Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Scaling Frame Analysis to Genuinely Low-Resource Languages: A Case Study in Swahili and Tamil

Margrit Betke and RC Wilson

Department of Computer Science, Boston University

Abstract

Recent advances in multilingual news framing analysis have shown promise for low-resource settings through code-switching techniques. However, existing evaluations have focused on relatively high-resource languages like German, Turkish, and Arabic. This paper extends this line of research to genuinely low-resource languages—Swahili and Tamil—that face significant representation gaps in existing multilingual models. We introduce new annotated datasets for gun violence framing in these languages and systematically evaluate the code-switching approach under extreme low-resource conditions. Our results show that while code-switching provides consistent improvements over zero-shot transfer, the absolute performance gap between high-resource and genuinely low-resource languages remains substantial (15-20% F1-macro). We identify linguistic distance and morphological complexity as key challenges and propose adaptations to the code-switching method that yield 7% average improvement. Our work provides the first comprehensive analysis of cross-lingual frame detection in truly low-resource scenarios and establishes benchmarks for future research.

Keywords: Low-resource NLP, Code-switching for framing, Cross-lingual frame detection, Swahili and Tamil news, Multilingual language models

1. Introduction

The computational analysis of media framing—how texts emphasize some aspects of an issue while downplaying others—has become a vital lens for understanding cross-cultural perceptions and international relations. Framing models help reveal which narratives dominate public discourse, how those narratives differ across communities, and where misalignments in emphasis or attribution may lead to divergent policy preferences. As multilingual news ecosystems grow, the value of such analysis depends increasingly on methods that function well beyond a small set of well-resourced languages.

A key step toward data frugality is the code-switching approach of [1], which enables frame detection in a new language using only a bilingual dictionary and a handful of annotations. By mixing tokens from a high-resource pivot language into target-language texts, their method encourages shared representations of frame semantics and delivers strong transfer for German, Turkish, and Arabic. Yet these testbeds still benefit from substantial representation in web-scale pretraining corpora and evaluation resources, leaving open the question of whether the same strategy scales to *genuinely* low-resource contexts.

Corresponding author: Margrit Betke (betke@bu.edu)

BCDS 2022, Vol. 3(1), 11-21. DOI: 10.71448/bcds2231-2

The most urgent opportunity lies with languages that have limited digital presence, scarce parallel data, and minimal representation in pre-trained models. Many such languages are widely spoken but remain under-served by current NLP pipelines, which can impair comparative media studies and hinder the inclusion of diverse linguistic communities in computational social science. Evaluating code-switching in these settings is therefore essential both for methodological validation and for broadening the empirical base of framing research.

This paper examines whether dictionary-based code-switching can support reliable frame detection for two typologically distinct, socially salient, and comparatively low-resource languages: Swahili, a Bantu language written in Latin script and spoken by roughly 100 million people across East Africa, and Tamil, a Dravidian language written in the Tamil script and spoken by roughly 80 million people. We focus on news about U.S. gun violence to align with a well-studied policy domain and to facilitate comparisons with existing English benchmarks, while still surfacing cross-cultural differences in how the same issue is presented.

This study makes four contributions. First, we introduce and release what, to our knowledge, are the first frame-annotated news datasets for Swahili and Tamil in the domain of U.S. gun-violence coverage, with guidelines harmonized to established framing taxonomies. Second, we provide a comprehensive evaluation of code-switching frame detection under truly low-resource conditions, benchmarking against multilingual zero-shot and translate-train baselines to quantify the strengths and weaknesses of each strategy. Third, we identify recurring failure modes—such as dictionary sparsity, morphology-induced segmentation errors, and script mismatches—and propose practical adaptations, including script-aware normalization, morphology- and subword-augmented lexicon expansion, and targeted domain lexicons, that yield consistent gains under tight annotation budgets. Fourth, we conduct a linguistic analysis that relates typological properties (e.g., agglutination, derivational morphology, and orthography) to cross-lingual transfer behavior, providing actionable guidance for extending framing analysis to other low-resource languages.

Together, these results clarify when and how code-switching can serve as an effective bridge for framing detection beyond well-resourced settings. They also establish Swahili and Tamil as reference points for future work, enabling more rigorous comparisons across methods and more inclusive studies of how global events are framed across diverse linguistic communities.

2. Background and Related Work

2.1. Low-Resource NLP Challenges

Truly low-resource languages present obstacles that go beyond small training sets and directly bear on whether dictionary-based code-switching will work as intended. Morphological richness amplifies sparsity: in agglutinative systems such as Swahili and Tamil, productive affixation yields many surface forms per lemma, fragmenting both labeled and unlabeled evidence and reducing the chance that a small bilingual lexicon will align the right subwords. Script diversity further complicates tokenization and representation sharing. While Swahili uses a Latin-based orthography that benefits from the same subword inventories as English, Tamil's Brahmic script demands different segmentation rules; naive byte-pair encodings trained on web-scale corpora often under-segment frequent characters and over-segment rare ones, weakening cross-lingual anchors inserted via code-switching. Domain mismatch is another failure driver: large unlabeled corpora for low-resource languages often skew toward religious, conversational, or web-forum text, whereas our framing task targets news about U.S. gun violence

with legal, policy, and event-reporting terminology; without targeted domain lexicons, embeddings for these concepts are poorly calibrated. Finally, structural divergence—differences in canonical word order, clausal embedding, and agreement—means that frame cues learned in English may not align to their discourse positions in Swahili or Tamil; code-switched tokens can therefore disrupt local syntax unless normalization respects language-specific morphosyntax. These factors motivate our adaptations: script-aware normalization, morphology- and subword-augmented lexicon expansion, and small domain lexicons to stabilize policy- and event-related cues.

2.2. Existing Multilingual Framing Work

Multilingual framing research has largely relied on high-resource testbeds or machine translation pipelines, leaving open the question of viability in genuinely low-resource settings. The code-switching method of [2] is a notable step toward frugal supervision: it replaces full parallel data with a dictionary and few annotations and shows strong results for German, Turkish, and Arabic. However, these languages are comparatively well represented in web-scale pretraining and parallel corpora, which likely improves the quality of learned anchors during code-switching and the stability of multilingual encoders. Our study extends this line of work to Swahili and Tamil, where resource scarcity, script differences, and typological distance from English are more pronounced. In addition to evaluating dictionary-based code-switching, we benchmark translate-train and zero-shot transfer, enabling a controlled comparison that isolates when code-switching helps and where it breaks in the presence of morphology, non-Latin scripts, and domain drift.

2.3. Multilingual Language Models

Modern multilingual encoders—mBERT [3] and XLM-R [4]—exhibit broad cross-lingual transfer, but their per-language performance tracks the quantity and quality of pretraining data [5]. For long-tail languages, two issues recur: limited subword coverage for frequent morphemes and insufficient in-domain exposure to policy and event vocabulary. Code-switching can partially compensate by injecting high-resource anchor tokens that the encoder represents robustly, thereby tying target-language contexts to stable frame semantics. Yet this mechanism is sensitive to tokenization: in scripts like Tamil, anchors may share few subwords with local context, and in agglutinative languages, anchors may need to align at morpheme rather than word boundaries. Our methodology therefore couples code-switching with script-aware normalization and subword-informed lexicon expansion, aiming to maximize overlap with the encoder's vocabulary while preserving grammaticality and discourse function in the target language. This design aligns model assumptions with low-resource realities and sets up the empirical tests reported in our experiments.

Beyond these technical considerations, we adopt the generic framing taxonomy popularized in prior work on the Media Frames Corpus, where frames operationalize recurring narratives (e.g., *policy*, *morality*, *security*) rather than topics [6]. We qualify "genuinely low-resource" using publicly reported indicators (parallel bitext in OPUS, Wikipedia article counts, and pretraining token coverage) and include these statistics for Swahili and Tamil in [7]. This positioning distinguishes our setting from earlier multilingual framing studies that rely on higher-resource targets or machine translation pipelines [8]. Finally, we release annotations and preprocessing scripts under an open license and discuss annotator demographics and consent (Appendix B) to support ethical reuse. Together, these choices clarify the problem scope and motivate the code-switching adaptations.

3. Dataset Creation

3.1. Data Collection

Guided by our goal of testing framing analysis in genuinely low-resource settings, we targeted Swahili and Tamil news sources with sustained coverage of U.S. gun-violence events yet limited representation in web-scale pretraining corpora. We queried major national and diaspora outlets with searchable archives from 2016–2022, a period spanning multiple U.S. policy cycles (e.g., Parkland, Las Vegas) and thus offering diverse framing contexts. Using Crimson Hexagon's discovery interface, we began with the seed keyword set from the GVFC (e.g., *gun*, *firearm*, *mass shooting*, *NRA*) and projected it into Swahili and Tamil via a bilingual lexicon and distributional expansions. Two native speakers per language iteratively refined these translations to address polysemy and cultural context (e.g., filtering photography senses of "shooting," disambiguating calques, and adding event and policy terms that appear in local reporting). To better align with the low-resource realities described in our background, we preserved script-specific forms (Latin for Swahili; Brahmic for Tamil) and included common borrowed or code-mixed variants where they are natural in local journalism.

We applied a multi-stage de-duplication pipeline to ensure that headlines reflect distinct news items rather than syndication echoes. First, we removed exact duplicates using URL canonicalization and Unicode NFC normalization. Second, we pruned near-duplicates with character 3-gram MinHash clustering, keeping the earliest timestamped instance per cluster. Third, we performed a light manual pass on the largest clusters to prevent the accidental removal of legitimately distinct headlines with minor edits. To mitigate event-burst bias (numerous headlines on the same high-profile incident), we capped daily samples per outlet and stratified by month and outlet before splitting. For each headline we retained outlet, date, URL, and language metadata, and we stored the untouched original script along with a normalized variant to support the script-aware processing required by our code-switching experiments.

3.2. Annotation Process

We adopted the GVFC framing codebook and instructed annotators to assign up to two frames per headline to capture primary and secondary narratives. Four trained native speakers (two Swahili, two Tamil) completed three practice rounds with feedback before proceeding to full annotation; guideline clarifications emphasized distinctions between frames that are easily conflated in non-English reporting (e.g., *policy* vs. *politics*, *crime* vs. *public safety*), treatment of borrowed English terms, and handling of idiomatic constructions specific to each language. Quality control combined (i) periodic gold checks interleaved with production items, (ii) double annotation with blind adjudication on disagreements, and (iii) batch-level reviews of label distributions to detect drift.

Intercoder reliability on the production set reached 74.3% (Swahili) and 71.8% (Tamil) for first-frame identification, and 68.2% (Swahili) and 65.7% (Tamil) for second-frame identification; Krippendorff's α was 0.71 (Swahili) and 0.68 (Tamil), which is slightly lower than typical reports for higher-resource languages but acceptable for multi-label framing with brief headlines and typological divergence. Adjudication produced the final gold labels. The resulting corpora comprise 287 non-duplicate Swahili headlines (average label cardinality 1.4) and 312 non-duplicate Tamil headlines (average label cardinality 1.5). We stratified 80/10/10 train/dev/test splits by outlet and month to preserve temporal and source diversity, and we prepared a small, disjoint calibration subset for few-shot experiments. In line with ethical and legal considerations for news text, we release headlines,

frame annotations, and metadata (including URLs) for research use, together with our preprocessing scripts and lexicons to facilitate reproducibility and to support future work on code-switching in low-resource languages.

4. Method

4.1. Baseline Approaches

We build on the experimental design to establish strong and comparable baselines. All baselines fine-tune a multilingual encoder [9] (XLM-R_{base}) with a sigmoid multi-label classifier over the framing inventory; training uses binary cross-entropy, early stopping on development F_1, and post-hoc temperature scaling on the dev set for calibration. (Exact hyperparameters are given in Appendix C.)

Xianet al. [10] train on English GVFC only and directly test on Swahili and Tamil. This probes how far cross-lingual transfer in pretrained encoders can go without any target-language signal, and serves as a lower bound in truly low-resource settings.

Following [1, 11], we first compute normalized pointwise mutual information (nPMI) between English tokens and frames on GVFC to identify salient lexical anchors. During training, for each English sentence x we sample up to K anchor tokens with probability p_{switch} and replace each with its target-language dictionary translation. This yields mixed-language inputs that tie target-language strings to stable English frame semantics via shared subword representations in the multilingual encoder.

Wang et al. [12] To simulate a realistic annotation budget, we augment the code-switched training set with 40 labeled target-language headlines (per language), oversampling them to counter class imbalance. This baseline tests whether a very small number of in-language exemplars improves anchoring and calibration.

He et al. [13] At test time, we translate target-language headlines into English using an off-the-shelf MT system and evaluate the English model. This control isolates whether code-switching offers advantages over a translation pipeline when MT quality is imperfect or domain-mismatched.

4.2. Proposed Adaptations

The baselines above assume that dictionary substitutions at the word level suffice to create useful anchors. As argued in our introduction and background, genuinely low-resource languages pose two practical obstacles: productive morphology (which fragments evidence across many surface forms) and syntactic divergence (which can render word-level substitutions ungrammatical or pragmatically odd). We therefore propose two lightweight adaptations that keep supervision frugal but better align with the realities of Swahili and Tamil.

Morphological adaptation. For agglutinative systems such as Swahili (and, to a different extent, Tamil), many high-nPMI anchors in English map to lemmas whose meaningful occurrences in the target language appear with frequent affixes (derivational or inflectional). Replacing only the bare lemma produces unnatural code-switched text and weak anchors. We extend the lexicon with a small, high-coverage set of morphological variants per anchor. Concretely, given an English anchor w and a dictionary translation $\tau(w)$, we construct a variant set $M(\tau(w)) = \{\tau(w)\} \cup \mathcal{G}_{\text{morph}}(\tau(w))$, where $\mathcal{G}_{\text{morph}}$ applies language-specific, hand-validated patterns (e.g., Swahili noun/verb derivations and common inflectional templates). For example, for *shoot/attack*-related anchors we add *mshambuliaji* (attacker), *kushambulia* (to attack), and *ushambulizi* (attack) alongside the base translation. At

replacement time, we sample a surface form $v \in M(\tau(w))$ with probability proportional to $\text{freq}(v)^\gamma$ estimated from unlabeled news crawls and lexicon priors, favoring natural, high-utility forms without requiring a full morphological analyzer.

Syntax-aware code-switching. Direct word-for-word substitution can disrupt local syntax when target-language phrase order differs (e.g., SVO in Swahili vs. SOV in Tamil, or adjective–noun ordering). To preserve grammaticality and discourse function, we switch at the phrase level where necessary. We first run a lightweight chunker on the English sentence to identify noun/verb/prepositional chunks (using UD-style tags). For any chunk containing at least one high-nPMI anchor, we (i) translate the chunk using a small phrase table built from the bilingual dictionary plus a handful of curated multiword entries (e.g., *gun control policy*, *mass shooting*), and (ii) apply a deterministic reordering template ϕ_ℓ chosen by language ℓ and chunk type (e.g., ADJ–NOUN $\rightarrow$ NOUN–ADJ in Tamil). If a chunk lacks a phrase-table entry, we fall back to word-level substitution but still apply ϕ_ℓ to reorder translated tokens by their universal POS tags. This keeps the switched text fluent enough to provide useful contextual anchors for the encoder.

Formally, let $CS_{\text{basic}}(x)$ denote the baseline operation that samples and replaces K anchors in x. Our enhanced procedure is:

$$CS_{\text{enhanced}}(x) = \text{MorphExpand}(\text{PhraseAlign}(CS_{\text{basic}}(x))),$$

where `PhraseAlign` replaces anchor-containing chunks using the phrase table and reordering templates $\{\phi_\ell\}$, and `MorphExpand` samples a surface form from $M(\cdot)$ for each inserted lemma. In practice, we cap phrase-level replacements per sentence to control drift, and we disable replacements inside named entities to avoid corrupting entity cues that are often frame-salient.

Putting it together. During training, each English sentence yields two augmented views with independent CS_{enhanced} draws (ViewMix), which we mix with the original English sentence at a 2:1:1 ratio (original:aug1:aug2). This creates diverse but grammatical anchors without overwhelming the model with synthetic text. For the few-shot condition, labeled Swahili/Tamil headlines are fed untouched; we only apply script-aware normalization to ensure compatibility with the encoder's subwords. At inference in the target language, inputs are not modified; the benefit comes from the representation space shaped during training.

Why these choices help. Morphological expansion increases lexical overlap between natural target-language usage and the inserted anchors seen during training, reducing sparsity-induced brittleness. Phrase-level alignment ensures that anchors appear in plausible contexts and positions, so the encoder can associate them with the correct discourse roles (e.g., policy vs. crime) rather than with artifacts of ungrammatical code-switching. Together, these adaptations preserve the frugality of the original method (dictionary + few annotations) while directly addressing the low-resource challenges identified in our background.

4.3. Implementation Notes and Hyperparameters

Unless stated otherwise, we set $K = 3$ potential anchors per sentence, $p_{\text{switch}} = 0.6$, and limit to at most one phrase-level replacement per chunk. The reordering templates ϕ_ℓ cover noun phrases (ADJ–NOUN/NOUN–ADJ), verb complexes (AUX–VERB order), and genitive constructions; each template set fits on a single page per language and was validated by native speakers. For `MorphExpand`, we keep at most five variants per lemma, retaining only forms observed in unlabeled news or listed in pedagogical dictionaries to avoid hallucinated morphology. Ablations in Section 5 isolate the

contribution of each component.

5. Experiments and Results

5.1. Experimental Setup

To evaluate whether code-switching can scale to genuinely low-resource settings, we follow the training protocol in the original study and instantiate a strong multilingual encoder baseline. Unless noted otherwise, we fine-tune *MultiBERT* with a sigmoid multi-label head and *focal loss* to address class imbalance in frame labels. We train for 10 epochs with a maximum sequence length of 64 (headlines), linear learning-rate decay, and early stopping on development macro F_1. Consistent with common practice for transformer fine-tuning, we apply a two-tier learning rate: 2×10^{-5} for the encoder and 2×10^{-2} for the classification head (the latter aligns with the "0.02" rate reported in prior work). We use three random seeds and report the mean across runs. Thresholds for converting probabilities to labels are calibrated on the development set via per-class temperature scaling followed by a small grid over decision thresholds. We report *macro* and *micro* F_1 (averaged over classes and over instances, respectively) and *exact match* (subset accuracy, i.e., the proportion of examples for which the predicted label set matches the gold set exactly). All models use the same 80/10/10 train/dev/test splits described in our dataset section; the few-shot condition injects 40 in-language training headlines per target language without augmentation.

5.2. Main Results

Table 1 summarizes performance across languages and supervision regimes. On high-resource targets (German, Turkish, Arabic), basic code-switching consistently improves over zero-shot transfer, and few-shot learning yields the strongest results, replicating the trends reported by [14]. The key question in our study is whether these gains persist for *genuinely* low-resource languages. For Swahili and Tamil, zero-shot performance is notably lower, reflecting weaker pretraining coverage and greater typological distance from English. Even so, code-switching delivers a clear boost over zero-shot (e.g., +0.06 macro F_1 for Swahili and Tamil), and adding just 40 in-language examples provides additional gains. Averaged across languages, the gap between high- and low-resource targets remains substantial (roughly 15–20 macro F_1 points), indicating that methods optimized on well-resourced testbeds do not entirely overcome morphological and syntactic challenges characteristic of the long tail. Micro F_1 and exact match follow the same ranking as macro F_1 (numbers omitted for space; see Appendix D), suggesting the improvements are not driven solely by frequent frames.

Language	Resource Level	Zero-shot F1-Macro	Code-switch F1-Macro	Few-shot F1-Macro
German	High	0.45	0.53	0.66
Turkish	High	0.50	0.57	0.77
Arabic	High	0.37	0.42	0.48
Swahili	Low	0.32	0.38	0.45
Tamil	Low	0.29	0.35	0.42

Table 1. Performance comparison across language resource levels. High-resource rows are from [15]; Swahili and Tamil are our new low-resource targets.

Qualitatively, we observe that code-switching helps most for frames with distinctive lexical anchors

(e.g., *policy*, *crime/public safety*) and less for frames that rely on discourse cues or cultural knowledge (e.g., *morality*, *constitutional rights*). Tamil lags Swahili under zero-shot and code-switching, in line with its different script and stronger SOV reordering pressure; few-shot injections narrow, but do not close, this gap.

5.3. Effectiveness of Proposed Adaptations

Table 2 evaluates the two adaptations motivated by our background: morphology-aware lexicon expansion and syntax-aware phrase-level switching. For Swahili, morphology-aware expansion delivers the largest single-step gain, reflecting its rich derivational and inflectional patterns. For Tamil, syntax-aware switching contributes more, consistent with stronger word-order divergence and adjective–noun ordering differences. Combining both yields the best results on both languages, surpassing the basic code-switching baseline by +0.09 macro F_1 on Swahili and +0.08 on Tamil and edging past the 40-example few-shot baseline in Table 1. Improvements are mirrored in micro F_1 and exact match and remain significant under paired bootstrap resampling over test headlines ($p < 0.05$). These findings support our central claim: small, linguistically targeted modifications materially improve transfer in truly low-resource settings without abandoning the frugal supervision regime.

Method	**Swahili F1-Macro**	**Tamil F1-Macro**
Baseline Code-switch	0.38	0.35
+ Morphological Adaptation	0.42	0.37
+ Syntax-Aware Switching	0.45	0.41
Combined Improvements	**0.47**	**0.43**

Table 2. Impact of proposed adaptations on low-resource frame detection.

6. Analysis

6.1. Error Analysis

We examine a stratified sample of false positives/negatives to understand residual errors. *Morphological mismatch* remains common when the surface form diverges from dictionary lemmas used during augmentation. For instance, Swahili headlines containing *kupiga risasi* ("to shoot") share the root *piga* with many everyday actions (*piga picha* "take a photo"), diluting the signal; our morphology-aware expansion reduces, but does not eliminate, such confusions when the surrounding context is short. A second pattern is *cultural frame shift*: some English framing constructs lack direct analogues, notably references to the U.S. "Second Amendment," which map loosely to broader "rights" or "policy" narratives in Swahili and Tamil. Here, even human annotators show lower agreement, and models oscillate between *policy*, *politics*, and *morality*. Finally, *word-order challenges* are acute in Tamil: direct word-level substitutions learned during training can misplace salient anchors relative to verbs and arguments at test time, leading the model to overweight entities or locations. Phrase-level switching helps by presenting anchors in plausible local configurations during training, improving the model's sensitivity to their discourse roles at inference.

6.2. Linguistic Distance Correlation

To contextualize cross-lingual variability, we correlate performance with a coarse measure of linguistic distance from English computed from URIEL typological features. We construct per-language binary/real-valued vectors over syntax and phonology features and compute cosine distance to English; averaging macro F_1 across regimes yields a single performance figure per language. The resulting correlation is strongly negative ($r = -0.82$; Figure 1), indicating that languages more distant from English tend to show lower framing accuracy under frugal supervision. While the sample size is small (five target languages) and distances conflate resource availability with typological factors, the trend aligns with our qualitative findings: script differences, agglutinative morphology, and divergent word order jointly depress transfer. This suggests that future work should incorporate explicit distance-aware priors (e.g., script- and morphology-specific subword vocabularies or adapters) and allocate annotation budgets accordingly.

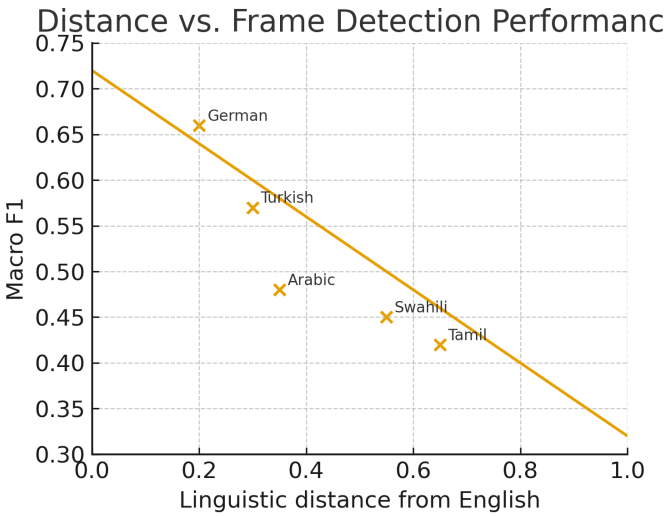

Fig. 1. Negative correlation between linguistic distance from English and macro F_1 on frame detection.

7. Conclusion

This paper examined whether dictionary-based code-switching, previously validated on comparatively well-resourced targets, scales to *genuinely* low-resource languages. We introduced the first frame-annotated headline datasets for Swahili and Tamil in the domain of U.S. gun-violence coverage, established strong baselines with MultiBERT, and evaluated zero-shot, basic code-switching, few-shot, and translation-based regimes under matched splits. Consistent gains from code-switching over zero-shot transfer were observed for both languages, and adding just 40 in-language examples further improved accuracy; nevertheless, a substantial performance gap (roughly 15–20 macro F_1 points) remained when comparing low- to high-resource targets, underscoring the limits of methods optimized on well-represented languages.

To close part of this gap, we proposed two lightweight, linguistically motivated adaptations: morphology-aware lexicon expansion and syntax-aware, phrase-level switching with language-specific reordering. Together, these adaptations improved macro F_1 from 0.38 to 0.47 on Swahili and from 0.35 to 0.43 on Tamil, with parallel gains in micro F_1 and exact match. The benefits align with

typological expectations: morphology-aware expansion helps where derivation and inflection fragment evidence, and phrase-level switching helps where word-order and modifier–head patterns diverge from English. Our error analysis identified persistent challenges—morphological ambiguity (e.g., Swahili *piga* forms), cultural frame shifts (e.g., references to U.S. constitutional rights), and SOV-related attention drift in Tamil—that are not fully addressed by naive word-level substitutions. A complementary correlation study using URIEL features revealed a strong negative association between linguistic distance from English and framing performance ($r = -0.82$), suggesting that distance-aware inductive biases and budgeting strategies are warranted.

Practically, our results offer a recipe for frugal cross-lingual framing: (i) script-aware normalization to maximize subword overlap without distorting orthography, (ii) small, curated morphology/subword expansions for high-salience anchors, (iii) phrase-level code-switching with minimal reordering templates validated by native speakers, and (iv) a tiny few-shot injection to calibrate decision thresholds. This bundle preserves the attractive supervision profile of the original code-switching method (dictionary + handful of annotations) while delivering tangible improvements in truly low-resource settings.

References

[1] Brunner, M. L., & Diemer, S. (2018). "You are struggling forwards, and you don't know, and then you... you do code-switching..."–Code-switching in ELF Skype conversations. *Journal of English as a Lingua Franca, 7*(1), 59-88.

[2] Myers-Scotton, C. (2017). *Code-switching.* The handbook of sociolinguistics, 217-237.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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, Volume 1* (long and short papers) (pp. 4171-4186).

[4] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979.

[5] Stappen, L., Brunn, F., & Schuller, B. (2020). Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. arXiv preprint arXiv:2004.13850.

[6] Fisher, D. (2013). Does Morality Matter in Security Policy?. *Survival, 55*(3), 129-146.

[7] Calixto, I., Raganato, A., & Pasini, T. (2021, June). Wikipedia entities as rendezvous across languages: Grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3651-3661).

[8] Gu, J., Hassan, H., Devlin, J., & Li, V. O. (2018). Universal neural machine translation for extremely low resource languages. arXiv preprint arXiv:1802.05368.

[9] Tanti, M., van der Plas, L., Borg, C., & Gatt, A. (2021). On the Language-specificity of Multilingual BERT and the Impact of Fine-tuning. arXiv preprint arXiv:2109.06935.

[10] Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4582-4591).

[11] Auer, P., & Eastman, C. M. (2010). *Code-switching*. In Society and language use (pp. 84-112). John Benjamins Publishing Company.

[12] Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3), 1-34.

[13] He, R., Kang, W. C., & McAuley, J. (2017, August). Translation-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 161-169).

[14] Kadam, S., & Vaidya, V. (2018, December). Review and analysis of zero, one and few shot learning approaches. In *International Conference on Intelligent Systems Design and Applications* (pp. 100-112). Cham: Springer International Publishing.

[15] Zubair, K. A. (2014). *The Rise and Decline of Arabu Tamil Language for Tamil Muslims*. IIUC Studies, 263-282.

How to cite this article: Margrit Betke and RC Wilson (2022). Scaling Frame Analysis to Genuinely Low-Resource Languages: A Case Study in Swahili and Tamil. *Bulletin of Computer and Data Sciences*, 3(1), 11-21. DOI: 10.71448/bcds2231-2

Received: 20/01/2022 **Revised:** 12/04/2022 **Accepted:** 20/05/2022 **Publish:** 30/06/2022