Phonetics Encode Language Identity More Than Grammar: Evidence from LOLO Typology Prediction

Anonymous ACL submission

Abstract

We test whether phonetic surface form alone can predict grammatical typology across languages. Using verse-aligned parallel Bible translations in 14 typologically diverse languages, we convert each verse to an International Phonetic Alphabet (IPA) character sequence via Epitran and predict two language-level WALS features: basic word order (SOV/SVO/VSO) and gender system size (None/Two/Three/Many). Because labels are constant within a language, random verse splits leak language identity and substantially overestimate generalization; we therefore adopt leave-one-language-out (LOLO) evaluation as the primary protocol. Across character n-gram TF–IDF baselines, phonological and phonotactic features, BiLSTM and Transformer encoders, and a gradient-reversal adversarial objective to suppress language-ID cues, random splits yield near-perfect accuracy, consistent with memorization. Under LOLO, performance is modest and highly variable across held-out languages, and representation analyses show embeddings cluster by language identity (and correlated genealogical and areal effects) more strongly than by typology. We release a reproducible IPA pipeline and offer an evaluation caution: in a phonetics-only setting, IPA robustly encodes language identity, while transferable signal for broad grammatical typology is weak and sensitive to label coverage.

1 Introduction

Human languages differ in how they *sound* and how they *structure* meaning. A long-standing question in linguistics is whether (and to what extent) phonological patterns are coupled to syntactic and morphological typology. In parallel, modern machine learning systems trained on raw sequences often appear to acquire non-trivial linguistic regularities, fueling claims of "emergent grammar" from distributional data. These two threads motivate a concrete empirical test: *if we erase meaning and orthography and retain only phonetic form, can we recover typological properties that are conventionally treated as grammatical?*

This paper studies whether phonetic representations alone carry predictive signal for two widely-used typological features: (i) *basic word order* (SOV, SVO, VSO) and (ii) *gender system size* (None, Two, Three, Many). We use verse-aligned parallel translations of the same source text across multiple languages, convert each verse to an IPA character sequence, and evaluate models under a *leave-one-language-out* (LOLO) protocol to measure cross-linguistic generalization rather than within-language memorization.

1.1 Motivation

The relationship between phonology and higher-level grammar is debated across linguistic theories. At the same time, sequence models can produce deceptively strong "generalization" when train/test splits allow leakage of language identity (e.g., the model sees the same language in both train and test, only with different sentences). This matters for typology prediction: if a model can identify the language from its phonetic inventory, it can trivially map language identity to typological labels without learning anything transferable across languages. A rigorous evaluation should therefore force *unseen-language* testing.

1.2 Research Questions

We formalize three research questions.

RQ1. Can IPA character sequences predict **basic word order** under LOLO evaluation?

RQ2. Can IPA character sequences predict **gender system size** under LOLO evaluation?

RQ3. Do learned representations cluster primarily by **language identity** rather than by grammatical labels (word order / gender)?

1.3 Contributions

1) We build a parallel, verse-aligned **IPA dataset pipeline** for cross-linguistic experiments (text → IPA), enabling phonetics-only modeling.
2) We show that **random within-language splits are misleading** for typology prediction, and motivate LOLO as the appropriate protocol for cross-linguistic generalization.
3) We provide a **comparative study** across feature-based baselines and neural encoders, including an adversarial language-identification objective designed to suppress language cues.
4) We analyze errors and representation structure, finding that **language identity dominates** the signal available in phonetics-only inputs, limiting reliable transfer to grammatical typology.

2 Related Work

2.1 Typology Prediction from Text and Multilingual Representations

Computational typology commonly predicts typological properties (e.g., WALS features) from textual distributions, using both hand-engineered surface statistics and learned representations Dryer and Haspelmath (2013); Dunn et al. (2011); Malaviya et al. (2017); Gutkin et al. (2020). Recent work shows that multilingual and multi-task language representations can encode typological regularities and support feature prediction under sparse supervision Malaviya et al. (2017); Bjerva and Augenstein (2018); Devlin et al. (2019); Conneau et al. (2020). However, such results can be inflated when train/test splits leak language identity, allowing models to exploit language-specific cues rather than cross-linguistic generalization Malaviya et al. (2017); Conneau et al. (2020). This has motivated evaluation protocols that explicitly test transfer to unseen languages, such as leave-one-language-out, rather than random splits that mix data from the same language.

2.2 Phonological Typology, Phonotactics, and Cross-Level Correlations

Phonological typology and phonotactic modeling study cross-linguistic regularities in sound systems and constraints on segment sequences Hayes and Wilson (2008); Prince and Smolensky (2004). Distributional patterns over phones can be learned and compared across languages and often reflect genealogical or areal structure Hayes and Wilson (2008). At the same time, classic interface theories do not posit a direct mapping from segmental phonology to core syntactic typology, instead emphasizing mediating levels such as prosody Selkirk (1984); Nespor and Vogel (1986). Accordingly, evidence that phonetic or phonotactic sequences alone robustly determine broad syntactic properties remains limited, and observed correlations must be interpreted cautiously.

2.3 Domain Adaptation and Adversarial Representation Learning

Adversarial representation learning is widely used to suppress nuisance factors such as domain, speaker, or language identity while preserving task-relevant signal. A standard approach is the Domain-Adversarial Neural Network (DANN), which uses a gradient reversal layer to encourage invariance to domain labels Ganin and Lempitsky (2015); Ganin et al. (2016). This framework has been applied extensively in cross-domain and cross-lingual transfer settings Ganin et al. (2016); Conneau et al. (2020). In our setting, adversarial training directly tests whether typological signal persists once language-identity cues are explicitly discouraged.

3 Data

3.1 Corpus and Parallel-Text Rationale

We use verse-aligned Bible translations as a controlled multilingual corpus. Each language version expresses (approximately) the same semantic content at the verse level, enabling a pragmatic control for topic and domain while varying only the target language. This parallel-text setting reduces spurious correlations that arise when languages are drawn from unrelated corpora (e.g., news vs. literature), where topic distributions can dominate learned representations. The source XML files are obtained from an open Bible corpus repository Christodoulopoulos (2025).

Why parallel text helps (and what it does not solve). Parallel text reduces semantic/domain shift across languages, but it does *not* eliminate stylistic translation artifacts. Translators differ in literalness, register, and conventionalized phrasing, and these choices may indirectly influence phoneme statistics. We treat this as an unavoidable confound and return to it in §7.

3.2 Language Set and Typological Labels

Our experiments use $L = 14$ languages spanning multiple language families and writing systems

(Table 1). For each language ℓ, we associate two typological targets:

1. **Basic word order** in $\{\text{SOV}, \text{SVO}, \text{VSO}\}$.
2. **Nominal gender system size** in $\{\text{NONE}, \text{TWO}, \text{THREE}, \text{MANY}\}$.

Labels are drawn from typological reference resources, using the widely adopted WALS categorization for word order and number of genders (Dryer and Haspelmath, 2013). We deliberately coarsen the label space to three word-order classes and four gender-size classes to (i) reduce sparsity and (ii) match the small-L setting where each class must be supported by multiple languages.

Label mapping decisions. For word order, WALS distinguishes additional patterns and mixed orders in some languages; we map to $\{\text{SOV}, \text{SVO}, \text{VSO}\}$ by selecting the dominant/basic order reported in typological references and excluding "mixed" or "both" categories. For gender, we use a coarse count-of-genders label (none/two/three/many), consistent with WALS-style grouping Dryer and Haspelmath (2013).

Coverage and structural pitfalls under LOLO. Leave-One-Language-Out (LOLO) creates a hard constraint: for any held-out language $\ell^\star$, all target classes must already appear in the remaining $L-1$ training languages. If a class appears in only a single language, then LOLO becomes structurally unfair for that class: when that language is held out, the model cannot learn the held-out label at all (it becomes an *unseen label* problem, not a modeling problem). In our current language set, the THREE-gender class is represented by only one language, making LOLO evaluation for that label inherently brittle. We therefore report per-language results and interpret failures with caution, distinguishing between genuine generalization failure and structural label absence.

3.3 IPA Conversion and Preprocessing

Each verse is converted to an International Phonetic Alphabet (IPA) character sequence International Phonetic Association (1999) using grapheme-to-phoneme mappings from Epitran Mortensen et al. (2018). We apply lightweight normalization to reduce irrelevant variation: Unicode normalization (NFC/NFKC as appropriate) and whitespace standardization; followed by removal of punctuation and verse markup artifacts; then lowercasing where applicable (for scripts with case); at last optional filtering of characters outside the IPA inventory plus spaces.

Table 1: Languages, families, and typological labels used in experiments. Each language contributes approximately 7,000 verses, totaling ∼98k examples.

ISO	Family	Word order	Gender
es	Indo-European	SVO	Two
fr	Indo-European	SVO	Two
ru	Indo-European	SVO	Three
hi	Indo-European	SOV	Two
te	Dravidian	SOV	None
tr	Turkic	SOV	None
zh	Sino-Tibetan	SVO	None
ja	Japonic	SOV	None
id	Austronesian	SVO	None
tl	Austronesian	VSO	None
mi	Austronesian	VSO	None
sw	Niger-Congo	SVO	Many
zu	Niger-Congo	SVO	Many
ar	Afro-Asiatic	VSO	Two

Conversion noise and script dependence. G2P transliteration is imperfect, and quality varies by language, orthographic depth, and the coverage of Epitran mappings Mortensen et al. (2018). This introduces a controlled but non-negligible noise source: the model may learn systematic transliteration artifacts. We treat this as part of the empirical setting: the pipeline approximates what a practitioner could do at scale without expert phonetic annotation.

3.4 Dataset Structure and the Need for LOLO

Let $\mathcal{V}_\ell = \{v_{\ell,1}, \ldots, v_{\ell,n_\ell}\}$ be the set of verses for language ℓ, and let $x_{\ell,i}$ be the IPA sequence for verse $v_{\ell,i}$. Each verse inherits *language-level* typological labels $y_\ell^{(wo)}$ (word order) and $y_\ell^{(gen)}$ (gender system size). Thus the dataset is hierarchically labeled:

$$\{(x_{\ell,i},\, y_\ell)\}_{\ell=1}^{L}, \quad i = 1, \ldots, n_\ell.$$

A random verse-level train/test split leaks language identity: the model can memorize language-specific phonotactics and then trivially recover the language-level label for that same language. To measure cross-linguistic generalization, we use

Leave-One-Language-Out (LOLO):

$$\text{Train on } \bigcup_{\ell \neq \ell^\star} \mathcal{V}_\ell, \qquad \text{Test on } \mathcal{V}_{\ell^\star}.$$

4 Experimental Setup

4.1 Tasks

We study two typology prediction tasks from IPA character sequences. Each training example is a verse-level IPA string, while the target label is constant within a language (language-level supervision).

T1: Word order prediction (3-way). Given an IPA sequence x from language ℓ, predict $y_\ell^{\text{WO}} \in \{\text{SOV}, \text{SVO}, \text{VSO}\}$.

T2: Gender system size prediction (4-way). Given an IPA sequence x from language ℓ, predict $y_\ell^{\text{GEN}} \in \{\text{NONE}, \text{TWO}, \text{THREE}, \text{MANY}\}$.

Auxiliary analysis (representation bias). In addition to grammar prediction, we explicitly measure how strongly representations encode *language identity*. This motivates the adversarial setup (Section 5, GRL) and explains why random verse splits can dramatically overestimate "generalization".

4.2 Train/Test Protocols and Leakage Control

A central challenge is that labels are *language-level*, but instances are *verse-level*. Therefore, naive random splits create label leakage: the model can learn to identify the language (from phonotactics/orthography-to-IPA artifacts) and then map language $\rightarrow$ label.

Random verse split (leaky baseline). We include a within-language random split only as a diagnostic baseline. Formally, we sample train and test verses i.i.d. from the pooled set $\mathcal{D}$, which mixes verses from the same language across train and test. This setting is *not* a valid cross-linguistic generalization test.

Leave-One-Language-Out (LOLO; primary). We evaluate generalization to unseen languages using LOLO. Let $\mathcal{L}$ be the set of languages. For each held-out language $\ell^\star \in \mathcal{L}$, we train on all verses from $\mathcal{L} \setminus \{\ell^\star\}$ and evaluate on all verses from $\ell^\star$:

$$\mathcal{D}_{\text{train}}(\ell^\star) = \{(x_i, y_{\ell(i)}) : \ell(i) \neq \ell^\star\}, \quad (1)$$

$$\mathcal{D}_{\text{test}}(\ell^\star) = \{(x_i, y_{\ell^\star}) : \ell(i) = \ell^\star\}. \quad (2)$$

We report (i) per-language held-out performance and (ii) an aggregate score computed by concatenating predictions across all held-out folds.

Structural label-coverage constraint. LOLO implicitly assumes every test label is present in training labels. If a class occurs in only one language (e.g., THREE appears only in a single language in some subsets), then for the fold holding out that language, the training set contains no examples of that class. In that fold, *perfect* prediction is impossible and some implementations will raise "unseen label" errors. We therefore (a) explicitly check per-fold label coverage, and (b) either (i) skip the invalid fold with a transparent note, or (ii) collapse labels to enforce coverage (reported separately when used).

4.3 Metrics

Because class imbalance and "minority" classes (notably VSO and THREE) are central failure modes, we report both accuracy and macro-averaged F_1. We also report confusion matrices for LOLO folds (normalized by gold-class counts) to visualize systematic collapse into majority classes.

4.4 Implementation Details

Tokenization and vocabulary. All neural models operate on character-level IPA symbols. We build a vocabulary from the training languages in each LOLO fold, add special symbols $\langle\text{PAD}\rangle$ and $\langle\text{UNK}\rangle$, and map each verse to an integer sequence. Across the full dataset the observed IPA inventory size is approximately $V \approx 1227$ (excluding specials), yielding $V \approx 1229$ including $\langle\text{PAD}\rangle, \langle\text{UNK}\rangle$.

Sequence length. Verses are truncated/padded to a fixed maximum length L (we use $L = 256$ in our main experiments) to form tensors of shape $[B, L]$ for batch size B.

Feature-based baselines. We train TF–IDF character n-gram models with logistic regression using scikit-learn Pedregosa et al. (2011). We treat these baselines as strong "surface" controls: they often achieve near-perfect random-split results, highlighting leakage, but degrade sharply under LOLO.

Neural encoders. We train compact sequence encoders to avoid "scale fixes everything" arguments: (i) BiLSTM (bidirectional LSTM) encoders (1997; 1997), (ii) a small Transformer encoder (2017). Optimization uses Adam (2015). All neural models are implemented in PyTorch (2019). Unless otherwise stated, we use a fixed number of epochs per LOLO fold (e.g., 5 for BiLSTM, 3 for the Transformer in our runs), and select the

Table 2: Model families and the granularity of their evaluation.

Approach	Granularity
TF–IDF + LogReg	verse-level
Inventory features	language-level
Phonotactic stats	language-level
Lang-level n-grams	language-level
BiLSTM	verse-level
Tiny Transformer	verse-level
Adversarial (GRL)	verse-level

best checkpoint on a held-out slice of the training languages (never using the held-out test language).

Adversarial language-invariance. To reduce language-ID leakage, we train an adversarial model with a shared encoder and two heads: (a) a grammar head (predicting word order or gender), and (b) a language-ID head trained through a gradient reversal layer (GRL) to encourage language-invariant representations (Ganin et al., 2016). Let L_g be grammar loss and L_ℓ be language loss; the optimization objective is:

$$\min_{\theta_e,\theta_g} \max_{\theta_\ell} \; L_g(\theta_e,\theta_g) - \lambda\, L_\ell(\theta_e,\theta_\ell), \qquad (3)$$

where $\lambda \geq 0$ controls the adversarial strength.

Reproducibility. We fix random seeds for data shuffling and model initialization, and report LOLO results across all held-out languages. Compute is the practical bottleneck: LOLO requires retraining once per language, so we emphasize compact models and limited epoch budgets in the primary study.

5 Modeling Approaches and Core Findings

5.1 Model families evaluated

We compare (i) strong surface-feature baselines, (ii) neural sequence encoders, and (iii) an adversarial language-invariance variant. Table 2 summarizes inputs and evaluation granularity.

5.2 TF–IDF character baselines (verse-level)

We featurize each verse with character n-grams over IPA symbols, then apply TF–IDF weighting and a multinomial logistic regression classifier.

TF–IDF. We use standard TF–IDF character n-grams with multinomial logistic regression.

Classifier. Given TF–IDF vector $v(d) \in \mathbb{R}^m$, logistic regression estimates

$$p(y = k \mid d) = \mathrm{softmax}(W v(d) + b)_k, \qquad (4)$$

trained by minimizing cross-entropy with ℓ_2 regularization.

Observed behavior under LOLO. These baselines exhibit *high variance across held-out languages*: some languages reach near-ceiling performance, while others collapse to near-zero, consistent with the model relying on language-specific phonotactics rather than grammar.

5.3 Language-level diagnostic featurizations (inventory, phonotactics, aggregated n-grams)

To probe whether grammar is recoverable at the *language* level (as opposed to verse level), we compute language-aggregated descriptors: (i) *inventory* vectors encoding presence/absence of phones, (ii) low-dimensional *phonotactic statistics* (e.g., vowel/consonant proportions, entropy-like summaries, and/or short-range transition counts), and (iii) aggregated character n-gram frequency profiles (Cavnar and Trenkle, 1994). These yield only 14 labeled instances, so they are *diagnostic* rather than competitive.

5.4 Neural sequence encoders

We evaluate two lightweight neural encoders over IPA character sequences: a bidirectional LSTM and a compact Transformer. In both cases, IPA symbols are embedded and encoded into contextual token representations, which are then pooled to obtain a fixed-dimensional verse representation. This representation is passed to a softmax classifier trained with cross-entropy loss. The BiLSTM captures bidirectional sequential dependencies, while the Transformer uses self-attention with positional encodings; aside from the encoder architecture, the training and pooling setup is identical across models.

Key empirical pattern. Neural capacity does *not* eliminate LOLO instability: the models still succeed for certain languages and fail catastrophically for others, implying the bottleneck is signal identifiability (and label coverage), not expressiveness.

Table 3: LOLO accuracy (%) across held-out languages (mean ± std). For adversarial gender, the Russian fold is structurally ill-posed (singleton `Three`) and is excluded.

Model	Word order	Gender
TF–IDF + LogReg	48.54 ± 43.66	53.29 ± 39.09
BiLSTM	37.30 ± 38.12	42.56 ± 37.96
Tiny Transformer	52.46 ± 42.87	45.50 ± 37.34
Adversarial (GRL)	56.36 ± 41.45	$39.78 \pm 41.58^{\dagger}$

5.5 Adversarial language-invariance with gradient reversal (GRL)

We implement a domain-adversarial objective inspired by GRL-based representation learning (Ganin et al., 2016). The encoder $f_\theta(x)$ feeds two heads: (i) a typology (grammar) predictor g_ϕ trained to predict y, and (ii) a language-ID discriminator h_ψ trained to predict the language ℓ.

Let $\mathcal{L}_{\text{GRAM}}$ be the typology cross-entropy and $\mathcal{L}_{\text{LANG}}$ the language-ID cross-entropy. Training seeks representations predictive of grammar while suppressing language identity:

$$\min_{\theta,\phi} \max_{\psi} \; \mathbb{E}\big[\mathcal{L}_{\text{GRAM}}(g_\phi(f_\theta(x)), y)\big] - \lambda\, \mathbb{E}\big[\mathcal{L}_{\text{LANG}}(h_\psi(f_\theta(x)), \ell)\big]. \tag{5}$$

In practice, the maximization over ψ is implemented by inserting a gradient reversal layer (GRL) between f_θ and h_ψ, which multiplies the gradient by $-\lambda$ during backpropagation.

What it changes empirically. Adversarial training can improve *average* LOLO accuracy on word order, but does not reliably fix minority-class collapse (e.g., VSO) and can degrade gender performance, especially under poor label coverage (e.g., singleton classes).

5.6 Core quantitative findings (LOLO)

Table 3 reports mean ± std accuracy across held-out languages for the verse-level models. Figure 1 visualizes per-language LOLO word-order accuracy for major models, exposing the "some languages work, others die" regime.

5.7 What these results mean

Two empirical facts dominate:

- **Random splits are misleading.** Near-perfect random-split accuracy primarily reflects language-ID leakage (phonetics ⇒ language ⇒ label), not cross-linguistic typology inference.

- **LOLO is unstable across held-out languages.** Even strong models exhibit extreme variance, and minority/rare typology classes (notably VSO; and `Three` in gender) are routinely crushed, consistent with weak identifiability under the current label coverage.

In short: the pipeline can learn *who* the language is from IPA extremely well, but that does not translate into a reliable predictor of *how* the language organizes grammar under cross-language generalization.

6 Analysis

6.1 What the Models Actually Learned: Identity Over Typology

A consistent pattern across all phases is the sharp contrast between (i) random verse splits and (ii) leave-one-language-out (LOLO). Under random splits, the model sees training and test verses from the *same* language, and since our targets are language-level typological labels (Section 3), the supervised objective becomes effectively equivalent to language identification.

Formally, let $\ell \in \mathcal{L}$ denote a language and let $t(\ell)$ be a deterministic mapping from language to a typology label (e.g., word order or gender class). Each verse x is sampled from a language-specific distribution $p(x \mid \ell)$, while its label is $y = t(\ell)$. Under a random verse split, training data includes examples from all $\ell \in \mathcal{L}$. Therefore, a classifier can achieve near-perfect accuracy by learning a proxy for $\arg\max_\ell p(\ell \mid x)$ and then outputting $t(\ell)$, without learning any cross-linguistic phonology–syntax mapping at all. This is consistent with the observed near-100% performance under within-language evaluation, which is a classic failure mode when language identity is available as a shortcut signal Jauhiainen et al. (2019).

Under LOLO, the shortcut collapses: the held-out language $\ell^\star$ is absent from training, so mapping $x \mapsto \ell$ is no longer helpful unless the representation captures *transferable* phonological structure shared across languages. In other words, LOLO forces the intended question: *is there a robust correlation*

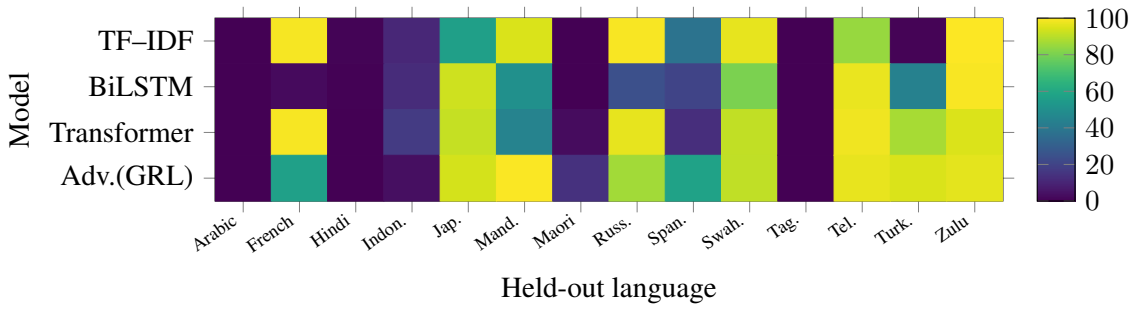

Figure 1: Per-language LOLO accuracy (%) for **word order**. The "checkerboard" structure (near-100% for some languages, near-0% for others) persists across models, supporting the interpretation that representations remain strongly language-specific.

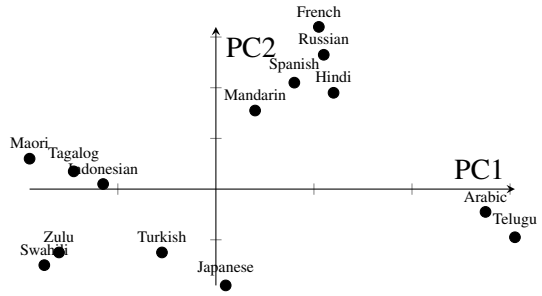

Figure 2: PCA of language-level centroid embeddings μ_ℓ (computed from IPA-based representations). In our runs, centroids cluster by language identity (and often by genealogy) more consistently than by typology label, supporting the interpretation that the dominant signal is language/family rather than grammar.

between phonetic surface patterns and typology that generalizes across languages?

6.1.1 Embedding-space evidence: clustering follows language/family

To validate what the models encode, we compute language-level centroids in representation space. Let $f(\cdot)$ be a feature map (e.g., TF–IDF vectors or a neural encoder's pooled embedding). For each language ℓ, define the mean embedding

$$\mu_\ell = \frac{1}{|D_\ell|} \sum_{x \in D_\ell} f(x), \tag{6}$$

and apply PCA to $\{\mu_\ell\}_{\ell \in \mathcal{L}}$. In our experiments, the dominant axes separate languages (and often genealogical groupings) more clearly than they separate typology labels, aligning with prior findings that learned language representations strongly encode genealogical/areal signals Bjerva and Augenstein (2018); Malaviya et al. (2017).

6.2 Why LOLO Is Hard in This Setup

LOLO is challenging here for structural reasons beyond model capacity.

Language-level targets. The supervision is constant within each language, which makes cross-lingual generalization an instance of *typology prediction* from phonological surface statistics rather than standard text classification. The desired mapping is closer to phonology $\rightarrow$ typology than to text $\rightarrow$ label, and any true signal is likely to be weak, indirect, and confounded by genealogy and contact effects.

Label coverage and "unseen class" pathology. A practical issue arises when a class appears in only one language (or very few languages). In that case, holding out that language yields a training set where the class is absent, making multi-class evaluation ill-posed: a label encoder literally cannot represent an unseen class, and a classifier has no evidence to learn it. This is not a mere implementation quirk; it is a structural limitation of LOLO with sparse typological coverage. A defensible write-up must state this explicitly and treat such labels with care (e.g., merging rare classes, reframing the task, or reporting a restricted-label evaluation).

Confounds from genealogy and translation style. Parallel text reduces semantic/domain variation, which is useful, but it can also amplify shared translation conventions within families or scripts. As a result, models may learn stable family/style signatures in the IPA stream that correlate with labels only because typology itself is genealogically clustered.

6.3 Adversarial Representation Learning: What GRL Can and Cannot Do

To reduce language-identification shortcuts, we apply domain-adversarial training with a gradient reversal layer (GRL). Let f_θ be a shared encoder, g_ϕ a grammar head, and d_ψ a language-ID discriminator. The objective is

$$\min_{\theta,\phi} \max_{\psi} \quad \mathcal{L}_{\text{gram}}(g_\phi(f_\theta(x)), y) - \lambda\, \mathcal{L}_{\text{lang}}(d_\psi(f_\theta(x)), \ell), \tag{7}$$

implemented by multiplying the gradient from $\mathcal{L}_{\text{lang}}$ by $-\lambda$ at the encoder via GRL Ganin et al. (2016); Ganin and Lempitsky (2015). Intuitively, the encoder is pushed toward representations informative for grammar while being uninformative for language identity, a common strategy in domain adaptation Ben-David et al. (2010).

Interpreting the observed gains (word order). In our LOLO results, adversarial training yields a modest aggregate improvement for word order (e.g., $\sim$56% vs. $\sim$52% for a non-adversarial transformer in the reported run). This suggests that there may exist weak, transferable phonotactic cues correlated with basic constituent order, and that suppressing some language-ID features can reduce overfitting to language-specific artifacts. However, improvements are not uniform: some languages improve, others degrade, consistent with the fragile and confounded nature of the signal.

Interpreting the observed drops (gender). For the gender task, adversarial training can *reduce* accuracy substantially in some runs. A principled interpretation is that, in the current dataset, gender class is strongly entangled with language identity and genealogy. If the only predictive evidence for gender is "which language family does this look like," then removing language signal removes most of the usable information. This is not a shortcoming of GRL; it is a diagnostic that the dataset provides limited cross-linguistic phonetic evidence for gender inventory size.

6.3.1 Per-language effect: adversarial delta plot

To make the adversarial effect concrete, we recommend reporting per-language deltas: $\Delta_\ell = \text{Acc}_\ell^{\text{adv}} - \text{Acc}_\ell^{\text{base}}$. Figure 3 is a compact visualization that reviewers actually like because it shows you understand heterogeneity rather than hiding behind a mean.

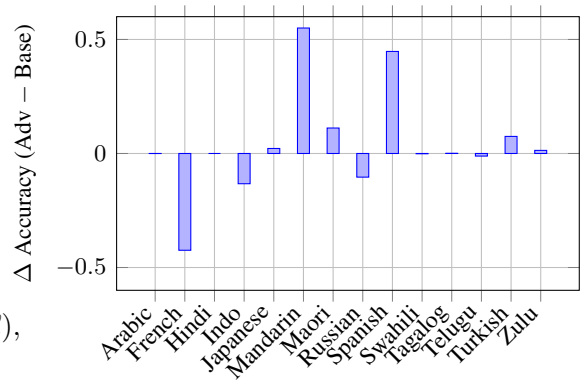

Figure 3: Per-language change in LOLO accuracy from adversarial training (GRL) relative to a non-adversarial baseline (e.g., TinyTransformer). Reporting Δ_ℓ highlights that adversarial training can help some languages while harming others, which is expected under confounding and label sparsity.

Takeaway. Across tasks, the analysis supports a conservative conclusion: IPA-derived representations encode language identity extremely well, and whatever typological signal exists is weaker, task-dependent, and easily dominated by genealogy and translation artifacts. Adversarial training can slightly improve word-order transfer in some settings, but it does not magically conjure typology from sound.

7 Conclusion

We show that, in a phonetics-only setting, IPA-derived sequences encode language identity robustly, while transferable signal for broad grammatical typology is weak under proper cross-language evaluation. Across surface features, neural sequence models, and adversarial objectives, random verse splits yield near-perfect performance that reflects language memorization rather than grammatical inference; under leave-one-language-out evaluation, performance becomes modest and highly variable across held-out languages and minority typology classes. These results indicate that the dominant learnable signal arises from language identity and correlated genealogical or areal effects, not language-universal phonology-to-typology mappings. Beyond this empirical finding, we contribute a reproducible verse-aligned IPA pipeline, demonstrate why random splits are misleading for language-level labels, and provide a comparative analysis of representation choices under strict cross-language generalization, offering a cautionary note for typology prediction from phonetic form alone.

Limitations

This document does not cover the content requirements for ACL or any other specific venue. Check the author instructions for information on maximum page lengths, the required "Limitations" section, and so on.

Corpus genre and translation effects

Our experiments use verse-aligned Bible translations as a parallel corpus. This controls for topical domain and coarse semantics across languages, but it also introduces genre and translation artifacts. Religious text tends to be stylistically conservative and may under-represent colloquial phonological processes. In addition, translations are shaped by translator norms (sometimes called "translationese"), which can flatten genuine cross-linguistic variation or introduce systematic preferences that are not native to the target language Baker (1993). These factors may either (i) inflate cross-language similarity in phonetic patterns or (ii) add translator-specific signals that are unrelated to typology.

IPA conversion noise and script coverage

IPA strings are produced via automatic grapheme-to-phoneme mapping (Epitran) (2018). While this provides a uniform representation, the output is an approximation of pronunciation, not a gold phonetic transcription. Conversion quality varies by language, orthography depth, and available mappings; some scripts/languages may be partially unsupported or require custom resources, yielding failures or systematic noise. Consequently, measured signals may partly reflect transliteration and normalization artifacts rather than phonology.

Typology label granularity and mapping decisions

We derive word-order and gender categories from typological inventories (e.g., WALS-style descriptors) Dryer and Haspelmath (2013). These labels are coarse by design. "Both" word order collapses multiple constructions into a single tag; "Many" gender conflates heterogeneous systems. Such discretization can obscure distinctions that matter for phonology, morphology, or syntax, and it may also produce ambiguous supervision when the underlying language reality is gradient or context-dependent.

Class coverage constraints under LOLO

Leave-one-language-out (LOLO) evaluation is structurally unforgiving when a class is represented by very few languages. If a label appears in only one (or effectively one) language, then holding that language out makes the class unseen during training, causing unavoidable failures (or forcing ad-hoc handling). Even with > 1 languages per class, heavy imbalance can lead to degenerate predictors that collapse minority classes (e.g., persistent failure on VSO or "Three") despite reasonable overall accuracy. Any LOLO-based claim therefore depends critically on balanced label coverage at the *language* level, not just at the *verse* level.

Non-independence of instances

Each verse is treated as an instance, yet the target labels (word order, gender system size) are language-level properties. This induces strong within-language dependence: N verses from the same language are not N independent samples for typology prediction. Verse-level micro-averaging can therefore exaggerate confidence and obscure the true unit of generalization (the language). A more statistically faithful view reports performance aggregated per held-out language (and, in larger studies, per family or area), rather than treating verses as i.i.d. samples.

Interpretability limits

Even when models achieve above-chance LOLO accuracy, the mechanisms remain unclear: they may exploit phonotactic cues, borrowing effects, transliteration regularities, or properties of the conversion pipeline. Without controlled ablations (e.g., randomized phoneme inventories, shuffled phonotactics, or matched-family controls), it is difficult to attribute success to genuine phonology–syntax coupling.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker and 1 others, editors, *Text and Technology*, pages 233–250. John Benjamins, Amsterdam.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typol-

ogy at different levels with language embeddings. *arXiv preprint arXiv:1802.09375*.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, pages 161–175.

Christos Christodoulopoulos. 2025. bible-corpus. GitHub repository. Accessed 2025.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473:79–82.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Yaroslav Ganin, Evgeniya Rustamov, Yaroslav Ganin, Yaroslav Ganin, Yaroslav Ganin, Yaroslav Ganin, and Yaroslav Ganin. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17(59):1–35.

Alexander Gutkin, Tatiana Merkulova, and Martin Jansche. 2020. Linguistic typology features from text: Inferring the sparse features of the world atlas of language structures. *arXiv preprint arXiv:2005.00512*.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Tommi Jauhiainen, Marco Lindén, and Krister Jauhiainen. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research (JAIR)*, 65:675–782.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision g2p for many languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Marina Nespor and Irene Vogel. 1986. *Prosodic Phonology*. Foris, Dordrecht.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research (JMLR)*, 12:2825–2830.

Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell, Malden, MA.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Elisabeth O. Selkirk. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.