Simulating the Evolution of Grammatical Gender from Latin to Old Occitan: A Computational Approach Using LSTM with Attention

Marinus Wiedner (Universität Freiburg) & *Matthias Schöffel* (Bayerische Akademie der Wissenschaften)

marinus.wiedner(at)romanistik.uni-freiburg.de, matthias.schoeffel(at)badw.de

Abstract

In this article, we present a first approach on simulating the change of gender from Latin to Old Occitan. The reduction of genders in the transition to the Romance languages is one of the most important developments in the nominal system. In order to simulate this language change, we used varied data, taken from the most exhaustive Old Occitan dictionary (DOM), from an edition of a juridical text as well as from the transcriptions of two manuscripts taken from COMETA (Wiedner 2025). The nouns with information on their etymon and the gender in Latin and Old Occitan were then used as input for the simulation. Building on previous simulation-based approaches (Polinsky and Van Everbroeck 2003), our model provides a form-based, data-driven perspective on how gender marking might evolve under the interaction of morphological cues. Despite its inherent simplifications, the model achieved good predictive accuracy and yielded interpretable tendencies in the distribution of attention, suggesting that even at the level of orthographic form, systematic regularities relevant to gender can be captured computationally.

Keywords

artificial intelligence, computer simulation, grammatical gender, language change, low resource language, Old Occitan

1 Introduction

One of the most important developments in the nominal system (alongside the reduction and loss of morphologically marked case) from Latin to the Romance languages is the reduction of the gender system from a tripartite gender system in Latin (with masculine, feminine, and neuter) to a binary gender system in most standard Romance languages, where the neuter has disappeared. With this study, we aim at simulating the development of grammatical gender from Latin to the Romance languages with the example of Old Occitan as a low resource language.

For the remainder of this study, we mean grammatical gender when using only gender. If we talk about other kinds of gender, like biological or social gender, we will make it explicit. We use the definition of gender as proposed by Hocket (1962: 231): “Genders are classes of nouns reflected in the behavior of associated words”. This means, that gender does not have to be something formally marked and thus visible or audible on the noun itself, but is mainly characterised by the agreement it triggers on other word classes, as e.g. adjectives or determiners. The

AILing

Wiedner, Marinus & Schöffel, Matthias. 2026.
Simulating the evolution of grammatical gender.
Special Issue: *Natural Language and AI.*Vol.3 No.1
DOI: 10.62408/ai-ling.v3i1.66.
AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070

nouns are therefore the *agreement controllers* while the other word classes which change according to the gender triggered by the noun are labelled *agreement targets* (see Corbett 2006: 4). The agreement targets that are relevant for this research, as they reflect gender in the Romance languages, are articles, adjectives, pronouns, and participles (or at least some representants of these word classes; see Loporcaro 2018: 8). This overview serves to contextualise the assumptions underlying the simulation and to clarify which empirical developments it is intended to model.

2 State of the art

Before going into more detail on the computer simulation that can be seen as the starting point for our project (see Polinsky and Van Everbroeck 2003), we will first give a short summary of the gender development from Latin to Old Occitan and on research on gender in the latter language.

2.1 Gender from Latin to Old Occitan

Latin is known to have a tripartite gender system, distinguishing between feminine, masculine, and neuter (see e.g. Pinkster 2015: 37–39). The three genders are distributed over the five declension classes and, to some extent, there is a correlation between declension class and gender. For instance, nouns from the first declension class with the nominative singular ending in *-a* are mainly feminine, while the nouns from the second and fourth declension class with the nominative singular ending in *-us* are mainly masculine. Nouns from the second declension class with the nominative singular ending in *-um* are mainly neuter. The gender assignment for nouns from the third and fifth declension class is more complex (see e.g. Panhuis 2015: 22; Wang 2023: 71–74), which contributes to the fact that nouns from these declension classes tend already in Classical Latin to be gender variable. This gender variability is then inherited into the Romance languages (see e.g. the forms derived from the Latin MARE ‘sea’: for instance Italian *mare* is masculine, while French *mer* is feminine, and Spa. *mar* triggers both genders; see e.g. Loporcaro 2018: 2; Wiedner to appear).

Superseding these morphological assignment rules, Latin also has a semantic core of gender assignment, which is mainly biological gender, or sex. This means that some nouns from the first declension class, which are majorly feminine, may also trigger masculine agreement, when denoting a male animate, e.g. AGRICOLA ‘farmer’. There are further semantic assignment rules, such as rivers or mountains being masculine, and islands or plants being feminine (see Pinkster 2015: 39). All of these semantic assignment rules are prior to the morphological assignment rules in Latin. The gender assignment is illustrated in Figure 1:

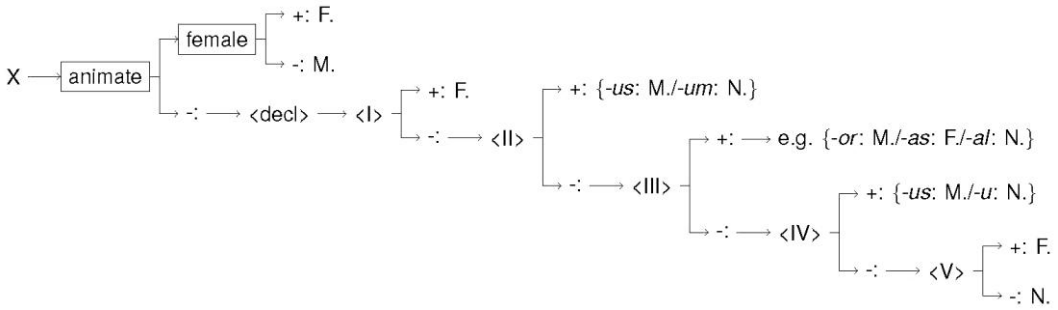

Figure 1: Gender assignment rules in Latin, adapted from Wang (2023: 73).

Regarding the gender in Old Occitan, information is scarce. There are some studies on the development of Greek loanwords ending in *-a*, such as e.g. *papa* 'pope' or *profeta* 'prophet' (see Jensen 1973; 1976: 79; Wiedner to appear), that show a certain gender variability for these nouns in Old Occitan, even when specifically denoting a man. Wiedner (to appear) could show gender ambiguity for more than 40 nouns, mostly inherited from the third declension class in Latin or loanwords from Greek. In this regard, Old Occitan behaves similarly to other Romance languages. It has only a binary gender distinction, the neuter does not exist anymore by the time of our data set. However, Loporcaro (2018: 204) lists some pronouns as neutral pronouns, e.g. *o*, *ço*, *ce* 'this', but only for default gender marking. Apart from this, no information on remnants of the neuter is available.

Chambon (2003: 362) tries to argue for a third gender for three nouns (carra 'cartload', *paira* 'one day of work by a pair of oxen owed as payment', and *semoia* 'half a bushel (a measure of capacity)'), but these three nouns cannot be regarded as belonging to a gender apart, as they only have a specific agreement in combination with the cardinal numbers *doa* 'two' and *tria* 'three', otherwise they behave as every other noun.

In general, the nouns inherited from Latin kept their gender in the transition to Old Occitan with the neuters being reassigned to either masculine or feminine gender: the former neuters from the second declension class in the singular (with the ending *-um*) were reassigned to masculine singular, while the neuter plurals ending in *-a* were mainly reanalysed as feminine singular, both due to morpho-phonological similarities (see Donaldson and Sibille 2024: 202).

2.2 Starting point – Polinsky and Van Everbroeck (2003)

The first study to our knowledge that takes a simulation-based approach to the development of gender is the study by Polinsky and Van Everbroeck (2003), who simulated the development from Latin to Old French with a connectionist model (see Hare and Elman 1995).

As their data basis, they used the 500 most frequent nouns from Late Latin texts. They added information on the phonetic and phonological representation

including the number of syllables, vowel quality and quantity as well as information on place of articulation, type of consonant, etc. Furthermore, information on type-token-frequency, the gender of the respective Gaulish noun (Gaulish being the Celtic substrate language), and phonetic similarities to other nouns, (i.e. Gaulish nouns) as well as information on animacy and relative frequencies of case-number-combinations were included.

They ran their simulations in ten generations, which each consisted of three epochs. After each generation, the model gave probabilities for the three genders as an output. The most likely gender was then used as the input for the subsequent generation and so forth.

Polinsky and Van Everbroeck (2003) ran different simulations, including all information as described before or only parts of the information, in order to assess the relevance of each variable for the accuracy of the model. This way, they could show different factors that play a key role for the outcome of the simulation, i.e. the frequency of nouns plays a role in the sense that more frequent neuter nouns led to a longer preservation of this gender. The second main result shows the importance of the substrate language Gaulish, a language with only two genders. Without the Gaulish data, neuter nouns persist much longer, whereas with Gaulish information, the model conforms more closely to historical reality and neuters disappear more quickly.

2.3 Similar studies

Like the simulation-based framework of Polinsky and Van Everbroeck (2003), other studies have applied computational and typological methods to model grammatical gender prediction and change. Hare and Elman (1995) used multi-generation neural networks to show that irregular verb patterns in Old English tend to regularise over time. Polinsky and Van Everbroeck (2003) extended this approach to gender systems, while Cotterell et al. (2018) further developed it with Long Short-Term Memories (LSTMs), demonstrating that high-frequency irregulars persist whereas low-frequency forms regularise, consistent with frequency-based learnability.

2.3.1 Gender prediction with machine learning

Several studies have focused on the synchronic modeling of gender assignment. Allassonnière-Tang and Basirat (2020) trained a neural classifier with word embeddings on Swedish nouns, showing that distributed lexical representations encode grammatical gender cues. Veeman and Basirat (2020) extended this approach cross-linguistically to Swedish, Danish, and Dutch, demonstrating that neural models can recover gender even without explicit syntactic cues such as articles. The work by Cucerzan and Yarowsky (2003) used weakly supervised learning to induce noun gender in French, combining contextual and morphological cues to achieve high accuracy with minimal labeled data. Similarly, orthographic

LSTM models for French achieve gender classification accuracies of around 92–94% with additional gains from semantic cues (Sahai and Sharma 2021), though these studies focus on synchronic prediction rather than diachronic change.

2.3.2 Typological and information-theoretic approaches

Williams et al. (2020) employed information-theoretic methods to quantify the extent to which phonological form and semantic meaning influence noun declension class membership in Czech and German, while controlling for grammatical gender. They found that both form and meaning provide significant, partly independent information about class, and that their interaction is also informative. The study introduces a quantitative framework for measuring cue strength and shows that the reliability of these cues varies across declension classes and languages. Wichers Schreur et al. (2022) trained statistical classifiers on Chechen and Tsova-Tush (five-gender Nakh languages) and found that while both morphological and semantic cues are informative, semantic features dominate prediction accuracy (https://typologyatcrossroads.unibo.it/). Similar tendencies have been reported for other typologically rich systems (Allassonnière-Tang et al. 2021), supporting the general hypothesis that meaning-based regularities play a primary role in gender organisation across languages.

2.3.3 Usage-based and analogical explanations

Beyond computational modeling, analogical and frequency-based analyses have also been proposed. Coker (2009) examined gender shifts in Ancient Greek and argued that changes (2nd-declension *-oς* nouns moving from feminine to masculine) are best explained by category reorganisation driven by frequency, saliency, and animacy rather than formal analogy rules.

3 Data overview

After these introductory chapters on gender in general and some computational studies on language change, the following section will focus on the data used for this simulation. In order to be able to accurately simulate the development of gender from Latin to Old Occitan, we need comparable data on both languages.

We started our data collection with the Old Occitan texts, as the available resources for this medieval language are scarce. Therefore, we first extracted all nouns listed in the *Dictionnaire de l'occitan médiéval en ligne* (DOMél) which gave us a total of 15,822 nouns. We then excluded doublets, but added graphical variants taken from the *digitales Dictionnaire de l'occitan médiéval* (dDOM), which is the digitised version of the card archive of the dictionary, where all sources, graphical variants, and additional information are listed.

To get data on Latin, we used the existing connection between the DOM and the *Französisches Etymologisches Wörterbuch* (FEW), where the Latin etyma

of the respective Occitan nouns are listed. The problem with this Latin data (and also with most of the other dictionaries) is that it is only available in the nominative even though most of the nouns inherited from Latin into the Romance languages are derived from the accusative. Furthermore, the nouns are listed in the FEW without any indication of gender in Latin. The first problem remains unsolved up until now (see Section 7.1; it would probably be necessary to manually add the accusative forms), while we solved the latter problem with an automatic annotation using the tagger *LatinCy* (Burns 2023). Another, not yet solved problem with this data, is the type of etymon: a considerable number of Old Occitan nouns are also derived from other word classes, which pose certain difficulties for the simulation, e.g. nouns built through deverbal conversion, which is interesting for a discussion on gender assignment in conversion processes (see Marzo and Wiedner 2025), but difficult to implement in the simulation, as we do not have a gender for the Latin verb. Moreover, loanwords from other languages, i.e. Frankish, pose the same problem. Therefore, we excluded all nouns that were derived from other word classes but Latin nouns, as well as all loanwords that were loaned directly into Old Occitan and not already in Latin.

To further enrich our dataset and enable the inclusion of context information in the future, we drew data from the COMETA corpus (Wiedner 2025). This corpus consists of 17 manuscripts written in the second half of the 13th century or in the 14th century in the regions of Languedoc and Provence. These texts were semi-automatically annotated using the Transkribus model for Old Occitan Handwriting (Wiedner 2023). The texts were then pre-annotated with LLMs according to the principles proposed in Schöffel et al. (2025a) and Schöffel et al. (2025b), and two manuscripts were manually post-corrected in order to obtain the gold-standard. Both manuscripts are from the *Bibliothèque nationale de* France, the one from the collection *Nouvelles Acquisitions Françaises* (NAF) 6195 (the *Vida de Sant Honorat*; in total about 46,000 tokens; about half of the nouns from this manuscript are annotated), the other one from the collection *Français* (shelf-number 25425; the *Chanson de la Croisade contre les Albigeois*; in total about 100,000 tokens). All 12,445 nouns from these two manuscripts were then manually connected with their respective etymon. Furthermore, information on the gender in the respective context as well as the gender of the respective noun was added.

The same was applied to an existing edition of the Old Occitan juridical text *Lo Codi*, where we used the edition by Derrer (1974), with a total of about 150,000 tokens. This text was annotated in the same way as the other texts, so we also have the combination of the Old Occitan noun connected with the Latin etymon and gender in both languages, which added another 4503 nouns. So, in addition to the lexicographic data from the DOM, we have a chronic and a hagiography directly taken from the manuscript as well as an edition of a juridical text. This allows us to have a certain range of variation in the data set.

Before using the data for our simulation, we extracted and excluded all doublets in order to not have one and the same noun multiple times in our corpus,

as this would lead to a strong bias in the data set. This leaves us with a total of 7047 Old Occitan nouns, for which we have information on the graphical form, the respective etymon as well as information on gender in both Old Occitan and Latin. Before including these nouns in the simulation, we identified and excluded all doublets as well as all nouns lacking a clearly marked gender value (i.e., nouns that appear in the manuscript without any agreement marker that would allow a reliable gender determination). This information is used for the simulation which will be presented in the following. Figure 2 shows the distribution of gender in both the Old Occitan and the Latin data set. It is clear that there is no neuter noun in the Old Occitan data, given that the neuter has disappeared in Occitan by that time. The former neuter nouns are thus reassigned to either feminine or masculine gender with a clear preference for the latter within our data set.

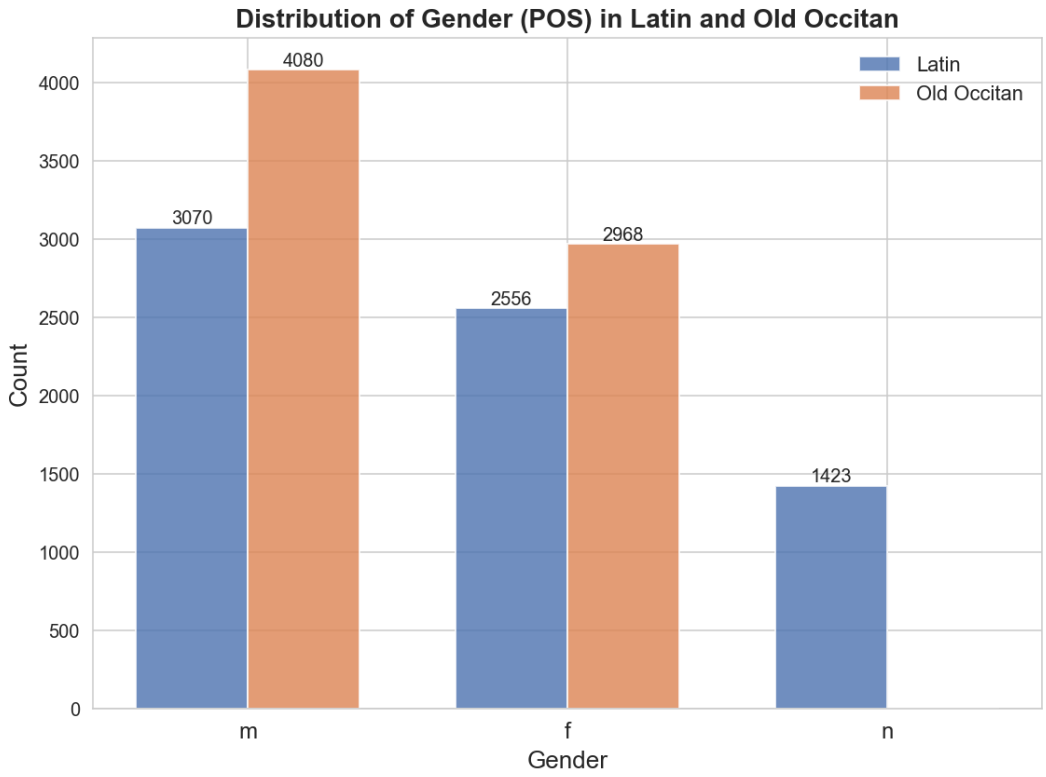

Figure 2: Distribution of Gender in the Latin and Old Occitan data set.

4 Model architecture: LSTM with attention

In order to model the development of gender from Latin to Old Occitan, we employ a LSTM network combined with an attention mechanism. LSTMs (Hochreiter and Schmidhuber 1997) are a type of recurrent neural network (RNN) specifically designed to capture long-range dependencies in sequential data. Unlike standard RNNs, which tend to suffer from vanishing or exploding gradients, LSTMs use a

system of gating units (input, forget, and output gates) to control the flow of information through the network. This architecture allows the model to retain relevant information over extended sequences and to discard less relevant details. Given that linguistic change is often shaped by patterns spanning across morphological contexts, LSTMs are particularly well-suited for our task.

The attention mechanism enhances the LSTM by allowing the model to focus selectively on the most informative parts of the input sequence when making predictions. Rather than compressing the entire input into a single fixed-length vector, the attention layer computes a weighted representation of all hidden states, highlighting which segments of a word are most relevant for determining its gender. In the implementation, the model processes both Latin and Old Occitan forms as sequences of characters, which are first mapped to embeddings and then enriched with relative positional encodings, which preserve the order-specific information crucial for identifying morphological features. Although LSTMs capture sequential information implicitly, relative positional encodings provide explicit order signals that are particularly useful for the attention layers, allowing the model to distinguish morphologically relevant positions and to support precise cross-lingual character alignment. The sequences are passed through bidirectional LSTMs (BiLSTMs), which is essential for this task as it allows the network to incorporate both preceding and following character contexts, thus leveraging cues distributed across stem and ending. The resulting hidden states are normalised and fed into two separate multi-head self-attention layers. Although multi-head self-attention already computes multiple subspace projections within a single layer, two separate Multi Head Self Attention (MHSA) modules are used to process the Latin and Old Occitan sequences independently. This separation allows the model to learn context-sensitive representations for each language before integration, supporting precise modelling of morphological patterns and cross-lingual correspondences. The attended vectors are then averaged and concatenated with the Latin gender feature (encoded as a 3-dimensional one-hot vector) to form the input to the final fully connected layers, which predict the gender of the Old Occitan form. Finally, the model integrates the attended sequence representations by averaging the hidden states and concatenating these two vectors with the Latin gender feature (encoded as a 3-dimensional one-hot vector). This combined vector is then passed through fully connected layers to predict the gender of the Old Occitan form.

This architecture allows the network to leverage both subword-level character patterns and cross-linguistic correspondences in predicting gender, capturing subtle diachronic shifts. Combining LSTMs with attention allows the model to capture both global sequential patterns and local morphological cues, such as the ending or suffix, relevant to gender assignment. This integration provides a flexible, probabilistic framework for modelling the diachronic evolution of gender from Latin to Old Occitan.

5 Implementation details

The simulation of gender evolution from Latin to Old Occitan was implemented in PyTorch, combining a character-level bidirectional LSTM with a multi-head attention mechanism (see Section 4). The model was trained to predict the gender of Old Occitan nouns given the Latin etymon, the Old Occitan reflex, and the Latin gender. The following sections provide a detailed description of data processing, model design, and training procedures.

5.1 Data encoding and overview

All characters occurring in both the Latin and Old Occitan word forms were extracted and encoded as numerical indices. Each word (e.g., NOMEN → *nom*) was then converted into a sequence of character indices, allowing the model to process morphological information at the subword level. This character-sequence-based encoding enables the model to generalise across unseen or orthographically variable forms.

5.2 Dataset construction and encoding

Each data point in the corpus consists of a Latin form (Lemma), its corresponding Old Occitan form, and the gender in both languages. The task is structured as a binary classification problem, where the target label represents the gender of the Old Occitan noun (masculine = 1, feminine = 0). A custom PyTorch dataset class manages these entries, converting the character sequences into index tensors. Crucially, the Latin gender is encoded as a 3-dimensional one-hot vector and integrated as an auxiliary input feature to inform the prediction. Because word lengths are variable, sequences are padded to the maximum length within each batch. The original, unpadded sequence lengths are stored and utilised throughout the training process to ensure that padding tokens are correctly masked and do not affect model performance. By retaining and applying the true sequence lengths during training, the model processes only the linguistically relevant portion of each input, while maintaining computational efficiency through batch-wise padding.

5.3 Model architecture and hyperparameter optimisation

The experimental procedure is designed for robust evaluation and hyperparameter search. Due to the limited size of the dataset, the data were split evenly into training and validation subsets. Although a 70:30 split is more common, the 50:50 division was chosen to obtain more reliable validation results under data-scarce conditions. To further increase robustness under data-scarce conditions, 5-fold cross-validation was employed. The dataset was partitioned into 5 equally sized folds, and the model was trained 5 times, each time k-1 folds were used for the training and the remaining

fold for validation. Final performance metrics were obtained by averaging results across folds. This procedure reduces variance associated with a single train-validation split and provides a more stable estimate of generalisation performance. Hyperparameter optimisation was managed using a WandB grid search, which systematically explores a predefined set of architectural and regularisation parameters, including the hidden size, number of layers, dropout rates, embedding dimension, and number of attention heads. Training was performed for a maximum of 20 epochs using the AdamW optimiser. The loss function is a weighted cross-entropy loss that leverages pre-computed class weights to compensate for the inherent imbalance between masculine and feminine target forms, thereby preventing training bias. To ensure stability, gradient clipping was applied during the backward pass to prevent exploding gradients. The training process incorporates two critical control mechanisms: first, a custom early stopping module monitors the validation loss and terminates training if no improvement is observed over patience=3 epochs; second, a ReduceLROnPlateau scheduler dynamically reduces the learning rate if the validation loss plateaus for patience=4 epochs for convergence reasons. After each epoch, performance was evaluated on the validation set using standard metrics, including accuracy, precision, recall, F1-score. Final evaluation includes the computation of the Confusion Matrix.

5.4 Evaluation

At the end of training, the best model was selected based on its macro F1-score, reflecting balanced performance across both gender classes. The validation results were monitored for each epoch, and additional diagnostic metrics (e.g., number of misclassified tokens) were logged to support interpretability and reproducibility.

6 Simulation

The model used for the experiments was a recurrent neural network with an LSTM backbone and multihead attention incorporating relative positional encoding. The model received two inputs per instance. The Latin and the corresponding Old Occitan form—together with the Latin grammatical gender. The task was to predict the Old Occitan gender (male or female). The total number of trainable parameters was computed for documentation purposes. Class imbalance was addressed by calculating class weights, which were passed to a weighted cross-entropy loss function. Training was performed using the AdamW optimiser with gradient clipping to prevent exploding gradients. A learning rate scheduler reduced the learning rate upon plateauing validation loss, and early stopping terminated training after three consecutive epochs without improvement (see Table 1).

Table 1: Overview of model parameters.

Parameter	Value
Hidden Size	4
Embedding dimension	64
Learning Rate	0.002
Dropout LSTM	0.4
Dropout Attention	0.2
Epochs	25
Weight Decay	0.01
Number of heads	2
Number of Layers	4
Optimiser	AdamW

Validation accuracy, precision, recall, and F1 score were computed (see classification report in Table 2). The model achieved an overall accuracy of 91.4%, demonstrating strong performance in gender classification. Precision and recall were well balanced, with F1-scores of 0.895 for female and 0.927 for male. The slightly lower recall for the female class suggests a minor bias toward male predictions, likely reflecting class imbalance in the dataset (2039 vs. 1485 instances). To assess robustness under data-scarce conditions, k-fold cross-validation was conducted additionally. Across folds, the model achieved a mean accuracy of 91.4% (± 3.7%) and a macro F1-score of 91.1% (± 3.6%). The close correspondence between the cross-validation averages and the single-split results indicate that the reported performance is not attributable to a favourable partition. While the observed variance across folds reflects some sensitivity to training data composition, the stable macro F1-score confirms balanced and reliable generalisation across classes despite the limited dataset size.

Table 2: Classification report.

	Precision	Recall	F1-score	Support
Female	0.926	0.865	0.895	1485
Male	0.906	0.949	0.927	2039
Accuracy	0.914	0.914	0.914	0.914
Macro avg	0.916	0.907	0.911	3524
Weighted avg	0.915	0.914	0.914	3524

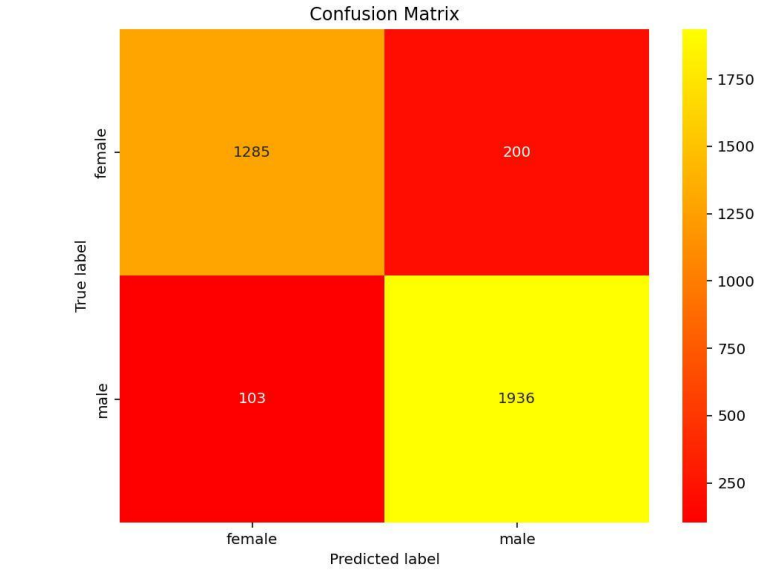

Figure 3: Confusion matrix.

The best-performing model was saved to disk once the minimum validation loss was reached. The confusion matrix is shown in Figure 3. When available, attention weight matrices were visualised as heatmaps for a subset of correct and incorrect examples, allowing a comparative inspection of attention patterns across word pairs. A quantitative analysis of attention distributions was also performed, including entropy, maximum attention, the Gini coefficient, and ratios of attention allocated to word prefixes and suffixes. The results are summarised and discussed in the subsequent sections.

6.1 Theoretical basis for the cross-lingual attention mechanism

To investigate how the model aligns characters between Latin lemmas and their corresponding Old Occitan forms, we analyse the cross-lingual attention matrices produced by the multihead attention mechanism. Each attention matrix $A \in R^{n \times m}$ represents the learned correspondence between *n* characters of the Occitan word and *m* characters of the Latin input. The value A_{ij} denotes the degree to which the model attends to the *j*-th Latin character when generating or encoding the *i*-th Occitan character.

This cross-lingual attention measure thus quantifies character-level alignment across languages, reflecting how the model internally relates source and target linguistic forms. High attention weights on specific character pairs indicate systematic correspondences (e.g., shared morphological endings), whereas more diffuse attention patterns suggest a weaker structural mapping.

Besides, the Gini coefficient measures how concentrated or evenly distributed the model's attention weights are within an attention matrix. To compute it, all attention values for a given example are first flattened into a one-dimensional array and sorted in ascending order. The cumulative sum of these sorted attention values is then used to calculate inequality in their distribution. The coefficient is derived from the relative differences between all attention values and reflects the deviation from perfect uniformity. A Gini value near zero points towards an attention that is distributed almost evenly across all character pairs, which means that the model assigns similar importance to many input positions. A higher Gini value, by contrast, signals that attention is concentrated on a smaller subset of positions, showing that the model focuses more narrowly on specific alignments between characters in the Latin and Old Occitan forms.

The maximum attention value represents the single highest attention weight within the attention matrix for a given word pair. It identifies the strongest connection the model forms between a specific character in the Old Occitan form and a character in the Latin source. A higher maximum attention value suggests that the model strongly associates certain character positions across the two languages, potentially reflecting points of high morphological or orthographic correspondence. Together, the Gini coefficient and the maximum attention value offer complementary perspectives: while the Gini coefficient captures the overall concentration of attention, the maximum attention highlights the single most dominant alignment learned by the model.

6.2 Discussion of results

The following Table 3 shows the main results regarding the model's accuracy in respect to the different variables, that will be explained and discussed in the following.

Table 3: Quantitative attention metrics.

Metric	Correct	Incorrect
Mean Max Attention	0.62	0.48
Mean Gini	0.68	0.59
Mean Suffix Attention[1]	1.60	1.37
Mean Prefix Attention	1.08	0.79
Mean Suffix/Prefix Ratio	1.75	2.24

The quantitative attention metrics reveal systematic differences between correctly and incorrectly classified examples, but they also require cautious

[1] Computed as the average of normalized attention weights assigned to the last three characters of each word, averaged across all words and attention heads.

interpretation. The higher mean maximum attention in correct predictions (0.62 vs. 0.48) suggests that accurate classifications are associated with stronger and more confident alignment between specific character pairs. This pattern points to a more decisive internal representation during successful predictions. However, a high maximum attention alone does not necessarily indicate linguistic relevance. It may also reflect overfitting to frequent character correspondences or noise that coincidentally aligns with the target label.

Similarly, the higher mean Gini coefficient for correct examples (0.68 vs. 0.59) implies that successful cases show a more concentrated attention distribution, while misclassifications are characterised by a broader, less selective focus. This may indicate that the model benefits from identifying and emphasising key morphological or orthographic features. Yet, a higher Gini value could also mean excessive focus on a few characters at the expense of other relevant cues, which might limit generalisation. Thus, while concentration appears beneficial in this dataset, overly sharp attention may reduce robustness in unseen examples.

The differences in suffix and prefix attention further reflect the linguistic structure of the data. Correct classifications show higher average attention toward both suffixes and prefixes, with particularly strong emphasis on word endings (suffix attention 1.60 vs. 1.37). This is consistent with the known morphological importance of suffixes in gender marking in Latin-derived languages (see e.g. Bauer 2010: 533–541). Nonetheless, attention values in the prefix region (1.08 vs. 0.79) also contribute meaningfully, suggesting that the model does not rely exclusively on suffix cues.

The suffix-to-prefix ratio (1.75 for correct vs. 2.24 for incorrect) highlights that overly strong reliance on suffixes may actually correlate with misclassification. This finding suggests that balanced attention across the word is more conducive to accurate gender prediction, while overemphasis on endings might lead the model to ignore other informative segments. However, because attention weights are not guaranteed to reflect true linguistic reasoning, these interpretations should be treated as indicative rather than conclusive.

The interpretability of attention mechanisms is debated because attention weights, while intuitively appealing as indicators of what the model focuses on (Jain and Wallace 2019) do not necessarily reflect the true causal importance of specific input elements for a model's decision. Several studies (e.g., Jain and Wallace 2019; Serrano and Smith 2019; Wiegreffe and Pinter 2019) have shown that modifying or randomising attention weights can sometimes leave the model's output unchanged, suggesting that attention does not always serve as a faithful explanation of model reasoning. Attention distributions are also influenced by factors such as normalisation, scaling, and layer interactions, which may distort their interpretive meaning. Moreover, different architectures or parameter settings can produce similar performance with very different attention patterns, indicating that multiple internal representations can lead to the same decision. Consequently, while attention visualisations and derived metrics (here, Gini coefficients, or positional

focus) can offer descriptive insights into model behavior, they should be interpreted with caution and complemented by other analytical methods (e.g., ablation studies) to more reliably assess how the model processes linguistic information.

6.3 Concrete example

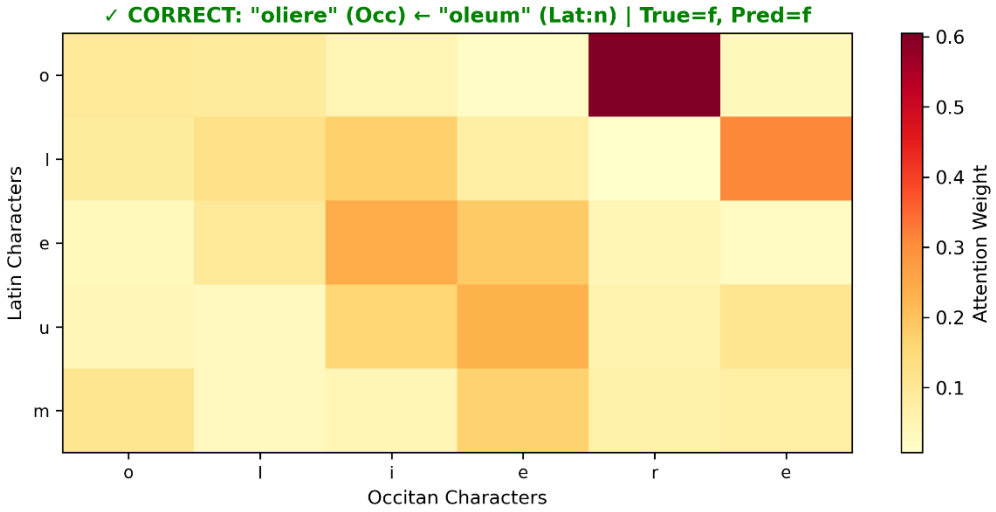

Figure 4: Cross-Lingual Attention Matrix for *oliere* (Occitan) and *oleum* (Latin).

The cross-lingual attention matrix for the correctly predicted pair OLEUM (Latin) > *oliere* (Old Occitan) (see Figure 4) illustrates how the model distributes its internal focus across graphemic correspondences between the two languages. The attention mechanism learns to associate Latin source characters with their Occitan reflexes, producing a soft alignment that encodes potential morphological and orthographic transformations. In this case, the model assigns moderate attention weights between the Latin vowels ***e*, *u*** and the Occitan ***i*, *e***, as well as distributed attention across internal character positions. These patterns suggest that the model captures aspects of non-linear morphological derivation rather than simple one-to-one morphological correspondence.

The highest attention weight, linking Latin *r* to Occitan *o*, lacks a straightforward linguistic basis and instead likely reflects the model's contextual weighting rather than genuine etymological correspondence. This underlines a known limitation of attention-based interpretability: attention weights express relative relevance for prediction, not guaranteed linguistic alignment. Particularly in multihead attention architectures, individual heads may attend to positional or structural cues rather than surface graphemic similarity.

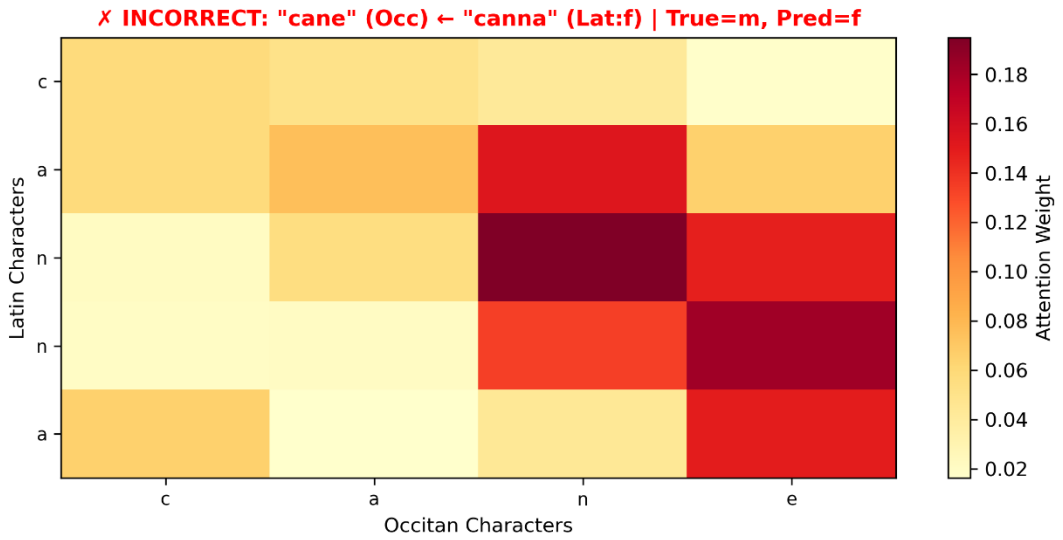

Figure 5: Cross-Lingual Attention Matrix for *cane* (Occitan) and *canna* (Latin).

The cross-lingual attention matrix for the misclassified pair CANNA (Latin) > *cane* (Old Occitan) shows a highly concentrated attention pattern centered on the internal characters of both words, particularly around the Occitan *a–n–e* and Latin *n–n–a* sequences. This distribution indicates that the model attempts to align the central segments of the two words, which share partial orthographic overlap. The attention mass is strongest around the Latin *n* and *a*, corresponding to the final syllable of *canna*, where the morphological gender marking in Latin (the *-a* ending) typically signals the feminine class.

The attention mechanism identifies the *a–n–a* substring as a strong alignment anchor, which likely biases the classifier toward the feminine class due to the statistical association between final *-a* endings and feminine gender in both Latin and Occitan. Consequently, the attention pattern mirrors a form-based generalisation bias. The network captures graphemic similarity but fails to differentiate between inherited gender patterns and chance orthographic resemblance.

Despite the misclassification, this example is informative. The visualisation reveals how the model internalises morpho-orthographic regularities (e.g., final *-a* = feminine). The error therefore does not appear to be random, but rather the result of a potentially systematic generalisation that does not have to correspond to a linguistically accurate interpretation. These findings should, however, be interpreted with caution: attention weights reflect relative relevance within the model's internal processing and do not necessarily correspond to explicit linguistic alignment. Further analyses, for example, through controlled probing or feature attribution methods, would be necessary to determine whether such attention patterns genuinely encode cross-linguistic structural relations or merely correlate with frequent surface cues.

7 Limitations

The following section describes the study's limitations, considering challenges arising from both the data and the modelling approach.

7.1 Limitations of the data

From a linguistic point of view, our data set has certain limitations, which will be discussed here, in order to better understand the outcome of our simulation and to be able to increase the authenticity for future simulations.

First of all, there are certain limitations in the Latin part of the data: as we did not yet treat the phonetic level, but rather took a grapheme-based approach, the vowel quantity was not taken into account. Latin is a language with a phonemic distinction according to the vowel length, which also plays a role in language change (see Loporcaro 2015: 9–12). Going hand in hand with this limitation, also the vowel quality in the transition from Latin to Old Occitan was not considered, i.e. differences in the close-mid vowels and open-mid vowels are not taken into account, given that their graphic representation in medieval texts is identic (diacritics to indicate these differences were applied consistently only later).

Further limitations with the Latin data set lie in the fact that we did not yet include graphical variation nor variation in gender as present in Latin. Graphical variation giving hints also on phonological variation certainly played a key role in language change. Moreover, variation in gender that was already present in Latin was often preserved into the Romance languages, leading to a certain number of gender variable nouns that either differ between languages (e.g. the nouns derived from MARE in the Romance languages, which is feminine in French and masculine in Italian) or even within varieties (e.g. *sanc*, which is masculine in Provencal and gender variable in the other Old Occitan varieties; see Wiedner to appear). In order to better display the variation in these domains that is inherited already from Latin, the information from the sections *scribitur* and *de genere* from the *Thesaurus Linguae Latinae* (ThLL) should be exploited.

A general problem with both the Latin and the Old Occitan data set is that case was not fully considered. The Latin data directly extracted from the FEW is only available in the nominative, even though it is well known in historical Romance linguistics that the majority of nouns inherited into the Romance languages are derived from the Latin accusative. Furthermore, the number value (singular versus plural) was not yet considered, even though the number in Latin plays an important role in reassignment processes of the Latin neuters from the second declension class in the transition to the Romance languages (see e.g. Chircu-Buftea 2011: 10). In Wiedner (to appear) it could be shown that, at least for the nouns studied there, neither case nor number matter significantly for gender variation in Old Occitan, but it would nevertheless be important to include and interconnect number-case combinations in both data sets.

A not yet resolved difficulty lies in the treatment of Old Occitan nouns that are derived from word classes other than nouns or that are loaned from other languages. For now, we excluded these problematic cases, but it would be important to include also these nouns in order to enrich our data set and to make it more varied. But it is not yet clear how to include these other word classes in our simulation, as we need a gender also for the etymon, since many word classes do not trigger gender agreement as they are no agreement controllers (and some, like verbs, are also no agreement target for gender). We therefore have a null-value in the respective slot that cannot be treated in our simulation.

The two factors that played an important role in the simulation by Polinsky and Van Everbroeck (2003) presented in Section 2.2 are also not yet included. We do not have any indications of absolute frequencies of the respective nouns so we could not yet include the respective information. To our knowledge, there is no dictionary or similar resource on overall frequencies of Old Occitan nouns. The other factor, the contact language Gaulish, is not included in our simulation, as we do not dispose of any Gaulish data, which is in general scarce (see Blom 2023: 130). As an additional contact language that was not used by Polinsky and Van Everbroeck (2003), we would like to include data on Medieval Latin for future simulations.

7.2 Limitations of the model

The limitations of the BiLSTM-Attention model for this study primarily stem from its character-level constraint and its inability to fully capture the contextual and semantic complexity of real-world gender systems. Operating only on character sequences means that the model implicitly learns morphological patterns, but lacks direct access to abstract linguistic units like morphemes, stems, or declension classes. As a result, its internal representations are difficult to interpret in traditional linguistic terms, making it challenging to relate learned patterns to known morphological structures.

Furthermore, the model is trained on isolated word forms rather than within sentential or even larger contexts, which limits its capacity to model gender according to its functional definition as agreement behaviour. It cannot account for the syntactic environments or agreement targets (articles, adjectives, or participles) that shape gender expression and variation within texts.

Additionally, the model must infer all outcomes from the form itself, as it lacks access to semantic features such as animacy, humanness, or referential gender, all of which played an important role in the historical development of gender assignment from Latin to Old Occitan. This absence of meaning-related information narrows the explanatory power of the simulation and prevents the model from capturing semantically motivated reassignments. Additionally, the use of a fixed categorical input for Latin gender oversimplifies the historical reality by assuming a stable source system. It ignores the variability of gender that was already present in Latin (see ThLL as a source).

Finally, since the model operates on a static mapping between Latin and Old Occitan forms without exposure to intermediate evolutionary stages, it effectively models end-point correspondences rather than the dynamic, multi-generational processes through which gender change occurred. Consequently, while the model offers insights into form-based correspondences, its diachronic realism and interpretability remain constrained by these structural and linguistic simplifications.

A further limitation concerns the interpretation of the cross-lingual attention mechanism. Although attention visualisations can be useful for exploring where the model tends to focus when relating Latin and Occitan forms, their linguistic significance remains uncertain. Attention weights quantify internal relevance within the model, but it is not clear to what extent they reflect genuine morphological correspondences as opposed to statistical or positional regularities in the data. In some cases, apparently structured alignment patterns may arise from architectural properties of the multihead attention mechanism or from training biases rather than from learned linguistic relationships. As a result, the heatmaps should be viewed as suggestive rather than as direct evidence of systematic cross-linguistic mapping. More targeted analyses (e.g., through probing, controlled ablations, or alignment-specific evaluation) would be needed to determine whether the observed attention patterns correspond to interpretable linguistic knowledge or simply reflect frequent co-occurrence cues in the dataset.

8 Conclusion

This study set out to explore the development of grammatical gender from Latin to Old Occitan through a computational lens, using a bidirectional LSTM model with attention to simulate the diachronic mapping of gender systems. Building on the previous simulation-based approach by Polinsky and Van Everbroeck (2003), our model provides a form-based, data-driven perspective on how gender marking might evolve under the interaction of morphological cues. Despite its inherent simplifications, the model achieved robust predictive accuracy and yielded interpretable tendencies in the distribution of attention, suggesting that even at the level of orthographic form, systematic regularities relevant to gender can be captured computationally.

The limitations of our approach underscore the complexity of modelling grammatical gender in its full functional sense. The model's purely character-level representation cannot capture the agreement-based nature of gender or the semantic and contextual forces that historically influenced its reorganisation. Likewise, the cross-lingual attention patterns should not be overinterpreted as direct evidence of linguistic alignment, although the plots are visually suggestive. They offer a heuristic view into the model's internal structure, pointing to areas where computational and linguistic reasoning may overlap but not necessarily coincide.

Methodologically, this study demonstrates the potential of neural architectures for historical-linguistic inquiry, while emphasising the need for

interpretive caution and linguistic grounding. The combination of quantitative evaluation and qualitative inspection opens promising directions for future work, including (i) the incorporation of intermediate historical stages to approximate gradual change, (ii) the inclusion of semantic and frequency information, and (iii) the extension of the model to contextualised input reflecting agreement behaviour.

The present results indicate that neural sequence models can capture elements of the morphological structure underlying gender change. The explanatory reach remains restricted without explicit linguistic and diachronic grounding. Advancing this line of research will require richer data and modelling strategies that connect form-based regularities with the functional and semantic dimensions of grammatical gender.

Conflicts of interest

The authors declare no conflicts of interest regarding the publication of this contribution.

References

Allassonnière-Tang, Marc & Basirat, Ali. 2020. Word embedding and neural network on grammatical gender – A case study of Swedish. *arXiv*. https://doi.org/10.48550/arXiv.2007.14222.

Allassonnière-Tang, Marc & Brown, Dunstan & Fedden, Sebastian. 2021. Testing semantic dominance in Mian gender: Three machine learning models. *Oceanic Linguistics* 60(2). 302–334. https://hal.science/hal-03509042v1/document (last accessed on 31/10/2025).

Bauer, Brigitte. 2010. Word Formation. In Maiden, Martin & Smith, John Charles & Ledgeway, Adam (eds): *The Cambridge History of the Romance Languages*, 532–563. Cambridge: Cambridge University Press.

Blom, Alderik H. 2023. Gaulish in the Late Empire (c. 200–600 CE). In Mullen, Alex & Woudhuysen, George (eds): *Languages and Communities in the Late-Roman and Post-Imperial Western Provinces*, 129–154. Oxford: Oxford University Press.

Burns, Patrick J. 2023. LatinCy: Synthetic trained pipelines for Latin NLP. *arXiv*. https://doi.org/10.48550/arXiv.2305.04365.

Chambon, Jean-Pierre. 2003. La déclinaison en ancien occitan, ou : comment s'en débarrasser ? : Une réanalyse descriptive non orthodoxe de la flexion substantivale. *Revue de Linguistique Romane* 67(267–268). 343–364.

Chircu-Buftea, Adrian. 2011. *Précis de morphologie romane*. Cluj-Napoca: Casa Cartii de Stiinta.

Coker, Amy. 2009. Analogical change and grammatical gender in ancient Greek. *Journal of Greek Linguistics* 9(1). 34–55.

https://www.sciencedirect.com/org/science/article/pii/S1566584409000026 (last accessed on 31/10/2025).

Corbett, Greville G. 2006. *Agreement*. Cambridge: Cambridge University Press.

Cotterell, Ryan & Kirov, Christo & Hulden, Mans & Eisner, Jason. 2018. On the diachronic stability of irregularity in inflectional morphology. *arXiv*. https://doi.org/10.48550/arXiv.1804.08262

Cucerzan, Silviu & Yarowsky, David. 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*. 40–47. https://aclanthology.org/N03-1006/ (last accessed on 31/10/2025).

dDOM = Stimm, Helmut & Stempel, Wolf-Dieter & Selig, Maria (eds). 1960–. *digitales Dictionnaire de l'occitan medieval*, München: Bayerische Akademie der Wissenschaften. https://dienste.badw.de:9999/dom/db (last accessed on 31/10/2025).

Derrer, Felix. 1974. *Lo codi : eine Summa codicis in provenzalischer Sprache aus dem XII. Jahrhundert; die provenzalische Fassung der Handschrift A; (Sorbonne 632); Vorarbeiten zu einer kritischen Textausgabe*. Zürich: Juris.

DOMél = Stimm, Helmut & Stempel, Wolf-Dieter & Selig, Maria (eds). 1960–. *Dictionnaire de l'occitan médiéval*. http://www.dom-en-ligne.de/ (last accessed on 31/10/2025).

Donaldson, Bryan & Sibille, Jean. 2024. Histoire interne de la langue. In Esher, Louise & Sibille, Jean (eds). *Manuel de linguistique occitane*, 193–230. Berlin, Boston: De Gruyter.

FEW = Wartburg, Walter von. 1922–2002. *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*, 25 vols. Basel: Zbinden. https://lecteur-few.atilf.fr/ (last accessed on 31/10/2025).

Hare, Mary & Elman, Jeffrey L. 1995. Learning and morphological change. *Cognition* 56. 61–98. https://www.sciencedirect.com/science/article/pii/0010027794006555 (last accessed on 31/10/2025).

Hochreiter, Sepp & Schmidhuber, Jürgen. 1997. Long short-term memory. In *Neural computation* 9(8). 1735–1780. https://www.bioinf.jku.at/publications/older/2604.pdf (last accessed on 31/10/2025)

Hockett, Charles Francis. 1962. *A Course in Modern Linguistics* (4th edition). New York: Macmillan.

Jain, Sarthak & Wallace, Byron C. 2019. Attention is not Explanation. 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)*. 3543–3556. https://doi.org/10.48550/arXiv.1902.10186

Jensen, Frede. 1973. Désaccord entre genre et flexion : Les substantifs masculins à desinence féminine en provençal. *Revue des langues romanes* 80. 393–404.

Jensen, Frede. 1976. *The Old Provençal Noun and Adjective Declension*. Odense: Odense University Press.

Loporcaro, Michele. 2015. *Vowel Length from Latin to Romance*. Oxford: Oxford University Press.

Loporcaro, Michele. 2018. *Gender from Latin to Romance: History, Geography, Typology*. Oxford: Oxford University Press.

Marzo, Daniela & Wiedner, Marinus. 2025. Remarks on grammatical gender in Romance. In Linzmeier, Laura & Teixera Kalkhoff, Alexander M. & Wiesinger, Evelyn (eds), *Parla, e sia breve e arguto. Festschrift für Maria Selig / Studies in Honor of Maria Selig* (ScriptOralia 147). 201–207. Tübingen: Narr.

Panhuis, Dirk. 2015. *Lateinische Grammatik*. Berlin/Boston: De Gruyter.

Pinkster, Harm. 2015. *The Oxford Latin Syntax Volume I*. Oxford: Oxford University Press.

Polinsky, Maria & van Everbroeck, Ezra. 2003. Development of gender classifications: Modeling the historical change from Latin to French. *Language* 79(2). 356–390. https://www.jstor.org/stable/4489422 (last accessed on 31/10/2025).

Sahai, Saumya & Sharma, Dravyansh. 2021. Predicting and explaining French grammatical gender. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*. 90–96. https://aclanthology.org/2021.sigtyp-1.9/ (last accessed on 31/10/2025).

Schöffel, Matthias & Wiedner, Marinus & Garcés Arias, Esteban & Ruppert, Paula & Heumann, Christian & Aßenmacher, Matthias. 2025a. Modern models, medieval texts: A POS tagging study of Old Occitan. In Hämäläinen, Mika & Öhman, Emily & Bizzoni, Yuri & Miyagawa, So & Alnajjar, Khalid (eds), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. 334–349. https://aclanthology.org/2025.nlp4dh-1.30/ (last accessed on 31/10/2025).

Schöffel, Matthias & Garcés Arias, Esteban & Wiedner, Marinus & Ruppert, Paula & Li, Meimingwei & Heumann, Christian & Aßenmacher, Matthias. 2025b. Unveiling factors for enhanced POS tagging: A study of low-resource medieval Romance languages. *arXiv*. https://doi.org/10.48550/arXiv.2506.17715.

Serrano, Sofia & Smith, Noah A. 2019. Is Attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2931–2951. https://aclanthology.org/P19-1282/ (last accessed on 31/10/2025).

ThLL = *Thesaurus linguae latinae*. Berlin/Boston: De Gruyter. https://publikationen.badw.de/de/thesaurus/lemmata (last accessed on 31/10/2025).

Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Łukasz & Polosukhin, Illia. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems* 30. 5998–6008.

Veeman, Hartger & Basirat, Ali. 2020. An exploration of the encoding of grammatical gender in word embeddings. *arXiv*. https://doi.org/10.48550/arXiv.2008.01946.

Wang, Ziwen. 2023. *Pérdida del género neutro del latín al hispanorromance altomedieval. Una reconstrucción panrománica*. Barcelona: Doctoral thesis. https://ddd.uab.cat/pub/tesis/2024/hdl_10803_691297/ziwa1de1.pdf

Wichers Schreur, Jesse & Allassonnière-Tang, Marc & Bellamy, Kate & Rochant, Neige. 2022. Predicting grammatical gender in Nakh languages: Three methods compared. *Linguistic Typology at the Crossroads* 2(2). 93–126. https://doi.org/10.6092/issn.2785-0943/14545.

Wiedner, Marinus 2023. Old Occitan handwriting. (Modell-Nr. 52822, CER=3,51%), PyLaia-Modell for handwritten Occitan from the 13th and 14th century. [Computer software]. https://app.transkribus.org/models/public/text/old-occitan-handwriting (last accessed on 31/10/2025).

Wiedner, Marinus (ed.). 2025. COMETA : *Corpus de l'occitan médiéval comparative et annoté: Provence et Languedoc.* https://zenodo.org/records/15300719 (last accessed on 31/10/2025).

Wiedner, Marinus. to appear. Doublons de genre en occitan médiéval : huit études de cas sur corpus. In Wissner, Inka & Dufter, Andreas (eds): *Aspects de la variation en diachronie. Regards sur la Galloromania.* (Beihefte zur Zeitschrift für romanische Philologie). Berlin/Boston: De Gruyter.

Wiegreffe, Sarah & Pinter, Yuval. 2019. Attention is not not Explanation. 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). 11–20. https://aclanthology.org/D19-1002/ (last accessed on 31/10/2025).

Williams, Adina & Pimentel, Tiago & Blix, Hagen & McCarthy, Arya D. & Chodroff, Eleanor & Cotterell, Ryan. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6682–6695. https://aclanthology.org/2020.acl-main.597/ (last accessed on 31/10/2025).