Lexique
ISSN : 2804-7397
Éditeur : Université de Lille

38 | 2026
Varia

Assessing the import of morphology in gender assignment: A computational study of French

Évaluer l'importance de la morphologie dans l'attribution du genre : une étude computationnelle du français

Nazanin Shafiabadi, Nina Nusbaumer et Olivier Bonami

https://www.peren-revues.fr/lexique/2262

DOI : 10.54563/lexique.2262

Référence électronique
Nazanin Shafiabadi, Nina Nusbaumer et Olivier Bonami, « Assessing the import of morphology in gender assignment: A computational study of French », *Lexique* [En ligne], 38 | 2026, mis en ligne le 01 juillet 2026, consulté le 09 septembre 2026.
URL : https://www.peren-revues.fr/lexique/2262

Droits d'auteur
CC-BY

Assessing the import of morphology in gender assignment: A computational study of French

Évaluer l'importance de la morphologie dans l'attribution du genre : une étude computationnelle du français

Nazanin Shafiabadi, Nina Nusbaumer et Olivier Bonami

PLAN

1. Introduction
2. Background
3. A computational model of gender assignment
   3.1. Data sources and corpus design
   3.2. Form representation
   3.3. Model and training procedure
4. Assessing model performance
   4.1. General performance
   4.2. Error analysis
      4.2.1. Model agreement
      4.2.2. Model disagreement
5. Assessing the time course of prediction
   5.1. Motivation
   5.2. The model learns highly predictive endings
   5.3. Highly predictive endings tend to be suffixes
   5.4. Interim conclusion
6. Conclusions

NOTES DE L'AUTEUR

Supplementary materials All the data and code reported on in this paper is available at: https://doi.org/10.17605/OSF.IO/HZAR9
Contributions The three authors contributed equally to this work.
Acknowledgements Aspects of this work were presented at the workshop on *Future directions in Word and Paradigm morphology* held as part of the *Première rencontre Langues et langage à la croisée des disciplines*, Paris, September 2024. We are grateful to workshop organizers, reviewers, and audience participants, as well as two anonymous reviewers from *Lexique*, for their valuable input. The work builds heavily, on the one hand, on an unpublished joint study of the last author together with Matías Guzmán Naranjo and Delphine Tribout (Bonami et al. 2019), and, on the other hand, on expertise developed by the first two authors when conducting

a related study as an *Advanced Morphology* class project at Université Paris Cité in 2023, together with Nickolas Dacelo.
Funding This work has received support from the investment programme "France 2030" as part of the IdEx programme (ANR-18-IDEX-0001) implemented by Université Paris Cité, under which the inIdEx project EFL is conducted.

TEXTE

1. Introduction

1 Grammatical gender is known to follow systematic patterns, yet its assignment is not always straightforward. While morphology provides transparent cues in many cases – such as suffixes that reliably mark masculine or feminine – exceptions, irregularities, or competing cues can make gender assignment more difficult. French, in particular, has a two-gender system – feminine and masculine – where the grammatical gender of a noun can often be inferred from its semantics (e.g., nouns referring to women are typically feminine), orthography and/or phonology (e.g., nouns ending in <ère> /ɛʁ/ are typically feminine), or morphology (e.g., nouns ending in the suffix *-euse* are always feminine). Although it has long been established that grammatical gender is not arbitrary (Corbett, 1991), its predictors are not always easy to disentangle: is the gender of *actrice*$_{\text{F}}$ 'actress' determined by its semantics or by the suffix *-rice*? Notably, about a third of French nouns with a unique gender end in a derivational suffix (Bonami, Guzmán Naranjo & Tribout, 2019), and most suffixes are compatible with only one gender (Bonami & Boyé, 2019). This raises the question of whether gender assignment is primarily predictable from morphology, phonology, or both.

2 Since the 2000s, machine learning techniques have provided new ways of addressing this question. The standard approach is to train a computational classifier to predict gender from a set of features and then interpret its accuracy on unseen data as an indicator of predictability. Previous studies (Sokolik & Smith, 1992; Eddington, 2002; Matthews, 2005; Plaster, Polinsky & Harizanov, 2013; Bonami et al., 2019) have applied such methods to a range of languages. However, these studies are limited in two important respects. First, they rely on relatively small datasets (600–3,683 datapoints), which restricts the

generalizability of their findings. Second, the model architectures used – such as Skousen's (1989) Analogical Modeling of Language (AML), perceptrons, or decision trees – fail to capture the sequential and hierarchical properties of word forms. Consequently, it remains unclear to what extent grammatical gender can be predicted from word form alone, and what features drive this predictability.

3 In this study, we address these limitations by investigating French gender assignment with Long Short-Term Memory (LSTM) networks trained on a large dataset (~20,000 nouns). By comparing different input configurations, we assess the extent to which gender is predictable from word form alone, without explicit morphological annotation. Specifically, we train models on orthographic and phonemic representations separately, using a character-by-character approach. This design allows us not only to measure overall predictability, but also to trace how predictions evolve as more of the word becomes available at each time step. Our findings show that gender is highly predictable from form alone, and that model performance is largely driven by derivational suffixes or suffix-like endings.

4 The remainder of the paper is structured as follows: Section 2 reviews previous work on gender assignment and prediction. Section 3 describes the implementation of our model. Section 4 presents our main results, including overall model performance and a detailed error analysis. Section 5 examines the time course of the models' predictions as characters are incrementally added. Section 6 concludes.

2. Background

5 Grammatical gender is a classification of nouns manifest in the different agreement patterns they trigger (Hockett, 1958). For instance, the two French nouns *roman*$_{\text{M}}$ 'novel' and *histoire*$_{\text{F}}$ 'story' belong to different genders, since they cooccur with different inflected forms of determiners and adjectives (1).

(1) a. *un* *beau* *roman*
IND.M.SG beautiful.M.SG novel[M].SG
a beautiful novel'
b. *une* *belle* *histoire*

IND.F.SG beautiful.F.SG novel[F].SG
'a beautiful story'

6 Although the grammatical gender of a noun is in principle arbitrary, it tends to be partially predictable from other properties of the noun.[1] This is manifest in the names traditionally given to genders, which tend to reflect semantic generalizations: for instance, where most nouns used to refer to women fall in the same gender, that gender is often called *feminine*; and *mutatis mutandis* for men and masculine gender. Broad semantic classes are far from being the only possible motivating factor; in many gender systems, gender assignment has been found to be motivated by a combination of phonological shape, morphological properties (inflection class, derivational history), and semantic factors (Corbett, 1991). Going back to French, a study of a lexicon of more than 30,000 nouns by Tucker, Lambert and Rigault (1977) provided an early overview of the relevant factors, uncovering striking generalizations, such as the fact that inanimate VN compounds (e.g. *casse-noix*$_{M}$ 'nutcracker', literally 'break-nut') are systematically masculine.

7 Early work in the vein of Corbett (1991) attempted to account for gender assignment in the form of a system of ordered rules, where the order is intended to capture the relative importance of different rules for one language. Building heavily on Tucker, Lambert & Rigault (1977), Corbett (1991, pp. 57-61) outlines the assignment rules of French as follows:

(2) a. Semantic assignment rules
- Sex-differentiable nouns denoting males are masculine.
- Sex-differentiable nouns denoting females are feminine.
- [...]

b. Morphological assignment rules
- Compound nouns formed from a verb plus some other element are masculine.
- [...]

c. Phonological assignment rules
- Nouns ending in /ʒ/ are masculine
- Nouns ending in /m/ are masculine
- [...]
- Nouns in /ezɔ̃/, /sjɔ̃/, /zjɔ̃/, /ʒjɔ̃/ and /tjɔ̃/ are feminine; remaining nouns in /ɔ̃/ are masculine.

8 Corbett argues for the ordered character of rules by pointing out that the semantic rules take precedence over the formal rules: for instance the noun *femme*$_{F}$ 'woman' is not subject to the phonological rule about /m/-final nouns, as it has already been dealt with by the semantic rules. A point he glosses over, however, is possible hesitations about phonological or morphological generalizations. For instance, the vast majority of nouns ending in /ʒ/ also end in the masculine-assigning derivational suffix *-age* (e.g. *ratage*$_{M}$ 'failure' from *rater* 'fail'); according to Bonami et al. (2019), the proportion of masculines falls to 50% among nonsuffixed nouns, so that arguably there is no phonological generalization to speak of. In the same vein, it is striking that the list of feminine assigning phonological endings in the last rule (/ezɔ̃/, /sjɔ̃/, /zjɔ̃/, /ʒjɔ̃/ and /tjɔ̃/) more or less coincides with the list of nouns ending in the suffixes *-aison* (e.g. *comparaison*$_{F}$ 'comparison' from *comparer* 'compare') and *-ion* (e.g. *division*$_{F}$ 'division' from *diviser* 'divide'), both ultimately descending from Latin *-io* and forming action nouns from verb, while most masculines in /ɔ̃/ end in the denominal suffix *-on* (e.g. *carton*$_{M}$ 'cardboard' from *carte*$_{F}$ 'card'). While the exact analysis of *-ion* suffixation is a contentious issue (see e.g. Bonami, Boyé and Kerleroux 2009), the extent to which Corbett's generalization should be presented as phonological rather than morphological is certainly subject to interpretation.

9 From a contemporary perspective, it is appropriate to recast gender assignment in terms of multivariate probabilistic prediction: when they need to make a prediction about the gender of a noun, speakers may rely on a combination of cues of various types, and there is no reason *a priori* to assume that some types of cues are more relevant than others. In a recent paper, Fedden, Guzmán Naranjo and Corbett (2025) provide a magistral demonstration of the relevance of such an approach. Relying on an annotated dataset of more than 30,000 German nouns derived from the CELEX lexical database (Baayen, Piepenbrock & Gulikers, 1995), Fedden and colleagues train computational classifiers to predict gender from a wide variety of predictors, both separately and in combination: phonological shape, broad lexical semantic features, fine syntactic and semantic properties as captured by distributional vectors, and inflection class. This allows them to reach a number of important conclusions, including

the following. First, German gender is largely predictable; the best combination of predictors, combining phonological shape, inflection class, and semantics, accurately predicts gender 96% of the time. Second, all five types of predictors have predictive value, as all classifiers considered do better than a random baseline. Third, phonological shape is the best single predictor, reaching an accuracy of 92% on its own, which is much higher than the second best predictor, inflection class, at 86%. Finally, derivational information has no extra predictive power once phonological shape is taken into account.

10 Fedden et al.'s (2025) study hence highlights the usefulness of computational modelling in providing definite answers to questions about the extent to which various properties of nouns participate in gender assignment, the extent to which they are predictive, and how they interact. In an older but yet unpublished study, Bonami et al. (2019) used essentially the same approach to examine gender assignment in French.[2] Relying on a preliminary version of the dataset later published as Échantinom (Bonami & Tribout, 2021), they trained classifiers to predict gender from phonological information (identity of the last 3 segments), information on word formation (broad word formation type, identity of the prefix if any, identity of the suffix if any), or both. When focusing on just suffixed nouns, they found that phonological shape was an excellent predictor (accuracy 94%) but was outperformed by knowledge of the identity of the suffix (97%). However, when considering the dataset as a whole, phonological shape was the best predictor (80%), and addition of morphological information did not lead to a measurable improvement in accuracy.

11 Bonami et al. (2019) speculate that these results follow from the fact that phonological shape is an excellent proxy for information about suffixes. In their reasoning, the true cause of gender predictability for suffixed nouns is that most suffixes are gender assigning: all nouns formed with such suffixes have the same gender. However, suffixed nouns create an imbalance in the relationship between phonological shape and gender: when some word-final phoneme sequence coincides with a productive suffix, all other things being equal, the vast majority of nouns ending in that sequence are formed with that suffix, and hence will have the gender associated with the suffix. As a result, a (human or computational) classifier will mostly get the

information they need to predict gender by looking at just phonology, although the ultimate cause of predictability is morphological.

12 As appealing as it is, this story at this point is largely speculative. It could very well be that causality runs in the other direction, and that suffixes happen to have a shape that make them fit into phonological classes of words for which gender is predictable. In addition, the model architecture used by Bonami et al. limits interpretability. First, it relies on a fixed-length window at the end of the word for prediction, not taking into account the fact that affixes have diverse lengths. Second, there is no simple interpretation of the internal states of the model, making it hard to make sense of whether it arrives at the same conclusion on two different data points for the same or different reasons.

13 The goal of the present study is to move beyond speculation, and provide evidence that morphology is the ultimate cause of gender predictability. To this end, we predict gender from shape using a type of recurrent neural network, as detailed in the next section. The appeal of this strategy is that the model adjusts its prediction as it progresses within the word character by character. This allows us to probe the extent to which words that are morphologically similar are processed in the same way.

3. A computational model of gender assignment

14 Building on the theoretical perspectives reviewed above, our study adopts an experimental approach designed to test whether morphological structure provides the decisive cues for gender assignment in French. More specifically, we ask to what extent grammatical gender can be predicted from surface form, and whether morphology – particularly suffixal morphology – provides the crucial information. We approach these questions through a computational framework that traces how predictions evolve along the word form. Our hypotheses are threefold. First, we expect the model to learn gender primarily from statistical associations between endings and gender. Second, we hypothesize that these associations largely arise from the presence of gender-assigning suffixes. Third, we anticipate that

errors will occur in cases where lexical items run counter to these generalizations. To lay the groundwork for our experiments, we first present the resources on which they rely (Section 3.1), then the representations of word forms used as input (Section 3.2), and finally the model architecture and training procedure (Section 3.3).

3.1. Data sources and corpus design

15 The primary resource for this study is *Flexique* (Bonami, Caron & Plancq, 2014), a large-scale inflectional lexicon of spoken French derived from *Lexique* version 3.70 (New, Pallier, Ferrand & Matos, 2001). Lexique is an open-source lexical database containing phonetic, lexical, morphosyntactic, and frequency information on 142,694 French words. Flexique extends Lexique by filling in the paradigms of all lexemes at least one of whose forms is found in Lexique, and providing normalized and carefully curated phonemic transcriptions. Flexique is organized into separate tables for nouns, adjectives, and verbs, and documents 31,005 noun lexemes. As the present study focuses exclusively on nominal gender, we extracted only the noun entries. Each noun is annotated with:

a. A unique lexeme identifier derived from the orthography of the citation form.
b. A list of orthographic variants of the citation form.
c. Gender information, encoded as 'm' for masculine, 'f' for feminine, and 'b' ('both') for common gender nouns with identical masculine and feminine forms (e.g. *secrétaire* 'secretary').
d. A list of inflected forms in quasi-IPA transcription.

16 For evaluation purposes, we relied on *Échantinom* (Bonami & Tribout, 2021), a partially overlapping lexicon that provides fine-grained morphological information and phonemic transcription for 5,000 French nouns. While Flexique offers broader lexical coverage, Échantinom provides morphological analysis, including explicit identification of affixes. This morphological information was not used during training, but rather serves as a resource for analyzing the models' predictions and identifying structural regularities in gender marking.

17 The distribution of gender across both (unprocessed) datasets is given in Table 1. To ensure a consistent binary classification task, we

excluded all nouns marked as 'b' from all subsets.

Gender	Flexique		Echantinom		
	#Total	Avg. Len	#Total	#Suffxed	Avg. Len
m	17,193 (≈56%)	8	2,875 (≈58%)	1,115 (≈39%)	8
f	12,219 (≈39%)	9	1,659 (≈33%)	764 (≈46%)	8
b	1,593 (≈5%)	9	466 (≈9%)[1]	201 (≈43%)	9
total	31,005	9	5,000	2,080 (≈42%)	8

Table 1. Distribution of gender across the noun lexicon in the Flexique and Echantinom datasets.

18 To avoid data leakage, we strictly separated the datasets as follows:

- Training set: 80% of the Flexique nouns not present in Échantinom and not marked as 'b' (both genders)
- Validation set: The remaining 20% of the same filtered subset
- Test set: All Échantinom nouns excluding those marked 'b'

19 An overview of the dataset partitions is provided in Table 2. This partitioning guarantees that no lexeme from the test set appears in the training or validation sets, thereby ensuring that evaluation is performed on entirely unseen lexical items.

Set	Size	Composition	Gender Distribution	
			Fem.	Mas.
Training	19,904	(F − E − *b*) x 80%	8,452 (≈42%)	11,452 (≈58%)
Validation	4,976	(F − E − *b*) x 20%	2,109 (≈42%)	2,867 (≈58%)
Test	4,532	(E − *b*)	1,659 (≈37%)	2,873 (≈63%)

Table 2. Overview of dataset partitioning and gender distribution. F = Flexique, E = Échantinom, b = nouns compatible with both genders.

3.2. Form representation

20 Each noun was represented as a sequence of characters. To investigate the contribution of different segments of the word to gender prediction, we employed two directional encodings:

1. Forward: characters are processed in their natural left-to-right order;
2.

Backward: characters are processed in reverse, from the final character to the first.

21 The backward encoding was designed to probe the hypothesis that the final segments of a noun – often corresponding to derivational suffixes – carry a disproportionate share of the predictive signal for gender assignment. Although the model has no explicit access to morphological boundaries, the prediction trajectory it produces can be aligned *post hoc* with the suffix boundaries annotated in Échantinom. This enables a fine-grained analysis of whether the point at which model confidence stabilizes corresponds to the morphological breakpoint.

22 We conducted parallel experiments using phonemic transcriptions in place of orthographic forms, drawing on the transcriptions available in Flexique. Since nasal vowels such as ɑ̃, ɔ̃, ɛ̃, and œ̃are represented as two-character sequences in the data, we applied a preprocessing step to re-encode each of them as a single atomic symbol (1–4). This ensured that the model treats each nasal vowel as a single, unified phonemic unit rather than a sequence of unrelated symbols. All textual inputs were already lowercased, so no further normalization or preprocessing was applied.

3.3. Model and training procedure

23 To model the sequential prediction of grammatical gender, we implemented a *Long Short-Term Memory* (LSTM) architecture (Hochreiter & Schmidhuber, 1997). This architecture is particularly well suited to modeling variable-length sequences and learning long-distance dependencies, making it an appropriate choice for our task of predicting grammatical gender from orthographic or phonemic input. The model processes each noun character by character and generates a probability distribution over gender values at every time step. This setup enables us not only to evaluate the model's final prediction but also to track the evolution of its confidence across the input sequence. In doing so, we can identify the point at which the model converges on a stable prediction and evaluate whether this point tends to coincide with known morphological suffix boundaries, thus offering indirect evidence regarding the salience of suffixes in gender assignment.

24 The architecture, illustrated in Figure 1, consists of the following components:

- Embedding layer: Maps each character in the input sequence to a vector , forming a sequence of character embeddings ;
- LSTM layer: processes the sequence of embeddings and returns a corresponding sequence of hidden states , where each ;
- Linear layer: projects each hidden state into a two-dimensional space, producing a logit vector representing the output classes (masculine vs. feminine);
- Output layer: applies a softmax function to each projected vector , yielding a probability distribution over the two classes at every character position.

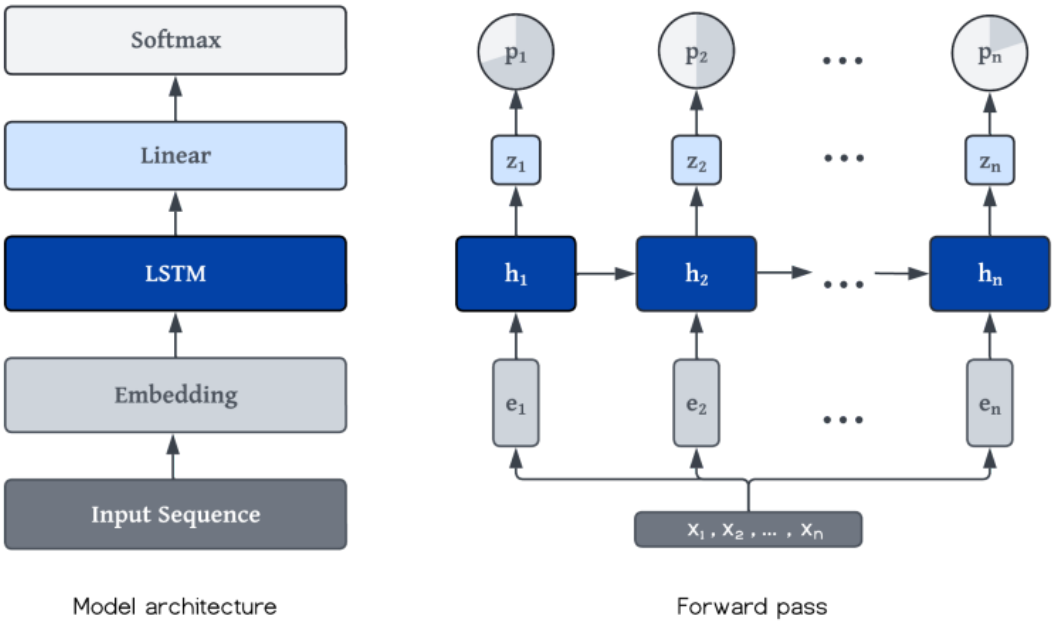


Figure 1. Model architecture (left) and forward pass (right). The left panel outlines the structural components of the model, while the right illustrates the flow of data and transformations applied at each stage.

25 During training, only the final output – after the complete word has been processed – is used to compute the cross-entropy loss. However, the intermediate outputs allow us to construct confidence curves that reflect the model's certainty at each step, which we analyze *post hoc*.

26 We trained separate models for each combination of representation type (orthographic vs. phonemic) and directionality (forward vs. backward), resulting in a total of four distinct models.[3] To establish a lower-bound baseline, we compared these models against a naïve majority-class classifier that always predicts the most frequent class in the training set (masculine).

27 This experimental setup provides a controlled framework for assessing the relative informativeness of orthographic versus phonemic representations, as well as forward versus backward encodings. By comparing the behavior of the four trained models with respect to their prediction curves and error patterns, we can directly evaluate whether morphological cues are indeed the underlying drivers of gender predictability. The following section presents the empirical outcomes of our experiments.

4. Assessing model performance

28 In this study, we examine the role of morphology in the prediction of gender of French nouns, focusing on models trained without explicit morphological knowledge. Gender assignment in French is challenging because multiple linguistic cues are closely intertwined: morphological, semantic, and phonological ones. Our goals are (i) to assess the extent to which gender can be inferred from form (orthographic or phonological) information alone, and (ii) to evaluate whether the model does so by identifying recurrent patterns of forms corresponding to suffixes. To this end, we analyze the predictions of four LSTM models that differ in the form of input (orthographic or phonemic) and in the direction in which the nouns are processed (forward or backward).

29 We begin by providing an overview of the models' predictions, whether correct or not, and assessing how closely they align with morphological boundaries. We then turn to instances of incorrect predictions to identify potentially systematic cues that would mislead the models' decisions. Finally, we analyse the time course of prediction while words are processed character by character, to better understand whether the model captures morphological regularities.

4.1. General performance

30 All four models reached similar accuracy levels, ranging from 80% to 90%, indicating that gender is highly predictable from surface form alone – whether orthographic or phonemic. To assess the extent to which the models have correctly learnt to recognize the patterns for gender assignment, we analyzed 4,532 lemmas from the test set, 37% of which are feminine nouns.

31 Subplot (a) in Figure 2, which shows individual model accuracies, indicates that both orthographic models outperform the phonemic ones, reaching above 90% accuracy compared to 81–82%. This suggests that orthographic form provides stronger cues for gender assignment than phonemic representations. Subplot (b), which breaks down the interaction between agreement and correctness, further reveals that in the majority of cases where all models agreed, the prediction was correct (n=3,211). Here, each bar corresponds to a combination of whether the four models agreed ("Yes"/_) and whether the prediction matched the gold label (_ /"Yes"). Thus, "Yes & Yes" indicates unanimous correct predictions, "Yes & No" unanimous errors, "No & Yes" cases with disagreement but at least one correct prediction, and "No & No" disagreement with no correct prediction. Whereas unanimous agreement was correct in the majority of cases, model disagreement was associated with higher error rates, as illustrated by the 818 cases where predictions diverged despite at least one model producing the correct label.

32 Subplot (c), presenting pairwise agreement across models, shows very high consistency between the two orthographic models (0.94) and slightly lower agreement between orthographic and phonemic models (≈0.83–0.86), suggesting that while both types of representation capture overlapping cues, their decision boundaries are not identical. Finally, subplot (d) shows overall agreement levels: full agreement across all four models occurs in 74% of cases, regardless of whether the prediction was correct, and majority agreement (≥3 models) rises to 91%, confirming that surface-form gender cues are largely robust and consistently detected across representations and directions.

33 From these comparisons we may conclude that there is a core of gender-relevant features that are robustly learned by all four models, irrespective of differences in model architecture.

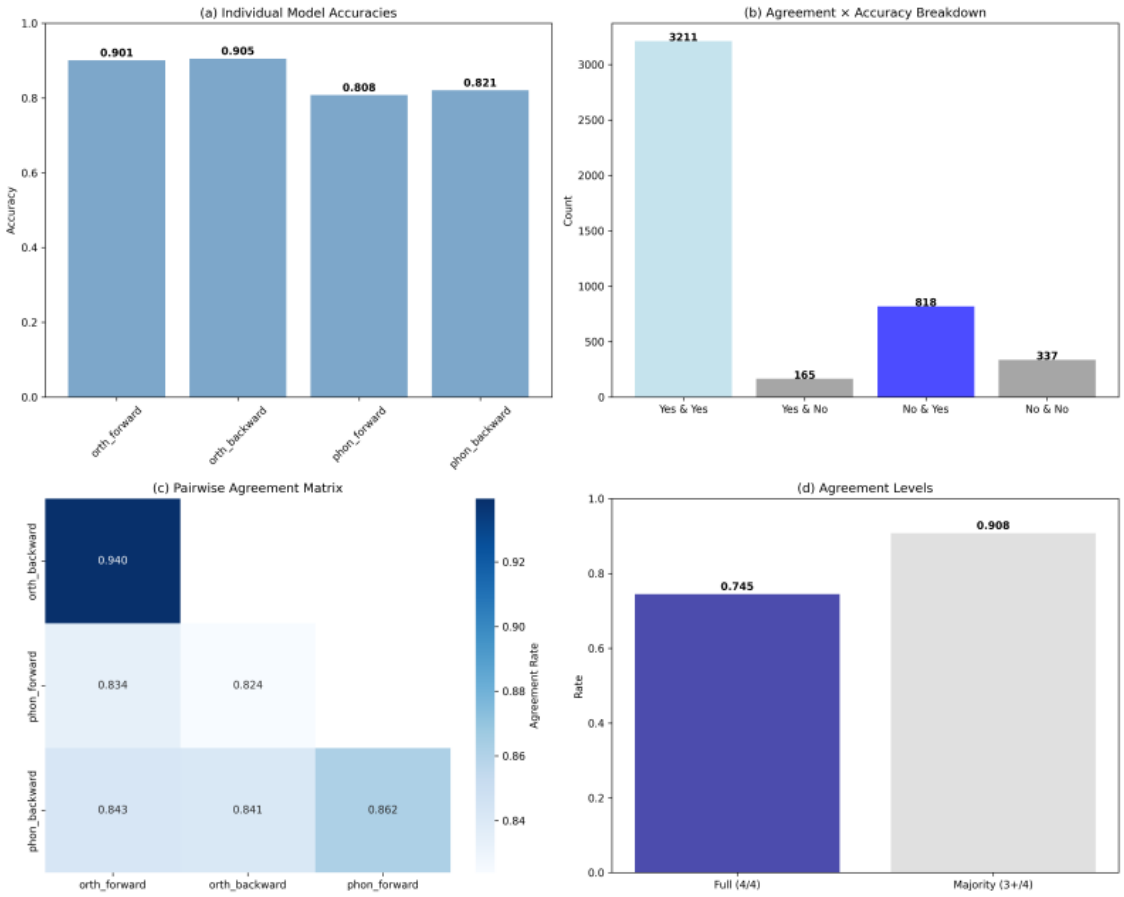


Figure 2. Model performance and agreement patterns across orthographic and phonemic representations

34 An interesting question is the extent to which the performance of the models varies with the token frequency of the nouns under consideration: since the Échantinom dataset filters out low frequency nouns, the statistical distribution of frequency is unbalanced between the training and test data, which could potentially lead to a drop in performance for higher frequency nouns.[4] To examine this, we computed frequency deciles for all nouns in the combined datasets, and report in Figure 3 the average inter-decile accuracy on the test data. It is notable that there is no sizable drop in accuracy as frequency increases in the orthographic models in the last frequency band, while there is a clear but moderate drop for the phonological models. This is unlikely to be due to a lack of evidence: while the

proportion of the data going to the test set increases with frequency, it is still below 30% in the last frequency band. Hence we speculate that the drop in accuracy is likely due to prediction of gender from phonology being genuinely harder for hard frequency nouns.

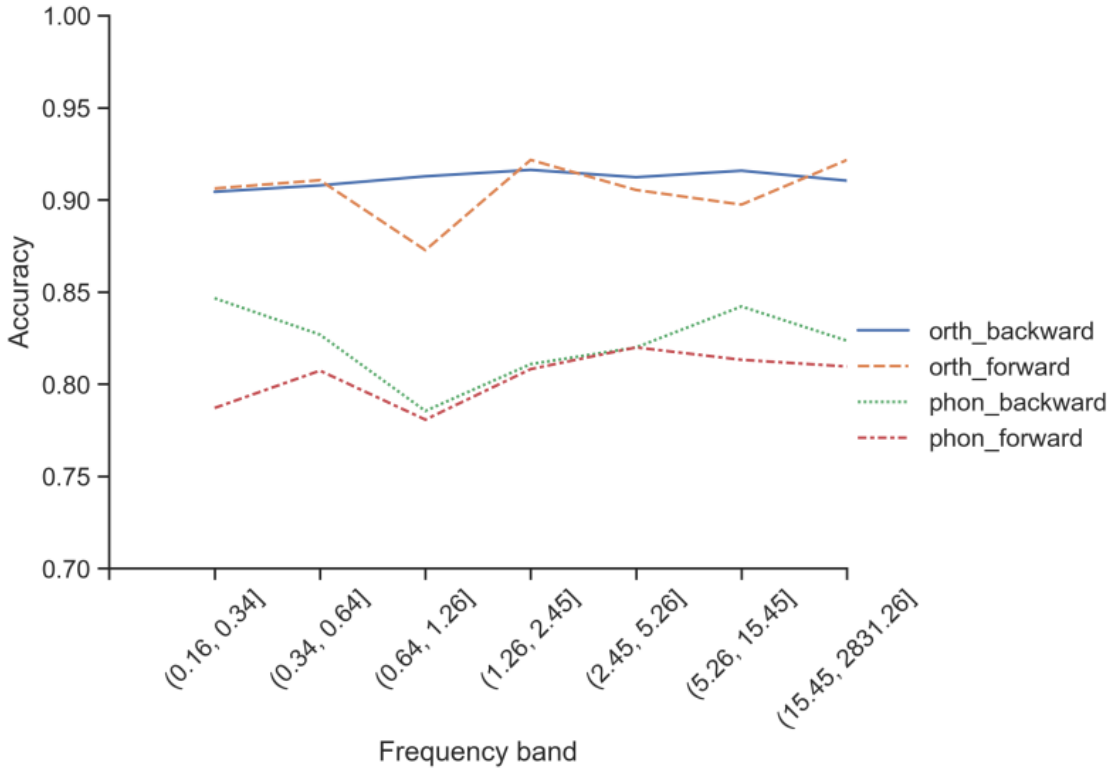


Figure 3. Average accuracy of the models in frequency bands corresponding to inter-decile intervals in the combined training, validation, and test datasets. Note that there is nothing to report below the third decile, as the test set only contains nouns with a frequency above 0.3.

4.2. Error analysis

35 An essential aspect of training artificial neural networks is minimizing the outcome variability across runs to guarantee reproducibility and strengthen the robustness of the conclusions. To this end, we compared each model's predictions over 10 runs and observed no significant differences in the number of errors (orthographic models: mean = 423, median = 421.5; phonemic models: mean = 902.5, median = 884). Of these, 42% (n=179) of the errors in the orthographic backward model and 38% (n=347) in the phonemic backward model were consistent across all runs. This stability justifies our decision to select one representative run at random for the analyses presented below.

36 Figure 4 summarizes the distribution of predicted genders among the errors for each model. Across all models, we observe a systematic bias: feminine nouns are more frequently misclassified as masculine than the reverse. Binomial tests indicate that this asymmetry is statistically significant for the phonemic models, where errors involving feminine nouns account for 65% of total errors in the forward model ($p < 0.001$) and 60% in the backward model ($p < 0.001$). In the orthographic models, the effect is weaker: 59% of errors involve F→M misclassifications in the forward model ($p = 0.003$), and 51% in the backward model ($p = 0.36$), the latter not reaching significance. This tendency may suggest less regular or exclusive gender-marking patterns for feminine nouns, especially in the case of phonemic form.

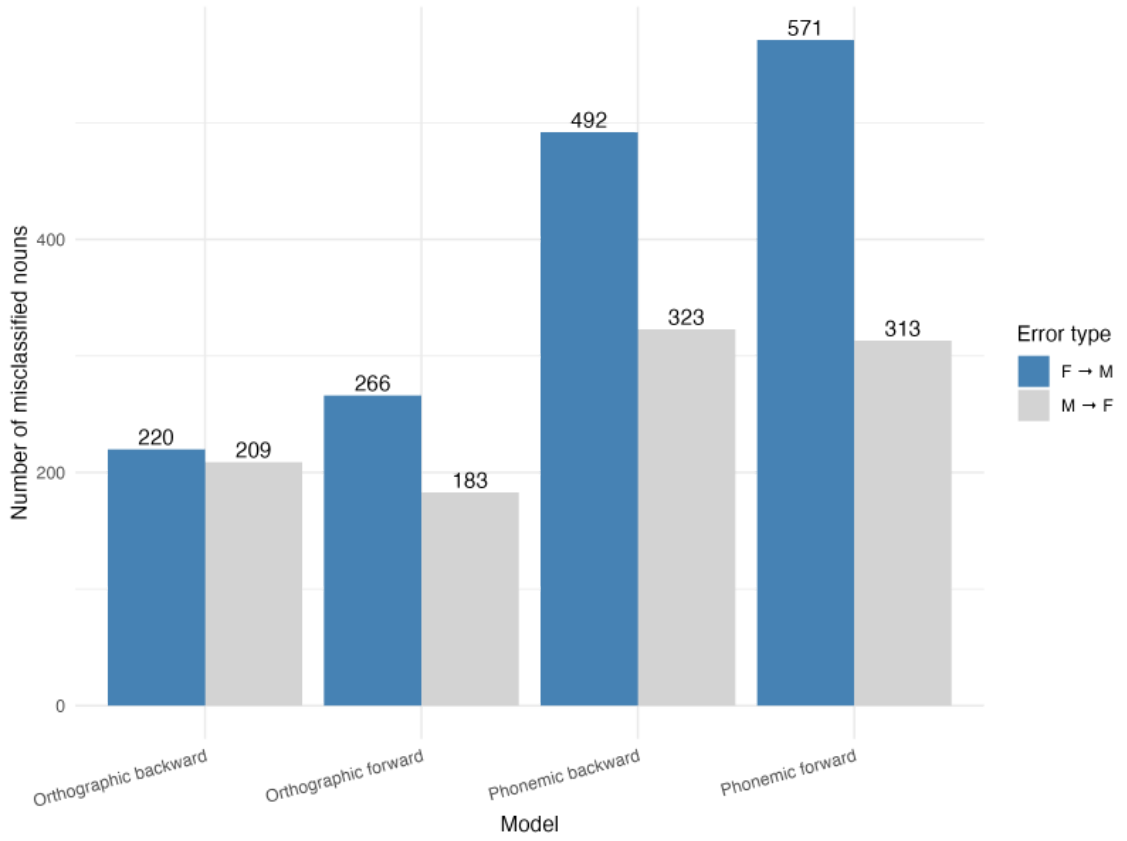


Figure 4. Asymmetry of gender misclassification errors across model configurations: true gender being misclassified as the other one.

4.2.1. Model agreement

37 To better understand the systematic sources of misprediction, we analyze the 242 nouns that are consistently misclassified by both the

orthographic and phonemic models. The errors produced by the different models are of different morphological types. These are largely led by simplexes, i.e. nouns that are not the result of a French [5] word formation process (see Table 3). While constituting 43% of the test set, these nouns account for 66% of the errors generated by all models, yielding a relative error proportion of 1.53× (+53%). Prefixed nouns, while representing a much smaller proportion of the errors, also show a moderate overrepresentation (1.5×, +50%), just like compounds that show only a slight inflation (1.2×, +20%). By contrast, suffixed nouns represent 11% of the errors, while making up 37% of the test set, making them substantially underrepresented in the errors (0.3×, −70%). The high relative proportion of errors of the words formed by nonconcatenative processes (e.g., *info*$_{F}$ 'piece of information', *mayo*$_{F}$ 'mayonnaise') can be explained by their weak representation among the training data, forcing the models to rely on other cues. This finding suggests that word-ending patterns play an important role in the gender marking of French nouns.

Process	Distribution in the test set	Proportion among errors	Relative proportion
Simplex	43%	66%	1.53× (≈ +53%)
Suffix	37%	11%	0.30× (≈ −70%)
Nonconcat	2%	7%	3.50× (≈ +250%)
Conversion	10%	7%	0.70× (≈ −30%)
Polylexical	5%	6%	1.20× (≈ +20%)
Prefix	2%	3%	1.50× (≈ +50%)

Table 3. Relative proportion in errors given each process' distribution in the test set and its proportion among the errors.

38 While misclassifications occur across all morphological processes, neither the process itself nor its relative frequency is sufficient to account for the observed errors. Instead, a more consistent explanation emerges from the endings of the nouns: errors systematically cluster around word-final sequences that are strongly associated with one gender, specifically in the relatively rare cases where individual nouns defy this dominant pattern (e.g. *fleur*$_{F}$ 'flower' vs. *professeur*$_{M}$ 'teacher', *chauffeur*$_{M}$ 'driver', *masseur*$_{M}$ 'masseur'). Crucially, this effect holds regardless of whether the ending corresponds to a productive suffix as in *nourrisson*$_{M}$ 'newborn' or merely resembles one in form as in *centurion*$_{M}$ 'centurion'. This conclusion is

further reinforced by the composition of the test set: among suffixed nouns, gender distribution is relatively balanced (43% of them are feminine) and a large proportion of non-suffixed nouns (61%) nevertheless display endings that resemble productive suffixes. In other words, the decisive factor seems not to be the morphological process by which the noun was formed, but the surface ending that the model interprets through the lens of gender regularities learned during training. For instance, feminine nouns in *-eur* (e.g. *couleur*$_{F}$ 'color') and masculine nouns in *-ère* (e.g. *frère*$_{M}$ 'brother') are particularly prone to errors because the dominant gender pattern for these endings points in the opposite direction. In such cases, morphological regularities are decisive, regardless of the morphological process: *-ère* as a suffix is consistently feminine (e.g., *fermière*$_{F}$ 'female farmer', *poussière*$_{F}$ 'dust', *matière*$_{F}$ 'matter'), while *-eur* overwhelmingly marks masculine nouns (e.g., *acteur*$_{M}$ 'actor', *chanteur*$_{M}$ 'singer', *tracteur*$_{M}$ 'liar'), with rarer feminine occurrences (e.g., *sœur*$_{F}$ '*sister*', *lueur*$_{F}$ 'glimmer', *ardeur*$_{F}$ 'eagerness'). Similar mismatches are observed with other ending patterns, such as /ɛn/, which is strongly associated with feminine nouns (e.g., *fontaine*$_{F}$ 'fountain', *africaine*$_{F}$ 'African woman', *européenne*$_{F}$ 'European woman'), leading the models to systematically misgender nouns like *capitaine*$_{M}$ ('captain'), where the surface ending conflicts with the predominant morphological tendency. In phonemic transcriptions, the contrast between genders associated with a given ending is sometimes less pronounced, as different endings can be realized identically. This reduces the model's bias toward one gender or the other. For example, both *gruyère*$_{M}$ and *fermière*$_{F}$ share the same phonemic ending /jɛʁ/, despite belonging to different genders.

39 Polylexical nouns are also noteworthy, as they sometimes create cases where the final element of a compound belongs to the opposite gender from the overall noun. For example, *serre-tête*$_{M}$ ('headband') ends with *tête*$_{F}$ ('head'), which has a typically feminine ending /ɛt/, and *croquemitaine*$_{M}$ ('bogeyman') shares its head with *mitaine*$_{F}$ ('mitten'), ending in the predominantly feminine suffix *-aine*.

40 As these examples illustrate, the factors driving misclassification operate rather independently of morphological categories. Instead, errors almost always stem from endings that are strongly associated

with the opposite gender, whether or not these endings align with morphological boundaries.

4.2.2. Model disagreement

41 Having identified the cues that led to the same errors in both the orthographic and phonemic models, we now examine the challenges in gender prediction specific to the orthographic and phonemic forms separately to better disentangle the role of morphology in gender prediction. In particular, we investigate the misleading cues specific to each representation by analyzing the errors that are unique to the orthographic models and those that are unique to the phonemic models. The backward models produced a sample of 719 lemmas, while the forward models yielded 767. Our analysis here focuses on the sample of non-overlapping errors between the two forms.

42 This forward sample consists of 767 lemmas, of which 166 correspond to orthographic errors and 601 to phonemic errors (see Table 4). We note the important disparity in representation between both forms, which suggests that phonemic forms may be less discriminative with respect to gender. Within this subset, half of it (55%) are feminine nouns, accounting for 34% of the orthographic errors and 60% of the phonemic errors. These results suggest that the morphology of feminine nouns is more ambiguous than that of masculine nouns from the perspective of phonemic forms, while it is the opposite for the orthographic form.

Category	**Error set (true gender)**	**Orthographic (wrong gender)**	**Phonemic (wrong gender)**
Feminine (F)	419 (55%)	57 (predicted as M - 34%)	362 (predicted as M - 60%)
Masculine (M)	348 (45%)	109 (predicted as F - 66%)	239 (predicted as F - 40%)
Total	**767**	**166**	**601**

Table 4. Gender prediction error counts by form compared to their proportion in the sample of non-overlapping errors between the two forms.

43 We aim to identify the specific challenges faced by each model. In the case of the orthographic model, 70% of the errors involve *simplex*

nouns – non-derived forms such as *axe*$_{M}$ 'axis' or *gala*$_{M}$ 'gala' – a majority of which are masculine nouns incorrectly classified as feminine. A likely explanation is that final <e> or <a> typically associated with feminine nouns, biases the model toward predicting the feminine gender when other informative bigram or trigram cues are lacking. It is worth noting that some of these nouns are borrowings (e.g., *drive*$_{M}$, *pulque*$_{M}$), which may further complicate the model's predictions due to their atypical morphology.

44 When examining word length, we find that shorter nouns are equally misclassified by gender (e.g., among 3-letter words). Importantly, feminine and masculine nouns are of comparable mean length in both the training and test sets (≈8 letters, see Table 1), indicating that any asymmetry cannot be attributed to overall length differences. For longer words, however, misclassification shows a clear bias: masculine nouns are increasingly wrongly predicted as feminine. This tendency becomes evident from 4-letter words onwards (58% of the 4-letter words are masculine nouns wrongly predicted as feminine ones) and remains pronounced for longer words, with rates above 60% for most word lengths and peaking at over 70% among nouns with eight or more letters. This pattern suggests that in shorter words, where less context is available, predictions are more balanced, but as word length increases, additional internal cues increasingly mislead the model's decision toward predicting feminine gender for masculine nouns. The high error rate for masculine noun endings further supports the hypothesis that the model uses suffix-based heuristics even when they conflict with the correct gender, a pattern exacerbated by the lack of additional internal cues or derivational morphemes.

45 As in previous observations, the orthographic model tends to misclassify nouns with endings that are typically associated with the opposite gender. This includes suffixes like *-ule* (e.g., *globule*$_{M}$) and *-aine* (e.g., *capitaine*$_{M}$). A similar pattern is found for nouns ending in -(s)son (e.g., *ourson*$_{M}$ 'bear cub', *nourrisson*$_{M}$ 'infant'), where masculine forms are often misclassified as feminine. By contrast, nouns ending in -çon (e.g., *contrefaçon*$_{F}$ 'counterfeit', *malfaçon*$_{F}$ 'defect') tend to be misclassified as masculine.

46 We now turn to the errors unique to the forward phonemic model. A clear disparity emerges between the number of errors unique to this model compared to the orthographic one, suggesting that the phonemic transcription of French nouns is less straightforward for gender assignment than their orthographic form. Consistently with previous observations, *simplex* nouns account for the majority of these errors (57%), while converted nouns account for 22%, and suffixed ones for 12%. Interestingly, many *simplex* nouns are borrowings, whose atypical phoneme sequences may lead the model into error.

47 In short, similar patterns as before support the idea that words whose endings go against the majority gender associated with that phonological form are more likely to be misclassified. Typically, the phonemic ending /ʁ/, which is most often associated with masculine nouns (e.g., /daʁ/ *dard*$_{M}$ 'sting'), may have driven 69% of feminine nouns ending in /ʁ/ to be incorrectly predicted as masculine, as in /kuʁ/ *cour*$_{F}$ 'court (law)', /paʁ/ *part*$_{F}$ 'share', and /fwaʁ/ *foire*$_{F}$ 'fair'. Another misleading cue is the final phoneme /p/, where 93% of feminine nouns ending in /p/ are misclassified as masculine (e.g., /ajp/ *hype*$_{F}$ 'hype'[6], /ɡɛp/ *guêpe*$_{F}$ 'wasp', /top/ *taupe*$_{F}$ 'mole'). Similarly, errors with endings like /ɛl/ (e.g., /ʁityɛl/ *rituel*$_{M}$ 'ritual', /plyʁjɛl/ *pluriel*$_{M}$ 'plural', /omosɛksyɛl/ *homosexuel*$_{M}$ 'homosexual') can be explained by the prevalence of feminine-ending nouns with this pattern (e.g., /pubɛl/ *poubelle*$_{F}$ 'trash can', /etɛ̃sɛl/ *étincelle*$_{F}$ 'spark', /madmwazɛl/ *mademoiselle*$_{F}$ 'young lady').

48 The notable difference in errors between the two models can be traced back to the representativity of word endings. A striking example involves the endings <ette> versus /ɛt/ and <ée> versus /e/. Nouns ending in the orthographic sequence <ette> are strongly biased toward the feminine gender (94% in Flexique). In contrast, the imbalance is less pronounced for nouns ending in the corresponding phonological sequence (84%), due to masculine nouns such as *poète*$_{M}$ 'poet', *prophète*$_{M}$ 'prophet', *squelette*$_{M}$ 'skeleton', *internet*$_{M}$, and others. The contrast is even more pronounced for nouns ending in <ée>: while 93% of words with this orthographic ending are feminine, only 50.2% of nouns with the equivalent phonemic ending /e/ are feminine, largely because of masculine nouns ending in *-ier* (e.g., *pompier*$_{M}$, *poirier*$_{M}$). As a result, the model sometimes predicts

masculine gender for words like /nwazɛt/ *noisette*$_F$ 'hazelnut' or /ane/ *année*$_F$ 'year'. Such observation also explains why ~63% of the phonemic models' errors are feminine nouns misclassified as masculine ones (see Figure 4). Other final phonemes, such as /s/, /l/, and /t/, do not show a clear gender association and result in rather random misclassifications. Representative examples are given in Table 5.

49 Regarding word length, the phonemic model disproportionately misclassifies feminine nouns as masculine for shorter words (between ~60% and ~70% for words of up to 5 phones). This tendency reverses for longer nouns: from six phones onward, errors skew toward misclassifying masculine nouns as feminine (56% at six phones, rising to 100% for 11- and 12-phone words), a similar behavior as the orthographic model.

Final phoneme	**Feminine nouns (misclassified as masculine)**	**Masculine nouns (misclassified as feminine)**
/s/	/ʃos/ *chausse*$_F$ 'shoe' /pas/ *passe*$_F$ 'pass' /glis/ *glisse*$_F$ 'slide'	/vis/ *vis*$_M$ 'screw' /sas/ *sas*$_M$ 'airlock' /stʁas/ *strass*$_M$ 'rhinestone'
/l/	/kul/ *coule*$_F$ 'flow' /pɛl/ *pelle*$_F$ 'shovel' /mal/ *malle*$_F$ 'trunk'	/val/ *val*$_M$ 'valley' /ʁol/ *rôle*$_M$ 'role' /fil/ *fil*$_M$ 'thread'
/t/	/lat/ *latte*$_F$ 'slat' /flɔt/ *flotte*$_F$ 'fleet' /taʁt/ *tarte*$_F$ 'pie'	/lyt/ *luth*$_M$ 'lute' /ʃit/ *chite*$_M$ 'Chitauri' (loanword) /byt/ *but*$_M$ 'goal'

Table 5. Examples of ambiguous final phonemes and model errors.

5. Assessing the time course of prediction

5.1. Motivation

50 Up to now we have examined the holistic performance of the models: for various families of words of interest, we have compared gender assignment predicted by a model with either the truth or the predictions of another model. In doing so, we have reduced the complex behavior of a model when treating an input string to a binary decision as to whether the gender is most likely to be masculine or feminine

after having examined the whole string. This ignores a lot of information on *how* the model arrives at that conclusion, that is of interest to us. By examining the time course of model predictions as the model progresses through the string, we may be able to reach conclusions not only as to whether the model is reaching the same final prediction for two words, but whether it reaches that conclusion in the same way. As a concrete example, consider the plot in Figure 5. We are examining the time course of predictions of the orthographic backward model, for a pair of words that end up being assigned feminine gender. The two curves start at the same point: without having seen any evidence, the model just assumes that the word is a little bit more likely to be masculine, reflecting a slight masculine bias in the training data. The two curves then diverge initially: this is certainly due to the fact that most words ending in <e> are feminine, while most words ending in <n> are masculine. The curve for *passerelle*$_{\text{F}}$ then quickly climbs up to a very high probability of feminine; for *saturation*$_{\text{F}}$, this only happens after the full trigram <ion> has been seen.

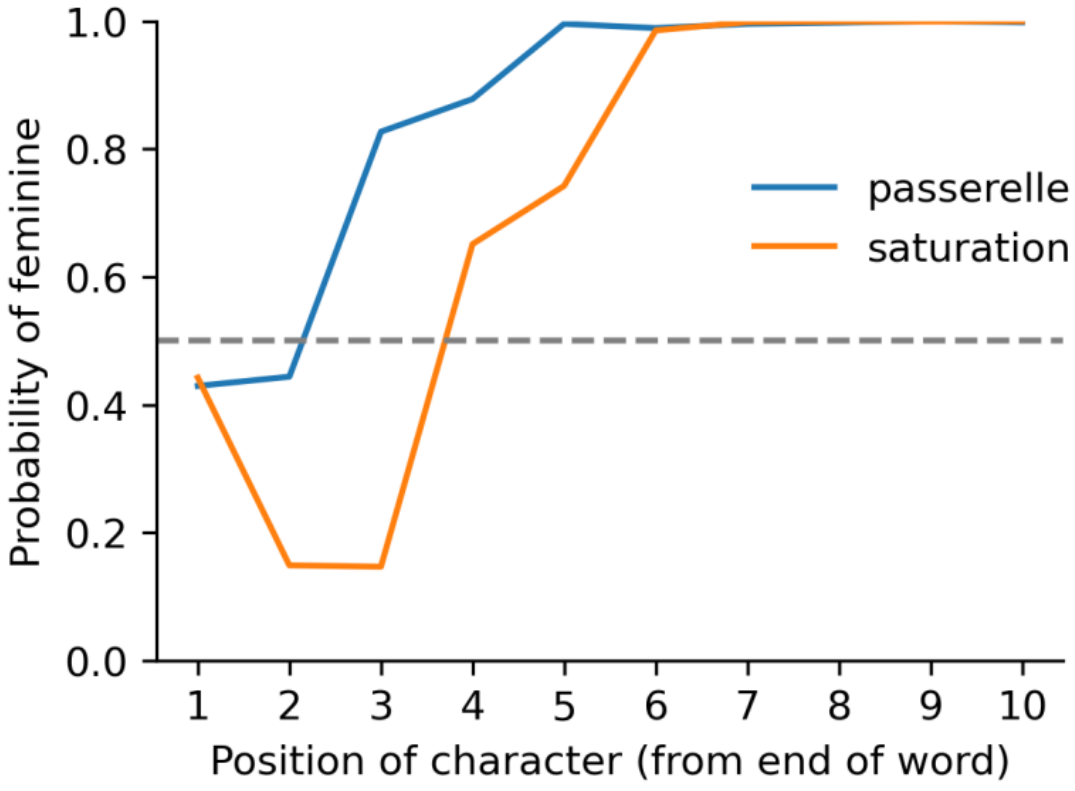


Figure 5. The time course of prediction of the backwards orthographic model for the two words *passerelle*$_{F}$ and *saturation*$_{F}$. The dashed gray horizontal line indicates the decision

boundary as 50% probability of feminine.

51 Now consider the situation in Figure 6: although they start and end at the same point as before, the two curves are much more similar to one another. This is evidence that the model paid most attention to the final substring <ion> that is common to the two words: once that substring has been examined, whatever remains in the string, the model is on its course to a strong prediction of feminine.

52 This contrasts with the situation shown in Figure 7. Here we are looking at two words that end in the gender-assigning suffixes *-ion* and *-on*, which happen to share the same final bigram <on>. Predictions are superposed initially but then diverge radically after the third character. This is clear evidence that the model has appropriately learned that, for words ending in <on>, the antepenultimate is crucial to gender prediction.

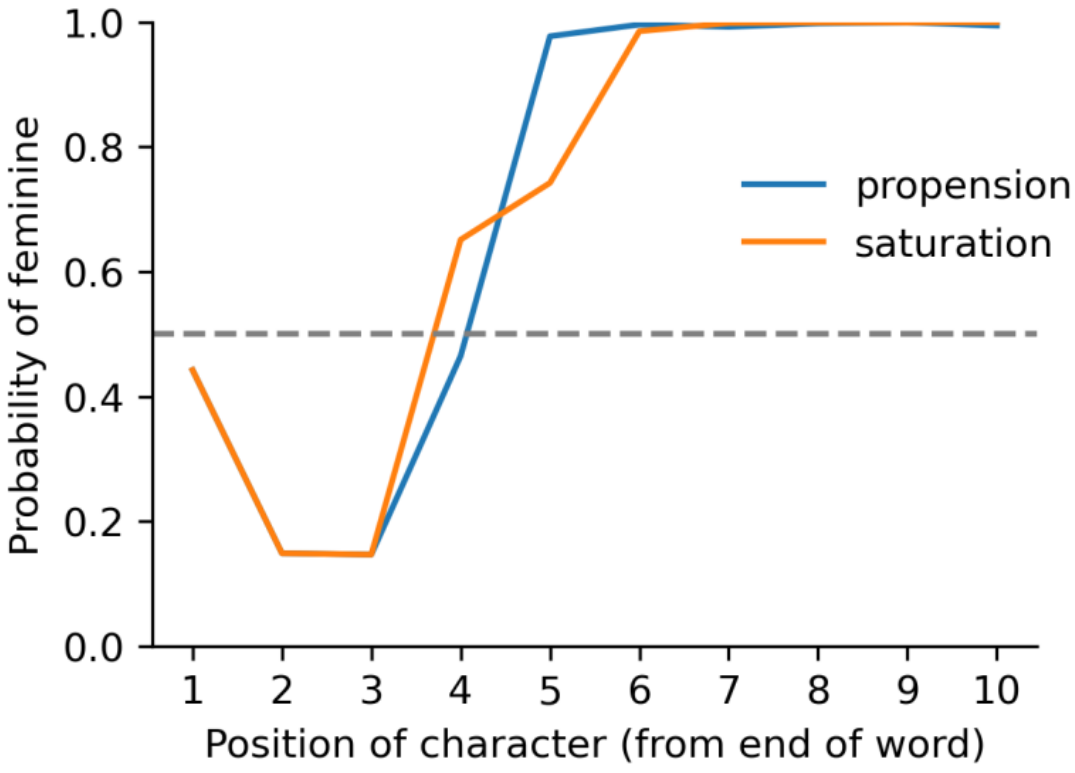


Figure 6. The time course of prediction of the backwards orthographic model for the two words *propension*$_F$ and *saturation*$_F$.

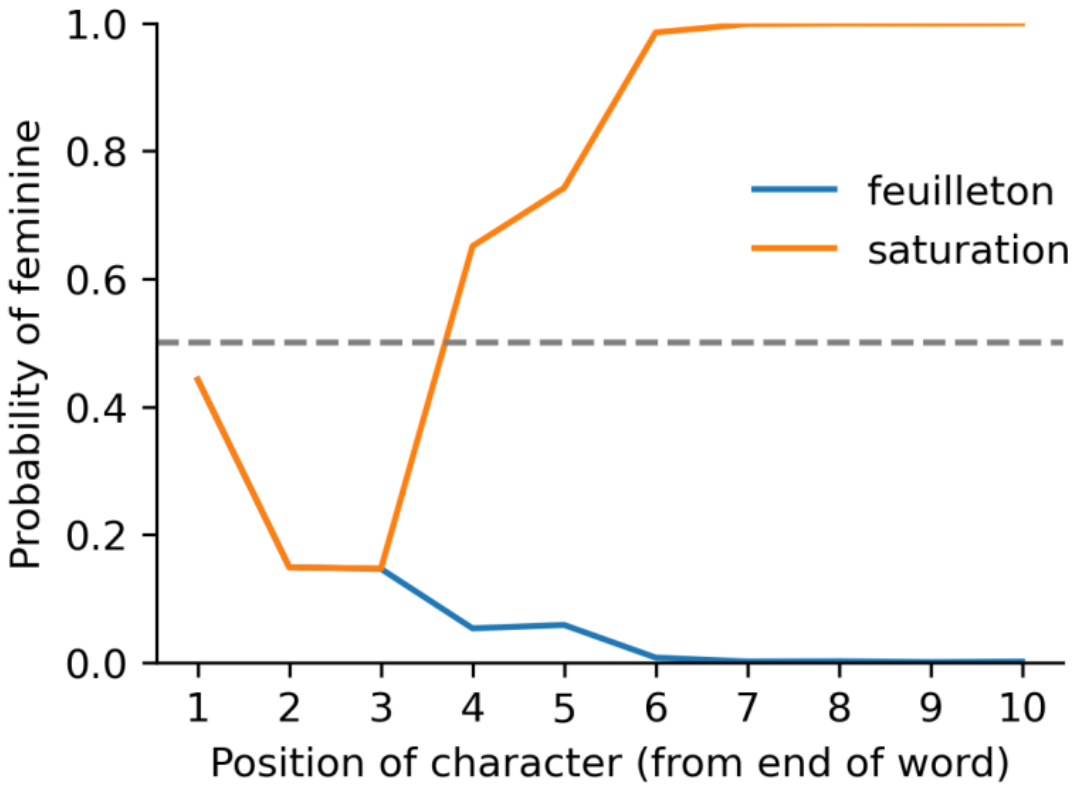


Figure 7. The time course of prediction of the backwards orthographic model for the two words *feuilleton*$_M$ and *saturation*$_F$.

53 As a final observation, let us examine how the time course of prediction with the forward model compares to that of the backward model. Figure 8 looks at the same two words as Figure 6 for the forward model.

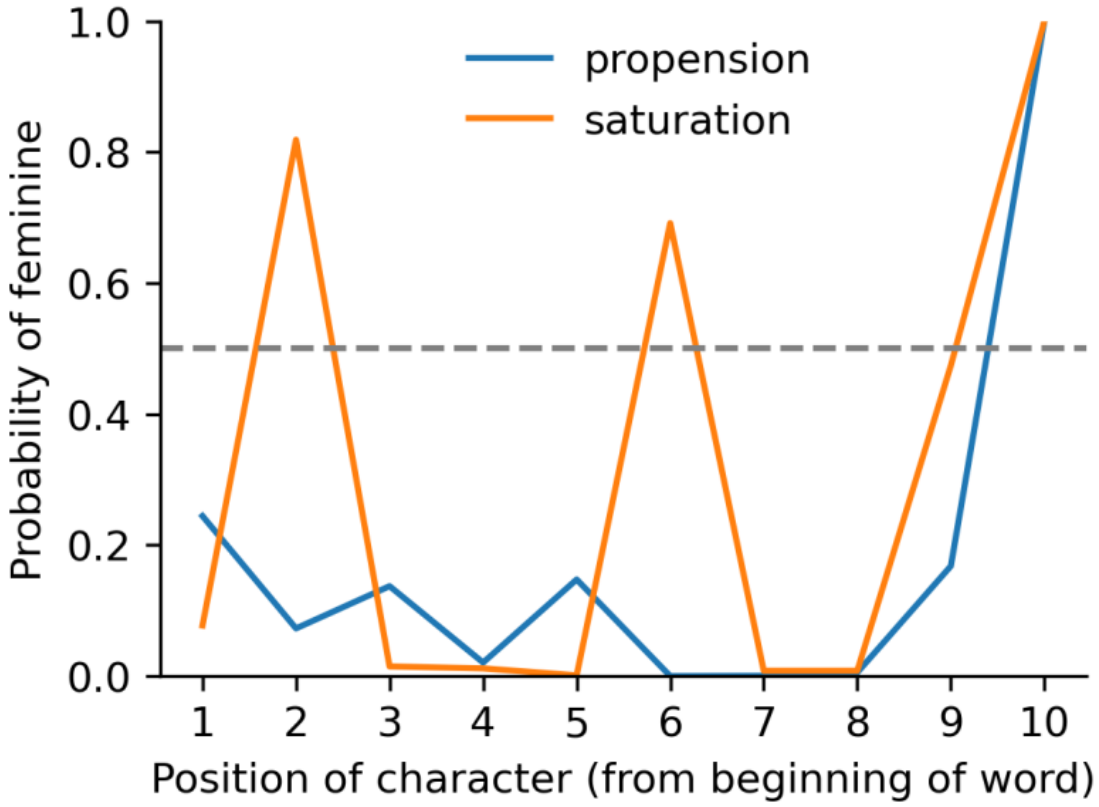


Figure 8. The time course of prediction of the forward orthographic model for the two words *propension*$_F$ and *saturation*$_F$.

54 Although the forward model also reaches a very confident prediction once it has seen the whole word, the time course of prediction is very different. This is reflective of the fact that only the end of words is predictive. Because of this, and because the model can't anticipate when the end of the word will be reached, it has learned to always assume it is at the end of the word (notice for instance how the prediction of feminine is boosted every time the letter <a>, a typically feminine ending, is seen). This pays off in the end, but it entails that strings which are similarly predictive of gender end up having very different time courses of prediction.

55 These circumstantial observations lead us to two hypotheses that would explain the contrast between Figures 5 and 6:

- Hypothesis 1: the more reliably a final ngram is associated with one gender, the more words ending in that ngram will be processed similarly by the backward model.
-

> Hypothesis 2: suffixed words are overrepresented among words that end in an ngram reliably associated with one gender.

We address these two hypotheses in turn.

5.2. The model learns highly predictive endings

56 We first need to operationalize what it means for a final ngram to be "reliably associated with one gender". Obviously there are two dimensions to this. On the one hand, the proportion of words ending in that ngram that are feminine must be either high or low; it should be far from 50%. On the other hand, there should be lots of words that end in that ngram, so that the model has an opportunity to learn.

57 The first dimension we operationalize as the Shannon entropy of the empirical distribution of genders in the test data. Entropy measures the amount of uncertainty associated with a random variable. In the special case where there are only two possible outcomes (as here, where we have two genders, it varies between 0 if one of the outcomes never occurs (in other words, we get only masculines or only feminines), and 1, if they are exactly as likely. Note that, importantly for our purposes, entropy is symmetric: we are measuring how far we are from a balanced distribution, not towards which of the two genders the distribution is leaning.

58 We operationalize the second dimension as the logarithm of the number of nouns in the test set that ends in the ngram of interest. Note that we are dealing with (type) frequency data, and applying a logarithmic transformation is a standard way of addressing the skewness of frequency (most things are very rare, but some are very frequent).

59 We also need to operationalize what it means for our gender prediction model to "process a set of words similarly". For any word w of length n in a set of words of interest W, we treat the sequence of probabilities of being feminine assigned by the model as a vector of length n. We first compute the average vector, the *centroid* of W, and then the average Euclidean distance of individual vectors to the centroid. This will be our measure of similarity of prediction, which we will call the *dispersion* of the predictions: if the model makes

exactly identical predictions for all words in W, their vectors will all coincide with the centroid, and the dispersion will be zero. The maximal possible value for dispersion is the square root of word length; in practice we did not observe average distances to the centroid higher than 1. [7]

60 With this in hand, our goal is now to measure whether dispersion is influenced by entropy and log frequency. To that extent, we fit a series of Tweedie Generalized Linear Models, henceforth Tweedie GLMs (see Dunn and Smyth 2018 for an introduction). Tweedie GLMs are a large family of GLMs that encompasses more well known types, including linear, Poisson, and Gamma regression. Tweedie GLMs with an index parameter ξ strictly between 1 and 2 are appropriate for modelling situations where the dependent variable is nonnegative but can be zero, as is the case here with dispersion. GLMs were fit using the Python stastmodels package (Seabold & Perktold, 2010).

61 All the GLMs we fit had the same structure: the dependent variable was dispersion, and the independent variables were (centered and standardized) entropy, log frequency, and their interaction. We fit separate GLMs for unigrams, bigrams, trigrams and quadrigrams, as these correspond to the common lengths of affixes. For each unigram size, we used log-likelihood tests to determine the optimal value of $1<\xi<2$, and we report on the optimal GLM.

62 Table 6 reports the coefficients of the optimal GLM at each n-gram size. [8] All four lead to the same conclusions, which is illustrated by the plot of predictions in the case of bigrams in Figure 9. At average log frequency, entropy has a positive effect on dispersion: the more balance between masculine and feminine, the more dispersion there is. This effect however is modulated by the log type frequency of the ngram: the effect of entropy on dispersion is dampened at lower frequency values, but boosted at higher frequency values.

ngram size	entropy	frequency	interaction
1	0.3502	0.2978	0.5888
2	0.3573	0.0392	0.3064
3	0.3706	0.0909	0.1636
4	0.1989	0.0761	0.0720

Table 6. coefficients of GLMs predicting dispersion from entropy, log frequency and their interaction, for different ngram sizes.

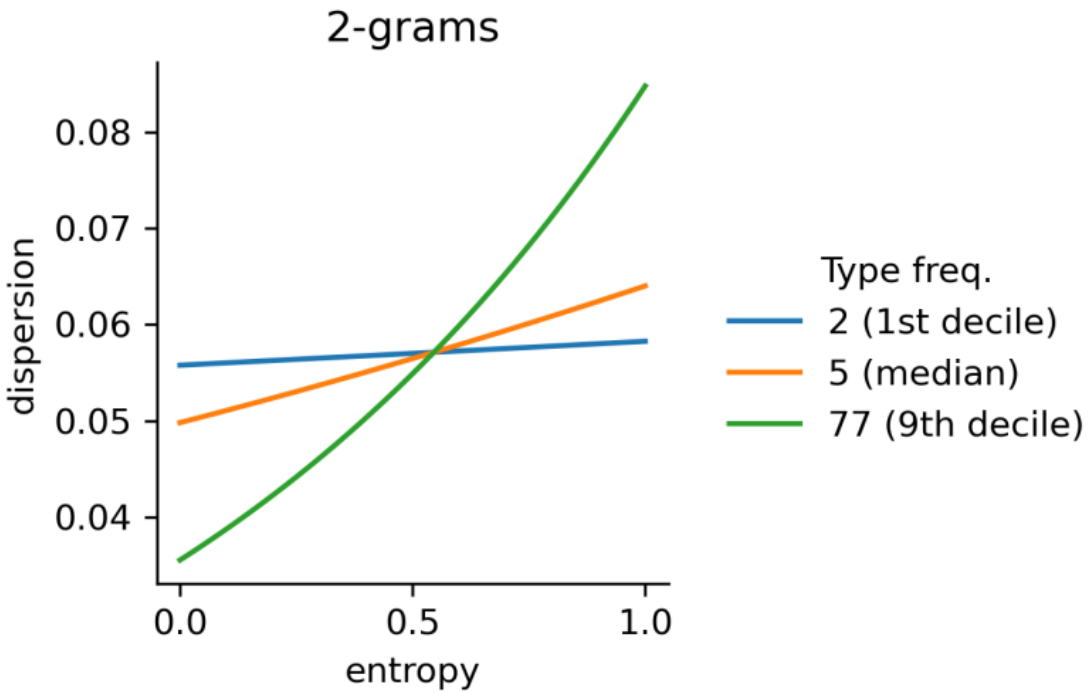
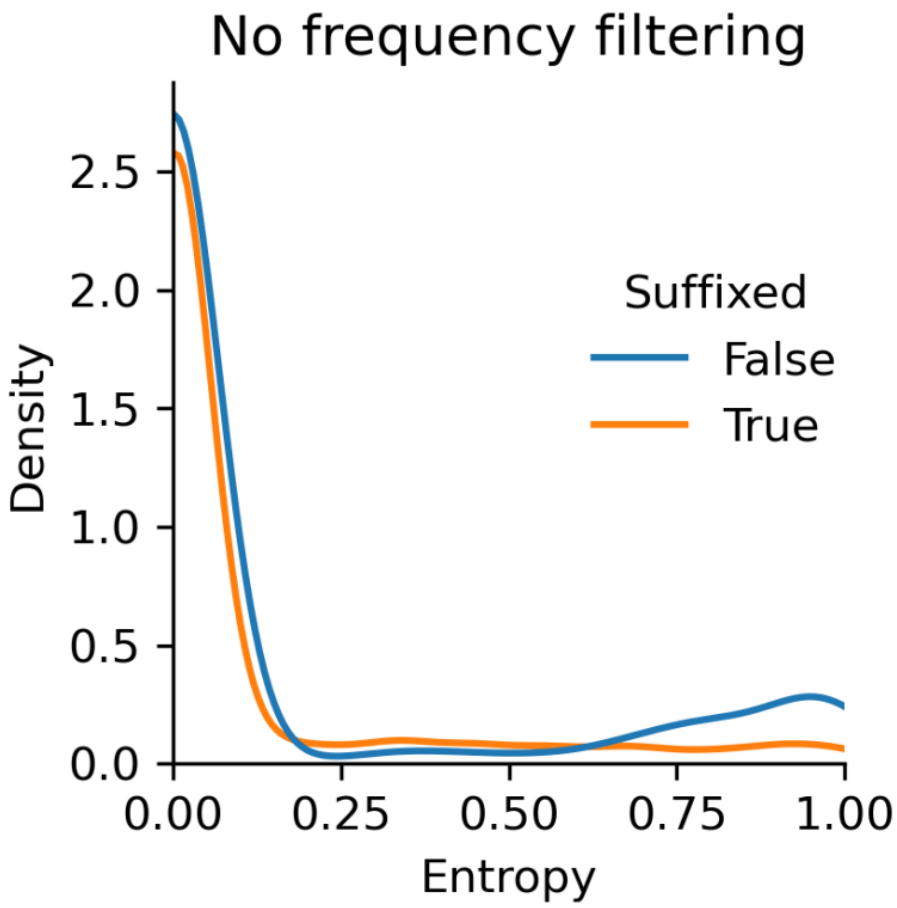


Figure 9. Predictions of the bigram Tweedie GLM. The plot displays the predicted effect of entropy on dispersion at type frequency values corresponding to the first, fifth and ninth decile in the distribution of bigram frequencies, as a way of visualizing the interaction between entropy and frequency.

5.3. Highly predictive endings tend to be suffixes

63 The results above provide clear evidence for Hypothesis 1: the way the backward LSTM orthographic model processes words that end with the same ngram becomes more similar as these ngrams are more frequent and more strongly associated with one of the two genders in the lexicon. This is a crucial piece of the puzzle, showing that it is possible to learn from form-based information alone and that final sequences of variable lengths at the end of words are strong predictors of gender. What is still missing is the link with morphology. Hypothesis 2 holds that the suffixed words are overrepresented among words ending in highly gender-predictive ngrams, that is, those ngrams Section 5.2 showed to lead to consistent behavior of the LSTM.

64 Figure 10 strongly suggests that Hypothesis 2 is borne out. Here we are looking at the distribution of gender entropy values of the final 4-grams of words in the Échantinom dataset.[9] The left panel, showing the overall distribution, suggests that there is overall a small imbalance: suffixed nouns are underrepresented among very high entropy values. This picture is somewhat misleading however, because of the influence of words ending in a rare 4-gram. Of the 4531 nouns in the dataset, 1347 end in a unique 4-gram. By definition the 4-gram is then seen only with one gender in the sample, leading to a very poor estimation of gender entropy, which has to be 0 on the basis of that trivial sample. More generally, the estimation of gender entropy is poor for those ngrams that are rare in the dataset. The right panel of Figure 10 shows the distribution of entropy values for those 1979 nouns ending in an ngram with a frequency of 10 or more in the dataset. Here we see two important effects emerging. First, suffixed nouns are overrepresented in general, making up 80% of the data; in other words, a noun ending in a frequent 4-gram is 4 times more likely to be suffixed than not. Second, this average behavior is driven by what happens for nouns with low gender entropy: high entropy suffixed and unsuffixed nouns are about equally rare, and the proportion of suffixed nouns steadily increases as the entropy decreases.


No frequency filtering
Density
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
Entropy
Suffixed
False
True

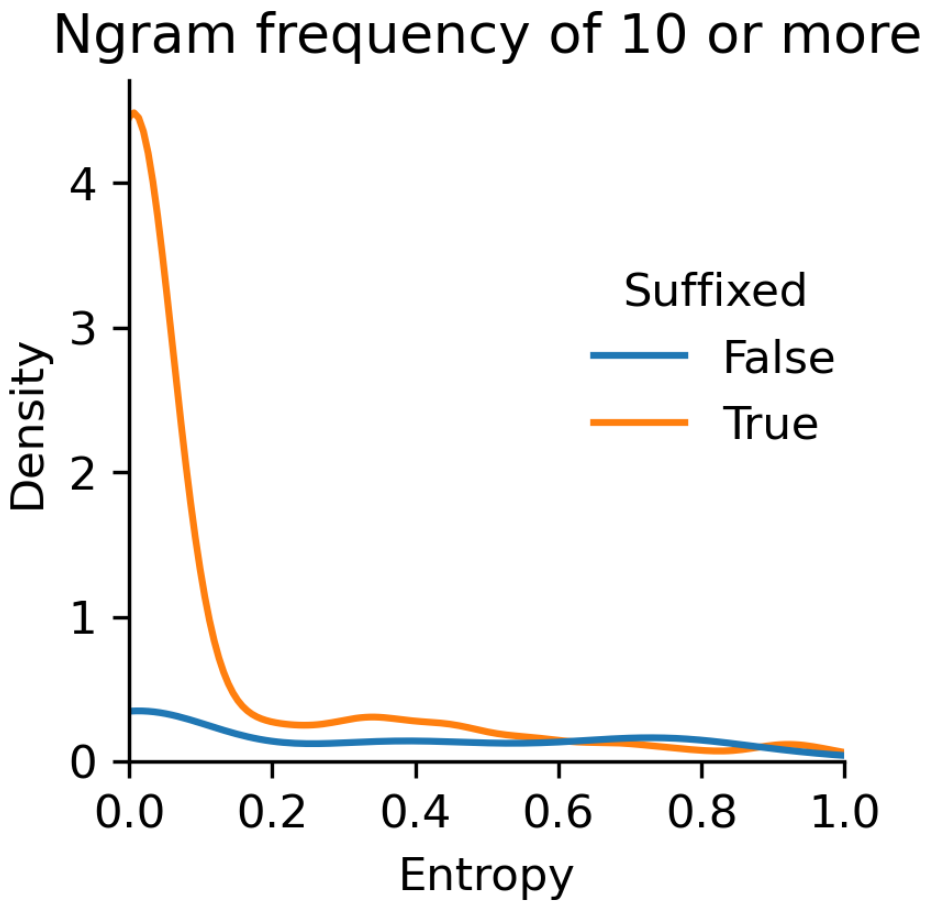


Figure 10. Density plots of final 4-gram gender entropy values for nouns in the Échantinom dataset. Left panel: all nouns. Right panel: nouns ending in a 4-gram with a type frequency of 10 or more in Échantinom.

65 While Figure 10 is highly suggestive, it does not account directly for the interaction between final ngram frequency and the distribution of entropy values, nor does it inform us on the statistical robustness of the effect. We thus fit a logistic regression to the data, predicting whether a noun is suffixed based on the gender entropy of its final 4-gram, the log frequency of that 4-gram, and the interaction between the two variables (coefficients provided in Table 7). A robust (p<0.001) effect of all 3 variables was confirmed: a higher final 4-gram log frequency leads to a higher probability of being suffixed, as does a lower entropy; and the effect of entropy is enhanced at higher frequencies.

	Coefficient	**Std. error**	**z value**	**P>|z|**	**95% C.I.**	
Intercept	-0.2190	0.040	-5.436	0.000	[-0.298,	-0.140]

Entropy	-0.3863	0.038	-10.082	0.000	[-0.461,	-0.311]
Logsize	1.5730	0.052	30.380	0.000	[1.472,	1.675]
Entropy:Logsize	-0.2644	0.065	-4.040	0.000	[-0.393,	-0.136]

Table 7. Coefficients of a logistic regression predicting suffixed status from properties of the final 4-gram of a word. Entropy and logsize were centered and standardized.

5.4. Interim conclusion

66 To sum up, we have found evidence that the backward orthographic LSTM model is more consistent in its predictive behavior when the final ngram is a more robust predictor of gender, as measured by the fact that the distribution of genders among words ending in that ngram is more imbalanced (i.e., the gender entropy is lower), and the number of nouns ending in that ngram is higher. This strongly suggests that Hypothesis 1 is borne out: the model pays little attention to the rest of the word when it has found a highly predictive final ngram. Next, we have found evidence that the very properties that lead to consistent behavior for the backward model (final ngram with low gender entropy and high type frequency) are much more represented among suffixed nouns than among nouns that do not end in a suffix, leading to a confirmation of Hypothesis 2: where the backward model makes consistent predictions, these are driven by suffixed nouns. Overall then, we may conclude that, although the generalizations the model exhibits having learned are purely based on orthography, learning is driven by the statistical prevalence of suffixed nouns and the imbalance in the shape-gender associations they lead to.

67 These results are broadly confirmed when applying exactly the same methods to the backward phonological model.[10] This indicates that, while prediction from phonology is less accurate than prediction from orthography, the form of suffixes and the statistical distribution of suffixed words play a similar role in shaping the predictability of gender from shape when considering orthography or phonology.

6. Conclusions

68 Since Corbett's (1991) seminal work, it is generally recognized that the grammatical gender of a noun is partially predictable by various

linguistic properties of that noun. Although Corbett's claims and analyses were not overtly quantitative, various modelling studies in the ensuing three decades, focusing on a variety of languages using a variety of methods, have confirmed that basic picture (Sokolik & Smith, 1992; Eddington, 2002; Matthews, 2005; Plaster et al., 2013; Guzmán Naranjo, 2019). Recent work (see in particular Fedden et al. 2025 on German) relying on large, deeply annotated datasets has slightly changed the picture, by showing not only that gender is highly predictable, but that attending solely to the form of a word (ignoring morphological and semantic properties) already achieves very high accuracy. Interestingly, additional predictors contribute at most minor improvements. This led Bonami et al. (2019) to speculate that, while the form of a word is highly predictive, that predictability is due to imbalances in the statistical distribution of phonological shapes due to morphology.

69 The goal of the present paper was to provide empirical evidence for that conjecture. To do so, we trained LSTMs, a type of recurrent neural network, on a large set of French nouns, and tested it on a smaller but morphologically annotated sample. The purpose of using a recurrent network was to be able to examine the model's gender prediction character by character as it progresses through a word, as a way of probing what type of information the model is relying on. We specifically trained four models, predicting gender from either orthographic or phonemic form, and progressing in the word from the beginning or the end.

70 The models reach comparable accuracy in the forward and in the backward direction, but higher accuracy when trained on orthographic than on phonological data. This is not a surprising result, since French orthography often makes distinctions with no phonological analogue, that may be informative of gender: see e.g. the contrast between orthographic endings <Vl>, where V is a vowel, typically masculine, and <Vlle>, typically feminine, but both usually corresponding to phonemic /Vl/. The accuracy of the phonemic models is very close to that reported by Bonami et al. (2019) using a simpler model architecture and a dataset an order of magnitude smaller. This result is hard to interpret in the absence of a human baseline, as humans are certainly not able to guess the gender of an unknown noun from its shape alone with perfect accuracy. Be that as

it may, optimizing accuracy was not a goal of the present paper. Error analysis reveals many interesting patterns, and generally suggests that the gender of simplex nouns is hardest to predict, in particular for those whose ending coincides with a suffix but which are of the other gender. This is broadly compatible with the idea that the morphological makeup of the lexicon has an influence on form-based gender prediction.

71 In the last section, we addressed our research question head on, by examining the time course of prediction of the backward models, i.e., those progressing in the word from the end. The intuition was that this would allow us to witness whether the model treated in the same way words that end in a sequence that happens to coincide with the shape of a suffix. We established that the backward models exhibited consistent behavior for collections of words ending with the same ngram, conditional on that ngram being both frequent in the lexicon and highly predictive of gender. We then provided evidence that words ending in these highly predictive ngrams were overwhelmingly suffixed words. Together, these two results lead to the conclusion that the models have learned to pay attention to final strings corresponding to gender-predictive suffixes, and to disregard the rest of the word's shape.

72 Extrapolating from these results to conclusions on gender assignment by humans is tricky. What we have shown is that the French nominal lexicon is structured in such a way that morphology, on the one hand, is predictive of gender, and, on the other hand, creates an imbalance in the distribution of word shapes leading to gender being largely learnable from those shapes. As a consequence, one can't assume that humans rely on morphological knowledge for gender prediction: it is entirely possible that, just like our models, they rely purely on form, and that their input data is structured in just the right way for that to be enough. This result is congruent with literature questioning the importance of morpheme boundaries for human learning and processing (see Baayen, Chuang, Shafaei-Bajestan & Blevins, 2019 and many others). Beyond that however, human experimentation would be needed to determine whether this is what humans do. We hope future research will address that important issue.

BIBLIOGRAPHIE

References

Baayen, R. H., Chuang, Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, 2019. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX*, second edition. Linguistic Data Consortium.

Bonami, O., & Boyé, G. (2019). Paradigm uniformity and the French gender system. In M. Baerman, O. Bond & A. Hippisley (Eds.), *Perspectives on morphology: Papers in honour of Greville G. Corbett* (pp. 171-192). Edinburgh niversity Press.

Bonami, O., Boyé, G., & Kerleroux, F. (2009). L'allomorphie radicale et la relation flexion-construction. In Fradin, B., Kerleroux, F., & Plénat, M. (Eds.), *Aperçus de morphologie du français* (pp. 103-125). Presses de l'Université de Vincennes.

Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *Actes du quatrième Congrès Mondial de Linguistique Française* (pp. 2583-2596).

Bonami, O., Guzmán Naranjo, M., & Tribout, D. (2019). *The role of morphology in gender assignment in French* [Conference talk]. *Second International Symposium on Morphology*, Université Paris Diderot.

Bonami, O., & Tribout, D. (2021). Échantinom: a hand-annotated morphological lexicon of French nouns. In *Proceedings of the International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)* (pp. 42–51).

Burnett, H., & Bonami, O. (2019). Linguistic prescription, ideological structure and the actuation of linguistic changes: Grammatical gender in French parliamentary debates. *Language in Society*, 49(1), 65-93. https://doi.org/10.1017/S0047404518001161

Burnett, H., & Pozniak, C. (2021). Political dimensions of gender inclusive writing in Parisian universities. *Journal of Sociolinguistics*, 25, 808-831. https://doi.org/10.1111/josl.12489

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Dunn, P. K., & Smyth, G. K. (2018). *Generalized linear models with examples in R*. Springer.

Eddington, D. (2002). Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics*, 9, 49-75.

Fedden, S., Guzmán Naranjo, M., & Corbett, G. G. (2025). Typology meets statistical modeling: The German

gender system. *Language*, 101(2), 251-290.

Guzmán Naranjo, M. (2019). *Analogical classification in formal grammar*. Language Science Press.

Gygax, P., Sato, S., Oetl, A., & Gabriel, U. (2021). The masculine form and its multiple interpretations: a challenge for our cognitive system. *Language Sciences*, 83.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9.

Hockett, C. F. (1958). A *course in modern linguistics*. MacMillan.

Matthews, C. A. (2005). French Gender Attribution on the Basis of Similarity: A Comparison Between AM and Connectionist Models. *Journal of Quantitative Linguistics*, 12, 262-296.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet : LEXIQUE. *L'Année Psychologique*, 101, 447-462.

Plaster, K., Polinsky, M., & Harizanov, B. (2013). Noun classes grow on trees: Noun classification in the North-East Caucasus." In B. Bickel, L. A. Grenoble, D. A. Peterson, & A. Timberlake (Eds.). *Language Typology and Historical Contingency. In honor of Johanna Nichols* (pp. 153-169). John Benjamins.

Pozniak, C., Corbeau, E., & Burnett, H. (2024). Contextual dilution in French gender inclusive writing: An experimental investigation. *Journal of French Language Studies*, 34, 273-292.

Richy, C. & Burnett, H. (2021). Démêler les effets des stéréotypes et le genre grammatical dans le biais masculin : une approche expérimentale. *GLAD!*, 10.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* (pp. 92-96).

Skousen, R. (1989). *Analogical modeling of natural language*. Kluwer.

Sokolik, M. E., & Smith, M. E. (1992). Assignment of gender to French nouns in primary and secondary language: a connectionist model. *Second Language Research* 8(1), 39-58.

Tucker, G. R., Lambert, W. E., & Rigault, A. A. (1977). *The French speaker's skill with grammatical gender: An example of rule-governed behavior*. Mouton De Gruyter.

NOTES

1 In this paper we are entirely focused on *gender assignment*, i.e., the factors that condition the grammatical gender that a noun is conventionally associated with at the lexical level. This question is orthogonal to other important issues about grammatical gender, including the interpretive effects of the decision to designate a referent using a noun of one or another gender (Gygax, Sato, Oetl & Gabriel., 2021; Richy & Burnett, 2021),

the evolution of conventions in the relationship between social and grammatical gender (Bonami & Boyé, 2019; Burnett & Bonami, 2019), or the existence and properties of strategies to avoid the association of a specific grammatical gender with a referent (Burnett & Pozniak, 2021; Pozniak, Corbeau & Burnett, 2024).

2 It should be stressed that, despite their similarity, the two studies cannot be used to compare quantitatively the predictability of the German and French systems. First, Fedden et al. used gradient boosting trees, a state of the art classification method, while Bonami et al. used a simple perceptron. Second, Fedden et al.'s study relies on an order of magnitude more data than Bonami et al.'s. Third, the two studies use different operationalizations of the similarity of phonological shapes. Hence it may well be that the apparently higher predictability of German gender is an artefact due to methodological differences.

3 The models were trained using the Adam optimizer, configured with the following hyperparameters: an embedding dimension of 128, an LSTM hidden size of 256, a batch size of 64, 20 training epochs, and a learning rate of 0.001. Early stopping was employed with a patience of 3 epochs to prevent overfitting.

4 We are indebted to a reviewer for raising this interesting issue.

5 French simplex nouns can include borrowings that may arise from morphological processes in other languages.

6 Note, however, that *hype* is an English borrowing, which means that other elements may interfere with prediction.

7 These definitions are approximative in that they do not account for the fact that words vary in length. For purposes of computing centroids, we aligned words from the end, and computed a vector with as many dimensions as there are positions in the longest word in the set; for each position we averaged over values for those words for which the position is instantiated. For purposes of Euclidean distance calculations, given two vectors v of length m and w of length $m>n$, we trimmed w to length n and computed distance between v and that trimmed vector.

8 We do not report on statistical significance, as this has no interpretation in the present context: our GLMs are trained on the whole sample space of possible ngrams, and hence there is no sense in asking whether observations can be assumed to hold outside of that sample space.

9 Although we might consider other ngram sizes, 4-grams are most relevant as the vast majority of suffixed words (98%) end in a suffix of 4 characters or less.

10 Exact numbers are of course different when considering phonological data, but all qualitative conclusions hold. Here are the noticeable differences. (i) When fitting a tweedie GLM predicting dispersion of prediction curves from gender entropy and log frequency of final ngrams, we found small negative coefficients for the interaction of entropy and frequency for unigrams, in contrast with the larger and positive coefficient found with orthographic ngrams (Table 6). All other coefficients were similar, at all ngram sizes. (ii) When examining the gender entropy of final ngrams of frequent nouns (see Figure 10, right panel), the discrepancy between suffixed and unsuffixed nouns was even more pronounced. (iii) Logistic regression documented a significant main effect of entropy and log frequency (p<0.001) on probability of being suffixed, but no significant interaction between the two.

RÉSUMÉS

Français

En français, le genre grammatical d'un nom est influencé par l'interaction complexe d'indices sémantiques, phonologiques et morphologiques. Les études précédentes ont montré que le genre est partiellement prédictible à partir de la forme des mots, mais elles ne tranchent pas la question de savoir si la prédictibilité est principalement de source phonologique ou morphologique. Cet article aborde la question à l'aide de réseaux de neurones récurrents entraînés sur un grand lexique de noms du français. Nous évaluons des modèles entraînés sur des représentations orthographiques ou phonémiques, et examinons la qualité des prédictions et le déroulement de la prise de décision. Nos résultats montrent que le genre peut être prédit avec fiabilité à partir de la forme seule et que l'orthographe fournit des indices plus robustes que la phonologie. L'analyse d'erreur montre que le genre des noms non-construits est plus difficile à prédire que celui des noms suffixés. Crucialement, les prédictions les plus fiables et les plus consistantes résultent de séquences finales de mots correspondant le plus souvent à des suffixes dérivationnels. Ces résultats suggèrent que la morphologie est la cause ultime de la fiabilité de l'assignation du genre, alors que la forme de surface est l'intermédiaire par lequel cette prédictibilité émerge.

English
Grammatical gender in French is shaped by a complex interplay of semantic, phonological, and morphological cues. While previous studies have shown that gender is partially predictable from word form, they leave open the question of whether predictability is primarily phonological or morphological in origin. In this paper, we address this issue using recurrent neural networks trained on a large lexicon of French nouns. We evaluate models on both orthographic and phonemic representations, tracking prediction accuracy and the time course of decision-making. Our results show that gender can be reliably inferred from form alone, with orthography offering stronger cues than phonology. Error analyses reveal that simplex nouns are the most challenging, while suffixed nouns are highly predictable. Crucially, we find that the most consistent and accurate predictions arise from frequent word-final sequences that overwhelmingly correspond to derivational suffixes. These findings support the view that morphology is the ultimate driver of gender assignment predictability, while surface form serves as a proxy through which this predictability emerges.

INDEX

Mots-clés
genre grammatical, français, morphologie computationnelle

Keywords
grammatical gender, French, computational morphology

AUTEURS

Nazanin Shafiabadi
Sorbonne Université, CNRS, ISIR, 75005, Paris, France
nazanin.shafiabadi@isir.upmc.fr

Nina Nusbaumer
Université Paris Cité, CNRS, LLF, 75013, Paris, France
nina.nusbaumer@u-paris.fr

Olivier Bonami
Université Paris Cité, CNRS, LLF, 75013, Paris, France
olivier.bonami@u-paris.fr