Structured Partial Predictability in Non-Concatenative Morphology: The Case of Tashlhiyt Berber

John Alderete, Hamza Sellami
Simon Fraser University, Canada & Mohammed V University, Rabat, Morocco
alderete@sfu.ca, hamza-sellami@um5r.ac.ma

Abstract

Non-concatenative morphology poses a persistent challenge for NLP, yet structured quantitative resources for Amazigh (Berber) languages remain scarce. We present the first large-scale computational study of Tashlhiyt Berber plural formation, drawing on a richly annotated dataset of 1,185 noun paradigms with phonological, morphological and semantic features. We decompose the plural system into macro-level word-formation strategies and micro-level stem mutations, and evaluate predictability across ten target domains using linguistic feature models, N-gram baselines, and Bi-LSTM neural models. Results reveal a structured split: linguistic features decisively outperform neural models on systematic macro-level strategies (e.g., +44.5pp F_1), while Bi-LSTMs better capture lexically idiosyncratic patterns. Rather than supporting a categorical rule/memory divide, this complementarity reveals gradient layers of regularity within a single morphological system. These findings demonstrate the value of linguistically informed annotation for probing morphological complexity in low-resource, typologically diverse languages. All data, code, and models are publicly available.

Keywords: non-concatenative morphology, Tashlhiyt Berber, Amazigh languages, structured linguistic annotation, morphological predictability, low-resource languages, grammar-memory interface

1. Introduction

While morphological models perform well on concatenative systems, non-concatenative morphology, which requires encoding grammar through internal stem modification, remains a major NLP challenge. Afroasiatic morphology utilizes complex vocalic ablaut, gemination, and templatic processes (McCarthy, 1979). Despite neural dominance in benchmarks (Cotterell et al., 2016; Vylomova et al., 2020), black-box architectures often obscure the grammar-memory interface: the degree to which forms are productively computed versus lexically stored (Kirov and Cotterell, 2018). Quantifying this predictability is essential for optimizing model architectures and probing morphological complexity in typologically diverse languages (Cotterell et al., 2019).

Amazigh languages offer a natural testing ground for these questions, and computational linguistics for these languages has seen meaningful progress over the past two decades, driven in part by IRCAM's standardization efforts[1] and a growing ecosystem of NLP resources and tools. This includes work on computer-assisted linguistic analysis (Jebbour, 1996; Taghbalout et al., 2015), automatic language detection (Adouane et al., 2016b,a; Lafkioui, 2008), machine translation (Taghbalout et al., 2015), annotated dataset development (Amri et al., 2017), morphological analysis tools (Raiss and Cavalli-Sforza, 2012), and deep learning approaches to sequence labeling tasks such as POS tagging (Bani et al., 2023).

Nevertheless, while research on Arabic has quantified partial predictability and templatic effects in plural formation (Dawdy-Hesterberg and Pierrehumbert, 2014), no comparable analysis exists for Amazigh varieties. Specifically, the question of where rule-governed grammar ends and lexically idiosyncratic memory begins has not been computationally investigated. Tashlhiyt, an Amazigh (Berber) language spoken by 7-8 million people in southern Morocco, provides an ideal test case, combining external suffixation and internal mutation in its plural system. Bridging this gap is critical for developing low-resource morphological tools (Khalifa et al., 2020) and for testing whether Afroasiatic findings generalize or reflect language-specific idiosyncrasies (cf. McCurdy et al. (2020) on parallel questions in German).

We address this complexity via a multi-layered framework, decomposing plurals into macro-level strategies (i.e., suffixation vs. stem mutations) and micro-level stem mutations (e.g., vowel ablaut and insertion), and evaluating predictability across ten target domains using linguistic feature models, N-gram baselines, and Bi-LSTM neural models.[2] We find a structured split: hand-crafted features decisively outperform Bi-LSTM baselines on macro-level tasks (e.g., +44pp F_1 on 3-way classification), while the Bi-LSTM better captures specific micro-level domains where explicit structural generalizations fail. Rather than reflecting a categor-

[1] https://tal.ircam.ma/talam/

[2] All data, code, and trained models are available at: github.com/aldo-git-bit/predicting-tashlhiyt-plural.

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 124–135
May 11, 2026.

ical rule/memory divide, this complementarity reveals gradient layers of regularity within a single morphological system — consistent with accounts that treat morphological productivity as gradient rather than categorical (Bybee and McClelland, 2005; Hay and Baayen, 2005; O'Donnell, 2015). High irreducible error alongside low idiosyncrasy severity further suggests that complexity stems from competition among overlapping regularities rather than arbitrary lexical listing. These results demonstrate that structured linguistic annotation provides a diagnostic tool for probing morphological complexity, with implications for hybrid modeling in low-resource, non-concatenative systems.

2. Related Research

Prior work on Amazigh plurals distinguishes three classes: external (primarily /-n/ suffixation), internal (vowel changes and insertion), and mixed (Basset, 1952; Saib, 1986; Jebbour, 1988; Lasri, 1991; Taine-Cheikh, 2006; Hasselbach, 2007; Ben Si Saïd, 2014, 2020). Several analyses argue that surface markers such as vowel insertion are phonological consequences of affixation rather than independent morphemes (Jebbour, 1996; Lasri, 1991), and Idrissi (2000) proposes that prosodic structure conditions plural class selection. Plural assignment is generally characterized as partially systematic, shaped by the interaction of prosodic and morphological factors with lexical idiosyncrasy, but no prior work has quantitatively evaluated the predictability of plural class from singular forms. We address this gap computationally.

3. Dataset

3.1. Data Source

We use 1,185 Tashlhiyt noun paradigms from (Alderete et al., 2025), which draws on and integrates noun paradigms from a comprehensive grammatical study (Jebbour, 1996) and a Tashlhiyt text corpus (Jebbour et al., 2021). Nouns follow a hierarchical schema: three macro-classes—External (primarily suffixation of /-n/), Internal (stem-internal vowel changes such as ablaut and vowel insertion), and Mixed (combining suffixation with stem modifications)—and specific micro-level patterns (Table 1)[3]. Predictive features cover morphology (gender, derivation, paradigm structure, stem augments, and loan status) and semantics (animacy, humanness, and a 22-category semantic field based on Buck (2008)). Full feature labels and vocabulary sizes are provided in Appendix A.

Consider the following concrete examples to illustrate the problem we are investigating. The singular nouns for 'speech' and 'finger' in Table 1 have the same consonant-vowel pattern in their stems, /wal/ and /dˤadˤ/, respectively. Thus, while the inputs to plural formation both have a CVC pattern, they have two very different outcomes: /wal/ receives a final vowel-glide sequence in *i-waliw-n*, while /dˤadˤ/ undergoes both vowel substitution and final vowel insertion in *i-dˤudˤa-n*. These micro-level differences are compounded by macro-level divergences in word-formation strategy. The CCVC stem /snus/ exhibits stem-internal ablaut: *i-snas* 'donkey', but the comparable stem /fraw/ for 'sheet' has no internal changes and instead receives a simple suffix in the plural: *i-fraw-n*. Our goal is to computationally quantify the factors that govern these differences, distinguishing rule-governed regularities from lexical idiosyncrasy.

Table 1: Plural formation patterns in Tashlhiyt

Pattern	n **(%)**	**Example (Sg. → Pl.)**
Macro-patterns (1,185 total)		
External	623 (52.8%)	*a-ʃddir* → *i-ʃddir-n* 'bush'
Internal	357 (30.3%)	*a-sds* → *i-sdas* 'trough'
Mixed	200 (17.0%)	*a-dmr* → *i-dmar-n* 'chest'
Micro-patterns (Only Internal and Mixed, 562 total)		
Medial A	63 (11.3%)	*a-sgdl* → *i-sgdal* 'block'
Final A	57 (10.2%)	*a-llun* → *i-lluna* 'tambourine'
Final Vw	27 (4.9%)	*a-wal* → *i-waliw-n* 'speech'
Final T	38 (6.8%)	*a-qaʃʃa* → *i-qaʃʃat-n* 'crest'
Ablaut	226 (40.6%)	*a-snus* → *i-snas* 'donkey'
Abl+Med.A	30 (5.4%)	*a-satm* → *i-sutam* 'niche'
Abl+Fin.A	14 (2.5%)	*a-dˤadˤ* → *i-dˤudˤa-n* 'finger'
Templatic	102 (18.3%)	*lkbbutˤ* → *lkbabtˤ* 'coat'

3.2. Phonological N-grams

To provide hand-crafted features and a baseline, we extracted edge-aligned phonological n-grams (i.e., contiguous sequences of 1–3 IPA segments anchored to word boundaries) from singular stems, yielding 2,265 macro-level and 1,356 micro-level features. We reduced dimensionality using LASSO with Stability Selection (100 bootstrap iterations, threshold ≥ 0.50) to define a static feature space of 2,019 and 1,149 features, respectively. Critically, while this procedure characterized the feature space on the full dataset, all model weights were learned strictly within training folds during 10-fold cross-validation to prevent data leakage (see Appendix B for details).

Per-task feature selection isolated informative n-gram subsets ranging from broad coverage for complex Ablaut (1,106 features) to compact signatures for Templatic patterns (92 features). This

[3] Ablaut = vowel quality change in the stem (e.g., /u/ → *a*); Medial A = insertion of *a* in a stem-medial position; Final A = insertion of *a* in final position; Final T and Final Vw = insertions of *t* and vowel+glide; Abl.Med.A denotes the cooccurrence of Ablaut and Medial A; likewise Abl.Fin.A is Ablaut + Final A; Templatic = the imposition of a fixed CV pattern on the stem

optimization filtered noise and prioritized informative, rare n-grams, significantly improving generalization; for example, F_1 for templatic mutations rose from 0.344 to 0.718 (a gain of 37.4pp).

3.3. Prosodic Features

We syllabified stems using the sonority-based system of Dell and Elmedlaoui (2002), parsing results into L/H syllables and metrical feet. To reduce dimensionality by 70% and enhance interpretability, raw strings were replaced with nine theory-driven features. For instance, `p_LH_ends_L` (final light syllables) identifies stems 7.22$\times$ more likely to undergo Medial A mutation ($+23.6$% effect), confirming that insertion satisfies prosodic weight requirements. Other features quantify moras, feet, and metrical residue. This setup ensures each mutation type is associated with $\geq$3 significant predictors ($p < 0.05$).

4. Computational Experiments

4.1. Multi-Level Prediction and Ablation

We implement a multi-level framework comprising ten independent experiments to address morphological complexity. Pluralization is decomposed into two tiers: a **macro-level** ($N = 1,185$) targeting broad strategies and a **micro-level** ($N = 562$) targeting specific mutation patterns (e.g., Ablaut, Templatic). This tiered approach isolates distinct dimensions of change, allowing for tailored feature selection and a granular mapping of the grammar-memory interface. Systematic ablation of morphological, semantic, and phonological subsets identifies core predictors and evaluates feature redundancy across all domains. These studies quantify the specific linguistic drivers of pluralization while facilitating critical baseline comparisons, such as probing the predictive value of abstract prosodic features over surface n-grams.

4.2. Experimental Models

To prioritize interpretability, we utilize three architectures: Logistic Regression (L_2-regularized), Random Forest, and Gradient Boosting (XGBoost). Our protocol emphasizes linguistic transparency over metric maximization; we therefore eschew exhaustive hyperparameter tuning in favor of fixed, conservative configurations across all domains (e.g., Logistic Regression $C = 1.0$; see Appendix B). By maintaining identical parameters across feature subsets, observed performance deltas reflect the informational quality of hand-crafted features rather than optimization artifacts.[4] This framework provides a consistent baseline for quantifying the relative contributions of morphological, phonological, and semantic features to plural formation.

4.3. Baseline Models

We utilize two baselines. An N-gram baseline uses Logistic Regression on edge-aligned phoneme sequences as a distributional superset. A bidirectional LSTM (Bi-LSTM) provides a neural ceiling for models trained from scratch at this data scale, capturing dependencies from raw characters without linguistic guidance. This architecture is deliberately matched to the data regime. Our task, that of classifying plural strategy from isolated singular forms, differs fundamentally from the seq2seq morphological inflection benchmarks like the SIGMORPHON shared tasks (Cotterell et al., 2016; Vylomova et al., 2020), which provide explicit input–output pairs and benefit from character-level copying. With 1,185 paradigms for 3- or 8-way classification, training from scratch is the only regime that avoids confounding the comparison with external pre-training data. Furthermore, we favor character-level probes because powerful pretrained models like mBERT lack sufficient Amazigh representation. In sum, these two baselines isolate the gain of our features: N-grams quantify the value of prosodic abstractions over surface phonotactics, while the Bi-LSTM identifies the idiosyncrasy gap that remains irreducible through sequence learning alone.

The Bi-LSTM was trained from scratch on the same 1,185 paradigms. Each singular form was represented as a sequence of IPA characters mapped to 32-dimensional embeddings. The model comprised a single bidirectional LSTM layer (64 hidden units per direction) with a fully connected classification head, trained with Adam, early stopping (patience = 10), and dropout of 0.3.

4.4. Data Imbalance and Over-fitting

To mitigate overfitting, we employ stratified 10-fold cross-validation. We address significant class imbalance (up to 19.8:1) through a combination of class weighting and SMOTE (Synthetic Minority Oversampling Technique; Chawla et al. (2002)) applied to training folds. Our primary evaluation metric is Macro-F_1, applied **uniformly** across all binary and multiclass domains. By equalizing class contributions—averaging F_1 scores across both categories in binary tasks—we ensure that performance on rare mutation patterns is not obscured by majority-class heuristics. This provides a more robust assessment of model quality than accuracy or positive-class-only metrics.

[4] A sensitivity check across $C \in 0.1, 1.0, 10.0$ confirmed that feature-set rankings were preserved in 9/10 domains at all three values (mean Macro-F_1 variation: 0.03), indicating that observed performance deltas reflect feature informativeness rather than the specific choice of C=1.0.

4.5. Residual Lexical Idiosyncrasy

To investigate the grammar-memory tradeoff, we quantify residual lexical idiosyncrasy, or the proportion of plural assignment resisting systematic prediction. The **computational ceiling** is the maximum performance of either model, representing a relative upper bound constrained by our architectures rather than an absolute human limit. **Irreducible error** is the percentage of forms where both models fail, while the **learnable proportion** marks the upper bound of predictable variance. To evaluate synergy between distributional and featural learning, we measure **model complementarity** as the ratio of non-overlapping errors to the learnable proportion. Finally, **idiosyncrasy severity** isolates high-confidence joint failures (forms where both models assign probability ≥ 0.70 to their predicted class yet misclassify), distinguishing learnable regularities from the irreducible, arbitrary core of the system.

5. Results

Within the hierarchical decomposition of the plural system, our results demonstrate an architecture characterized by structured partial predictability, where the divide between grammatical rules and lexical memory is clearly reflected in model performance across different tiers of analysis.

5.1. Macro-Level Systematicity

Macro-level pluralization is highly systematic and governed by rule-based dependencies. Hand-crafted features decisively outperform neural sequence learning, exceeding the Bi-LSTM by $+0.445$ in 3-way classification and $+0.378$ for stem mutations (Table 2). Paired t-tests confirm M+P significantly outperforms the Bi-LSTM across all macro-level tasks ($p < 0.001$). This performance gap suggests that macro-level morphological choices rely on abstract categories like gender or phonological patterns. While interpretable features effectively capture these rules, our Bi-LSTM baseline, which is trained from scratch on 1,185 paradigms, does not induce the abstract regularities that hand-crafted features encode directly.[5] This mirrors findings in Arabic, where coarse-grained phonological representations capture generalizations that surface distributional patterns miss (Dawdy-Hesterberg and Pierrehumbert, 2014).

[5]The only published Bi-LSTM result for Amazigh (Bani et al. (2023): 97% accuracy) addresses POS tagging on 60k tokens of running text with rich sentential context — a fundamentally different task from predicting plural strategy class from isolated singular forms. No prior work applies neural models to Tashlhiyt plural strategy classification, and no Amazigh variety appears in the SIGMORPHON shared tasks (2016–2024).

5.2. Phonological Primacy

Phonological features emerge as the primary predictive driver across nearly all domains. Although the combined M+P model often achieves the highest Macro-F_1, morphological features provide only marginal refinements to a predominantly phonological signal. While the N-gram baseline is competitive, hand-crafted features provide a consistent boost, particularly in multiclass tasks like 3-way macro-classification ($\Delta\text{Ngr} = +0.068$). By encoding explicit prosodic constraints, these features demonstrate that Tashlhiyt pluralization is sensitive to metrical structure rather than surface phonotactics alone, suggesting that morphological modeling for low-resource templatic languages should consider engineering prosodic abstractions.

5.3. Micro-Level Variation

At the micro-level, performance is more balanced between grammar and memory. Systematic patterns like Templatic and Ablaut plurals demonstrate high predictability (*Ceil* $= 0.907$ and 0.761, respectively), mirroring macro-level systematicity. In imbalanced domains (e.g., Medial A, Final A), hand-crafted features effectively capture rare mutations when evaluated under Macro-F_1 weighting. Three micro-level tasks (Medial A, Ablaut, Templatic) show no significant difference between M+P and the Bi-LSTM, suggesting these patterns are equally learnable via explicit prosodic features or distributional sequence learning. Ultimately, specific phonological patterns, rather than data volume, appear to be the primary driver of micro-level learnability.

5.4. Quantifying the Idiosyncrasy Gap

Our residual analysis provides a conservative lower bound for the memory component of the system. While Irreducible Error remains high in multiclass tasks (27.7% for 8-way mutation), idiosyncrasy severity—the proportion of high-confidence misclassifications—remains below 5% in most domains. This suggests that Tashlhiyt plural formation is not dominated by true lexical exceptions. Instead, the "unpredictability" often cited in traditional grammars likely reflects model uncertainty between competing, phonologically plausible patterns rather than arbitrary lexical listing. The high complementarity observed in domains like Ablaut (0.520) further suggests that distributional memory and grammatical rules provide non-redundant signals, supporting the use of hybrid architectures for non-concatenative morphological analysis.

		Hand-Crafted Feature Models					Baseline Comparisons				Residual Analysis			
	Domain	Sem	Morph	Phon	M+P	All	Ngr	LSTM	△Ngr	△LSTM	Ceil	IrrErr%	Compl	Sev%
Macro	Has Suffix	0.500	0.541	0.785	**0.797**	0.781	0.760	0.735	0.036	0.062	0.797	12.2	0.421	3.8
	Has Mutation	0.566	0.567	0.771	**0.777**	0.769	0.754	0.399	0.023	0.378	0.777	8.6	0.524	0.0
	3-way	0.402	0.374	0.658	**0.683**	0.666	0.614	0.238	0.068	0.445	0.683	17.6	0.462	0.0
Micro	Medial A	0.495	0.503	0.670	**0.715**	0.688	0.678	0.673	0.037	0.042	0.715	7.3	0.324	2.7
	Final A	0.464	0.476	0.457	0.593	**0.617**	0.660	0.458	−0.067	0.135	0.593	7.4	0.260	3.3
	Final Vw	0.522	0.550	**0.678**	0.645	0.671	0.609	0.519	0.037	0.126	0.645	2.2	0.204	0.5
	Ablaut	0.582	0.659	**0.775**	0.761	0.748	0.739	0.736	0.021	0.025	0.761	10.9	0.520	2.7
	Final T	0.480	0.515	0.570	0.636	**0.648**	0.663	0.512	−0.027	0.124	0.636	4.9	0.245	0.8
	Templatic	0.605	0.788	0.889	0.907	**0.910**	0.872	0.848	0.034	0.059	0.907	0.5	0.025	0.3
	8-way	0.172	0.192	0.439	**0.469**	0.464	0.418	0.086	0.051	0.383	0.469	27.7	0.500	0.0

Table 2: Plural prediction and residual analysis (10-fold CV) in Macro-F_1 or %; M+P is the reference for all Δ and residuals. Baselines: Ngr (N-gram); LSTM (Bi-LSTM); Δ (M+P $-$ baseline). Residuals: Ceil (max F_1); IrrErr (joint failure); Compl (Complementarity); Sev (high-confidence exceptions). Significance: Appendix Table 14.

6. Discussion

6.1. Main Findings

Tashlhiyt pluralization follows a tiered dependency structure rather than a monolithic rule set. At the macro-level, hand-crafted features decisively outperform neural sequence learning—notably by +0.445 F_1 in 3-way classification—confirming that high-level strategies rely on systematic abstractions that Bi-LSTMs struggle to induce from low-resource data. At the micro-level, M+P significantly outperforms the Bi-LSTM in 4/7 mutation-specific tasks, while three domains (Medial A, Ablaut, Templatic) show no significant difference ($p < 0.05$), suggesting these patterns are equally accessible to either approach. High complementarity further indicates that feature-based and memory-based models capture distinct error patterns, supporting hybrid architectures for non-concatenative morphological analysis.

6.2. The Grammar-Memory Interface

Our findings are consistent with gradient productivity accounts in which computation and storage occupy a continuum rather than categorical routes (Bybee and McClelland, 2005; Hay and Baayen, 2005; O'Donnell, 2015). M+P dominates 7/10 tasks, indicating Tashlhiyt plural formation is more systematic than traditionally assumed. High complementarity in domains like Ablaut (0.520) reveals non-redundant signals: hand-crafted features isolate prosodic structure, while the Bi-LSTM captures patterns where explicit generalizations fail. In this low-resource setting, the Bi-LSTM cannot induce the abstract regularities that features encode directly, explaining its macro-level failure; its micro-level advantages reflect sensitivity to less productive, item-level patterns. High irreducible error (27.7%) alongside low idiosyncrasy severity (<5%) suggests complexity stems from competition among overlapping regularities, consistent with the low conditional entropy conjecture (Ackerman and Malouf, 2013; Cotterell et al., 2019), rather than arbitrary lexical listing.

6.3. Implications for Morphological NLP

Hand-crafted features decisively outperform neural baselines in macro-level tasks (>40pp F_1), demonstrating that sequence models trained from scratch on limited data struggle to induce abstract Amazigh morphological regularities without explicit structural bias. This confirms that explicit linguistic resources are a technical necessity for under-resourced Afroasiatic varieties (Khalifa et al., 2020). Ablation results also suggest that data augmentation for Tashlhiyt must respect prosodic structure, unlike unconstrained generation common in low-resource inflection (Anastasopoulos and Neubig, 2019). For low-resource non-concatenative systems, linguistically informed features are as critical as model architecture.

These limitations on neural models, however, need not be permanent. The annotation schema developed here (e.g., plural class labels, prosodic weight features, foot structure, derivational categories) can serve as a reusable template for related varieties like Central Atlas Tamazight and Tarifit, which share non-concatenative processes but differ in pattern inventories. The systematic singular–plural correspondences could also bootstrap pre-annotation tools that suggest candidate classes for new items, easing community-driven documentation. As structured datasets grow, neural models may acquire sufficient signal to learn regularities they currently fail to induce from raw forms (Kann and Schütze, 2016), making hand-crafted features a bridge to, rather than a substitute for, data-driven approaches.

7. Limitations

Dataset Scope and Speaker Bias. While our dataset of 1,185 nouns is substantial for a low-resource language, its composition introduces potential biases. A significant portion of the paradigms is derived from Jebbour (1996), which reflects the internal grammar of a single native speaker. While these data are supplemented by corpus resources, they may not fully capture the range of cross-dialectal and intra-speaker variation known to exist in Amazigh plural marking. Consequently, our findings on lexical idiosyncrasy may partially reflect speaker-specific patterns rather than language-wide irregulars.

Depth of Phonotactic Modeling. Our feature engineering focuses on prosodic weight, syllable structure, and edge-aligned n-grams. However, we do not explicitly model more abstract phonological interactions, such as vowel-to-vowel co-occurrence constraints or CV-templates (e.g., the root-and-pattern logic dominant in Semitic). Prior work on Arabic suggests that these templatic configurations are important to plural learnability (Dawdy-Hesterberg and Pierrehumbert, 2014). The absence of these features may explain some of the residual variance in our micro-level models, particularly in the prediction of vowel ablaut.

Refining Residual and Error Analysis. Our quantification of residual lexical idiosyncrasy currently defines idiosyncrasy as the intersection of failure across distributional and featural models. A more comprehensive approach would incorporate feature-space minimal pair analysis, identifying items with near-identical feature vectors but divergent class labels; the density of such local ambiguities provides a more precise diagnostic of irreducible idiosyncrasy than global accuracy alone. Furthermore, a qualitative error analysis is needed to determine if misclassifications cluster around specific phonological environments (e.g., specific vowel combinations or consonant clusters) not currently captured in our feature set. Identifying such clusters would distinguish between true lexical exceptions and systematic patterns overlooked by the current model.

Methodological Constraints. To ensure a fair comparison across feature sets, we employed fixed hyperparameters across all models. While this protocol minimizes optimization artifacts, it likely results in sub-optimal performance metrics. Furthermore, our Bi-LSTM baseline represents what character-level sequence models can achieve when trained from scratch at this data scale. More powerful approaches, including pretrained multilingual models, cross-lingual transfer from Afroasiatic languages, or data augmentation techniques (Anastasopoulos and Neubig, 2019), could narrow the performance gap. Our results therefore establish a lower bound for neural approaches rather than an absolute ceiling on sequence model capacity. However, adopting such methods would shift the experimental question from what can be induced from the data alone to what can be transferred from external resources, confounding the controlled comparison that is the paper's central contribution.

8. Ethics Statement

Linguistic Representation and Data Use. This research focuses on Tashlhiyt, a low-resource Amazigh language. Our primary goal is to improve the computational visibility of underrepresented Afroasiatic languages. The dataset used in this study is derived from published lexicographic sources and native speaker intuitions; it contains no personally identifiable information (PII) or sensitive data. By releasing the annotated dataset and feature sets, we aim to provide a public resource for the Amazigh community and researchers to foster further developments in indigenous language technologies.

Annotator Labor. The manual annotation and adjudication performed in this study were conducted by the second author, who was compensated for his labor consistent with local labor standards and academic research norms. No external crowdsourcing or unpaid labor was used in the production of the feature sets.

AI Assistance. We utilized AI assistants (specifically Anthropic's Claude) to generate, integrate, and orchestrate some Python scripts used for experimental analysis, as well as to provide editorial suggestions for prose clarity. All code logic, data processing pipelines, and textual revisions were manually verified and authorized by the human authors, who bear full responsibility for the methodology and content.

Environmental Impact. The computational experiments described in this study, including 10-fold cross-validation and feature selection across 10 domains, were conducted on consumer-grade hardware (Apple M4 Pro). The total training time was under four hours. Given the efficiency of the models used (Logistic Regression, Random Forest, and Gradient Boosting), the carbon footprint of this research is negligible compared to the training of large-scale language models.

9. Acknowledgements

This article has benefited from comments and suggestions from Karim Bensoukas, Abdelkrim Jebbour, Rachid Ridouane, and three anonymous SLiDE reviewers. It was supported by a Social Science and Humanities Research Council of Canada grant (435-2020-0193).

10. Bibliographical References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Wafia Adouane, Nasredine Semmar, and Richard Johansson. 2016a. Romanized berber and romanized arabic automatic language identification using machine learning. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 53–61.

Wafia Adouane, Nasredine Semmar, Richard Johansson, and Victoria Bobicev. 2016b. Automatic detection of arabicized berber and arabic varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 63–72.

Samir Amri, Lahbib Zenkouar, and Mohamed Outahajala. 2017. Build a morphosyntaxically annotated amazigh corpus. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, pages 1–7.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. *arXiv preprint arXiv:1908.05838*.

Rkia Bani, Samir Amri, Lahbib Zenkouar, and Zouhair Guennoun. 2023. Deep neural networks for part-of-speech tagging in under-resourced amazigh. *Revue d'Intelligence Artificielle*, 37(3):611.

André Basset. 1952. *La langue berbère*. Oxford University Press.

Samir Ben Si Saïd. 2014. *De la nature de la variation diatopique en kabyle: Étude de la formation des singulier et pluriel nominaux*. Doctoral dissertation, Université Nice Sophia Antipolis.

Samir Ben Si Saïd. 2020. La morphologie nominale et les segments flottants en kabyle. *Multilinguales*, 13.

Carl Darling Buck. 2008. *A dictionary of selected synonyms in the principal Indo-European languages*. University of Chicago Press.

Joan Bybee and James L McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review*, 22.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The sigmorphon 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology*, pages 10–22.

Lisa Garnand Dawdy-Hesterberg and Janet Breckenridge Pierrehumbert. 2014. Learnability and generalisation of Arabic broken plural nouns. *Language, cognition and neuroscience*, 29(10):1268–1282.

François Dell and Mohamed Elmedlaoui. 2002. *Syllables in Tashlhiyt Berber and in Moroccan Arabic*, volume 2. SKluwer Academic Publishers.

Rebecca Hasselbach. 2007. External plural markers in Semitic: A new assessment. In Cynthia L. Miller, editor, *Studies in Semitic and Afroasiatic Linguistics Presented to Gene B. Gragg*, pages 123–138. The Oriental Institute of the University of Chicago.

Jennifer B Hay and R Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in cognitive sciences*, 9(7):342–348.

Ali Idrissi. 2000. On Berber plurals. In Jacqueline Lecarme, Jean Lowenstamm, and Ur Shlonsky, editors, *Research in Afroasiatic Grammar*, pages 101–124. John Benjamins.

Abdelkrim Jebbour. 1988. Process of formation of the nominal plural in Tamazight (Tachelhiyt of Tiznit, Morocco): Non-concatenative approach. D.E.S. thesis, Mohamed V University, Faculty of Letters, Rabat.

Abdelkrim Jebbour. 1996. *Morphologie et contraintes prosodiques en berbère (Tachelhit de*

Tiznit) — Analyse linguistique et traitement automatique. Doctorat d'état dissertation, Mohammed V University, Faculty of Letters and Human Sciences, Rabat.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70.

Salam Khalifa, Nasser Zalmout, and Nizar Habash. 2020. Morphological analysis and disambiguation for Gulf Arabic: The interplay between resources and methods. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3895–3904.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Mena B Lafkioui. 2008. Dialectometry analyses of berber lexis. *Folia Orientalia*, 44:71–88.

Ahmed Lasri. 1991. *Aspects de la phonologie nonlinéaire du parler berbère chleuh de Tidli*. Doctoral dissertation, Université Paris 3.

John J McCarthy. 1979. *Formal problems in Semitic phonology and morphology*. Routledge.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: limitations of encoder-decoder neural networks as cognitive models for German plurals. *arXiv preprint arXiv:2005.08826*.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Hanae Raiss and Violetta Cavalli-Sforza. 2012. Anmorph: Amazigh nouns morphological analyzer. In *Press in Proceedings of the 5th Int. Conf. on Amazigh and ICT*.

Jilali Saib. 1986. Noun pluralization in Berber: A study in internal reconstruction. *Languages and Literatures*, 5:109–133.

I Taghbalout, F Ataa Allah, and M El Marraki. 2015. Amazigh noun inflection in the universal networking language. *International Journal of Education and Information Technology*, 9:122–128.

Catherine Taine-Cheikh. 2006. Alternances vocaliques et affixations dans la morphologie nominale du berbère : le pluriel en zénaga. In Dymitr Ibriszimow, Rainer Vossen, and Harry Stroomer, editors, *Études berbères III : Le nom, le pronom et autres articles*, pages 253–267. Rüdiger Köppe.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, et al. 2020. Sigmorphon 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39.

11. Language Resource References

Alderete, John and Agarwal, Piyush and Holubowsky, Kaye and Jebbour, Abdelkrim. 2025. *Tashlhiyt Noun Paradigm Database: 1914 Paradigms with 50 Linguistic Attributes*. Database of 1914 noun paradigms with phonological, morphological, and semantic attributes for gender and plural prediction.

Jebbour, Abdelkrim and Boukous, Ahmed and El Alami, Abdelkrim and Li, Jane S. Y. and Ridouane, Rachid and Alderete, John. 2021. *Nine Tashlhiyt Texts: Structured Representations of 18,000 Words (First Release)*. Dataset.

Appendices

A. Features

Table 3: Overview of all target variables and features used in the Tashlhiyt plural prediction experiments. Target variables (y) are divided into macro-level strategies and micro-level mutation patterns. Features (m, s, p) include both intrinsic database properties and engineered prosodic/distributional markers.

Feature Type	Feature Label	Variable Type	Vocab Size
Target (y) (Macro)	`y_macro_suffix`	binary	2
	`y_macro_mutated`	binary	2
	`y_macro_3way`	multi-class	3
Target (y) (Micro)	`y_micro_medial_a`	binary	2
	`y_micro_final_a`	binary	2
	`y_micro_final_vw`	binary	2
	`y_micro_ablaut`	binary	2
	`y_micro_final_t`	binary	2
	`y_micro_templatic`	binary	2
	`y_micro_8way`	multi-class	8
Morphological	`m_gender`	binary	2
	`m_deriv_cat`	categorical	6
	`m_paradigm_struc`	categorical	3
	`m_r_augment`	categorical	3
	`m_loan_status`	binary	2
Semantic	`s_animacy`	binary	2
	`s_humanness`	binary	2
	`s_semantic_field`	categorical	22
Phonological (Prosodic & Distributional)	`p_ngrams_macro`	multi-category	2,019
	`p_LH_ends_L`	binary	2
	`p_LH_initial_weight`	binary	2
	`p_LH_less_2_syllables`	binary	2
	`p_LH_all_heavy`	binary	2
	`p_LH_count_heavies`	binary split	2
	`p_LH_count_moras`	continuous	1–10
	`p_foot_count_feet`	continuous	0–3
	`p_foot_residue_right`	binary	2
	`p_foot_residue`	binary	2

B. Reproducibility

Computing Infrastructure and Software. All experiments were conducted on a single Apple M4 processor (macOS 24.3.0). The total compute time for the full 10-domain suite—including feature selection, ablation studies, and stratified 10-fold cross-validation—was under 4 hours. The implementation utilized Python 3.13.2 with the following library versions: `scikit-learn` 1.3.0, `xgboost` 2.0.3, `imbalanced-learn` 0.14.1, and `numpy` 1.24.3.

Fixed Hyperparameter Specifications. To ensure that performance deltas reflect feature quality rather than optimization bias, we utilized identical hyperparameters across all 10 domains. All models used a fixed seed (`random_state`=42) for reproducibility.

- **Logistic Regression:** Inverse regularization strength $C = 1.0$; penalty: L_2; solver: `lbfgs`; `max_iter`=1000; `class_weight='balanced'`.
- **Random Forest:** $n\_estimators = 100$; `max_depth`=10; `min_samples_split`=2; `max_features='sqrt'`; `class_weight='balanced'`.
- **XGBoost:** $n\_estimators = 100$; `learning_rate`=0.1; `max_depth`=10; `subsample`=1.0; `colsample_bytree`=1.0. The `scale_pos_weight` was calculated per domain as $N_{\text{neg}}/N_{\text{pos}}$.
- **Bi-LSTM (Baseline):** Embedding dimension: 32; Hidden units: 64 (bidirectional); Dropout: 0.3; Optimizer: Adam; Early stopping patience: 10 epochs.

Class Imbalance and Validation. Models were evaluated using stratified 10-fold cross-validation. To address extreme class imbalance (minority class $< 10\%$), we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training folds of the `final_a`, `final_vw`, and `insert_c` domains. SMOTE was configured with `sampling_strategy='auto'` and an adaptive k-neighbors parameter set to $\min(5, N_{\text{minority}} - 1)$. Crucially, over-sampling was restricted to training folds to prevent data leakage, ensuring validation folds remained representative of the original data distribution.

Dataset and Task Domains. The full feature set and target variable mapping are detailed in Table 3. Macro-level domains include $N = 1,185$ nouns, while micro-level domains are restricted to the $N = 562$ nouns exhibiting internal or mixed plural patterns. All target variables were encoded as integers, and categorical features were one-hot encoded for the Logistic Regression baseline. The anonymized annotated dataset is included as supplementary material with this submission.

Feature Selection and Data Leakage Mitigation We acknowledge that performing n-gram feature selection on the full dataset before cross-validation introduces mild information leakage, as feature selection uses target labels from test folds. However, several factors mitigate this concern: (1) feature selection is performed *once* to define a stable feature space, not iteratively tuned per fold; (2) all model weights are learned exclusively on training data; (3) the selected features form a domain-general master set used across all 10 tasks, not task-specific tuning; and (4) our scientific claims focus on *relative* performance between feature sets (which is unaffected by this design choice) rather than absolute performance estimates. Future work could validate this approach by performing nested feature selection within each training fold.

C. Detailed Model Performance

Tables 4–13 report mean percentage ± standard deviation from stratified 10-fold cross-validation. Model architectures: LogReg (Logistic Regression), RanFor (Random Forest), and XGB (XGBoost). Feature sets: Sem (Semantic), Morph (Morphological), Phon (Phonological), M+P (Morphological and Phonological), All F's (All features). Notation: † denotes SMOTE application; ‡ denotes high cross-fold variation ($\sigma \geq 15.0$). See the Reproducibility section for fixed hyperparameter specifications. Tables 4–6 cover macro-level domains ($N = 1,185$); Tables 7–13 cover micro-level domains ($N = 562$). Table 14 reports significance tests.

Table 4: Performance: **Macro: Has Suffix**

Set	Model	Acc	F_1	AUC
Sem	LogReg	52.2 ± 3.9	60.2 ± 5.0	55.2 ± 4.2
	RanFor	56.1 ± 5.7	63.5 ± 7.2	60.5 ± 5.1
	XGB	58.2 ± 6.3	66.4 ± 6.5	60.4 ± 6.2
Morph	LogReg	55.5 ± 4.0	62.1 ± 4.0	57.1 ± 3.9
	RanFor	62.3 ± 8.1	70.1 ± 8.5	63.3 ± 5.9
	XGB	60.0 ± 6.8	67.5 ± 8.0	64.0 ± 6.4
Phon	LogReg	81.9 ± 4.0	87.1 ± 2.7	87.8 ± 4.9
	RanFor	69.9 ± 4.4	76.4 ± 3.9	79.0 ± 4.2
	XGB	73.4 ± 2.8	79.8 ± 2.2	81.2 ± 4.8
M+P	LogReg	82.8 ± 4.4	**87.6 ± 3.0**	88.9 ± 4.9
	RanFor	73.8 ± 4.5	79.9 ± 3.7	80.8 ± 4.2
	XGB	77.0 ± 3.6	82.9 ± 2.3	84.7 ± 4.5
All F's	LogReg	81.5 ± 3.9	86.7 ± 2.6	88.2 ± 4.8
	RanFor	75.8 ± 4.2	81.8 ± 3.4	81.5 ± 4.1
	XGB	77.4 ± 2.3	83.0 ± 1.6	84.3 ± 4.0
N-grams	LogReg	79.6 ± 3.2	85.2 ± 2.2	84.4 ± 5.4
	RanFor	74.5 ± 3.2	80.8 ± 2.2	80.6 ± 4.5
	XGB	73.9 ± 3.6	79.9 ± 2.8	80.9 ± 4.9

Table 5: Performance: **Macro: Has Mutation**

Set	Model	Acc	F_1	AUC
Sem	LogReg	56.7 ± 3.2	56.2 ± 3.3	60.9 ± 3.1
	RanFor	58.9 ± 3.0	63.5 ± 2.6	62.0 ± 3.0
	XGB	58.1 ± 2.2	59.8 ± 2.5	62.0 ± 2.4
Morph	LogReg	56.9 ± 3.3	59.5 ± 2.5	58.7 ± 5.2
	RanFor	57.0 ± 3.1	58.4 ± 3.8	61.1 ± 5.0
	XGB	56.5 ± 3.8	56.9 ± 5.9	61.0 ± 5.2
Phon	LogReg	77.1 ± 4.1	76.6 ± 3.9	84.5 ± 3.7
	RanFor	71.1 ± 4.8	68.3 ± 6.1	79.5 ± 4.1
	XGB	75.3 ± 3.4	74.7 ± 3.2	83.6 ± 3.7
M+P	LogReg	77.7 ± 3.4	**76.9 ± 3.2**	84.9 ± 3.8
	RanFor	73.8 ± 5.1	71.8 ± 5.3	81.3 ± 3.6
	XGB	76.3 ± 3.9	75.3 ± 3.7	85.2 ± 3.0
All F's	LogReg	77.0 ± 2.1	76.3 ± 1.7	84.5 ± 3.4
	RanFor	74.0 ± 3.7	71.6 ± 4.6	80.8 ± 3.6
	XGB	75.6 ± 2.9	74.6 ± 2.8	84.6 ± 2.8
N-grams	LogReg	75.4 ± 3.6	74.7 ± 3.7	82.9 ± 3.4
	RanFor	71.1 ± 3.4	74.5 ± 2.4	79.5 ± 4.3
	XGB	71.5 ± 4.2	73.7 ± 3.7	80.7 ± 3.4

Table 6: Performance: **Macro: 3-way Class.**

Set	Model	Acc	F_1	AUC
Sem	LogReg	41.9 ± 4.8	40.2 ± 4.4	60.6 ± 4.8
	RanFor	43.7 ± 6.7	42.1 ± 6.6	62.1 ± 6.0
	XGB	52.5 ± 3.6	31.7 ± 3.5	62.8 ± 6.0
Morph	LogReg	39.9 ± 4.5	37.4 ± 4.6	60.0 ± 5.3
	RanFor	45.7 ± 5.3	44.0 ± 4.9	63.6 ± 4.3
	XGB	52.8 ± 4.1	32.2 ± 3.8	63.9 ± 5.0
Phon	LogReg	70.4 ± 5.3	65.8 ± 5.5	83.5 ± 4.9
	RanFor	61.6 ± 5.9	57.4 ± 6.8	78.9 ± 4.5
	XGB	68.2 ± 5.0	62.1 ± 6.0	82.1 ± 5.1
M+P	LogReg	72.3 ± 5.0	**68.3 ± 5.4**	84.8 ± 4.9
	RanFor	64.8 ± 5.8	60.4 ± 6.1	80.4 ± 4.4
	XGB	69.6 ± 5.1	64.7 ± 5.5	85.1 ± 4.4
All F's	LogReg	70.5 ± 5.8	66.6 ± 5.8	84.3 ± 5.5
	RanFor	63.6 ± 5.5	59.5 ± 5.7	79.5 ± 4.8
	XGB	68.9 ± 6.8	63.9 ± 8.1	84.4 ± 5.1
N-grams	LogReg	67.3 ± 5.6	61.4 ± 5.4	81.2 ± 5.0
	RanFor	60.2 ± 6.2	57.1 ± 5.9	76.9 ± 5.4
	XGB	65.6 ± 6.4	55.3 ± 7.8	79.1 ± 4.5

Table 7: Performance: **Micro: Medial A**

Set	Model	Acc	F_1	AUC
Sem	LogReg	60.2 ± 5.2	26.6 ± 7.9	56.8 ± 11.5
	RanFor	54.1 ± 5.4	30.8 ± 6.6	57.3 ± 9.3
	XGB	30.8 ± 5.7	28.1 ± 3.6	54.6 ± 8.1
Morph	LogReg	55.2 ± 8.4	35.4 ± 8.1	64.8 ± 9.5
	RanFor	55.4 ± 7.9	36.3 ± 7.7	63.4 ± 9.7
	XGB	55.5 ± 8.3	37.7 ± 8.8	64.6 ± 10.6
Phon	LogReg	73.1 ± 4.2	52.7 ± 6.2	84.3 ± 5.2
	RanFor	72.1 ± 5.7	50.7 ± 7.4	81.1 ± 7.5
	XGB	66.0 ± 12.6	48.2 ± 10.2	81.7 ± 6.6
M+P	LogReg	81.3 ± 5.5	**54.9 ± 12.6**	85.1 ± 4.3
	RanFor	78.3 ± 3.8	51.6 ± 10.0	83.9 ± 5.5
	XGB	77.9 ± 4.3	53.7 ± 8.5	84.2 ± 6.2
All F's	LogReg	81.8 ± 3.5	48.7 ± 8.2	83.1 ± 5.9
	RanFor	79.2 ± 3.6	49.9 ± 9.4	82.0 ± 6.6
	XGB	77.2 ± 6.8	51.3 ± 10.3	81.6 ± 8.6
N-grams	LogReg	78.8 ± 4.8	49.0 ± 10.2	82.7 ± 7.8
	RanFor	71.0 ± 5.1	44.2 ± 6.1	78.1 ± 6.4
	XGB	66.0 ± 6.8	44.9 ± 8.9	78.6 ± 8.9

Table 8: Performance: **Micro: Final A** (†)

Set	Model	Acc	F_1	AUC
Sem	LogReg	58.7 ± 7.0	21.4 ± 8.4	55.3 ± 9.5
	RanFor	54.6 ± 8.5	22.4 ± 4.8	55.7 ± 9.9
	XGB	35.1 ± 5.9	23.6 ± 3.3	57.1 ± 9.4
Morph	LogReg	53.0 ± 5.7	31.0 ± 3.9	66.3 ± 5.8
	RanFor	52.5 ± 5.3	29.6 ± 3.7	65.7 ± 6.6
	XGB	52.3 ± 5.2	29.9 ± 3.5	65.5 ± 6.1
Phon	LogReg	52.3 ± 5.4	27.1 ± 5.0	68.1 ± 7.7
	RanFor	49.8 ± 7.1	27.7 ± 5.1	68.4 ± 6.6
	XGB	29.4 ± 2.8	26.1 ± 1.5	71.3 ± 7.8
M+P	LogReg	76.0 ± 4.2	**33.3 ± 11.5**	75.7 ± 7.9
	RanFor	66.2 ± 4.9	34.6 ± 2.5	76.1 ± 5.9
	XGB	63.9 ± 4.7	34.0 ± 3.1	75.6 ± 7.4
All F's	LogReg	84.3 ± 4.2	32.3 ± 16.5‡	74.6 ± 9.2
	RanFor	70.8 ± 6.2	32.5 ± 5.1	73.5 ± 7.2
	XGB	69.9 ± 4.4	30.8 ± 8.0	71.5 ± 8.1
N-grams	LogReg	81.5 ± 4.6	43.1 ± 10.3	74.6 ± 8.2
	RanFor	69.0 ± 8.6	32.1 ± 12.0	69.7 ± 10.9
	XGB	63.3 ± 5.7	32.9 ± 7.6	69.6 ± 8.8

Table 9: Performance: **Micro: Final Vw** (†)

Set	Model	Acc	F_1	AUC
Sem	LogReg	76.3 ± 4.9	18.3 ± 11.9	72.2 ± 19.3‡
	RanFor	73.0 ± 4.9	19.1 ± 7.7	73.9 ± 18.1‡
	XGB	54.6 ± 4.2	12.6 ± 5.2	67.5 ± 18.1‡
Morph	LogReg	79.5 ± 7.2	21.9 ± 13.6	61.9 ± 22.4‡
	RanFor	83.3 ± 7.0	26.1 ± 15.6‡	73.3 ± 19.1‡
	XGB	80.3 ± 9.6	23.6 ± 15.3‡	71.3 ± 19.7‡
Phon	LogReg	91.6 ± 2.5	40.1 ± 16.8‡	79.5 ± 17.3‡
	RanFor	73.7 ± 9.2	17.8 ± 9.3	72.7 ± 18.1‡
	XGB	39.3 ± 11.0	11.3 ± 4.1	72.1 ± 16.8‡
M+P	LogReg	92.3 ± 2.4	**33.2 ± 20.5**‡	82.5 ± 17.7‡
	RanFor	91.6 ± 3.7	44.1 ± 25.1‡	79.3 ± 18.3‡
	XGB	88.8 ± 5.3	35.0 ± 17.7‡	72.8 ± 20.1‡
All F's	LogReg	93.2 ± 2.6	37.9 ± 24.1‡	82.8 ± 16.1‡
	RanFor	90.7 ± 4.3	40.3 ± 22.6‡	78.1 ± 18.6‡
N-grams	LogReg	89.7 ± 3.0	27.3 ± 21.5‡	81.3 ± 12.7
	RanFor	82.4 ± 3.9	23.4 ± 15.4‡	81.6 ± 13.3
	XGB	68.9 ± 6.8	18.3 ± 7.6	80.9 ± 11.3

Table 10: Performance: **Micro: Ablaut**

Set	Model	Acc	F_1	AUC
Sem	LogReg	58.4 ± 5.0	59.2 ± 4.4	64.3 ± 7.3
	RanFor	60.3 ± 6.0	58.7 ± 5.6	63.8 ± 7.4
	XGB	61.0 ± 5.8	60.3 ± 5.4	64.4 ± 6.8
Morph	LogReg	66.4 ± 6.9	62.9 ± 8.1	70.7 ± 6.8
	RanFor	64.6 ± 6.1	62.5 ± 7.4	70.8 ± 7.1
	XGB	63.9 ± 5.3	62.3 ± 6.9	70.9 ± 6.9
Phon	LogReg	77.6 ± 4.4	76.5 ± 4.6	83.1 ± 5.5
	RanFor	71.9 ± 5.0	70.9 ± 5.1	80.7 ± 6.2
	XGB	75.2 ± 5.7	74.4 ± 5.9	81.9 ± 7.0
M+P	LogReg	76.1 ± 6.3	**75.6 ± 5.7**	83.9 ± 5.9
	RanFor	72.2 ± 5.4	72.5 ± 5.4	81.9 ± 6.3
	XGB	78.6 ± 8.0	77.6 ± 8.7	85.8 ± 5.7
All F's	LogReg	74.9 ± 7.0	73.8 ± 6.9	83.2 ± 6.7
	RanFor	71.9 ± 5.8	72.4 ± 4.8	81.0 ± 6.6
	XGB	77.2 ± 8.5	76.2 ± 8.7	84.7 ± 6.6
N-grams	LogReg	74.0 ± 4.9	73.5 ± 4.4	82.0 ± 5.0
	RanFor	72.4 ± 4.7	71.3 ± 4.7	80.3 ± 5.6
	XGB	74.0 ± 4.9	72.9 ± 5.1	80.8 ± 5.7

Table 11: Performance: **Micro: Final T** (†)

Set	Model	Acc	F_1	AUC
Sem	LogReg	74.4 ± 5.0	11.1 ± 10.3	55.2 ± 15.7‡
	RanFor	61.2 ± 6.8	12.3 ± 7.0	53.6 ± 14.5
	XGB	36.3 ± 9.6	13.2 ± 6.1	56.3 ± 17.2‡
Morph	LogReg	69.9 ± 6.3	21.8 ± 4.7	66.1 ± 8.3
	RanFor	75.1 ± 5.6	26.0 ± 8.0	69.0 ± 11.2
	XGB	75.5 ± 5.6	28.6 ± 7.6	71.0 ± 13.9
Phon	LogReg	73.3 ± 6.5	30.7 ± 8.0	86.9 ± 7.1
	RanFor	64.4 ± 10.3	25.2 ± 7.0	86.9 ± 8.0
	XGB	56.1 ± 6.3	22.2 ± 4.5	87.0 ± 7.5
M+P	LogReg	85.9 ± 4.0	**35.0 ± 17.8**‡	88.0 ± 6.1
	RanFor	80.1 ± 5.9	32.4 ± 7.3	87.2 ± 4.8
	XGB	81.1 ± 4.1	31.2 ± 8.4	85.5 ± 8.1
All F's	LogReg	89.5 ± 3.2	35.3 ± 13.7	84.7 ± 6.7
	RanFor	84.7 ± 4.9	37.3 ± 12.2	87.0 ± 7.0
	XGB	86.5 ± 5.3	40.4 ± 14.9	87.7 ± 6.9
N-grams	LogReg	87.0 ± 3.9	39.9 ± 13.4	89.5 ± 6.8
	RanFor	79.2 ± 5.7	29.8 ± 17.2‡	88.1 ± 7.1
	XGB	71.5 ± 5.3	29.2 ± 8.0	86.0 ± 7.9

Table 12: Performance: **Micro: Templatic**

Set	Model	Acc	F_1	AUC
Sem	LogReg	68.7 ± 6.5	42.5 ± 10.9	77.1 ± 7.8
	RanFor	70.8 ± 6.6	44.0 ± 12.1	75.5 ± 8.1
	XGB	72.1 ± 7.1	44.3 ± 12.0	76.0 ± 7.6
Morph	LogReg	83.6 ± 4.5	68.7 ± 6.4	92.7 ± 3.5
	RanFor	84.2 ± 4.3	69.2 ± 6.4	92.2 ± 4.1
	XGB	85.1 ± 4.4	70.4 ± 6.4	92.8 ± 2.6
Phon	LogReg	93.2 ± 2.8	81.9 ± 7.5	94.1 ± 4.4
	RanFor	91.1 ± 3.3	75.3 ± 10.8	91.9 ± 4.5
	XGB	91.6 ± 2.9	77.1 ± 9.4	93.0 ± 4.4
M+P	LogReg	94.1 ± 1.9	**85.0 ± 4.6**	97.9 ± 1.5
	RanFor	92.7 ± 3.3	83.3 ± 6.8	97.6 ± 1.7
	XGB	95.9 ± 2.1	89.5 ± 5.3	97.5 ± 1.9
All F's	LogReg	94.3 ± 2.0	85.5 ± 4.9	97.8 ± 1.3
	RanFor	92.2 ± 3.6	82.1 ± 7.6	97.7 ± 1.6
	XGB	94.0 ± 2.4	84.4 ± 5.8	97.2 ± 1.6
N-grams	LogReg	92.2 ± 3.6	79.3 ± 9.3	93.9 ± 6.0
	RanFor	92.2 ± 3.3	78.3 ± 8.7	92.1 ± 5.9
	XGB	92.4 ± 2.5	79.7 ± 6.2	92.0 ± 6.7

Table 13: Performance: **Micro: 8-way Class.**

Set	Model	Acc	F_1	AUC
Sem	LogReg	22.2 ± 4.6	17.2 ± 6.0	65.5 ± 6.1
	RanFor	22.2 ± 5.1	17.2 ± 5.6	67.0 ± 4.9
	XGB	45.4 ± 2.3	14.2 ± 2.1	66.8 ± 6.7
Morph	LogReg	31.3 ± 8.7	19.2 ± 5.2	71.1 ± 6.2
	RanFor	28.6 ± 3.1	22.3 ± 2.7	73.4 ± 4.5
	XGB	50.4 ± 4.5	19.9 ± 4.8	73.9 ± 1.8
Phon	LogReg	60.1 ± 6.3	43.9 ± 6.4	81.7 ± 4.2
	RanFor	48.6 ± 5.5	37.0 ± 4.9	78.1 ± 4.8
	XGB	62.5 ± 6.3	39.8 ± 9.9	82.9 ± 2.7
M+P	LogReg	64.8 ± 9.0	**46.9 ± 8.9**	85.1 ± 6.3
	RanFor	56.4 ± 7.1	42.2 ± 6.2	82.2 ± 5.0
	XGB	67.6 ± 5.4	45.8 ± 10.4	87.2 ± 2.6
All F's	LogReg	64.4 ± 6.9	46.4 ± 7.5	86.4 ± 6.5
	RanFor	56.4 ± 8.0	44.7 ± 10.1	86.0 ± 4.9
	XGB	66.7 ± 7.3	44.9 ± 12.5	88.1 ± 3.6
N-grams	LogReg	57.7 ± 7.0	41.8 ± 6.7	80.8 ± 4.1
	RanFor	45.0 ± 6.3	35.0 ± 7.2	77.6 ± 5.9
	XGB	58.9 ± 6.5	38.6 ± 8.9	81.5 ± 3.7

Domain	M+P vs. N-gr		M+P vs. LSTM	
	Δ	p	Δ	p
Macro-Level				
Has Suffix	+0.036*	0.018	+0.062***	<0.001
Has Mutation	+0.023*	0.044	+0.378***	<0.001
3-way	+0.068***	<0.001	+0.445***	<0.001
Micro-Level				
Medial A	+0.037	0.067	+0.042	0.426
Final A	−0.067*	0.048	+0.135***	<0.001
Final V/W	+0.037	0.156	+0.126**	0.005
Ablaut	+0.021	0.197	+0.025	0.180
Insert C	−0.027	0.547	+0.124*	0.013
Templatic	+0.034	0.054	+0.059	0.486
8-way	+0.051**	0.006	+0.383***	<0.001

Table 14: Significance tests (paired two-tailed t-test) comparing M+P against baselines. Δ values denote mean difference in Macro-F_1 across 10 folds. Significance: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.