

Paradigms over Surface Cues: Inflection Class Drives Hindi Gender Assignment in Expert-Annotated Noun Paradigms

John Alderete¹ Yogya Agrawal¹ Tra My Nguyen¹ Puja Shah¹ Aanchan Mohan²

¹Simon Fraser University, Canada

²Northeastern University, Vancouver, Canada

alderete@sfu.ca

Abstract

Grammatical gender assignment is governed by interacting phonological, semantic, and morphological factors whose relative weight varies across languages. We present the first large-scale computational analysis of gender in Hindi using a novel dataset of over 6,000 noun paradigms annotated for morphological, phonological, semantic, and lexical strata features. We tested inflection class against competing predictors with interpretable classifiers and neural networks and found that it dominates, predicting gender at $\approx 97\%$ accuracy, while other features contribute negligibly. The dependence is asymmetric: inflection class nearly determines gender, whereas gender is only a partial predictor of inflection class. Inflection class is itself recoverable in part from phonological and etymological cues without reference to gender, which suggests a resolution to the bootstrapping problem: form-based cues scaffold inflection class, which in turn scaffolds gender. These findings establish Hindi as an extreme case of morphologically aligned gender and expose a limitation of language-agnostic approaches: methods that treat semantic and surface-form cues as broadly predictive of gender may miss Hindi’s near-deterministic dependence on inflection class. Our dataset enables probing whether multilingual models encode this dependency and whether it survives cross-lingual transfer.

1 Introduction

Grammatical gender pervades the morphology and syntax of many languages, driving morpho-syntactic agreement and shaping inflectional paradigms. Understanding gender assignment, the mapping from noun attributes to gender categories, is a core question for linguistic theory (Corbett and Fraser, 2000; Kramer, 2020). It is also a prerequisite for robust NLP: grammatical gender shapes distributional representations in language models (Stańczak et al., 2024), propagates bias in machine

translation (Stanovsky et al., 2019; Savoldi et al., 2024), and conditions morphological agreement in generation (Vanmassenhove et al., 2018).

Despite this shared interest, linguistic and computational approaches differ in the features they use to model gender assignment. Inflection class, which organizes nouns into paradigmatic classes, is central to morphological theory (Aronoff, 1993; Carstairs-McCarthy, 1994; Stump, 2001) and has been shown to be a critical factor in gender assignment in several unrelated languages (Aronoff, 1991; Kürschner and Nübling, 2011; Corbett, 1991). In Russian, for example, it predicts gender for all declinable nouns except sex-differentiable ones (Corbett, 1982). Computational models, by contrast, emphasize semantic features and observable phonological cues (Allasonnière-Tang et al., 2021; Sahai and Sharma, 2021). Although word-edge character n -grams are often used as proxies for inflectional morphology (Bojanowski et al., 2017), they capture surface regularities distinct from the abstractions encoded by inflection class. We bridge these two approaches by testing whether inflection class improves gender assignment.

Gender assignment in Hindi, a language with over 600 million speakers, has not been studied quantitatively. Hindi has five inflection classes and two genders, so predicting gender from inflection class is an empirical question, not a tautology. Prior work suggests that, as in Russian, Hindi gender aligns closely with inflection class (Singh and Sarma, 2010; Sinha, 2018). However, this claim has not been tested against competing predictors proposed for Hindi: semantic features (Williams et al., 2019) and the stem-ending phonology typical of descriptive grammars. These grammars state gender tendencies in terms of citation-form stem endings; canonically, $\bar{a}$ -final nouns are masculine and $\bar{i}$ -final feminine (Kachru, 2006; Agnihotri, 2022). Such heuristics are silent for the consonant-final nouns that form the bulk of the lexicon and admit many ex-

ceptions even where they apply (Singh and Sarma, 2010). Inflection class is a distinct, more abstract construct, defined over the full paradigm rather than the citation form, and its relationship to gender has not yet been quantified. Hindi thus provides a test case for inflection class in gender assignment.

This analysis also bears on multilingual representation learning. Cross-lingual transfer assumes that shared representations capture regularities that generalize across languages, but the relative weight of those regularities varies typologically. Hindi is an informative test case: models that transfer gender knowledge from languages where semantic or phonological cues are more influential may underrepresent the decisive role of inflection class. Our dataset makes this dependency explicit and supports future probing of multilingual representations in a morphologically rich Indo-Aryan language.

We present the first large-scale machine learning analysis of Hindi gender assignment, testing inflection class against phonological, morphological, semantic, and lexical features via ablation. Our contributions are as follows:

- **Dataset:** We release¹ the first large-scale Hindi noun dataset (over 6,000 paradigms) with rich morphological, phonological, semantic, and etymological features, usable both as a linguistic resource and as a probing benchmark for whether multilingual representations encode inflection class and grammatical gender.
- **Predictors of gender:** Inflection class dominates all other predictors of gender in Hindi, including semantic and phonological features; we also document a systematic gender bias in the lexicon.
- **Interdependence:** Neural network analysis reveals an asymmetry: feature combinations yield substantial predictive boosts for inflection class but negligible boosts for gender, consistent with inflection class being learned first and scaffolding gender.

2 Related Research

Gender assignment is a rule-based system that predicts noun gender from formal and semantic properties (Corbett, 1991, 2007; Kramer, 2020). Computational linguistics operationalizes this framework as a supervised classification task, using labeled

noun datasets to learn the mapping from noun features to gender categories. Studies have explored diverse feature types, including phonological *n*-grams and semantic embeddings in French (Sahai and Sharma, 2021), semantic and phonological categories in Mian [mpt] (Allasonnière-Tang et al., 2021) and Nakh languages [che, inh, bbl] (Schreur et al., 2022), phonological and morphological form in Tashlhiyt [shi] (Alderete et al., 2025), cross-linguistic semantic structure via canonical correlation analysis (Williams et al., 2019), and mutual information over agreement patterns (Williams et al., 2021). The consistent goal is to determine the relative importance of noun form versus meaning in gender prediction.

Rigorously testing the inflection-class hypothesis presents challenges: definitive testing requires large-scale datasets to model complex interactions and avoid spurious correlations, yet many prior studies rely on fewer than 1,000 nouns. Second, it requires a rich feature set to ensure inflection class is not simply a proxy for phonological cues. However, existing studies often use opportunistic feature sets from databases lacking annotation for abstract morphological categories. We overcome these limitations by developing a large, richly annotated dataset, enabling the first computational test of inflection class in gender assignment.

Supervised classification treats inflection class as a predictor of gender, but a distinct question concerns their interdependence. The two are tightly correlated in some languages (e.g., Russian), posing a chicken-and-egg problem for the learner (Tesar and Smolensky, 2000; Pinker, 1984): assigning gender presupposes inflection class, and vice versa, yet a child, or a model, learning Hindi starts with neither. We quantify predictability in both directions: gender from inflection class, and inflection class from gender. This analysis yields asymmetric evidence about which category may be learned first.

3 Dataset and Features

Hindi nouns inflect for number (singular/plural) and case (direct/oblique), yielding four-cell paradigms (Table 1). Grammatical gender (masculine/feminine) is not marked with a transparent set of affixes; rather it is an abstract lexical attribute that controls agreement on adjectives and finite verbs. This section describes the construction, annotation, and composition of our dataset’s paradigms.

¹<https://github.com/aldo-git-bit/hindi-gender-assignment-ML>

Table 1: Schema for Hindi inflection classes

	0 (F/M)	1 (F)	2 (M)	3 (F)	4 (M)
sg.dir	X	X-ī	X-ā	X	X
sg.obl	X	X-ī	X-e	X	X
pl.dir	X	X-iyā	X-e	X-ē	X
pl.obl	X	X-iyō	X-ō	X-ō	X-ō
sg.dir	pyār	khiṛk-ī	bhānj-ā	rāt	lekhak
sg.obl	pyār	khiṛk-ī	bhānj-e	rāt	lekhak
pl.dir	pyār	khiṛk-iyā	bhānj-e	rāt-ē	lekhak
pl.obl	pyār	khiṛk-iyō	bhānj-ō	rāt-ō	lekhak-ō
	‘love’	‘window’	‘nephew’	‘night’	‘writer’

3.1 Data Sources and Preprocessing

Our dataset of 6,040 unique Hindi noun paradigms was extracted from the Hindi Treebank (16,649 sentences, 352K tokens in the UD release; [Bhat et al. 2017](#)). We parsed the corpus’s CoNLL-U files to extract the lemma, wordform, part-of-speech, and morphological features (gender, number, case) for each wordform. After filtering for nouns, we grouped wordforms by their lemma and gender to construct paradigms. Most paradigms (93%) were partial because the treebank did not attest wordforms for all four cells, necessitating the manual annotation described below. We resolved encoding-based duplicates and adjudicated multi-gender lemmas to distinguish homophones from gendered pairs (e.g., *laṛkā* ‘boy’ vs. *laṛkī* ‘girl’).

3.2 Feature Annotation

We expanded the feature set with a combination of algorithmic processing and manual annotation by two trained Hindi native speaker analysts.

3.2.1 Morphological Features

Following [Singh and Sarma \(2010\)](#) and [Sinha \(2018\)](#), we annotate inflection class using a five-class system that unifies nominal and adjectival inflection, improving on traditional stem-ending accounts (Table 1).

Two native speakers annotated all 6,040 nouns for inflection class using Table 1 as a template, resolving partial paradigms via linguistic intuition. English loans with plural morphology (e.g., *रिंग्स* *rīṅsa* (pl) ‘rings’) formed a sixth minor class (class 5). Since compounding and derivation often influence gender assignment ([McGregor, 1995](#)), we also annotated each noun for compound status, derived status, and derivational category (90+ values).

Although inflection classes generalize over sets of affixes, our schema does not introduce circularity. An inflection class is an arbitrary label defined by a noun’s distribution across the full four-

cell paradigm (number $\times$ case), not by any single gender-correlated ending. Furthermore, classes 3 (feminine) and 4 (masculine), which together account for 72% of the data, share null singular forms and diverge only in the plural direct form (Table 1). Gender is predicted near-categorically even for these consonant-final stems, which carry no gender-diagnostic ending. The competing form-based predictors are likewise independent of the class-defining exponents: phonological n -grams are extracted from stems after inflectional endings are stripped (§3.2.3), and derivational affixes are stem-internal, distinct from the inflectional suffixes that identify class.

3.2.2 Semantic Features

We annotated binary animacy (entities capable of independent action) via HindiWordNet hypernym lookup (e.g., ANIMT, INANI; [Jain et al. 2013](#)), validated against a 362-noun gold set (99.7% accuracy); 1,986 nouns absent from HWN or with ambiguous senses were hand-annotated.

To derive a set of broad, cross-linguistically comparable semantic fields, we used the UCREL Semantic Analysis System (USAS) ([Rayson et al., 2004](#)). Hindi lemmas were first translated to English glosses via the Google Translate API ([Google, 2025](#)). We then applied the PyMUSAS tagger ([Piao et al., 2022](#)) to these glosses and extracted the top-level semantic domain from the resulting hierarchical tag (e.g., “F” for Food and Drink), yielding a 21-category coarse semantic feature. Although using English glosses introduces translation noise (see Limitations), it enables USAS categorization; the second scheme below avoids this issue altogether.

Finally, to create a classification scheme based on native Hindi lexical structure, we extracted ontological categories directly from HindiWordNet. For each lemma, we queried the HWN API ([CFILT, IIT Bombay, 2025](#)) and recorded its highest-level hypernym as a fine-grained semantic feature. For the 1,389 nouns not found in HWN (primarily loanwords), we used a two-stage fallback: common abbreviations were manually tagged, and the remainder were categorized using a large language model prompted to select from the existing set of HWN categories.

Intersecting fine-grained HWN features with animacy reveals a systematic lexical gender bias (see §6.3). Of nine “animate-centric” WordNet categories (>50% animate; $n=824$), Wilson 95% confidence intervals identified three with significant

masculine skew: Occupation (n=48, 89.6% masc.), Mythological Character (n=27, 85.2% masc.), and Person (n=673, 84.4% masc.).

3.2.3 Phonological Features

We extracted n -grams from noun stems rather than lemmas or wordforms (which contain the structure we seek to predict). As automated Hindi stemmers lacked required accuracy, we manually derived stems by stripping inflectional endings from singular-direct forms. Stems were represented as both Devanagari strings and phonetic transcriptions. For example, removing the sg.dir inflection from भान्जा *bhānjā* yields the stem भान्ज [bʱandʒ]. Phonetic transcriptions were subsequently normalized for phenomena like schwa deletion using a rule-based process.

We extracted character and phoneme n -grams (lengths 1–3) from both stem representations. This process targeted word edges, which are crucial for morphological processing and predictive of gender (Tucker et al., 1977). This process yielded 5,385 script and 4,265 phonetic n -grams, which were winnowed down via feature selection. Given that Devanagari transparently reflects sound structure, script and phonetic n -grams were highly correlated. We utilized script n -grams for all analyses as they showed slightly higher predictive power.

3.2.4 Lexical Strata

To account for the influence of language contact on gender assignment, we annotated each noun for its lexical stratum. Our primary source was McGregor’s (1993) *Oxford Hindi-English Dictionary*, from which we automatically extracted etymological tags. This process yielded four mutually exclusive classes: Sanskrit, English, Arabic-Persian, and Native. We treat “Native” as a catch-all category for non-loanwords, acknowledging that this conflates *Tadbhava* (evolved from Sanskrit) and *Deśhaj* (indigenous) strata. Tags for Arabic and Persian were merged into a single class due to their low frequency and shared historical context; post-hoc analysis confirmed comparable distributions across the two strata. Because dictionary coverage of modern loanwords is often incomplete, especially for English, two native speaker analysts systematically reviewed the corpus to identify and tag all English loans not listed in the source. Remaining untagged nouns were classified as Native.

3.3 Annotation Quality Control

To ensure high-quality manual annotations for the most complex features (inflection class and deriva-

Table 2: Dataset overview

Category	Count	% of Total	
Total nouns	6040	100.00%	
Fully declinable	5497	91.01%	
Zero/Irregular	543	8.99%	
Compounds	669	11.08%	
Derived nouns	870	14.40%	
Animate nouns	881	14.59%	
Nouns in gendered pairs	546	9.04%	

Gender	Count	% of Total	
Masculine	4010	66.39%	
Feminine	2030	33.61%	

Inflection Class	Count	% Masc.	% Total
1	716	0.42%	11.85%
2	432	99.31%	7.15%
3	1110	0.36%	18.37%
4	3239	98.52%	53.63%
Zero inflection (class 0)	483	69.57%	8.00%
Irregular (class 5)	60	78.33%	0.99%

Etymology	Count	% Masc.	% Total
Arabic-Persian	576	58.33%	9.54%
English	1512	73.81%	25.03%
Sanskrit	1308	70.18%	21.66%
Native	2644	62.03%	43.77%

tional categories), we employed a collaborative adjudication process. For each task, the two native speaker analysts first co-annotated a development set of 200 nouns. During this initial phase, all disagreements were resolved through discussion and reference to published dictionaries, and the resulting decisions were formalized into a detailed annotation rubric. This rubric was then used to guide the independent annotation of the remaining data, with all ambiguous or edge cases being reserved for joint review and a final consensus decision. As a final verification step, we performed a consistency check on the completed dataset. For example, all nouns assigned to a given inflection class were grouped and reviewed together to ensure homogeneity and correct any remaining outliers.

3.4 Dataset Overview

Table 2 details the distribution of gender, inflection class, and lexical strata. As expected, the fully declinable classes (1-4) exhibit near-categorical gender alignment, whereas indeclinable and irregular forms show greater variation; we retain these latter cases to ensure rigorous hypothesis testing. We address observed data imbalances, specifically a 2:1 masculine skew and the prevalence of Inflection Class 4, in our experimental setup.

While Table 2 shows that the fully declinable classes (1-4) are closely aligned with gender, these descriptive statistics leave several unanswered questions. First, they cannot quantify the relative contribution of inflection class against competing pre-

dictors, particularly semantic features, which prior investigation of Hindi identifies as a viable signal for gender assignment (Williams et al., 2019). Second, the 543 nouns in zero-inflection and irregular classes show substantial gender variation, and it is unclear whether this residual variance is systematically recoverable from other linguistic features. Third, frequency counts cannot reveal non-linear interactions between feature types relevant to understanding the interdependence of gender and inflection. The computational experiments in the following sections address each of these questions through systematic ablation and interaction analysis.

4 Experiments

4.1 Computational approach

We formulate the prediction of gender and inflection class as supervised classification tasks. To investigate the interdependence of these attributes, we alternate the target variable while utilizing the remaining linguistic factors as input features. Our analysis prioritizes interpretability: we first employ a suite of tree-based (“interpretable”) classifiers to quantify feature importance, followed by neural networks to probe complex, non-linear interactions between features. We probe the importance of individual features, like inflection class, as well as entire feature sets to distinguish the traditionally important linguistic factors of form and meaning.

4.2 Feature Correlation

We analyzed feature redundancy using Pearson correlation for binary indicator variables and Cramér’s V for nominal categorical features (as the latter accounts for the non-linear associations in multi-class data). We found no significant correlations among the majority of attributes. However, we observed a high association between animacy and fine-grained semantic features (HWN) (Cramér’s $V = 0.863$), necessitating the removal of one to prevent multicollinearity. We excluded animacy based on three factors: predictiveness, as WordNet categories predict binary animacy with 96.2% accuracy; category purity, where 76% of WordNet categories contain >95% exclusively animate or inanimate nouns; and information content, as the 108-value WordNet feature preserves fine-grained distinctions while implicitly encoding the binary animacy signal.

4.3 Feature Encoding and Selection

We encoded binary attributes as boolean indicators and low-cardinality categorical variables as one-hot

vectors. To address high dimensionality in phonological n -grams and morphological derivational categories, we applied a statistical feature selection strategy based on the Wilson score interval (Wilson, 1927), which provides robust confidence estimates for binomial proportions even with the small sample sizes typical of rare n -grams. For each variable value (e.g., a specific n -gram or derivational category), we computed the 95% binomial confidence interval of its associated masculine proportion. A feature was retained only if its confidence interval strictly excluded the global masculine baseline (66.4%). This ensured that only features significantly deviating from the prior class distribution were selected, removing values that provide no information beyond the baseline rate. This yielded the important feature subsets, reducing the phonological n -gram vocabulary from 5,385 to 160 and the derivational categories from 90 to 17.

4.4 Data Imbalance and Regularization

We mitigate the effects of class skew—moderate in gender (2:1) and severe in inflection class (53:1)—through stratified sampling across all train, validation, and test splits. Given these imbalances, our evaluation prioritizes Macro F1 and Area Under the Receiver Operating Characteristic curve (AUC-ROC) over accuracy to ensure adequate performance on minority classes.

To address overfitting, particularly in high-dimensional settings such as those involving phonological n -grams, we apply distinct regularization strategies tailored to each model class. For tree-based models, we impose strict topological constraints to prevent the memorization of training data. This proved critical: preliminary experiments without these constraints exhibited train-test performance gaps exceeding 22%, reduced to under 3% following regularization. For neural networks, we counteract imbalance directly within the optimization objective using inverse class frequency weighting, scaling each example’s loss contribution to penalize misclassification of minority classes, alongside standard regularization such as dropout.

4.5 Interpretable Models

To balance predictive performance with interpretability, we employed four distinct classifiers: Logistic Regression (as a linear baseline), Decision Trees, Random Forests, and Gradient Boosting. Our experimental protocol prioritized comparative linguistic analysis over absolute metric maximization. Consequently, we eschewed extensive

hyperparameter tuning (e.g., grid search) for fixed, conservative configurations across all feature sets. This ensures performance differences reflect feature quality rather than optimization or model-capacity artifacts.

To validate this fixed-parameter methodology, we performed an exhaustive grid search (540 combinations) on the Semantic/Random Forest configuration. The tuned model yielded only a +0.83% improvement (67.88% $\rightarrow$ 68.71%) over our fixed configuration, confirming that the substantial performance gaps observed between feature sets are driven by the data, not the tuning. Models were trained on a stratified 80% partition and evaluated on the held-out 20%, with the statistical significance of results validated using McNemar’s tests and bootstrap confidence intervals for AUC. Exact search parameters and configurations are provided in Appendix B for reproducibility.

Finally, to ensure the robustness of our findings despite the severe class imbalance in inflection prediction (where Class 5 constitutes only 0.99% of the corpus, $n = 60$), we repeated all experiments using 5-fold stratified cross-validation. This procedure ensured that approximately 60 total test samples for Class 5 were evaluated across folds, rather than the ≈ 12 samples available in a single split. The full cross-validation results (Appendix A) demonstrate high stability, with single-split accuracy for the primary feature sets (Morphological and All Features) within 1.3 percentage points of the cross-validated means, further attributing performance differences to feature quality rather than sampling variance.

4.6 Neural Network Models

We designed a Multilayer Perceptron (MLP) to capture non-linear interactions between linguistic factors. High-cardinality categorical features are mapped to dense embeddings, with dimensions scaled logarithmically relative to vocabulary size to balance representational capacity and compactness. To ensure computational feasibility, phonological inputs are limited to the top-20 most frequent n -grams. We explicitly model feature interdependence via an interaction module: selected feature pairs are combined via an outer product to capture co-occurrences, then compressed via linear projection to a 128-dimensional representation. This interaction signal is concatenated with the base features and processed through a pyramidal hidden structure (128 $\rightarrow$ 64 $\rightarrow$ 32) using Batch Normalization and ReLU activations.

Models were trained on a stratified 70/15/15 data split using the Adam optimizer with an initial learning rate of 10^{-3} , regulated by a ten-epoch warmup and a schedule that halves the learning rate when validation performance plateaus for five epochs. To prevent optimization from defaulting to the majority class in the highly imbalanced inflection task, we utilized validation Macro F1 as the primary metric for early stopping (patience=30), whereas validation accuracy was employed for the binary gender task. Generalization was further enforced through weight decay (10^{-4}) and progressive dropout (0.2–0.3) across the hidden layers.

5 Results

5.1 Predicting Gender

Figure 1 (Left) illustrates the primacy of inflectional morphology in predicting gender. Feature sets containing inflection class (Morphological and All Features) achieve near-perfect classification (Accuracy $\approx 96.6\%$, AUC ≈ 0.98), drastically outperforming the majority baseline (66.4%). Crucially, adding phonological, semantic, and etymological features to the inflectional set yields no statistically significant improvement (McNemar’s test, $p > 0.44$). This indicates that while other linguistic factors correlate with gender, their predictive signal is largely redundant given the inflectional paradigm.

Ablation reveals a distinct hierarchy among the remaining features. Removing inflection class precipitates a sharp drop in accuracy to $\approx 73\%$ (AUC ≈ 0.72), a significant decrease yet still well above the random baseline. Within this non-inflectional space, phonology provides the strongest independent signal (Accuracy $\approx 71\%$), whereas semantic and etymological features offer negligible gains (Accuracy $< 68\%$, AUC < 0.62). The ≈ 26 -point gap between stem phonology and inflection class ($\approx 97\%$) is telling: it shows the near-categorical alignment is a property of the abstract paradigm, not of the stem-ending cues that descriptive grammars associate with gender. Importantly, this holds even for the HWN semantic features, a gold-standard native representation derived without translation. This suggests the weak predictive power of semantics is intrinsic to the Hindi gender system rather than an artifact of the cross-lingual tagging pipeline. This hierarchy is corroborated by feature importance analysis, where inflectional attributes account for 77.5% of predictive power, far exceeding all other linguistic factors.

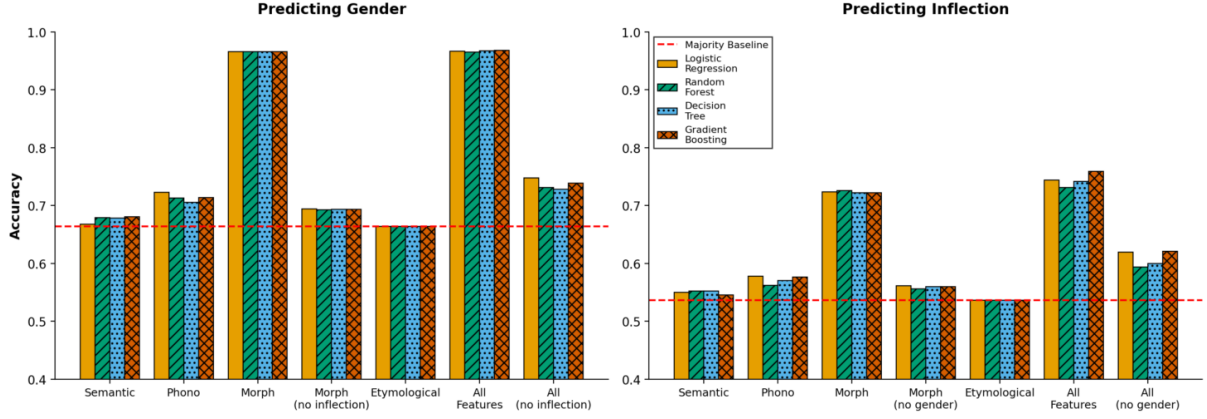


Figure 1: Prediction accuracy across different feature sets for (Left) Gender and (Right) Inflection Class using tree-based classifiers. The red dashed line represents the majority class baseline (66.4% for gender, 53.6% for inflection). Trends in Macro F1 and AUC-ROC (see Appendix A) were consistent with accuracy, confirming that high performance is not an artifact of class imbalance.

Table 3: Performance impact of feature interactions in the neural network model **trained without primary predictors (gender and inflection class)**. The top row shows the baseline model (no interaction module). Subsequent rows show the performance when specific pairwise feature interactions are explicitly modeled; bolded values have $\Delta > +0.5\%$.

Interaction	Gender				Inflection Class			
	Acc	F1	AUC	Δ Acc	Acc	Macro F1	AUC	Δ Acc
Baseline (no interaction)	74.50%	83.35%	0.7414	—	41.17%	35.85%	0.7611	—
Etymology $\times$ Derivation	75.06%	83.60%	0.7503	+0.55%	43.49%	37.21%	0.7635	+2.32%
Etymology $\times$ Phonology	74.28%	83.39%	0.7409	-0.22%	44.70%	37.92%	0.7652	+3.53%
Etymology $\times$ Semantic	74.61%	83.41%	0.7341	+0.11%	40.29%	35.05%	0.7700	-0.88%
Phonology $\times$ Semantic	72.30%	82.19%	0.7232	-2.21%	37.75%	33.02%	0.7378	-3.42%
Derivation $\times$ Semantic	74.17%	83.33%	0.7411	-0.33%	38.08%	34.75%	0.7608	-3.09%
Derivation $\times$ Phonology	74.06%	83.15%	0.7369	-0.44%	41.94%	36.03%	0.7567	+0.77%

5.2 Predicting Inflection

Predicting inflection class proves less robust than gender, with the best-performing set (All Features) achieving an accuracy of $\approx 74.4\%$ against a 53.6% baseline (Figure 1, Right). As anticipated, gender acts as the primary predictor. It dominates feature importance analysis (53.1%), scoring roughly 12 $\times$ higher than the next most informative feature. Consequently, ablating gender precipitates substantial performance drops: accuracy falls by 16.4 percentage points for the Morphological set and 13.6 points for All Features, with even sharper declines in AUC-ROC ($\Delta -0.20$ and $\Delta -0.11$, respectively).

However, unlike the gender prediction task, gender is not the sole determinant of inflection class. Predictive signals outside of the gender-morphology relationship are statistically significant. The All Features set outperforms Morphological alone (Accuracy $\Delta +2.1\%$, AUC $\Delta +0.10$, McNemar’s $p < 0.001$), indicating that phonological and semantic features contribute unique information. Furthermore, even after ablating gender, the All (no gender) model maintains an accuracy of

$\approx 60.8\%$, significantly above the 53.6% baseline, driven largely by phonology, which stands as the second-strongest individual feature class. In sum, while gender and inflection class are interdependent, the relationship is asymmetrical: inflection class nearly determines gender, whereas gender is a dominant but partial predictor of inflection class.

5.3 Feature Interactions

To isolate the contribution of non-dominant features, we evaluated neural network models where the primary predictors (inflection class for gender prediction; gender for inflection prediction) were removed (see Table 3). As expected, the baselines for these restricted models are significantly lower than the full interpretable models (Gender: 74.5% vs 96.6%; Inflection: 41.2% vs 74.4%), confirming the centrality of the excluded features.

Against these restricted baselines, explicit feature interactions reveal an asymmetry between the two tasks. For gender prediction, pairwise interactions provide negligible gains: only the Etymology $\times$ Derivation interaction yields a positive improvement ($\Delta +0.55\%$), while all others are neu-

tral or detrimental (e.g., Phonology $\times$ Semantic: $\Delta -2.21\%$). In contrast, predicting inflection class benefits substantially from feature cross-products. Three of the six interaction pairs yield accuracy boosts exceeding 0.7%, the strongest being Etymology $\times$ Phonology ($\Delta +3.53\%$) and Etymology $\times$ Derivation ($\Delta +2.32\%$). This suggests that while gender assignment is relatively impervious to secondary feature combinations, inflection class is partially recoverable through the interaction of etymology with phonological and derivational form.

6 Discussion

6.1 Hindi as an Extreme Case

Morphological features, specifically inflection class, predict noun gender with near-perfect accuracy ($\approx 97\%$, AUC-ROC 0.98). This surpasses recent computational analyses of other gender systems trained on comparable form and meaning features, though cross-study figures are only indicative given differences in dataset size, feature sets, and modeling choices. French approaches this threshold (94%; [Sahai and Sharma, 2021](#)), but others exhibit significantly lower predictability, from moderate accuracy in Tashlhiyt (86%; [Alderete et al., 2025](#)) to performance only marginally above baseline in Mian (68%; [Allasonnière-Tang et al., 2021](#)) and Nakh languages (57–61%; [Schreur et al., 2022](#)). We attribute this to the intrinsic regularity of Hindi’s assignment system and the granularity of our dataset.

This dominance bears directly on claims of a universal “semantic core” in gender systems. [Williams et al. \(2019\)](#) use canonical correlation analysis to find significant gender–semantics correlations across 18 languages, including Hindi. Our ablation methodology provides a finer-grained, language-specific test: while semantic features carry some predictive signal ($\approx 67\%$), this contribution is redundant once inflection class is available ($\approx 97\%$). Hindi thus represents an extreme case of inflection-driven gender assignment, exceeding even Russian, where inflection class is dominant but subordinate to semantic rules for animates ([Corbett, 1982](#)).

6.2 The Bootstrapping Problem

Our results reveal a marked asymmetry in the interdependence of gender and inflection. While inflection class predicts gender nearly categorically, the inverse prediction is far less robust; predicting inflection relies on a distributed signal composed of gender, phonological n -grams, and specific etymological interactions. This structure suggests a

possible resolution to the bootstrapping problem. Since inflection class is partially recoverable from form-based and etymological cues independent of gender, it serves as the logical anchor point. This independent recoverability ($\approx 60.8\%$, §5.2) shows inflection class is a generalization a learner could extract, not an artifact of the annotation schema.

This anchoring role is consistent with a learning trajectory in which inflectional paradigms are acquired first, scaffolded by phonology, and subsequently serve as the primary predictor for grammatical gender. While our form-based features alone yield moderate accuracy for inflection, human acquisition likely leverages additional distributional cues. Factors such as semantic natural gender, token frequency, and phonological neighborhood density, central to discriminative learning models ([Bybee and Slobin, 1982](#); [Baayen et al., 2011](#)), likely supplement the phonological cues identified here to facilitate robust acquisition of inflection paradigms.

6.3 Lexical Roots of the Gender Bias

Our expert annotations surface a systematic masculine skew invisible to corpus statistics alone: among animate-centric categories, 84.2% of nouns are masculine versus a 66.4% baseline. This structurally explains gender biases in Hindi NLP systems: [Pujari et al. \(2019\)](#) found Hindi embeddings project neutral concepts onto masculine subspaces, [Gupta et al. \(2021\)](#) observed Hindi-English MT defaulting to masculine pronouns, and [Gupta et al. \(2025\)](#) report GPT-4o producing gender-biased outputs at 87.8% in Hindi versus 33.4% in English. Such behaviors are typically attributed to corpus-level distributional patterns, but our results identify a deeper source: the lexicon itself encodes a roughly 9:1 masculine skew for occupations that both static embeddings and generative models inherit. Because this bias is morphologically codified in inflection class membership, debiasing methods that operate only at the embedding level ([Pujari et al., 2019](#)) may be insufficient. Given the scarcity of bias-evaluation datasets for Indic languages ([Hada et al., 2024](#)), we suggest bias metrics move beyond global accuracy to report performance on high-skew subsets such as animate-centric categories.

6.4 Implications for Multilingual Transfer

From a multilingual perspective, our results identify a typology-conditioned transfer risk. Gender-prediction strategies learned from languages in

which semantic or phonological cues are informative may transfer poorly to Hindi, where inflection class dominates those cues. Conversely, since multilingual models encode morphosyntax cross-lingually (Stańczak et al., 2022), encoding Hindi gender should entail encoding nominal inflectional paradigms.

The near-categorical inflection–gender association therefore suggests that Hindi NLP systems should prioritize inflectional paradigms over semantic embedding spaces, particularly for the out-of-vocabulary terms on which morphological analyzers and taggers struggle; inflection class can itself be induced from phonological n -grams. This bears directly on Machine Translation (MT). Hindi-to-English translation of the epicene pronoun *vah* (he/she/it) requires the antecedent’s grammatical gender, which Stanovsky et al. (2019) note is often lost in neural representations; the antecedent’s inflectional morphology provides a more reliable signal for its recovery than semantic representations alone.

For English-to-Hindi generation, where gender governs agreement on dependent verbs and adjectives, the same dependence motivates testing whether conditioning generative models on inflectional paradigms reduces the “gender hallucination” of neural systems (Vanmassenhove et al., 2018). Such abstract categories surface only under controlled annotation and interpretable modeling: end-to-end systems trained on raw text capture surface correlates of agreement but not inflection class itself. Recent probing locates grammatical-gender information in a shared cross-lingual representation, attributing it largely to semantic features (Schröter and Basirat, 2025); the Hindi case shows this signal can instead be carried by inflectional form, which semantic probes do not isolate. Our dataset thus serves as a probing benchmark for whether multilingual models encode these regularities or rely on shallow heuristics, and as ground-truth validation when such models generate synthetic data for morphologically complex languages.

7 Conclusion

We presented the first large-scale computational analysis of Hindi gender assignment using a dataset of 6,040 expert-annotated noun paradigms. Inflection class predicts gender near-categorically ($\approx 97\%$), while phonological, semantic, and etymological cues contribute negligibly once it is available. The relationship is asymmetric, since inflec-

tion class is itself only partially recoverable from form. This establishes Hindi as an extreme case of morphologically aligned gender and suggests inflectional paradigms, rather than surface or semantic cues, as the primary representation for gender in Hindi NLP. Our data release can support probing of these regularities in multilingual models.

Limitations

Dataset Scope and Register. While our dataset of 6,040 paradigms allows for a detailed morphological analysis, it represents only a fraction of the full Hindi noun lexicon based on cross-linguistic vocabulary size estimates (Brysbaert et al., 2016). Furthermore, as the data were derived from the Hindi Treebank (news and formal text), our findings reflect the gender distribution of Standard Modern Hindi. Consequently, our results may not fully generalize to colloquial speech, social media text, or regional dialects (e.g., Braj, Awadhi) where gender agreement rules can be relaxed.

Static Labeling and Sociolinguistic Variation. Our target labels rely on treebank data and dictionary adjudication, effectively modeling grammatical gender as a static, single-label property of the lemma. Our annotators’ training also reflected Standard Modern Hindi, so their judgements may also underrepresent dialect variation. This “frozen” approach is necessary for a paradigm-based study but abstracts away from the sociolinguistic variation and “gender flux” present in natural usage (Masica, 1993). It does not capture nouns that exhibit variable agreement based on regional dialects or specific semantic contexts (word sense), nuances which would require a token-level analysis to resolve.

Lexical Sampling and Markedness. Our dataset relies on the Hindi Treebank, which consists primarily of formal news text. This sampling strategy introduces two specific biases. First, the prevalence of male-centric topics in news media likely results in a lexicon that over-represents masculine professional lemmas. Second, in Standard Modern Hindi, the masculine form often serves as the “unmarked” default for occupations (e.g., *lekhak* ‘writer’). Consequently, our finding of a 90% masculine skew in professions accurately reflects the formal, standardized lexicon but likely undercounts explicit feminine forms that may appear more frequently in other registers.

Feature Extraction Artifacts. One of our semantic features relies on cross-lingual projection, translating Hindi lemmas into English to utilize the USAS tagger. This pipeline introduces translation noise and may obscure cultural nuances or mis-categorize polysemous words, potentially underestimating the predictive power of semantics relative to morphology. Additionally, our etymological

annotation groups all non-borrowed terms into a single “Native” category. This conflates *Tadbhava* words (inherited from Sanskrit via Prakrit) with *Deshaj* words (indigenous/unknown origin), obscuring finer-grained lexical strata that may interact with phonological patterns.

Multilingual Transfer. Although our findings have implications for multilingual representation learning, the present experiments are monolingual and do not directly evaluate cross-lingual transfer or pretrained multilingual models. Future work should test whether multilingual encoders preserve the inflection–gender dependency identified here, and whether transfer from typologically different languages weakens this representation.

Modeling Constraints. Finally, our neural network analysis explicitly models only pairwise interactions (e.g., Etymology $\times$ Phonology). Linguistic systems may involve complex, multivariate dependencies (e.g., Etymology $\times$ Phonology $\times$ Animacy). While our approach highlights key asymmetries between gender and inflection prediction, it does not probe these higher-order interactions, which may contain additional predictive signal.

Ethics Statement

This work involves the creation of a dataset and models based on Standard Modern Hindi. We identify four primary ethical considerations:

Gender Binaries. Although this study concerns grammatical gender, a morphosyntactic property of nouns, our computational modeling relies on the strict masculine/feminine binary inherent to standard Hindi grammar. We acknowledge that, when applied to animate nouns, this binary intersects with social questions of gender identity. NLP systems built on this premise may misgender individuals or enforce normative standards that do not reflect all speakers (Larson, 2017).

Bias Propagation. Our semantic analysis (Section 3.2.2) quantifies a severe masculine skew in terms denoting high-agency and professional roles. Researchers using this dataset to train generative models should be aware that it encodes these historical biases, potentially leading to representational harms if not mitigated via techniques such as counterfactual data augmentation.

Annotator Labor. The manual annotation for this study was conducted by three of the authors,

who were compensated for their labor consistent with local labor standards.

AI Assistance. We utilized AI assistants (specifically Anthropic’s Claude) to plan and orchestrate some Python scripts used for experimental analysis, as well as to provide editorial suggestions on our original drafts. All code logic, data processing, and textual revisions were manually verified and authorized by the authors, who bear full responsibility for the methodology and content.

Acknowledgements

We gratefully acknowledge Rajesh Bhatt for his comments and questions and Joti Balani for organizational support. This research was supported by a Social Sciences and Humanities Research Council of Canada grant (435-2020-0193).

References

- Rama Kant Agnihotri. 2022. *Hindi: An essential grammar*. Routledge.
- John Alderete, Piyush Agarwal, Kaye Holubowsky, and Abdelkrim Jebbour. 2025. The Gender System of Tashlhiyt: Using Supervised Learning to Predict Noun Gender. *Linguistic Analysis*, 44:3–54.
- Marc Allasonnière-Tang, Dunstan Brown, and Sebastian Fedden. 2021. Testing semantic dominance in Mian gender. *Oceanic Linguistics*, 60(2):302–334.
- Mark Aronoff. 1991. Noun classes in Arapesh. In *Yearbook of morphology 1991*, pages 21–32. Springer.
- Mark Aronoff. 1993. *Morphology by itself: Stems and inflectional classes*, volume 22. MIT press.
- R Harald Baayen, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological review*, 118(3):438.
- Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. The Hindi/Urdu Treebank project. In *Handbook of linguistic annotation*, pages 659–697. Springer.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant’s age. *Frontiers in psychology*, 7:1116.
- Joan L Bybee and Dan I Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language*, 58(2):265–289.
- Andrew Carstairs-McCarthy. 1994. Inflection classes, gender, and the principle of contrast. *Language*, pages 737–788.
- CFILT, IIT Bombay. 2025. [Hindi WordNet \(HWN\) web interface](#). Accessed November 2025.
- Greville G Corbett. 1982. Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics*, 6(2):197–232.
- Greville G Corbett. 1991. *Gender*. Cambridge University Press.
- Greville G Corbett. 2007. Gender and noun classes. In *Language Typology and Syntactic Description: III: Grammatical categories and the lexicon*, pages 241–279. Cambridge University Press.
- Greville G Corbett and Norman M Fraser. 2000. Gender assignment: a typology and a model. *Systems of nominal classification*, 4:293–325.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. [Datasheets for datasets](#). *Communications of the ACM*, 64(12):86–92.
- Google. 2025. [Cloud Translation API](#). Accessed November 2025.
- Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23.
- Ishita Gupta, Ishika Joshi, Adrita Dey, and Tapan Parikh. 2025. “Since lawyers are males..”: Examining implicit gender bias in Hindi language generation by LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 3254–3264.
- Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Sesshadri, and 1 others. 2024. Akal badi ya bias: An exploratory study of gender bias in Hindi language technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1926–1939.
- Sambhav Jain, Naman Jain, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Dipti Misra Sharma. 2013. Exploring semantic information in Hindi WordNet for Hindi dependency parsing. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 189–197.

- Yamuna Kachru. 2006. *Hindi*, volume 12 of *London Oriental and African Language Library*. John Benjamins, Amsterdam.
- Ruth Kramer. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of linguistics*, 6(1):45–66.
- Sebastian Kürschner and Damaris Nübling. 2011. The interaction of gender and declension in Germanic languages. *Folia linguistica*, 45(2):355–388.
- Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11.
- Colin P Masica. 1993. *The Indo-Aryan languages*. Cambridge University Press.
- Ronald Stuart McGregor. 1993. *The Oxford Hindi-English Dictionary*. Oxford University Press, Oxford.
- Ronald Stuart McGregor. 1995. *Outline of Hindi grammar*. Oxford University Press.
- Scott Piao, Paul Rayson, Andrew Moore, and Mahmoud El-Haj. 2022. *PyMUSAS: A Python library for the UCREL semantic analysis system*. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6547–6554, Marseille, France. European Language Resources Association.
- Steven Pinker. 1984. *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA.
- Arun K Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased Hindi words with word-embedding. In *Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence*, pages 450–456.
- Paul Rayson, Dawn E Archer, Scott L Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, in association with LREC-04*, pages 7–12. European Language Resources Association.
- Saumya Sahai and Dravyansh Sharma. 2021. Predicting and explaining French grammatical gender. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 90–96.
- Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerbero-f-Arenas, and Luisa Bentivogli. 2024. *What the harm? Quantifying the tangible impact of gender bias in machine translation with a human-centered study*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.
- Jesse Wichers Schreur, Marc Allasonnière-Tang, Kate Bellamy, and Neige Rochant. 2022. Predicting grammatical gender in Nakh languages: Three methods compared. *Linguistic Typology at the Crossroads*, 2(2):93–126.
- Andrea Schröter and Ali Basirat. 2025. *Universal patterns of grammatical gender in multilingual large language models*. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 34–46, Suzhou, China. Association for Computational Linguistics.
- Smriti Singh and Vijayanthi M Sarma. 2010. Hindi noun inflection and distributed morphology. In *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*, pages 307–321.
- Yash Sinha. 2018. Hindi nominal suffixes are bimorphic: A distributed morphology analysis. *Proceedings of the Linguistic Society of America*, 3:20–1.
- Karolina Stańczak, Kevin Du, Adina Williams, Isabelle Augenstein, and Ryan Cotterell. 2024. *The causal influence of grammatical gender on distributional semantics*. *Transactions of the Association for Computational Linguistics*, 12:1672–1685.
- Karolina Stańczak, Edoardo Ponti, Lucas Torroba Hen-nigen, Ryan Cotterell, and Isabelle Augenstein. 2022. *Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models*. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. *Evaluating gender bias in machine translation*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Gregory T Stump. 2001. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press.
- Bruce Tesar and Paul Smolensky. 2000. *Learnability in optimality theory*. MIT Press.
- G Richard Tucker, Wallace E Lambert, and André Albert Rigault. 1977. *The French speaker’s skill with grammatical gender: An example of rule-governed behavior*. Mouton.
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.
- Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods*

in *Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*, 9:139–159.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Appendices

A Model Performance

Tables 5 and 6 below give complete performance metrics for all interpretable models predicting gender and inflection, cross-classified by feature sets, both with and without cross-fold validation (see Section 4.5). When feature sets lack sufficient discriminative power to differentiate among outcome classes, models may converge to a trivial strategy of predicting only the majority class. In such cases, all models, regardless of algorithmic differences, will exhibit identical performance at the majority baseline, as seen with etymological features for both gender prediction (66.4% accuracy; all models predict masculine) and inflection prediction (53.6% accuracy; all models predict class 4). This uniform performance indicates that the features do not provide adequate signal to justify predictions beyond the baseline, rather than reflecting a computational error.

B Reproducibility

Computing Infrastructure. All neural network models were trained on a single Apple M4 Pro chip (20-core integrated GPU). The total training time for the full suite of experiments, including the exhaustive grid search and cross-validation folds, was less than 3 hours.

Hyperparameter Search Space. The models described in Section 4.5 used two search configurations. The baseline comparisons used fixed hyperparameters:

Logistic Regression $C = 0.1$

Random Forest $n_estimators=100,$
 $max_depth=15, min_samples_leaf=5$

Decision Tree $max_depth=10,$
 $min_samples_leaf=5$

Table 4: Grid search ranges for the Semantic/Random Forest configuration.

Parameter	Values
<code>n_estimators</code>	{100, 200, 500}
<code>max_depth</code>	{10, 15, 20, 30, None}
<code>min_samples_leaf</code>	{1, 2, 5, 10}
<code>min_samples_split</code>	{2, 5, 10}
<code>max_features</code>	{sqrt, log2, 0.5}

Gradient Boosting $n_estimators=100,$
 $max_depth=5, min_samples_leaf=5,$
 $learning_rate=0.05$

To validate robustness, we ran an exhaustive grid search on the Semantic/Random Forest configuration over the ranges in Table 4. This search ($3 \times 5 \times 4 \times 3 \times 3 = 540$ combinations) yielded only a +0.83 percentage-point improvement (67.88% $\rightarrow$ 68.71%), confirming that the observed performance primarily reflects feature quality rather than optimization artifacts.

C Datasheet for the Dataset

This datasheet follows the framework proposed by Geburu et al. (2021) to document the motivation, composition, collection process, and recommended uses of the Hindi Noun Paradigm dataset.

C.1 Motivation

- **For what purpose was the dataset created?** The dataset was created to enable the first large-scale computational analysis of grammatical gender assignment in Hindi. It aims to fill a gap in low-resource NLP by providing high-quality annotations for inflectional morphology, phonology, semantics, and etymology, facilitating research into morphological processing and gender bias.
- **Who created the dataset?** The dataset was created by the authors of this paper.
- **Who funded the creation of the dataset?** The creation was funded by a research grant held by the first author from the Social Sciences and Humanities Research Council of Canada.

C.2 Composition

- **What do the instances that comprise the dataset represent?** Each instance represents a unique Hindi noun paradigm (lemma).
- **How many instances are there in total?** There are 6,040 unique noun paradigms.

Table 5: Consolidated performance metrics for Gender prediction (binary) and Inflection Class prediction (6-way) using interpretable models. The “Ablated” condition refers to removing the primary predictor: inflection class is removed for gender prediction, and gender is removed for inflection class prediction.

Feature Set	Model	Predicting Gender			Predicting Inflection Class		
		Acc	F1	AUC	Acc	F1	AUC
Semantic	LogReg	66.80%	52.12%	0.6178	54.97%	20.02%	0.6426
	RandForest	67.88%	48.59%	0.6256	55.22%	17.86%	0.6489
	DecTree	67.80%	50.73%	0.6119	55.22%	19.21%	0.6008
	GradBoost	68.05%	50.44%	0.6224	54.55%	17.98%	0.6317
Phonological	LogReg	72.27%	62.13%	0.7165	57.78%	26.06%	0.6599
	RandForest	71.27%	57.18%	0.6855	56.21%	19.54%	0.6440
	DecTree	70.53%	56.62%	0.6437	57.04%	24.42%	0.6204
	GradBoost	71.36%	61.38%	0.6793	57.62%	24.78%	0.6493
Morphological	LogReg	96.61%	96.12%	0.9835	72.35%	32.96%	0.7688
	RandForest	96.61%	96.12%	0.9806	72.60%	30.75%	0.7713
	DecTree	96.61%	96.12%	0.9838	72.19%	32.67%	0.7698
	GradBoost	96.61%	96.12%	0.9834	72.19%	32.67%	0.7704
Morphological (Ablated)	LogReg	69.45%	50.54%	0.5773	56.13%	18.68%	0.5751
	RandForest	69.29%	50.11%	0.5763	55.63%	17.76%	0.5731
	DecTree	69.37%	50.32%	0.5553	55.96%	18.34%	0.5647
	GradBoost	69.37%	50.32%	0.5763	55.96%	18.34%	0.5736
Etymological	LogReg	66.39%	39.90%	0.5503	53.64%	11.64%	0.6747
	RandForest	66.39%	39.90%	0.5503	53.64%	11.64%	0.6732
	DecTree	66.39%	39.90%	0.5503	53.64%	11.64%	0.6732
	GradBoost	66.39%	39.90%	0.5503	53.64%	11.64%	0.6747
All Features	LogReg	96.69%	96.24%	0.9791	74.42%	45.68%	0.8642
	RandForest	96.52%	96.04%	0.9795	73.10%	35.91%	0.8779
	DecTree	96.77%	96.33%	0.9795	74.17%	43.15%	0.8696
	GradBoost	96.85%	96.42%	0.9856	75.91%	43.61%	0.8965
All Features (Ablated)	LogReg	74.75%	68.65%	0.7537	61.92%	35.98%	0.7708
	RandForest	73.10%	60.34%	0.7250	59.35%	26.43%	0.7759
	DecTree	72.85%	63.64%	0.6926	60.02%	31.19%	0.7379
	GradBoost	73.84%	64.73%	0.7269	62.09%	33.18%	0.7916

- **Does the dataset contain all possible instances or is it a sample?** It is a sample derived from the Hindi Treebank (Bhat et al., 2017), representing nouns found in news and formal text. It does not cover the entire Hindi lexicon (estimated at >20,000 nouns).
- **What data does each instance consist of?** Each instance includes (see Table 7 for more detail):
 - **Record Info:** ID, Lemma (Script), Stem (Script/Phonetic), English Gloss.
 - **Morphological:** Gender (M/F), Inflection Class (0–5), Derivational Category (90+ types), Compound status.
 - **Semantic:** Animacy, USAS Category (coarse), WordNet Category (fine).
 - **Etymological:** Origin (Native, Sanskrit, Persian-Arabic, English).
 - **Phonological:** Character and phoneme n -grams.
- **Are there any known errors, sources of noise, or redundancies?** The semantic features derived via translation (USAS) may contain noise due to polysemy or translation errors.
- **Is there anything missing from individual instances?** Some noun paradigms are partial in the sense that they lack some wordforms in the four cell paradigm. However, these can be recovered with the inflection class labels provided with this dataset.
- **Are there recommended data splits (for example, training, development/validation, testing)?** No. We used standard data splits in our computational experiments (80/20, 5-fold cross-validation, 70/15/15), but they are not part of the dataset.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause**

Table 6: Cross-validation performance metrics (mean $\pm$ std across 5 folds) for Gender prediction (binary) and Inflection Class prediction (6-way) using interpretable models. The “Ablated” condition refers to removing the primary predictor: inflection class is removed for gender prediction, and gender is removed for inflection class prediction. Cross-validation demonstrates high stability across folds (accuracy std $\leq 1.1\%$) and validates single-split results for the primary feature sets (Morphological and All Features), whose accuracies fall within 1.3 percentage points of cross-validated means, confirming findings are robust and not artifacts of sampling variance.

Feature Set	Model	Predicting Gender			Predicting Inflection Class		
		Acc	F1	AUC	Acc	F1	AUC
Semantic	LogReg	67.35% $\pm$ 0.30%	53.61% $\pm$ 0.79%	0.6468 $\pm$ 0.0160	54.57% $\pm$ 0.87%	19.10% $\pm$ 0.99%	0.6506 $\pm$ 0.0161
	RandForest	67.86% $\pm$ 0.41%	48.34% $\pm$ 0.74%	0.6430 $\pm$ 0.0186	54.83% $\pm$ 0.39%	16.70% $\pm$ 0.86%	0.6510 $\pm$ 0.0153
	DecTree	67.72% $\pm$ 0.40%	50.12% $\pm$ 0.99%	0.6149 $\pm$ 0.0173	54.69% $\pm$ 0.65%	18.91% $\pm$ 1.23%	0.6112 $\pm$ 0.0147
	GradBoost	68.03% $\pm$ 0.31%	50.20% $\pm$ 0.98%	0.6383 $\pm$ 0.0174	54.52% $\pm$ 0.55%	18.24% $\pm$ 1.03%	0.6331 $\pm$ 0.0123
Phonological	LogReg	70.43% $\pm$ 1.03%	58.55% $\pm$ 2.21%	0.6785 $\pm$ 0.0218	57.53% $\pm$ 0.56%	26.07% $\pm$ 0.83%	0.6591 $\pm$ 0.0141
	RandForest	66.36% $\pm$ 0.07%	39.89% $\pm$ 0.02%	0.6316 $\pm$ 0.0152	56.26% $\pm$ 0.56%	19.37% $\pm$ 1.48%	0.6500 $\pm$ 0.0109
	DecTree	67.57% $\pm$ 0.81%	47.83% $\pm$ 1.59%	0.5438 $\pm$ 0.0076	56.74% $\pm$ 0.77%	23.04% $\pm$ 1.25%	0.6150 $\pm$ 0.0146
	GradBoost	67.80% $\pm$ 0.78%	47.82% $\pm$ 1.14%	0.5965 $\pm$ 0.0201	57.32% $\pm$ 0.40%	23.72% $\pm$ 1.02%	0.6555 $\pm$ 0.0134
Morphological	LogReg	96.49% $\pm$ 0.35%	95.97% $\pm$ 0.41%	0.9836 $\pm$ 0.0032	72.88% $\pm$ 0.53%	32.25% $\pm$ 0.79%	0.7682 $\pm$ 0.0132
	RandForest	96.47% $\pm$ 0.40%	95.95% $\pm$ 0.46%	0.9825 $\pm$ 0.0051	73.01% $\pm$ 0.35%	31.46% $\pm$ 0.53%	0.7726 $\pm$ 0.0110
	DecTree	96.36% $\pm$ 0.36%	95.82% $\pm$ 0.42%	0.9839 $\pm$ 0.0035	72.76% $\pm$ 0.45%	32.27% $\pm$ 0.92%	0.7717 $\pm$ 0.0106
	GradBoost	96.49% $\pm$ 0.35%	95.96% $\pm$ 0.41%	0.9841 $\pm$ 0.0034	72.76% $\pm$ 0.43%	32.43% $\pm$ 0.88%	0.7732 $\pm$ 0.0111
Morphological (Ablated)	LogReg	69.35% $\pm$ 0.26%	50.33% $\pm$ 0.98%	0.5733 $\pm$ 0.0097	56.49% $\pm$ 0.25%	18.96% $\pm$ 0.54%	0.5835 $\pm$ 0.0054
	RandForest	68.92% $\pm$ 0.28%	49.09% $\pm$ 1.13%	0.5703 $\pm$ 0.0083	56.09% $\pm$ 0.21%	18.24% $\pm$ 0.45%	0.5820 $\pm$ 0.0053
	DecTree	69.07% $\pm$ 0.25%	49.59% $\pm$ 1.06%	0.5543 $\pm$ 0.0057	56.27% $\pm$ 0.28%	18.56% $\pm$ 0.63%	0.5732 $\pm$ 0.0059
	GradBoost	69.12% $\pm$ 0.25%	49.72% $\pm$ 1.01%	0.5704 $\pm$ 0.0084	56.34% $\pm$ 0.25%	18.67% $\pm$ 0.57%	0.5821 $\pm$ 0.0053
Etymological	LogReg	66.39% $\pm$ 0.00%	39.90% $\pm$ 0.00%	0.5680 $\pm$ 0.0074	53.63% $\pm$ 0.03%	11.64% $\pm$ 0.00%	0.6759 $\pm$ 0.0083
	RandForest	66.39% $\pm$ 0.00%	39.90% $\pm$ 0.00%	0.5680 $\pm$ 0.0074	53.63% $\pm$ 0.03%	11.64% $\pm$ 0.00%	0.6747 $\pm$ 0.0085
	DecTree	66.39% $\pm$ 0.00%	39.90% $\pm$ 0.00%	0.5680 $\pm$ 0.0074	53.63% $\pm$ 0.03%	11.64% $\pm$ 0.00%	0.6748 $\pm$ 0.0087
	GradBoost	66.39% $\pm$ 0.00%	39.90% $\pm$ 0.00%	0.5680 $\pm$ 0.0074	53.63% $\pm$ 0.03%	11.64% $\pm$ 0.00%	0.6752 $\pm$ 0.0090
All Features	LogReg	96.69% $\pm$ 0.40%	96.24% $\pm$ 0.45%	0.9830 $\pm$ 0.0050	73.20% $\pm$ 1.08%	44.75% $\pm$ 1.31%	0.8657 $\pm$ 0.0130
	RandForest	96.47% $\pm$ 0.43%	95.95% $\pm$ 0.50%	0.9838 $\pm$ 0.0060	72.91% $\pm$ 0.61%	36.25% $\pm$ 1.52%	0.8867 $\pm$ 0.0062
	DecTree	96.74% $\pm$ 0.39%	96.27% $\pm$ 0.45%	0.9812 $\pm$ 0.0040	72.93% $\pm$ 0.86%	39.89% $\pm$ 0.80%	0.8561 $\pm$ 0.0186
	GradBoost	96.97% $\pm$ 0.43%	96.53% $\pm$ 0.50%	0.9856 $\pm$ 0.0043	74.77% $\pm$ 0.53%	42.42% $\pm$ 0.98%	0.8910 $\pm$ 0.0066
All Features (Ablated)	LogReg	72.95% $\pm$ 0.92%	66.25% $\pm$ 1.07%	0.7426 $\pm$ 0.0129	60.12% $\pm$ 0.59%	34.30% $\pm$ 0.86%	0.7667 $\pm$ 0.0144
	RandForest	69.25% $\pm$ 0.43%	48.97% $\pm$ 1.34%	0.6983 $\pm$ 0.0082	59.01% $\pm$ 0.94%	24.80% $\pm$ 2.27%	0.7823 $\pm$ 0.0104
	DecTree	70.73% $\pm$ 0.21%	56.47% $\pm$ 1.18%	0.6584 $\pm$ 0.0123	59.88% $\pm$ 0.74%	31.43% $\pm$ 1.97%	0.7487 $\pm$ 0.0139
	GradBoost	71.19% $\pm$ 0.52%	57.18% $\pm$ 1.20%	0.7018 $\pm$ 0.0075	61.01% $\pm$ 0.67%	32.11% $\pm$ 1.86%	0.7874 $\pm$ 0.0081

anxiety? The dataset contains terms related to sensitive topics (e.g., crime, violence) inherent to news corpora, but they are presented as lexical entries without graphic context.

- **Does the dataset reflect specific social biases?** Yes. As detailed in Section 3.2.2 and the Discussion, the dataset reflects a systemic gender bias where high-agency semantic categories (e.g., Occupations) are 89.6% masculine, significantly higher than the corpus baseline.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?** It is self-contained in the sense that all of the records are complete and do not require external resources for data processing. However, the data was derived and engineered using external resources, such as the Hindi Treebank and WordNet, as described in detail above.

C.3 Collection Process

- **What mechanisms or procedures were used to collect the data?** The core word list was extracted from the Hindi Treebank. Etymological tags were extracted from McGregor’s Oxford Hindi-English Dictionary (McGregor, 1993). Semantic tags were derived via the Hindi WordNet API and the PyMUSAS tagger (via Google Translate).
- **Who was involved in the data collection process and how were they compensated?** Manual annotation and adjudication were performed by two native Hindi speakers and a non-native speaker data analyst who are also co-authors of this paper.
 - **Recruitment:** The analysts were recruited through a call for NLP and linguistics students posted on job boards at Simon Fraser University.
 - **Selection:** Candidates underwent two interview rounds assessing Hindi proficiency and mathematical/linguistic aptitude.

Table 7: Overview of all data fields and features available in the dataset. This includes record-level information (identifiers), linguistic structures (lemmas, stems, glosses), as well as specific morphological, semantic, etymological, and phonological features used for analysis. The vocabulary size indicates the number of unique values for categorical variables.

Feature Type	Feature	Variable Type	Vocab Size
Record	r_id	string	–
	r_lemma_script	string	–
	r_stem_script	string	–
	r_stem_phonetic	string	–
	r_gloss_english	string	–
	r_same_lemma_type	categorical	3
Morphological	m_gender	categorical	2
	m_inflection	categorical	6
	m_deriv_cat	categorical	90
	m_deriv_cat_important	categorical	17
	m_deriv_cat_topk	categorical	10
	m_compound	binary	2
	m_derived	binary	2
Semantic	s_animacy	binary	2
	s_category_usas	categorical	21
	s_category_wnet	categorical	108
Etymology	e_english	binary	2
	e_persian_arabic	binary	2
	e_sanskrit	binary	2
Phonology	p_ngrams_script	multi-category	5385
	p_ngrams_phonetic	multi-category	4265
	p_ngrams_script_important	multi-category	160
	p_ngrams_phonetic_important	multi-category	151
	p_ngrams_script_topk	multi-category	20
	p_ngrams_phonetic_topk	multi-category	20

- **Compensation:** Analysts were compensated for their data collection and analysis work at standard university hourly rates. They were offered co-authorship contingent on their contribution to model development and results analysis, duties which they fulfilled.

- **Over what timeframe was the data collected?** The data collection and annotation process took place one year prior to the submission of this work.

C.4 Preprocessing, Cleaning, and Labeling

- **Was any preprocessing/cleaning/labeling of the data done?** Yes. We resolved duplicate lemmas arising from encoding inconsistencies and adjudicated homophones (e.g., distinguishing true homophones from gendered pairs like *laṛkā* ‘boy’ vs. *laṛkī* ‘girl’). Inflection classes were manually annotated based on a five-class schema derived from authoritative grammars, extended with a sixth class for English loan plurals.
- **Is the software used to preprocess/clean/label the data available?** Yes, the code for fea-

ture extraction and preprocessing is released with the project github repo.

C.5 Uses

- **Has the dataset been used for any tasks already?** It has been used in this paper to train interpretable classifiers and neural networks for gender and inflection prediction.
- **What (other) tasks could the dataset be used for?** It is suitable for training morphological analyzers, improving gender agreement in Machine Translation and NLG, and studying sociolinguistic bias in the Hindi lexicon.
- **Are there tasks for which the dataset should not be used?** The dataset relies on a strict binary gender formulation. It should not be used as a normative standard to enforce binary gender agreement in contexts requiring sensitivity to non-binary or gender-fluid identities. Furthermore, generative models trained on this data should employ mitigation strategies to avoid propagating the masculine skew found in high-agency professional terms.

C.6 Distribution

- **Will the dataset be distributed to third parties?** Yes.
- **How will the dataset be distributed?** The dataset is available at the github repo.
- **When will the dataset be distributed?** Upon publication.
- **What license (if any) is it distributed under?**
The dataset is released under the **CC-BY 4.0 International** license, allowing for sharing and adaptation with appropriate credit.
- **Who will be supporting/hosting/maintaining the dataset?** The first author.
- **Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?** It is possible that the dataset will be updated if errors and new instances are found. These will be announced on the GitHub project.