# SPICE: Synthetic Polyglot Injection for Cross-lingual Evaluation

**Angela Chen** **Kat Dou** **Kayli Requenez** **Joel Manu**

## Abstract

Multilingual language models (MLLMs) exhibit strong performance for high-resource languages but continue to struggle with low-resource and typologically distant ones. Prior work shows that increasing linguistic diversity during training can improve cross-lingual generalization, yet real-world data remain uneven, scarce, and difficult to scale. We explore whether synthetic languages can serve as a controllable source of structural diversity for multilingual fine-tuning. Using a stabilized extension of ConlangCrafter, we generate corpus-scale synthetic languages and translate English training data into them. We then fine-tune mT5-base jointly on English and these synthetic-language corpora to evaluate whether synthetic structural variation can improve cross-lingual transfer, particularly for low-resource languages. Our findings offer the first systematic assessment of synthetic languages as an alternative or complement to real low-resource data.

## 1 Introduction

Large Language Models (LLMs) have dramatically improved natural language processing, yet their performance remains uneven across languages. High-resource languages such as English benefit from abundant, diverse training data, while speakers of low-resource or typologically distant languages face degraded reliability, unnatural responses, and even harmful errors. This imbalance limits equitable access to AI technology, shaping who can meaningfully participate in an increasingly AI-mediated world.

Multilingual LLMs (MLLMs) aim to address this gap through cross-lingual transfer. However, transfer quality strongly depends on the diversity of languages seen during training. Prior work shows that typologically diverse training signals improve multilingual generalization: Stap and Monz (2025) demonstrate that models fine-tuned on more structurally diverse languages achieve better transfer, including to unseen language pairs. BUFFET (Asai et al., 2023) further highlights that performance varies widely across languages, particularly in low-resource settings, and uses a unified fine-tuning format to systematically benchmark this variation.

These observations raise a natural question: If linguistic diversity helps MLLMs, can we synthetically create the kinds of structural diversity that low-resource languages lack?

Synthetic languages—artificial but linguistically coherent systems—offer a way to inject structural diversity without requiring real low-resource data. If constructed languages mimic key typological properties, they may serve as a form of "virtual low-resource languages" that expand a model's exposure to unfamiliar morphosyntactic patterns.

In this work, we develop a scalable stabilization framework for generating synthetic corpora and investigate whether fine-tuning multilingual models on synthetic languages can improve cross-lingual transfer. We introduce SPICE, a method that fine-tunes mT5-base jointly on English and synthetic-language translations of BUFFET tasks. Our experiments evaluate both diverse and target-guided synthetic languages, testing whether (i) typologically diverse synthetic languages improve general cross-lingual transfer, (ii) target-guided synthetic languages improve performance for specific low-resource languages, and (iii) typological similarity or diversity correlates with transfer effectiveness.

## 2 Related Work

Existing approaches to minimizing this performance gap across languages have unique tradeoffs. There are small, language-specific models as well as large Multilingual LLMs (MLLMs). The advantage of single-language models is that there is no risk of representation imbalance or negative trans-

fer, in which the inclusion of one language might hurt the model's performance in others. Unfortunately, widespread adoption of numerous small models is unlikely since it is far less efficient than maintaining one single large model. Moreover, performance in low-resource languages can still benefit from cross-lingual transfer in MLLMs, where the model is able to leverage similarities between languages.

Extending on this, David Stap and Christof Monz investigate how expanding the range of languages used during fine-tuning affects performance—a topic that has produced conflicting results in prior research (Stap and Monz, 2025). They provide clarity by fine-tuning translation models and evaluating performance on translation directions (source-to-target pairs). They find that the more language directions included during fine-tuning, the better the performance for both seen and unseen pairs. Surprisingly, they also observe that models trained on the most diverse sets outperform specialized models. They attribute this trend to "language-agnostic representations in middle layers" of the most diverse models, meaning there are more shared, universal features that generalize across languages (Stap and Monz, 2025).

This surprising result aligns with broader evidence that diversity, rather than sheer data volume, is often the primary driver of strong cross-lingual performance. Recent work shows that multilingual models exposed to typologically varied training signals develop more language-agnostic internal representations and exhibit stronger zero-shot transfer to unseen language (Johnson et al., 2017). In these studies, expanding the range of linguistic structures leads to more stable and universal features in the model's middle layers. Motivated by this concept, our project explores existing benchmarks for fine-tuning beyond translation. Researchers at the University of Washington, Google DeepMind, and the Allen Institute for AI developed BUFFET (Benchmark of Unified Format Few-shot Transfer Evaluation) to benchmark few-shot cross-lingual transfer [1]. It contains 15 datasets covering 54 typologically diverse languages, unified to the same text-to-text format to make comparisons across models and languages consistent. The benchmark assesses the LLM's ability to perform a wide range of tasks, including classification, generation, extraction, and structured prediction.

To meaningfully test hypotheses about the role of linguistic diversity in cross-lingual transfer, however, researchers need ways to introduce controlled diversity in language. One recent line of work is ConlangCrafter, a system designed to algorithmically generate coherent artificial languages through a multi-stage pipeline (Johnson et al., 2017). The framework decomposes language construction into modular components – phonology, morphology, syntax, lexicon generation, and translation – mirroring the structure of natural linguistic systems. At each stage, the model injects controlled randomness to encourage typological diversity while incorporating self-refinement mechanisms to preserve internal consistency.

ConlangCrafter provides a useful foundation for our work because it offers fine-grained control over linguistic features while ensuring that conlangs remain consistent and interpretable. Importantly, this framework is particularly promising for low-resource contexts: many underrepresented languages are documented primarily through written grammars rather than large corpora, and the modular design of ConlangCrafter aligns naturally with this form of linguistic specification.

Building on these ideas, our work investigates whether synthetic languages can act as scalable substitutes for missing linguistic diversity in multilingual datasets. Whereas prior work has focused on using conlangs as proxies for typological study or for modeling human language learning, we extend the concept into the domain of multilingual fine-tuning. Specifically, we examine whether fine-tuning on conlangs can improve performance on tasks within BUFFET's unified-format evaluation suite, particularly for low-resource natural languages.

## 3 Methodology

Our methodology consists of four components: (1) constructing a persistent, coherent synthetic language, (2) translating BUFFET task data into that language to create a synthetic training corpus, (3) extracting typological feature vectors and computing similarity/diversity metrics, and (4) fine-tuning multilingual models on the resulting datasets.

### 3.1 Synthetic Language Construction

ConlangCrafter is an LLM-based system capable of generating individual components of a constructed language—phonology, lexicon, grammar, and translations—depending on which steps are requested. However, all chosen components are pro-

duced in a single model call, and the system cannot refine or extend the language across calls. To support corpus-scale generation, we restructure ConlangCrafter into a persistent, multi-stage pipeline that separates language creation from translation and enables gradual expansion of the linguistic specification.

We first generate the initial language specification: a phonological system, a seed lexicon, and a preliminary grammar sketch. These outputs are stored as phonology.txt, lexicon.csv, and grammar.txt and serve as the persistent state of the synthetic language. At this stage we do not generate translations; the language is allowed to evolve through subsequent steps. Then, we create two classes of synthetic languages depending on the experimental condition.

#### 3.1.1 Diverse Language Class

First, we generate unconstrained random languages, where typological settings are sampled from ConlangCrafter's built-in feature checklist. This produces a diverse set of synthetic languages intended to maximize structural variation.

#### 3.1.2 Target-Guided Language Class

Second, we generate target-guided synthetic languages designed to mimic the typological profile of a specific low-resource language in BUFFET. To do so, we extract a set of core structural features—such as word order, morphological alignment, clause structure, and nominal morphology—from the WALS database (Dryer and Haspelmath, 2013). These features correspond to the subset identified by the ConlangCrafter authors as capturing the most linguistically consequential dimensions of cross-linguistic variation (**?**). We then pass a custom constraint through all language-generation prompts, requiring the resulting synthetic language to match these features.

#### 3.1.3 Pre-Stabilization and Feezing

Both types of languages are then refined using the same pre-stabilization translation process. During pre-stabilization, the model translates small batches of English sentences while the lexicon and grammar remain editable. The specialized pre-stabilization prompt encourages reuse of existing grammatical resources and limits rule creation to cases where no existing construction suffices. Newly introduced lexical items or grammar rules are extracted from each batch and appended to the persistent language files, allowing the language to reach the expressive depth needed for the downstream tasks. The language is considered stable once the introduction of new grammar rules becomes negligible across a sliding window of batches. At this point, we freeze grammar.txt and proceed with full dataset translation.

### 3.2 Synthetic Dataset Construction

Once the language is stabilized, we construct the synthetic dataset by translating the BUFFET NLI and PAWS-X tasks into the synthetic language while preserving their seq2seq formats. Each English input-output pair receives a corresponding conlang version, and all translations adhere to the frozen grammar. The final dataset consists of this translated corpus, the expanded lexicon, and the stabilized grammar file.

### 3.3 Typological Feature Extraction and Distance Computation

After constructing and stabilizing each synthetic language—and completing full dataset translation so that no new lexical items or grammatical rules will be introduced—we extract a typological feature vector representing the structural properties of the final language. To obtain this vector, we feed the persistent language files into a dedicated prompt that asks ConlangCrafter to summarize the language's core typological settings. The prompt returns a WALS-style feature vector aligned with the subset of structural features mentioned in 3.1.2.

We compute typological distances between languages using Hamming distance, defined as the number of feature mismatches between two vectors. Distances are normalized by the length of the feature vector and converted into a similarity score:

$$\text{sim}(l, s) = 1 - \frac{\text{Ham(l,s)}}{N}$$

where $\text{Ham}(l, s)$ is the Hamming distance between real language $l$ and synthetic language $s$, and $N$ is the number of typological features. This similarity measure corresponds to the percentage of shared typological features between the two languages.

To characterize how typologically varied a set of synthetic languages is, we compute a diversity score following the definition used in prior ConlangCrafter work. Given synthetic languages $s_1, .., s_k$, we compute the mean pairwise normalized Hamming distance:

$$D_{\text{mean}} = \frac{2}{k(k-1)} \sum_{i<j} \text{Ham}(s_i, s_j)$$

which reflects the average structural dissimilarity across the synthetic-language set. Higher values of $D_{\text{mean}}$ indicate greater diversity.

Both the similarity measure $\text{sim}(l, s)$ and the diversity score $D_{\text{mean}}$ are used in Experiments 1 and 2 to analyze how typological structure relates to cross-lingual transfer performance.

### 3.4 Fine-tuning Setup

We fine-tune the mT5-base model (580M parameters), a multilingual encoder–decoder architecture covering over 100 languages. We apply LoRA adapters to the self-attention projection matrices, enabling low-rank parameter updates while keeping the pretrained backbone frozen. This approach substantially reduces compute cost and mitigates catastrophic forgetting: because the underlying multilingual representations remain fixed, language-specific drift is confined to the LoRA parameters, preserving mT5's pretrained cross-lingual alignment throughout fine-tuning.

For each experiment, we translate the full English training split of XNLI or PAWS-X into the relevant synthetic languages and fine-tune mT5 jointly on English and these synthetic-language translations. We follow BUFFET and prior multilingual fine-tuning work in including English in the mixture since high-resource languages act as anchors for cross-lingual alignment (Conneau et al., 2020; Arivazhagan et al., 2019; Hu et al., 2020).

Because these training sets remain large—consisting of tens to hundreds of thousands of examples per language—we follow BUFFET's large-scale English fine-tuning protocol and train the model for 3 epochs. This provides a consistent and appropriate training budget while avoiding the extreme overfitting associated with the 200–300-epoch few-shot schedules. We use AdamW with a constant learning rate and a short warm-up phase, and select checkpoints based on English dev-set accuracy to avoid biasing toward any particular synthetic language.

## 4 Experiments

### 4.1 Research Questions

Our experiments investigate whether synthetic languages can improve multilingual transfer. We focus on four questions: (1) whether fine-tuning on synthetic languages enhances cross-lingual transfer; (2) whether such improvements are concentrated in low-resource languages; (3) how synthetic-language fine-tuning compares to the few-shot target-language fine-tuning strategy used in BUFFET; and (4) how typological factors—such as similarity to target language or diversity among synthetic languages—correlate with transfer effectiveness.

### 4.2 Experimental Setup

**Model** All experiments use the mT5-base model; full fine-tuning details are provided in Section 3.3.

**Tasks and Datasets** We evaluate on two multilingual classification benchmarks containing both English and low-resource test splits:

- **XNLI**: a three-way natural language inference task (entailment, contradiction, neutral).
- **PAWS-X**: a binary paraphrase-detection task determining whether two sentences express the same meaning.

Both datasets are established cross-lingual transfer benchmarks and demonstrated sizable improvements in BUFFET, making them well suited for examining synthetic-language effects.

**Evaluation Metrics** We report overall accuracy (averaged across all languages), low-resource accuracy (average across low-resource languages), and per-language accuracy. To analyze transfer mechanisms, we additionally compute correlations between model performance and (1) typological similarity between synthetic and real languages, and (2) diversity within the synthetic-language set.

### 4.3 Baselines

We compare synthetic-language fine-tuning to two baselines.

- **Random classification**: yields approximately 0.33 accuracy for XNLI and 0.50 for PAWS-X across all metrics (overall, low-resource, per-language), providing a sanity-check lower bound.
- **BUFFET (English + Target FT)**: fine-tuning on English followed by few-shot examples in each target language achieves $\sim$0.52 accuracy on XNLI and $\sim$0.78 on PAWS-X.

We reproduce BUFFET results for all languages, additionally computing accuracy for low-resource languages and per-language, enabling a direct comparison to synthetic-language methods.

### 4.4 Experiment 1: Diverse Synthetic Languages

This experiment tests whether typological diversity in synthetic languages promotes general cross-lingual transfer without requiring any real low-resource data. We construct five synthetic languages designed to span a wide typological space using the generation pipeline described in Section 3.1 and fine-tune the model jointly on English and these diverse synthetic languages (see Section 3.4).

We evaluate the resulting model on all languages in XNLI and PAWS-X and compare its performance to both baselines. To analyze transfer mechanisms, we examine two typological factors:

**Similarity to synthetic languages** For each real language $l$ and synthetic language $s$, we summarize similarity using:

- Max similarity: the highest $\text{sim}(l, s)$ across the five synthetic languages, capturing whether at least one synthetic language is a close match to $l$.
- Mean similarity: the average $\text{sim}(l, s)$ across all five synthetic languages, capturing how well the synthetic set collectively covers $l$.

We correlate these metrics with per-language accuracy improvements over the random classification baseline.

**Diversity of the synthetic set** Following the ConlangCrafter definition, we compute the mean pairwise normalized Hamming distance between all synthetic-language pairs $D_{mean}$ (Section 3.3). We then test whether higher diversity is associated with stronger overall transfer.

### 4.5 Experiment 2: Target-Guided Synthetic Languages

This experiment tests whether synthetic languages tailored to a specific target language can leverage shared language features for cross-lingual transfer. For each low-resource target language (e.g. Swahili, Urdu, Thai, Vietnamese, Hindi), we generate a synthetic language constrained by its typological profile (Section 3.1.2). This approach leverages the fact that structural and typological features of low-resource languages are far easier to obtain than real annotated corpora. Moreover, typological features exist for thousands of languages—including extremely low-resource ones—via databases such as WALS (Dryer and Haspelmath, 2013), while most languages lack even basic digital corpora (Joshi et al., 2020; Mager et al., 2018; O'Horan et al., 2016). Each model is fine-tuned jointly on English and the guided synthetic language corresponding to its target (Section 3.4).

Whereas Experiment 1 examines whether incidental typological similarity predicts transfer gains, Experiment 2 examines the effect of deliberately engineered similarity. Because guided synthetic languages are constructed to match their target's feature vector, similarity is near-maximal by design. We therefore use the similarity score (Section 3.3) as a sanity check.

Evaluation mirrors Experiment 1 but focuses on per-target comparisons. For each target language $l$, we compare accuracy from its guided-synthetic model against (1) the diverse-synthetic model from Experiment 1 and (2) the BUFFET English+Target few-shot baseline. This allows us to test whether enforced typological similarity can function as a cost-effective alternative to obtaining real few-shot examples.

Finally, we examine spillover effects by testing whether non-target languages that share typological features with the guided synthetic language also benefit. Because each model here is trained on only one synthetic language, spillover patterns are much easier to attribute than in Experiment 1, where multiple synthetic languages influence outcomes

## 5 Results Interpretation

Across both experiments, improvements over the random classification baseline would indicate that synthetic corpora provide meaningful supervision rather than noise. Comparisons against BUFFET determine whether synthetic languages can approach or replace real few-shot target-language data.

Experiment 1: Correlation between per-language accuracy and both max and mean similarity suggest that incidental typological overlap between synthetic and real languages can support transfer. Finally, diversity analyses will indicate whether broader typological coverage within the synthetic set is associated with stronger performance.

Experiment 2: Gains over the Experiment 1 model suggest that typological targeting is more effective than diversity-driven synthetic-language construction. With only one synthetic language per model, correlations between accuracy and similarity per language offer a clean test of whether shared typological features drive transfer.

## References


Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. Massively multilingual neural machine translation in the wild. In *Proceedings of ACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

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

Jiajie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual benchmark for cross-lingual generalization. In *Proceedings of ICML*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeff Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Pratik Joshi, Danish Pruthi, and Chris Dyer. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of ACL*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technology for low-resource indigenous languages. In *Proceedings of LREC*.

Helen O'Horan, Geeticka Chauhan, Ivan Vulić, Ryan Cotterell, and Diana McCarthy. 2016. A survey of typological information for natural language processing. In *Proceedings of COLING*.

David Stap and Christof Monz. 2025. The effect of language diversity when fine-tuning large language models for translation. *Preprint*, arXiv:2505.13090.