_{p2} \leftarrow p_{\text{mask}} - p_{\text{word}}$
  $\Delta_{p1}^{\log} \leftarrow \log p_{\text{ctx}} - \log p_{\text{word}}$ ▷ log-deltas
  $\Delta_{p2}^{\log} \leftarrow \log p_{\text{mask}} - \log p_{\text{word}}$
  ▷ Δ_{p1}: context vs word-only; Δ_{p2}: masked-context vs word-only
  Record deltas and predicted labels for summary
end for
Report mean(Δ_{p1}), mean(Δ_{p2}) ▷ context induction (prob.)
Report mean$(\Delta_{p1}^{\log})$, mean$(\Delta_{p2}^{\log})$ ▷ context induction (log)
Report accuracy and macro F1 for M_{word}, M_{ctx}, and M_{mask} classification ▷

C.3 Algorithm 3: Estimating the Impact of PoS Tags on Grammatical Gender Prediction

Algorithm 3 Estimating the Impact of PoS Tags on Grammatical Gender Prediction

Require: Sentences S with PoS tags (see Algorithm 1)
Ensure: Influence of PoS Tags

for all sentence $s \in S$ **do**
  Retrieve PoS tags $P = (p_1, p_2, \ldots, p_T)$ for s
  ▷ e.g., attention-, gradient-, or perturbation-based scores
  Construct mapping between tokens and PoS tags
  for $t = 1$ to T **do**
    Mask token x_t and recompute model confidence ▷ occlusion
    Record confidence change Δc_t ▷ per token
  end for
  Aggregate token scores (a_t) and/or confidence changes (Δc_t) by PoS tag for sentence s
end for
Aggregate tag-wise statistics across all sentences
return PoS-tag contributions to gender prediction

D Model Architecture and Hyperparameters for the Experiments

D.1 Lemma Experiment

Embedding	Model	Best hyperparameters
FastText	Logistic Regression	$C = 0.0027$; solver = `liblinear`; class_weight = `None`
FastText	Random Forest	n_estimators = 400; max_depth = 23; min_samples_split = 6; min_samples_leaf = 5; class_weight = `balanced`
FastText	XGBoost	n_estimators = 100; max_depth = 7; learning_rate = 0.0210; subsample = 0.7344; colsample_bytree = 0.8147; gamma = 0.6742; min_child_weight = 5
FastText	Feedforward NN (FFN)	hidden_dim = 64; dropout = 0.3934; lr = 0.0024; batch_size = 16; optimizer = `Adam`
FastText	BiLSTM	hidden_dim = 64; num_layers = 1; dropout = 0.3552; lr = 0.00969; batch_size = 16; optimizer = `Adam`
mBERT	Logistic Regression	$C = 0.0005$; solver = `liblinear`; class_weight = `None`
mBERT	Random Forest	n_estimators = 200; max_depth = 41; min_samples_split = 8; min_samples_leaf = 5; class_weight = `balanced`
mBERT	XGBoost	n_estimators = 150; max_depth = 9; learning_rate = 0.0164; subsample = 0.5431; colsample_bytree = 0.6104; gamma = 4.3585; min_child_weight = 5
mBERT	BiLSTM	hidden_dim = 256; num_layers = 3; dropout = 0.1298; lr = 0.00025; batch_size = 32; optimizer = `AdamW`
ByT5	Logistic Regression	$C = 0.0733$; solver = `lbfgs`; class_weight = `None`
ByT5	Random Forest	n_estimators = 400; max_depth = 37; min_samples_split = 9; min_samples_leaf = 4; class_weight = `None`
ByT5	XGBoost	n_estimators = 200; max_depth = 3; learning_rate = 0.0266; subsample = 0.8787; colsample_bytree = 0.9173; gamma = 2.8813; min_child_weight = 2
ByT5	BiLSTM	hidden_dim = 128; num_layers = 1; dropout = 0.4417; lr = 0.00053; batch_size = 32; optimizer = `AdamW`

Table 13: Best hyperparameter settings for all models and embedding families.

D.2 Context-Level Experiment

Hyperparameter	Exp1	Exp2	Exp3
Model Architecture			
Model Type	FieldsOnlyModel	ContextReaderModel	ContextSimpleModel
Base Encoder	BERT-base-multilingual-cased (768 dim)		
Encoder Frozen	Yes	No	No
Input Features	Word-level only	Context + Word	Masked Context + Word
Attention Mechanism	None	Multi-head (8 heads)	None
Attention Dimensions	N/A	$d_k = 128$, $d_v = 128$	N/A
Relative Position Bias	N/A	Yes (window=64)	N/A
No Self-Peek	N/A	Yes	N/A
Feature Dimensions	768	2560	2304
MLP Hidden Size	512	512	512
Training Hyperparameters			
Epochs	20	20	20
Batch Size	128	128	128
Learning Rate	2×10^{-5}	1×10^{-5}	2×10^{-5}
Weight Decay	0.01	0.01	0.01
Warmup Ratio	0.06	0.06	0.06
Optimizer	AdamW		
LR Schedule	Linear warmup + linear decay		
Gradient Clipping	0.5 (max norm)		
Dropout	0.1	0.1	0.1
Early Stopping Patience	3 epochs		
Early Stopping Metric	Validation loss		
Class Weights	Balanced (sklearn)		
Data Configuration			
Max Sentence Length	128 tokens		
Max Fields Length	32 tokens		
Cross-Validation	Group K-Fold (3 folds)		
Random Seed	13		
Experiment-Specific Details			
Context Usage	None	Full sentence	Masked sentence
Masking Strategy	N/A	N/A	Noun replaced with `NOUNTOKEN`
Fields Format	`[OCC]{word} [LAT]{lemma} [LG]{gender}`		

Table 14: Hyperparameters and training configuration for experiments 1, 2, and 3.

E PoS Tagger in the Study

The PoS tagger used in this study is (Manjavacas et al., 2019). A key limitation is that PoS tags are automatically predicted for the full Occitan corpus, and downstream analyses (including our occlusion-based PoS importance estimates) inherit tagging errors. To quantify tagger quality, we manually annotated a 60,000-token subset and evaluated the tagger against this gold data, obtaining 71.31% overall accuracy. Performance varies by tag: ADJ shows the lowest accuracy, while our primary tag of interest, NOUN, achieves 70.32%. We therefore interpret PoS-conditioned results as informative but potentially biased by tagging noise.

F Statistical Tests of the Experiments

F.1 Statistical Significance for Lemma Experiment

To complement the fold-averaged results in Table 3, we directly compare mBERT and ByT5 under a matched downstream architecture using paired bootstrap resampling over *out-of-fold* (OOF) predictions from the same 10 CV splits. Out-of-fold (OOF) predictions are obtained from the same lemma-grouped CV splits; therefore, each OOF prediction is made on a held-out lemma and is free of lemma-level leakage. This analysis is slightly different from Table 3: there, Macro-F1 is reported as the mean across folds, whereas here we compute a single OOF Macro-F1 over all 4,444 held-out predictions to obtain a more robust paired comparison at the item level. Under the shared 2×BiLSTM + MHSA classifier, mBERT

yields a higher OOF Macro-F1 than ByT5 (0.7608 vs. 0.7213; $\Delta = +0.0395$). The paired bootstrap confirms that this advantage is reliable in the present setup: the 95% confidence interval remains strictly above zero, and no bootstrap resample yields $\Delta \leq 0$.

System	OOF Macro-F1	Notes
mBERT	0.7608	
ByT5	0.7213	
Δ (mBERT $-$ ByT5)	+0.0395	95% CI [+0.0250, +0.0543]
p-value		$p < 10^{-6}$

Table 15: Paired bootstrap comparison between mBERT and ByT5 under the same 2×BiLSTM + MHSA architecture and identical 10-fold CV splits. Unlike Table 3, which reports mean Macro-F1 across folds, this table reports Macro-F1 computed over out-of-fold predictions for a paired item-level comparison. Because Table 3 reports fold means, we additionally perform a paired bootstrap on pooled out-of-fold predictions under the same architecture to test whether the mBERT–ByT5 difference is reliable at the instance level.

F.2 Statistical Significance for PoS Tags

To assess whether the observed PoS-wise occlusion effects are reliably different from zero, we perform a *sign-flip permutation test* for each PoS tag independently. For a given tag p, let $\delta_{i,p}$ denote the per-sample occlusion score, i.e., the change in confidence for the gold label when tokens with tag p are masked. Under the null hypothesis that the mean effect of p is zero, the sign of each $\delta_{i,p}$ is exchangeable; we therefore generate a null distribution by randomly flipping the sign of the per-sample scores over 10,000 permutations and recomputing the mean. We report the observed mean effect together with the resulting p-value from a two-sided test of whether the mean effect differs from zero, which asks whether a PoS tag provides a reliably positive or negative contextual contribution. The full numerical results are reported in the main text in Table 7.

G Explainability through Attention Heads on Contextual Cues Experiments

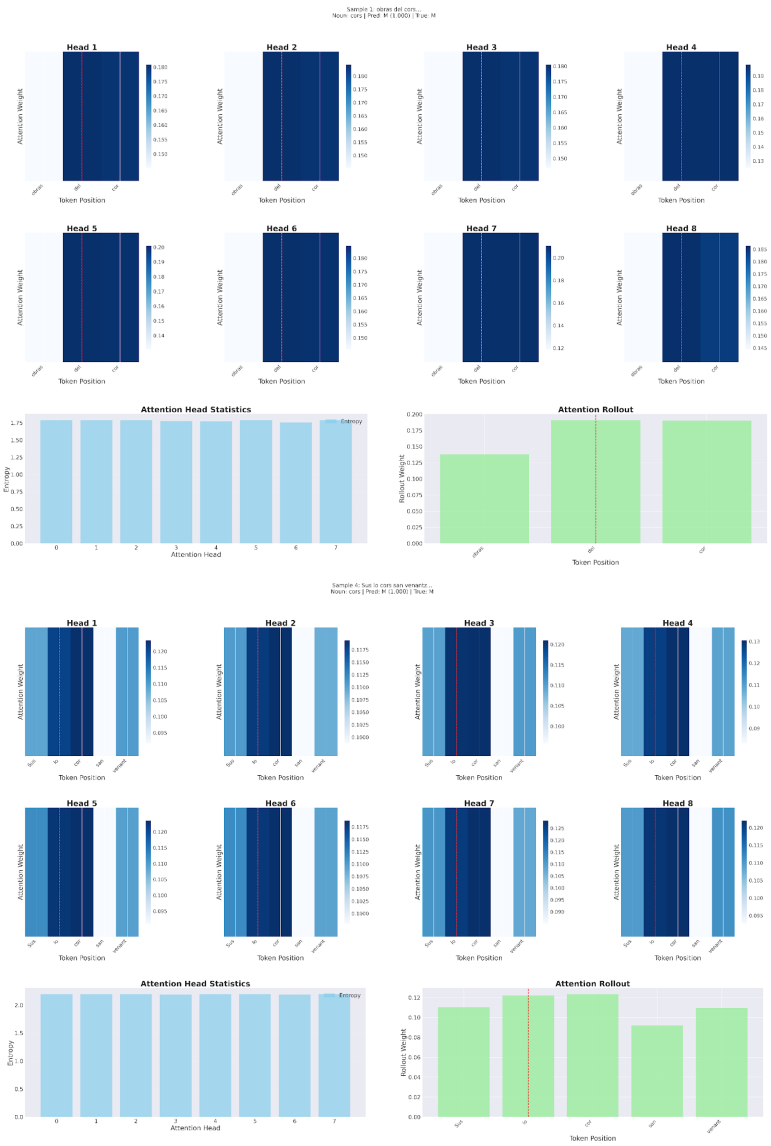

Figure 8: Attention-based contextual evidence for grammatical gender prediction shown for two representative Occitan sentences (top and bottom panels; target noun: *cors*). For each example, we visualize the 8 MHSA heads when using the noun representation as the query and the full sentence as keys/values; the dashed red line marks the target noun position. Across heads, attention concentrates on the noun and nearby agreement-bearing tokens (often including determiners/articles), consistent with morpho-syntactic cues for gender assignment. The per-head entropy plot (left) indicates broadly distributed head behavior, and attention rollout (right) summarizes aggregate token attribution across the sentence.

H Ablation Study on Contextual Cues Experiment

Ablation: removing Latin features. To quantify the contribution of Latin etymological information in the contextual setting, we ablate the Latin lemma and Latin gender from the input and retrain/evaluate the same context model under 3-fold lemma-grouped cross-validation. Table 16 reports mean performance (± std across folds). While the context model remains effective without Latin features, the contextual Δ gains in confidence are substantially reduced: with Latin lemma and gender, context increases the gold-class probability by $\sim$0.28 relative to the word-only baseline, whereas without Latin this increase drops to $\sim$0.09–0.11 (about 3$\times$ smaller). This indicates that Latin features provide critical complementary signal that amplifies the benefit of context.

Setting	**Accuracy**	**Macro-F1**
Context + full noun (no Latin)	0.961 ± 0.018	0.879 ± 0.036

Table 16: Contextual ablation without Latin lemma/gender. Results are mean $\pm$ std over 3-fold lemma-grouped cross-validation.

I Error Analysis

We analyse the 294 misclassifications made by the BiLSTM+Attention model across all folds using a SHAP-based surrogate approach. We train an XGBoost error predictor on 57 interpretable features capturing morphology (Latin/Occitan suffix cues, length, vowel ratio), frequency, sentence properties (length, noun position), and local syntax (PoS fractions and neighbouring tags) and explain its decisions with TreeSHAP (5-fold CV; ROC-AUC = 0.62). The strongest error drivers fall into three groups: (i) *context sparsity*—sentences with fewer agreement-bearing categories, especially adjectives in the immediate right context, yield more errors; (ii) *morphological ambiguity*—errors are more common when nouns occur at either sentence boundary; and (iii) *length/frequency effects*—errors are more common in shorter contexts and for mid-frequency items, consistent with a regime where very frequent forms are memorised and rare regular forms generalise more easily. Overall, masculine items exhibit a higher error rate than feminine ones, though the difference is very low.

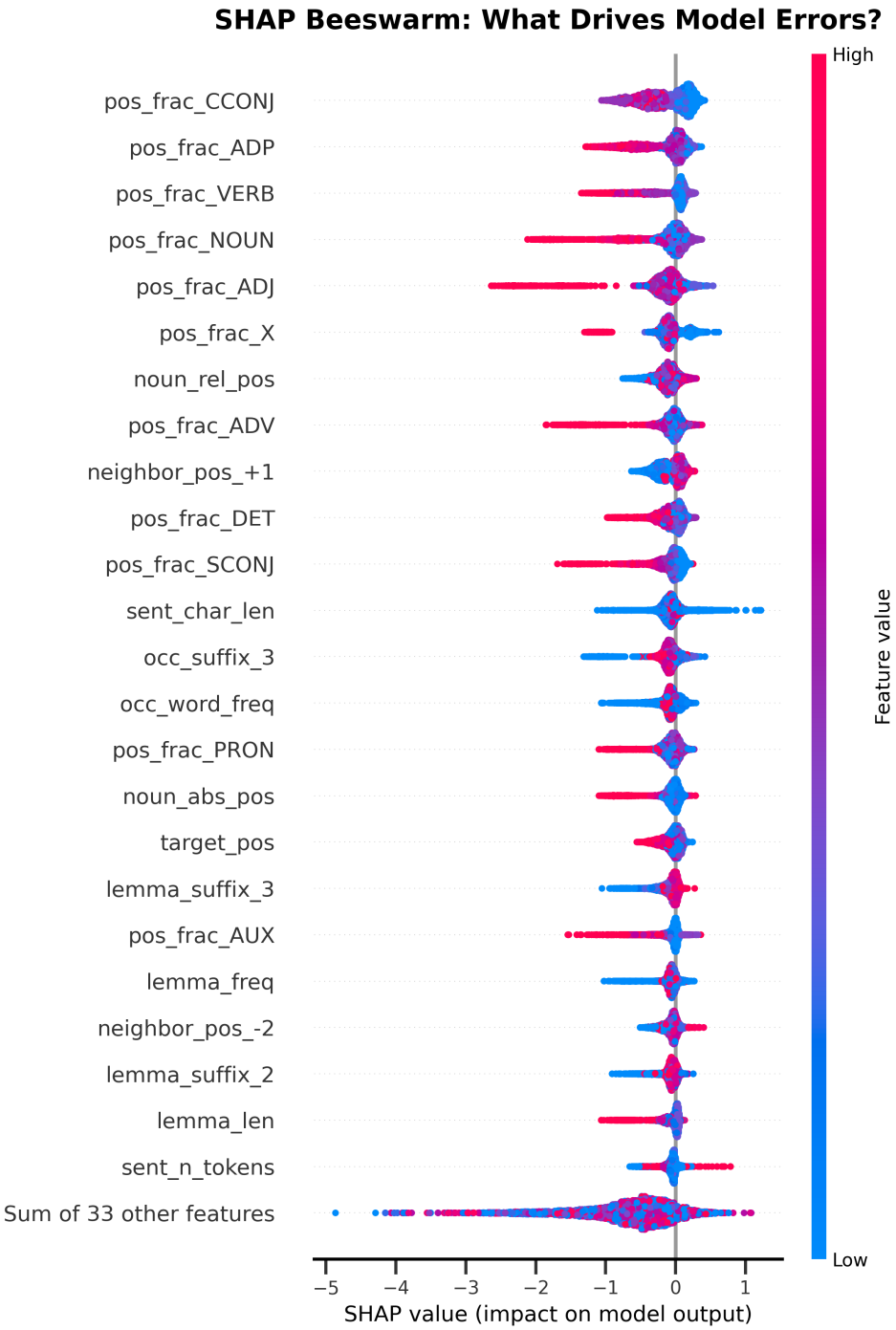

Figure 9: SHAP beeswarm plot showing feature contributions to model error prediction. Each dot represents a sample; the x-axis indicates the SHAP value (positive = pushes toward error, negative = toward correct), and colour encodes feature value (red = high, blue = low). The top five drivers are all POS composition features — fraction of coordinating conjunctions, adpositions, verbs, nouns, and adjectives in the sentence — indicating that syntactically sparse contexts lacking agreement-bearing words are the primary source of errors. Morphological features (Occitan suffix, Latin lemma suffix) and contextual factors (noun position, sentence length, word frequency) contribute secondarily. The immediate right-neighbour POS tag (rank 9) confirms that local agreement context, particularly an adjacent adjective, is a key disambiguating signal.