Published in AJSE, Vol: 25, Issue:2
Received on 7th August 2026
Revised on 8th October 2026
Published on 31st October 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
**ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)**
https://doi.org/10.53799/0efxvm96

# A Context-Free Framework for Evaluating Morphological Generalization in Lexical POS Classification

Lilibeth P. Coronel, Emmylou A. Emperador, and Gergie A. Ambato

***Abstract*—Part-of-speech (POS) classification typically uses sentence context, making it difficult to isolate grammatical information available from lexical form alone. This study evaluates context-free lexical POS classification and morphological generalization using 370,100 English word types annotated with Penn Treebank-style POS labels. A majority-class baseline and four morphology-sensitive approaches—rule-based suffix, statistical suffix, character n-gram, and character-level BiLSTM models—were evaluated on coarse- and fine-grained classification tasks. Morphological transfer was examined separately through suffix hold-out experiments in which selected suffix families were excluded from training. Under standard stratified evaluation, the BiLSTM achieved the highest Macro-F1 in both coarse-grained (0.8323) and fine-grained (0.8412) classification. Pairwise randomization tests with Holm correction showed significant differences among the prespecified adjacent model comparisons. Hold-out performance, however, depended strongly on suffix-family composition. Near-perfect scores for *-tion* and *-ness* largely reflected the fact that both held-out families contained only NOUN instances, whereas the heterogeneous *-ly* and *-ing* families yielded substantially lower Macro-F1 scores. For these two families, the character n-gram model obtained the highest point estimates. Strong performance under standard lexical classification therefore did not necessarily correspond to transfer to unseen morphological families. Morphological generalization should be evaluated separately from standard classification and interpreted in relation to the grammatical composition of the held-out family.**

***Index Terms*— Lexical POS classification, Computational morphology, Morphological generalization, Character-level modeling, Suffix hold-out evaluation**

## I. Introduction

PART of speech (POS) tagging is a fundamental task in natural language processing (NLP) and supports downstream applications such as syntactic parsing, machine translation, information extraction, and question answering [1]. Traditionally, POS tagging has been formulated as a context-dependent sequence-labeling problem in which grammatical categories are inferred from lexical identity and surrounding words. Advances in statistical learning and deep neural networks have substantially improved tagging accuracy, with contemporary neural architectures achieving strong performance through contextual representations learned from large-scale corpora [2], [3], [4]. As a result, contextual information has become central to computational POS classification.

Less attention has been given to how much grammatical information can be inferred from lexical form alone. Linguistic theory has long recognized morphology as a source of grammatical information. Derivational and inflectional processes constrain lexical categories and can provide cues to grammatical interpretation even without sentence context [5]. Lexical morphology may therefore contribute directly to grammatical classification rather than simply complementing contextual information.

This contribution is difficult to evaluate under conventional experimental settings. Standard POS tagging benchmarks are designed for sequence labeling, where lexical, syntactic, semantic, and contextual information are available simultaneously during prediction [6]. High classification performance therefore does not necessarily show that a model has learned transferable morphological representations, because successful prediction may also depend on contextual dependencies or lexical regularities observed during training. The contribution of word-internal structure consequently remains difficult to isolate.

A more controlled evaluation requires lexical information to be separated from sentence context. Rather than treating POS prediction exclusively as a sequence-labeling task, grammatical classification can be formulated as a context-free lexical classification problem in which each input consists of a single isolated word. Restricting model input to morphological and orthographic characteristics allows the contribution of lexical structure to grammatical classification to be examined directly.

This study therefore treats context-free lexical POS classification and morphological-family transfer as related but distinct evaluation problems. Using 370,100 isolated English word types, the study compares a majority-class baseline with four morphology-sensitive models: rule-based suffix,

Lilibeth P. Coronel is with the Department of Information Technology, Mindanao State University at Naawan, Philippines (e-mail: lilibeth.coronel@msunaawan.edu.ph).

Emmylou A. Emperador is with the Department of Secondary and Elementary Education, Mindanao State University at Naawan, Philippines (e-mail: emmylou.emperador@msunaawan.edu.ph).

Gergie A. Ambato is with the Department of Social Sciences and Humanities, Mindanao State University at Naawan, Philippines (e-mail: gergie.ambato@msunaawan.edu.ph).

statistical suffix, character n-gram, and character-level BiLSTM classifiers. These models are evaluated under both coarse- and fine-grained classification settings. A suffix hold-out protocol is then used to exclude entire morphological families from training, providing a stricter test of whether predictive patterns learned from other lexical forms transfer to words belonging to an unseen suffix family.

The study makes four contributions. First, it provides a controlled framework for measuring POS information available from lexical form without sentence context. Second, it compares explicit morphological rules, learned suffix statistics, distributed character n-grams, and character-level neural representations under the same input constraints. Third, it distinguishes standard lexical classification from morphological-family transfer through suffix hold-out evaluation. Fourth, it incorporates bootstrap uncertainty estimates and paired randomization tests with Holm correction, allowing descriptive performance differences to be separated from statistically supported differences among models.

## II. Literature Review

### A. Evolution of POS Tagging

Part-of-speech tagging has traditionally been formulated as a context-dependent sequence-labeling task in which grammatical categories are assigned using lexical identity and contextual dependencies. Early probabilistic approaches, including Hidden Markov Models and Maximum Entropy Markov Models, modeled sequential word–tag relationships to estimate the most probable grammatical category [7]. Conditional Random Fields (CRFs) subsequently improved sequence prediction by optimizing label assignments globally while incorporating richer feature representations [8]. Although these approaches differed algorithmically, they shared the assumption that contextual information is important for resolving grammatical ambiguity.

Deep learning shifted POS tagging from manually engineered features to automatically learned representations. Contemporary neural models exploit sentence-level dependencies to encode grammatical regularities and support syntactic disambiguation [9]. As a result, less attention has been given to determining how much grammatical information can be inferred directly from lexical form.

### B. Computational Models for Morphological-Based Classification

Beyond sequence labeling, computational approaches have also used subword information for lexical analysis. Character-level representations can capture prefixes, suffixes, and orthographic patterns associated with grammatical categories without relying entirely on whole-word identity. Linear classifiers based on character *n*-grams provide representations of local subword patterns, whereas recurrent neural networks, particularly Bidirectional Long Short-Term Memory (BiLSTM) models, can learn distributed character-level representations and nonlinear relationships among subword features [10], [11], [12].

These approaches show that grammatical information can be recovered from lexical structure. However, many character-level models for POS prediction are evaluated within sequence-labeling frameworks in which contextual information remains available during prediction. Under such conditions, the contribution of morphological representations cannot be fully separated from information derived from context.

### C. Morphology as an Independent Source of Grammatical Information

Morphology provides systematic cues to grammatical category. Derivational and inflectional processes constrain lexical category membership and can provide indicators of syntactic function [5]. Experimental studies in psycholinguistics further show that speakers can infer grammatical categories from morphological structure even in the absence of contextual information [13], [14], [15]. These findings support the view that lexical form itself can provide grammatical information.

Computational studies likewise show that character-based neural models can encode aspects of morphological structure [16], [17], [18], [19], [20]. However, these findings are often obtained in settings where morphological information is evaluated together with lexical, semantic, or syntactic information. The independent contribution of morphology therefore remains difficult to isolate.

### D. Evaluation Gap

Existing POS research leaves two evaluation issues insufficiently separated. First, morphology-sensitive models are commonly assessed within contextual sequence-labeling settings, making it difficult to determine how much predictive information is available from lexical structure alone. Second, performance under conventional random train–test partitions does not establish whether a model can transfer learned morphological regularities to morphological families absent from training. A classifier may perform well because related lexical or subword patterns are represented across both training and test partitions.

The present study addresses these issues through two complementary evaluations. Standard context-free classification measures how effectively different models recover coarse- and fine-grained POS categories from isolated lexical forms. A separate suffix hold-out protocol evaluates transfer when complete suffix families are excluded from training. These evaluations are treated separately because lexical classification performance and morphological-family generalization do not necessarily measure the same model capability.

## III. Methodology

This study used a context-free lexical classification framework to assess how much grammatical category information can be inferred from the orthographic and morphological structure of isolated English words. Unlike conventional POS tagging, which uses surrounding words and sentence-level dependencies, each input in the present framework consisted of a single lexical form. The design therefore restricted the available evidence to information encoded within the word itself. In addition to stratified train–validation–test evaluation, a suffix hold-out protocol was used to examine model performance when selected morphological families were absent from training. Figure 1 summarizes the experimental framework.

### A. Dataset Source and Construction

The experiments used the 370k English Words Corpus distributed through Kaggle [21]. The resource contains English lexical forms with POS labels automatically assigned using the Natural Language Toolkit (NLTK). The resulting dataset comprised 370,100 word types annotated with Penn Treebank-style POS categories [22].

Before analysis, the dataset was audited for missing values, duplicate lexical forms, duplicate word–label pairs, and nonalphabetic entries. Word forms were stripped of surrounding whitespace and converted to lowercase. The audit identified no missing observations, duplicate lexical forms, duplicate word–label pairs, or nonalphabetic entries. All 370,100 normalized word forms therefore remained unique, and normalization did not reduce the dataset size.

Each lexical form in the source resource is associated with a single automatically assigned POS label. This single-label structure should not be interpreted as evidence that English word forms are grammatically unambiguous, since many words can assume different POS categories depending on syntactic context. Accordingly, the present study evaluates model predictions against the label supplied by the source resource and does not attempt to represent all possible contextual uses of each lexical item.

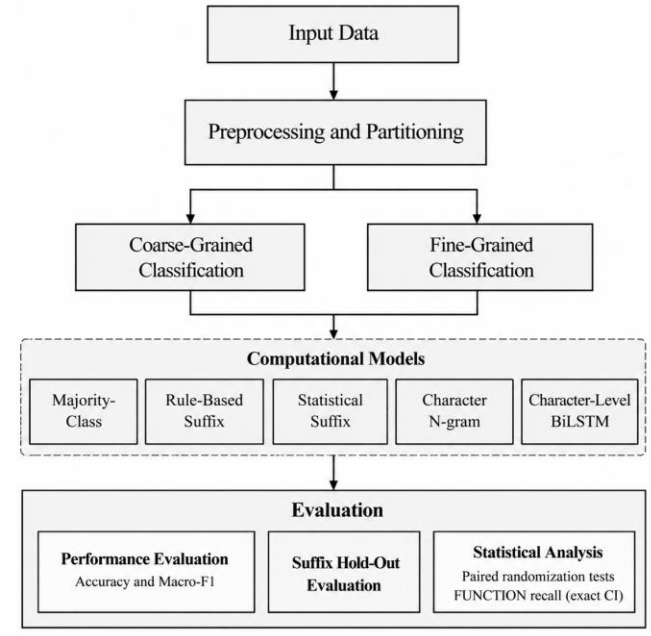


Fig. 1. Framework for Context-Free Lexical POS Classification and Morphological Evaluation.

### B. Experimental Tasks

Two standard lexical classification tasks and one suffix hold-out evaluation were defined.

#### 1) Coarse-Grained Classification

Penn Treebank-style POS tags were consolidated into five broad grammatical categories: NOUN, VERB, ADJ, ADV, and FUNCTION. Nominal tags were mapped to NOUN, verbal tags to VERB, adjectival tags to ADJ, and adverbial tags to ADV. Remaining closed-class categories, including determiners, pronouns, conjunctions, prepositions, and modal verbs, were grouped under FUNCTION. This representation reduced label sparsity while preserving broad grammatical distinctions relevant to morphology-based classification.

The resulting coarse-grained dataset contained 279,501 NOUN, 33,806 VERB, 39,572 ADJ, 17,038 ADV, and 183 FUNCTION instances. The substantial class imbalance, particularly the limited representation of FUNCTION, was considered in the selection of evaluation metrics and uncertainty estimates.

#### 2) Fine-Grained Classification

A second task retained the original Penn Treebank-style POS categories. To reduce instability associated with extremely sparse labels, only categories represented by at least 50 observations were retained. This criterion resulted in ten fine-grained categories: NN, NNS, JJ, RB, VBG, VBN, VB, VBD, JJS, and IN. The fine-grained task provided a more demanding classification setting because models were required to distinguish POS categories with potentially overlapping morphological forms.

#### 3) Suffix Hold-Out Evaluation

A separate suffix hold-out evaluation examined whether performance under conventional random partitioning transferred to morphological families absent from model training. Four suffix families were selected for the primary evaluation: *-tion*, *-ness*, *-ly*, and *-ing*.

Each lexical item was assigned to its longest matching suffix family. For each experiment, all words assigned to the target suffix family were reserved exclusively for testing. The remaining observations were divided into training and validation subsets using an 85:15 stratified split with random seed 42.

For the rule-based suffix model, the rule corresponding to the target suffix was explicitly removed during evaluation of that family. This Rule-Masked condition prevented the rule-based baseline from having direct access to the morphological pattern withheld from training. A separate full-rule condition that retained the target suffix rule was computed as a diagnostic reference but excluded from the primary model comparison because it incorporated explicit prior knowledge of the held-out pattern.

Macro-F1 for each suffix family was calculated only over grammatical classes represented in that family's ground-truth

test subset. Classes with zero ground-truth support were excluded rather than assigned an F1-score of zero. Family-level results were reported separately rather than averaged across suffix families because the families differed substantially in size and POS composition.

**Algorithm 1**. Suffix Hold-out Evaluation

**Input:** Dataset $D$, target suffix families $S$ = {-tion, -ness, -ly, -ing}
**Output:** Family-specific classification metrics
**Procedure:**
1. Normalize lexical forms and assign each word to its longest matching suffix family.
2. For each target suffix $s \in S$:
    2.1 Define all observations assigned to $s$ as the held-out test set.
    2.2 Define all remaining observations as the non-held-out pool.
    2.3 Stratify the non-held-out pool into 85% training and 15% validation data.
    2.4 Remove rule \(s\) from the Rule-Masked baseline.
    2.5 Train or construct each model using only the non-held-out training data.
    2.6 Select learned-model checkpoints using validation Macro-F1.
    2.7 Predict the POS labels of all observations in the held-out family.
    2.8 Determine the POS classes represented in the held-out ground truth.
    2.9 Compute Macro-F1 over represented classes only, together with accuracy.
3. Report results separately for each held-out suffix family.

### C. Data Preprocessing and Partitioning

Following the normalization and dataset audit described in Section III.A, the standard coarse- and fine-grained datasets were partitioned using stratified sampling into 70% training, 15% validation, and 15% test sets, with random seed 42. Because each normalized lexical form occurred only once in the source dataset, identical lexical forms could not occur across partitions, thereby preventing exact lexical overlap among the training, validation, and test sets.

For the suffix hold-out evaluation, all lexical forms belonging to the target suffix family were excluded from the training and validation sets and reserved exclusively for family-specific testing.

### D. Computational Models

The experiments included a majority-class baseline and four morphology-sensitive computational approaches: a rule-based suffix baseline, a statistical suffix model, a character n-gram linear model, and a character-level BiLSTM. All predictions were made from isolated word forms without sentence context.

#### 1) Majority-Class Baseline

A majority-class baseline assigned every test observation to the most frequent label in the corresponding training partition. This baseline provided a reference for interpreting accuracy under the highly imbalanced label distribution and for distinguishing model performance from that attributable primarily to majority-class prevalence.

#### 2) Rule-Based Suffix Baseline

The rule-based classifier used manually specified suffix-to-POS mappings representing common English morphological patterns. Examples included *-tion* → NOUN, *-ness* → NOUN, *-ly* → ADV, and *-ing* → VERB for the coarse-grained task. Fine-grained mappings included *-tion* → NN, *-ness* → NN, *-ly* → RB, and *-ing* → VBG. When multiple suffixes matched a word, the longest matching suffix was used. If no rule applied, the classifier returned the majority category of the corresponding training partition. The complete rule inventory was fixed before evaluation.

For suffix hold-out evaluation, the rule corresponding to the target suffix was removed from the rule inventory, producing the Rule-Masked baseline used in the primary comparison.

#### 3) Statistical Suffix Model

The statistical suffix model learned empirical suffix–label frequencies directly from the training partition. Suffixes of one to five characters were extracted, and predictions used the longest suffix observed during training that matched the test word. The predicted category was the label with the highest empirical frequency for the matched suffix. No smoothing was applied. When no learned suffix matched, prediction reverted to the majority class of the training partition.

Unlike the rule-based suffix baseline, the statistical suffix model received no manually specified mappings between suffixes and grammatical categories.

#### 4) Character N-gram Linear Model

The character n-gram linear model represented each lexical form using hashed character n-grams of lengths 2–5. Feature hashing used $2^{18}$, or 262,144, dimensions, with *alternate_sign*=False and L2 normalization. Classification was performed using stochastic gradient descent with logistic loss, L2 regularization ($\alpha=10^{-4}$), the optimal learning-rate schedule, balanced class weighting, and batches of 50,000 observations. Training was allowed for a maximum of 20 epochs, with validation Macro-F1 used for model selection and an early-stopping patience of three epochs.

The feature space contained 261,660 unique character n-grams, of which 165,593 occupied distinct hash bins. This resulted in 96,067 hash collisions, corresponding to an observed collision rate of 36.71%. Validation Macro-F1 was monitored across epochs and stabilized after the initial training iterations in both coarse- and fine-grained tasks. The corresponding convergence curves are available in the public reproducibility archive ( https://doi.org/10.5281/zenodo.23227900).

#### 5) Character-Level BiLSTM

The character-level BiLSTM represented each word as a sequence of characters. Words were encoded to a maximum length of 30 characters using 64-dimensional character embeddings. The BiLSTM contained 128 hidden units in each direction, and the final forward and backward hidden states were combined to form the lexical representation used for classification. Packed variable-length sequences were used so that padding did not contribute to the sequence representation.

The model was trained using class-weighted cross-entropy loss and the Adam optimizer with a fixed learning rate of $10^{-3}$ and a batch size of 256. Training was allowed for a maximum of 20 epochs with an early-stopping patience of three epochs. Training loss, validation loss, and validation Macro-F1 were monitored across epochs. The retained model corresponded to the checkpoint with the highest validation Macro-F1, while

early stopping limited further training after validation performance ceased to improve. The learning curves showed continued reduction in training loss, stabilization of validation Macro-F1, and increasing validation loss at later epochs, indicating the onset of overfitting. The corresponding convergence curves are available in the public reproducibility archive (https://doi.org/10.5281/zenodo.23227900).

### E. Evaluation Metrics and Statistical Analysis

Macro-F1 was used as the primary evaluation metric because the grammatical categories were substantially imbalanced and equal weighting of minority and majority classes was required. Accuracy was reported as a complementary measure. For the standard coarse- and fine-grained experiments, Macro-F1 was calculated over the corresponding task label set. For the suffix hold-out experiments, Macro-F1 was calculated only over POS categories represented in the ground-truth test subset for the evaluated family; categories with zero support were excluded.

Uncertainty in test-set performance was assessed using 2,000 bootstrap repetitions to obtain 95% confidence intervals. Because the FUNCTION category contained only 27 observations in the coarse-grained test partition, its recall was additionally reported with an exact Clopper–Pearson 95% confidence interval.

Differences in Macro-F1 between prespecified adjacent model pairs were evaluated using paired randomization tests with 2,000 repetitions. The comparisons were Rule-Based Suffix versus Statistical Suffix, Statistical Suffix versus Character N-gram, and Character N-gram versus BiLSTM. Raw $p$-values were adjusted using the Holm procedure to control for multiple comparisons, with statistical significance evaluated at $\alpha=0.05$.

### F. Reproducibility

The experimental configuration documented the preprocessing decisions, label mappings, model parameters, random seeds, evaluation procedures, statistical tests, and suffix hold-out definitions. The standard experiments used random seed 42. The software environment consisted of Python 3.9.12, NumPy 1.21.5, pandas 1.4.2, scikit-learn 1.1.3, and PyTorch 2.8.0 running on CPU.

The preprocessing scripts, model implementations, suffix definitions, experimental configurations, evaluation procedures, convergence figures, and reproducibility outputs are publicly available in the accompanying Zenodo archive (https://doi.org/10.5281/zenodo.23227900). The original lexical dataset is not redistributed and remains available from its source repository.

## IV. Results and Discussion

### A. Data Distribution

The lexical POS dataset exhibited pronounced class imbalance in both the coarse-grained and fine-grained classification tasks. The NOUN category dominated the coarse-grained dataset with 279,501 instances, whereas the FUNCTION category contained only 183 observations. A similar imbalance was observed in the fine-grained setting, where the NN tag accounted for 230,960 word types, while several Penn Treebank categories contained fewer than 100 observations.

Class imbalance affects classifier evaluation because overall accuracy can be dominated by performance on frequent classes. Under these conditions, a model may achieve high accuracy despite poor performance on minority categories. Macro-F1 was therefore selected as the primary evaluation metric because it assigns equal weight to each grammatical class regardless of frequency, while accuracy was retained as a complementary measure of overall classification performance.

### B. Coarse-Grained Lexical Classification

Under coarse-grained classification, the majority-class baseline achieved an accuracy of 0.7552 but a Macro-F1 of only 0.1721, reflecting the substantial label imbalance (Table 1). The BiLSTM produced the highest point estimates, with an accuracy of 0.9887 and a Macro-F1 of 0.8323. The statistical suffix model achieved a Macro-F1 of 0.7412, followed by the character n-gram model at 0.7199 and the rule-based suffix model at 0.6811. All morphology-sensitive models therefore exceeded the majority-class baseline in Macro-F1, with the BiLSTM achieving the highest standard-split performance.

TABLE I
MODEL PERFORMANCE ON COARSE-GRAINED POS CLASSIFICATION

| Model | Accuracy | Macro-F1 (95% CI) |
|---|---|---|
| Majority-class baseline | 0.7552 | 0.1721 [0.1717, 0.1726] |
| Rule-based suffix | 0.8985 | 0.6811 [0.6780, 0.6841] |
| Statistical suffix | 0.9496 | 0.7412 [0.7271, 0.7673] |
| Char N-gram | 0.9232 | 0.7199 [0.7070, 0.7350] |
| BiLSTM | 0.9887 | 0.8323 [0.8135, 0.8510] |

### C. Fine-Grained Lexical Classification

Fine-grained classification produced a similar model ordering but larger differences in Macro-F1 (Table 2). The majority-class baseline obtained a Macro-F1 of 0.0769. The rule-based suffix model reached 0.4011, followed by the character n-gram model at 0.6162 and the statistical suffix model at 0.6516. The BiLSTM again achieved the highest point estimates, with an accuracy of 0.9795 and a Macro-F1 of 0.8412.

TABLE II
MODEL PERFORMANCE ON FINE-GRAINED POS CLASSIFICATION

| Model | Accuracy | Macro-F1 (95% CI) |
|---|---|---|
| Majority-class baseline | 0.6243 | 0.0769 [0.0766, 0.0772] |
| Rule-based suffix | 0.7331 | 0.4011 [0.3966, 0.4054] |
| Statistical suffix | 0.9437 | 0.6516 [0.6306, 0.6761] |
| Char N-gram | 0.8593 | 0.6162 [0.6057, 0.6265] |
| BiLSTM | 0.9795 | 0.8412 [0.8267, 0.8543] |

### D. Statistical Uncertainty and Pairwise Model Comparison

The 95% bootstrap confidence intervals for Macro-F1 are reported in Tables 3 and 4 and quantify the uncertainty

associated with the standard test-set performance estimates. Because the FUNCTION category contained only 27 observations in the coarse-grained test set, its recall was examined separately using exact 95% confidence intervals. The majority-class and rule-based suffix baselines did not correctly classify any FUNCTION instances. The statistical suffix, character n-gram, and BiLSTM models achieved FUNCTION recalls of 0.037 (95% CI [0.001, 0.190]), 0.222 (95% CI [0.086, 0.423]), and 0.593 (95% CI [0.388, 0.776]), respectively. The wide confidence intervals reflect the small class support and indicate that FUNCTION-specific estimates should be interpreted cautiously.

TABLE III
COARSE-GRAINED PAIRED SIGNIFICANCE TESTS

| Models | Δ Macro-F1 | *p*_Raw | *p*_Holm |
|---|---|---|---|
| Rule-based suffix–Statistical suffix | 0.0601 | 0.0005 | 0.0015 |
| Statistical suffix – Char n-gram | -0.0213 | 0.04998 | 0.04998 |
| Char n-gram – BiLSTM | 0.1124 | 0.0005 | 0.0015 |

TABLE IV
FINE-GRAINED PAIRED SIGNIFICANCE TESTS

| Models | Δ Macro-F1 | *p*_Raw | *p*_Holm |
|---|---|---|---|
| Rule-based suffix – Statistical suffix | 0.2505 | 0.0005 | 0.0015 |
| Statistical suffix – Char n-gram | -0.0355 | 0.0005 | 0.0015 |
| Char n-gram – BiLSTM | 0.2250 | 0.0005 | 0.0015 |

Paired randomization tests were used to evaluate differences in Macro-F1 between prespecified adjacent model pairs, with Holm correction applied for multiple comparisons. For the coarse-grained task (Table 3), the statistical suffix model exceeded the rule-based suffix baseline by 0.0601 Macro-F1 (Holm-adjusted $p$ = .0015) and the character n-gram model by 0.0213 (adjusted $p$ = .04998). The BiLSTM exceeded the character n-gram model by 0.1124 (adjusted $p$ = .0015). For the fine-grained task (Table 4), the statistical suffix model exceeded the rule-based suffix baseline by 0.2505 Macro-F1 (adjusted $p$ = .0015) and the character n-gram model by 0.0355 (adjusted $p$ = .0015), while the BiLSTM exceeded the character n-gram model by 0.2250 (adjusted $p$ = .0015). All prespecified adjacent comparisons remained statistically significant after Holm correction at $\alpha$ = .05, although the coarse-grained statistical suffix–character n-gram comparison was close to the significance threshold.

### *E. Morphological Generalization under Suffix Hold-Out Evaluation*

Figure 2 and Table 5 present model performance when each target suffix family was excluded from training. Performance varied substantially across suffix families. For *-tion* and *-ness*, the majority-class, Rule-Masked, and statistical suffix models each obtained a Macro-F1 of 1.0000, while the BiLSTM achieved 0.9999 for both families. The character n-gram model obtained 0.9731 for *-tion* and 0.7523 for *-ness*. These high values should not be interpreted as direct evidence of morphological generalization because both held-out families contained only NOUN instances. High performance could therefore be achieved by predicting the single grammatical category represented in the test subset without demonstrating transfer of the withheld suffix pattern.

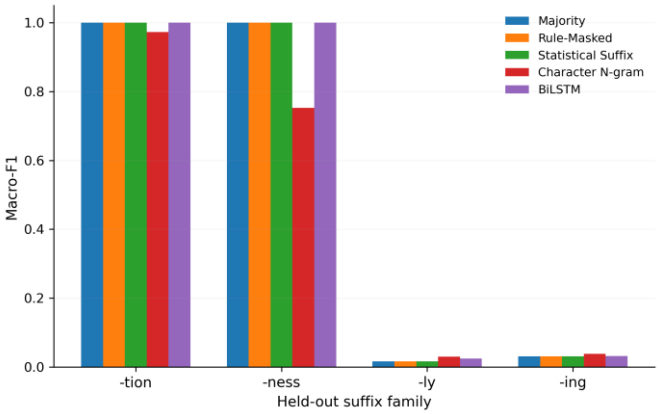

Fig. 2. Macro-F1 performance across held-out suffix families.

A different pattern emerged for *-ly* and *-ing*, both of which contained multiple POS categories. Under the *-ly* hold-out, the character n-gram model obtained the highest Macro-F1 point estimate (0.0305), followed by the BiLSTM (0.0247), whereas the majority-class, Rule-Masked, and statistical suffix models each obtained 0.0162. Under the *-ing* hold-out, the character n-gram model again achieved the highest Macro-F1 point estimate (0.0382), followed by the BiLSTM (0.0322), while the remaining models each obtained 0.0311. Thus, the strong BiLSTM performance under the standard random split did not

TABLE V
PERFORMANCE UNDER SUFFIX HOLD-OUT EVALUATION

| Held-out suffix | Test *n* | Classes present | Majority-Class | Rule-Masked | Statistical Suffix | Char N-gram | BiLSTM |
|---|---|---|---|---|---|---|---|
| *-tion* | 7,245 | NOUN | 1.0000 | 1.0000 | 1.0000 | 0.9731 | 0.9999 |
| *-ness* | 9,564 | NOUN | 1.0000 | 1.0000 | 1.0000 | 0.7523 | 0.9999 |
| *-ly* | 16,689 | ADJ, ADV, NOUN, VERB | 0.0162 | 0.0162 | 0.0162 | 0.0305 | 0.0247 |
| *-ing* | 18,119 | ADJ, FUNCTION, NOUN, VERB | 0.0311 | 0.0311 | 0.0311 | 0.0382 | 0.0322 |

*Note: Macro-F1 includes only classes present in each held-out test set. The corresponding suffix rule was excluded from the Rule-Masked baseline.*

correspond to the highest Macro-F1 for the heterogeneous *-ly* and *-ing* hold-outs.

The contrast among suffix families was associated with their grammatical composition. The *-ly* test subset was dominated by ADV instances but also contained NOUN, ADJ, and VERB forms, whereas the *-ing* subset was dominated by VERB instances with smaller numbers of NOUN, ADJ, and FUNCTION forms. Because Macro-F1 assigns equal weight to each represented class, performance on sparsely represented categories can substantially affect the family-level score. The low Macro-F1 values for *-ly* and *-ing* should therefore be interpreted in relation to their uneven within-family class distributions rather than as a simple indication of overall classification failure.

The hold-out results show that performance under a standard lexical split does not necessarily transfer to an unseen morphological family. The near-perfect results for *-tion* and *-ness* were largely attributable to their homogeneous NOUN composition, whereas the heterogeneous *-ly* and *-ing* families provided a more demanding test of transfer across POS categories. Within these heterogeneous families, the character n-gram model produced the highest Macro-F1 point estimates.

### F. Discussion and Implications

The standard and suffix hold-out experiments measured different aspects of model behavior. Under random lexical partitioning, the BiLSTM achieved the highest coarse- and fine-grained performance, indicating that character sequences contained substantial information about the supplied POS labels. The statistical suffix model also performed strongly, showing that empirical suffix–label distributions accounted for a substantial proportion of lexical predictability.

These results do not by themselves establish morphological generalization. When entire suffix families were excluded from training, performance depended strongly on the grammatical composition of the held-out family. The near-perfect scores for *-tion* and *-ness* do not provide clear evidence of suffix transfer because both families contained only NOUN instances. In contrast, *-ly* and *-ing* contained multiple POS categories and yielded very low Macro-F1 values across models. The character n-gram model, rather than the BiLSTM, produced the highest point estimates for both heterogeneous families.

The contrast shows that high performance under a standard lexical split can coexist with weak transfer under a morphology-based distribution shift. Suffix-family composition is therefore part of the evaluation problem: a homogeneous held-out family permits high classification performance without demonstrating recovery of the withheld morphological pattern, whereas a heterogeneous family requires the model to distinguish grammatical categories within the unseen family. Claims about morphological generalization should therefore be based on controlled hold-out evaluation rather than standard classification performance alone.

### G. Limitations

This study has several limitations. First, it evaluates context-free lexical POS classification rather than conventional contextual POS tagging. The source dataset assigns a single automatically generated POS label to each lexical form, although English words may assume different grammatical categories depending on context. Substantial class imbalance, particularly among sparsely represented POS categories, may also affect the stability of performance estimates.

Second, the suffix hold-out evaluation was limited to four English suffix families with markedly different POS compositions. The results therefore represent a specific test of morphological transfer and should not be generalized to all morphological patterns or languages. Future work may extend the framework to additional and more balanced morphological families, multilingual datasets, and lexical resources that explicitly represent grammatical ambiguity.

## V. Conclusion

This study evaluated the grammatical information available from isolated English word forms and distinguished standard lexical POS classification from transfer to unseen morphological families. Under stratified coarse- and fine-grained evaluation, the character-level BiLSTM achieved the strongest performance, while the statistical suffix model showed that learned suffix–POS associations accounted for substantial predictive information.

The suffix hold-out experiments produced a different pattern. Near-perfect performance on *-tion* and *-ness* largely reflected their homogeneous NOUN composition, whereas the heterogeneous *-ly* and *-ing* families yielded substantially lower Macro-F1 scores. Standard classification performance therefore should not be interpreted as direct evidence of morphological-family generalization.

Context-free lexical classification and suffix hold-out evaluation provided complementary evidence about model behavior. The former measured how effectively grammatical labels were recovered from word-internal information within the observed lexical distribution, whereas the latter tested whether performance transferred when a morphological family was absent from training. Treating these evaluations separately provides a clearer basis for assessing morphological generalization.

## Acknowledgment

The authors thank Mindanao State University at Naawan for the support provided during the conduct of this study.

## Author Contributions

Lilibeth Coronel contributed to the conceptualization, methodology, software, investigation, formal analysis, data curation, visualization, validation, supervision, and preparation of the original manuscript draft. Emmylou Emperador and Gergie Ambato contributed to validation, resources, and review and editing of the manuscript. All

authors reviewed and approved the final manuscript and accepted responsibility for its content.

## AI Tool Usage Acknowledgment

The authors used ChatGPT (OpenAI) to assist with manuscript organization and language refinement. All scientific content, analyses, interpretations, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

## References


[1] Manning, C., et al. *The Stanford CoreNLP Natural Language Processing Toolkit*. 2014. Baltimore, Maryland: Association for Computational Linguistics.

[2] Hou, L. and D. Smith. *Detecting Syntactic Change with Pre-trained Transformer Models*. 2023. Singapore: Association for Computational Linguistics.

[3] Ion, R., et al. *A Cross-model Study on Learning Romanian Parts of Speech with Transformer Models*. 2024. Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

[4] Qi, Y., S. Ali, and A. Murat, *Symmetric Affix–Context Co-Attention: A Dual-Gating Framework for Robust POS Tagging in Low-Resource MRLs.* Symmetry, 2025. 17(9): p. 1561.

[5] Lieber, R., *Introducing morphology*. 2021: Cambridge University Press.

[6] Rogers, A., O. Kovaleva, and A. Rumshisky, *A Primer in BERTology: What We Know About How BERT Works.* Transactions of the Association for Computational Linguistics, 2020. 8: p. 842-866.

[7] Ratnaparkhi, A. *A Maximum Entropy Model for Part-Of-Speech Tagging*. 1996.

[8] Lafferty, J.D., A. McCallum, and F.C.N. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, in *Proceedings of the Eighteenth International Conference on Machine Learning*. 2001, Morgan Kaufmann Publishers Inc. p. 282–289.

[9] Pathak, D., S. Nandi, and P. Sarmah. *AsPOS: Assamese Part of Speech Tagger using Deep Learning Approach*. in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*. 2022.

[10] Hochreiter, S. and J. Schmidhuber, *Long Short-Term Memory.* Neural Computation, 1997. 9(8): p. 1735-1780.

[11] Huang, Z., W. Xu, and K. Yu, *Bidirectional LSTM-CRF Models for Sequence Tagging.* ArXiv, 2015. abs/1508.01991.

[12] Ma, X. and E. Hovy. *End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF*. 2016. Berlin, Germany: Association for Computational Linguistics.

[13] Zhu, L., et al., *Learning morphology from cross-situational statistics.* Studies in Second Language Acquisition, 2025. 47(5): p. 1250-1276.

[14] Manouilidou, C., *Experimental approaches to morphology.* University of Ljubljana Publishing House, 2025.

[15] Lago, S., E. Oltrogge, and K. Stone, *Does productive agreement morphology increase sensitivity to agreement in a second language?* Glossa Psycholinguistics, 2025.

[16] López-Otal, M., et al., *Linguistic Interpretability of Transformer-based Language Models: a systematic review.* ArXiv.org, 2025.

[17] Someya, T., et al. *Derivational Probing: Unveiling the Layer-wise Derivation of Syntactic Structures in Neural Language Models*. 2025. Vienna, Austria: Association for Computational Linguistics.

[18] Aduah, W. and F. Meyer. *Designing and Contextualising Probes for African Languages*. 2025. Vienna, Austria: Association for Computational Linguistics.

[19] Aljaafari, N., D.S. Carvalho, and A. Freitas, *TRACE for Tracking the Emergence of Semantic Representations in Transformers.* arXiv.org, 2025.

[20] Millière, R., *Language Models as Models of Language.* arXiv (Cornell University), 2024.

[21] Bhatia, R. *370k English words corpus*. 2021; Available: https://www.kaggle.com/datasets/ruchi798/part-of-speech-tagging.

[22] Marcus, M.P., B. Santorini, and M.A. Marcinkiewicz, *Building a Large Annotated Corpus of English: The Penn Treebank.* 1993. Computational Linguistics, 19(2): p. 313-330.

**Lilibeth Coronel** is a Professor in the Department of Information Technology, Mindanao State University at Naawan, Philippines. She holds a Doctor of Information Technology. Her research interests include image processing, deep learning, natural language processing, and bioinformatics.

**Emmylou Emperador** is an Associate Professor in the Department of Secondary and Elementary Education, Mindanao State University at Naawan, Philippines. She holds a Doctor of Philosophy in Language Studies. Her research interests include English language studies, applied linguistics, language research, professional education, and educational research.

**Gergie Ambato** is an Associate Professor in the Department of Social Sciences and Humanities, Mindanao State University at Naawan, Philippines. She holds a Master of Arts in English Language Studies. Her research interests include English language studies, applied linguistics, discourse analysis, and research in language education.