Harnessing Linguistic Dissimilarity for Language Generalization on Unseen Low-Resource Varieties

Jinju Kim[ㄱ,ㄴ] **Haeji Jung**[ㄴ,ㄷ] **Youjeong Roh**[ㄴ,ㄹ]
Jong Hwan Ko[ㄱ] **David R. Mortensen**[ㄴ]
[ㄱ]Department of Electrical and Computer Engineering, Sungkyunkwan University
[ㄴ]Language Technologies Institute, Carnegie Mellon University
[ㄷ]Department of Computer Science, University of British Columbia
[ㄹ]Electronics and Telecommunications Research Institute
Correspondence: perla0328@g.skku.edu dmortens@cs.cmu.edu

Abstract

Low-resource language varieties used by specific groups remain neglected in the development of Multilingual Language Models. A great deal of cross-lingual research focuses on inter-lingual language transfer which strives to align allied varieties and minimize differences between them. However, for low-resource varieties, linguistic dissimilarity is also an important cue allowing generalization to unseen varieties. Unlike prior approaches, we propose a two-stage Language Generalization framework that focuses on capturing variety-specific cues while also exploiting rich overlap offered by high-resource source variety. First, we propose TOPPing, a source-selection method specifically designed for low-resource varieties. Second, we suggest a lightweight VAÇAÍ-Bowl architecture that learns variety-specific attributes with one branch while a parallel branch captures variety-invariant attributes using adversarial training. We evaluate our framework on structural prediction tasks, which are among the few tasks available, as proxy for performance on other downstream tasks. Using VAÇAÍ-Bowl with TOPPing yields an average 54.62% improvement in the dependency parsing task, which serves as a proxy for performance on other downstream tasks across 10 low-resource varieties.[1]

1 Introduction

Cross-lingual transfer methods overwhelmingly assume that aligning representations across languages is universally beneficial. We show it is not. Multilingual Language Models (MLMs) have extended language technologies to more than one hundred languages (Pires et al., 2019; Conneau et al., 2020), yet excelling in a language offers little meaning if a model cannot generalize to its variants (e.g., regional dialects). Crucially, a large portion of real-world communication is carried out in language

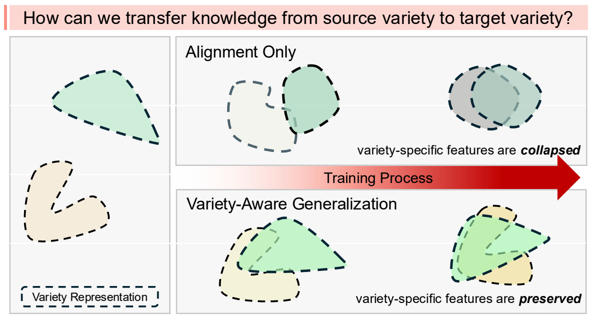

Figure 1: Visualization of training in the embedding space, comparing alignment-based only method and variety-aware generalization method.

varieties not represented among the roughly one hundred languages that dominate MLM training data. Decades of research in linguistics show that languages lay on a continuum of similarity, each with its own continuous domain (Lin et al., 2019), and when confronted with such intra-linguistic variants, existing models deliberately fail (Faisal et al., 2024). We find that alignment-only methods exacerbate this problem by collapsing variety-specific structure that models need to generalize (Figure 1).

Intra-linguistic variation is at least as pervasive as inter-linguistic variation. Linguists often refer to intra-language variants as "dialects" or "sociolects". In this paper, we avoid the term *dialect* for two reasons: (i) it can carry pejorative connotations, and (ii) it is unduly restrictive, implying mutual intelligibility with other variants of a language. We therefore adopt the more neutral and inclusive term *language variety* (or *variety*), that broadly denotes variants shaped by regional, social, and cultural distinctions of its speaker community (Chambers and Trudgill, 1998).

Prior research aiming to extend NLP to low-resource varieties, including interlingual transfer, has primarily centered on exploiting cross-variety similarities to align representations (e.g., Figure 1's variety-aligned only method) to facilitate knowl-

[1]Work done while the authors Jinju, Haeji and Youjeong were visiting Carnegie Mellon University.

edge transfer (Yang et al., 2022). However, alignment methods risk losing variety-specific information by collapsing distinct features. This overlooks the linguistic variations that naturally arise in real-world contexts. In this paper, instead, we propose a variety-aware Language Generalization framework that performs generalization without any training on the target low-resource variety. By learning not just *how it is similar to a high-resource variety*, but *how it is different*, the model learns to disentangle and strategically combine linguistic features to perform in unseen settings. We also propose an improved automatic method for identifying high-resource varieties most relevant for training a model targeting a particular low-resource variety without any usage of labels, annotation, or parallel dataset. Together, these methods achieve better results than all baselines on structural tasks including dependency parsing (DEP) and part-of-speech tagging (POS), which we believe to be an informative proxy for performance on other downstream tasks.

Our key contributions are as follows:

- This paper introduces the Language Generalization framework, focusing on making a model robust to unseen language variations.
- We introduce TOPPing, a method for selecting source varieties to generalize on a target low-resource variety without annotations or parallel dataset.
- We propose VAÇAI-Bowl, a novel and lightweight architecture to not only align, but also distinguish varieties.

2 Related Work

2.1 Low-Resource Varieties

The disparity of MLMs performing significantly worse in low-resource varieties arises even when the variety is typologically close to, and partially represented in, the training corpus. Through empirical studies, drops in performance when data shifts to a low-resourced variant have been proven to be biased towards dominant varieties (Blasi et al., 2022; Blaschke et al., 2024, 2023; Faisal et al., 2024; Srivastava and Chiang, 2025; Lin et al., 2025; Üstün et al., 2020).

Given this limitation, recent approaches leverage closely related high-resource varieties to support lower-resourced counterparts (Snæbjarnarson et al., 2023; Bafna et al., 2024). Bafna et al. (2025) train on artificially generated variants or adapt inputs at inference time, while Nguyen et al. (2025) apply test-time adaptation to bridge the dialectal gap.

2.2 Zero-shot Cross-lingual Transfer

Prevailing methods to improve cross-lingual transferability without explicit training on the target variety focus on aligning representations via token-level self-augmentation (Wang et al., 2023), manifold mixup of parallel sentences (Yang et al., 2022), graph-based embedding re-parameterization (Wu and Monz, 2023), or robust training with randomized smoothing (Huang et al., 2021). All treat variety-specific information as noise to be suppressed rather than a potential source of transfer knowledge.

Another common strategy is to train models on source languages that are linguistically similar to the target language. Prior work investigating the factors that influence transfer performance has shown that linguistic similarity tends to correlate with better cross-lingual transfer (Eronen et al., 2023a,b; Lauscher et al., 2020; Dufter and Schütze, 2020). This has motivated efforts to identify optimal source languages using various linguistic similarity metrics. For example, de Vries et al. (2022) examine part-of-speech tagging across diverse source-target language pairs and suggest optimal pairs for certain languages, and Lin et al. (2019) proposes a ranking method based on multiple linguistic similarity features, offering a more systematic framework with quantitative features. Rice et al. (2025) utilizes typological and dataset-dependent features to conclude the best rankings for cross-lingual performances. However, these rely on pre-annotated language information, making it inapplicable for varieties that lack annotated datasets.

Collectively, these methods operate under the implicit assumption that full alignment is universally beneficial. However, this assumption is rarely examined, and existing work provides limited justification for why suppressing variety-specific structure should be the default strategy. Our work directly investigates this gap.

2.3 Domain Generalization

Domain Generalization (DG) aims to ensure model performance on domains inaccessible during training (Blanchard et al., 2011; Muandet et al., 2013; Zhou et al., 2023), typically by learning domain-invariant features (Muandet et al., 2013; Li et al.,

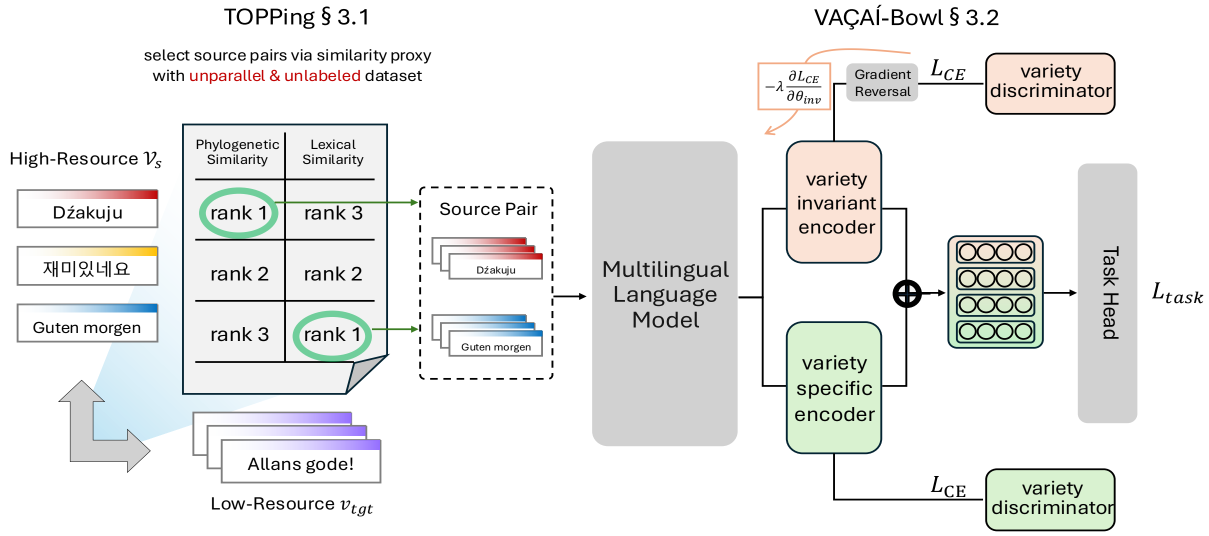

Figure 2: Overall framework of this paper. Using TOPPing, with just unparallel and unlabeled datasets, we can select source varieties with not only shared but also distinctive features to capture relationship between varieties. From the obtained source variety pair, VA learns the semantic differences of neighboring source varieties and learns to generalize on the target low-resource variety under unseen setting.

2018a,b), an approach also adopted in NLP (Wang et al., 2024, 2021; Li et al., 2024).

In multilingual NLP, prior work treats each language as a separate domain (Jung et al., 2024), ensembles adapters for zero-shot generalization (Rathore et al., 2023), or leverages typological features (Adilazuarda et al., 2024). Most focus on invariant features (Ganin et al., 2016; Ngo and Nguyen, 2024; Tahery et al., 2024); we aim to capture domain-specific signals to help the model recognize linguistic differences within similar varieties—thereby improving generalization to unseen varieties.

3 Methods

In this paper, we propose a novel framework that leverages both variety-invariant and variety-specific information to generalize to unseen low-resource variety, and a method to carefully select exploitable high-resource source variety to pair with the low-resource target variety.

3.1 TOPPing

For low-resource language varieties, obtaining sufficient labeled training data remains expensive and labor-intensive (Blasi et al., 2022; Faisal et al., 2024). When the low-resourced target variety has high-resourced neighbors in terms of linguistic similarity, utilizing the latter can be relatively cheaper. To mitigate the constraints of low-resource varieties, we introduce **TOPPing**: **T**oken-**O**verlap & **P**roximal embedding **P**air**ING**, a simple yet effective method that does not require annotations and allows preservation of varietal diversity for selecting source varieties. This allows a wide and general usage even when the variety is unseen, unannotated, and unlabeled. Previous works define language distances in various dimensions (Littell et al., 2017; Rama et al., 2020). In this work, we rely on two signals that can be computed automatically from raw text and capture complementary aspects of similarity. As illustrated in Figure 2, to encourage diversity in source selection, we compute these similarity signals independently rather than combining them into a single joint score. This preserves the room for variety-specific information training which serves as a key for generalization.

Let v_{tgt} be the target low-resource variety, and $\mathcal{V}_{\text{src}}$ the set of high-resource source varieties. For each variety v, let $\mathbf{X}_v = \{x_i^v\}_{i=1}^{N_v}$ denote its dataset. First, we obtain a source variety by proxying the phylogenetic distance via embedding distance (Rama et al., 2020). Let $f_{\text{CLS}_2}(x) \in \mathbb{R}^d$ denote the "[CLS]" representation of input x from the second layer of a MLM. We aim to obtain a source variety $v_{\text{sim}} \in \mathcal{V}_{src}$ such that:

$$v_{\text{sim}} = \arg\min_{v \in \mathcal{V}_{src}} \left\| \mu_v - \mu_{v_{tgt}} \right\|_2 , \quad \mu_v = \frac{1}{N_v} \sum_{x \in \mathbf{X}_v} f_{\text{CLS}_2}(x). \tag{1}$$

We use the second-layer "[CLS]" rather than the final layer because lower-layer representations have been shown to capture more typological and morpho-syntactic information, which aligns better to proxy phylogenetic structure (Hewitt and Manning, 2019; Mousi et al., 2024; Bakos et al., 2025).

Second, we use token overlap (Blaschke et al., 2025) between two varieties as a proxy for lexical distance. Here, we introduce **token-length weighted Jaccard Similarity** as lexical overlap calculation tailored to low-resource varieties. Unlike highly represented varieties, lexical items are often fragmented into shorter sub-tokens, which diminishes the discriminative power of a standard Jaccard similarity measure. Weighting overlaps by token length mitigates this bias, preventing varieties from being erroneously conflated based on scripts. The aim is to obtain a source variety $v_{\text{overlap}} \in \mathcal{V}_{src}$ such that :

$$v_{\text{overlap}} = \underset{v \in \mathcal{V}_{src}}{\arg\max}\ \mathrm{TJ}(\mathbf{X}_v, \mathbf{X}_{v_{tgt}}), \tag{2}$$

where $\mathrm{TJ}(\mathbf{X}_v, \mathbf{X}_{v_{tgt}})$ is token-length weighted Jaccard similarity.

$$\begin{aligned} \mathrm{TJ}(\mathbf{X}_a, \mathbf{X}_b) &= \frac{\sum_{tok \in T_a \cap T_b} \omega(tok)}{\sum_{tok \in T_a \cup T_b} \omega(tok)}, \\ \omega(tok) &= \max(1, \mathrm{len}(tok) - 1). \end{aligned} \tag{3}$$

The pair $\langle v_{\text{sim}}, v_{\text{overlap}} \rangle$, selected by independent ranking, leaves room for diversity in source pair selection.

3.2 VAÇAI-Bowl

In Figure 2, we illustrate our approach for Language Generalization : **V**ariety **A**ligned and Spe**C**(**Ç**)ific **A**ttr**I**butes **B**lending for **LOW**-resouce **L**anguage Varieties. This framework leverages both variety-invariant and variety-specific knowledge from high-resource varieties to effectively model representations for an unseen, low-resource variety.

In order to model variety-invariant and variety-specific features, we use a Multilingual Language Model that produces a "[CLS]" embedding for every input sentence. We implement two independent 2-layer MLP encoders :

- Variety-invariant encoder f_{inv} is trained adversarially to align varieties and learn invariant features. We obtain $h_{\text{inv}} = f_{\text{inv}}([CLS])$.
- Variety-specific encoder f_{spc} is trained normally to emphasize variety-specific features. We can also obtain $h_{\text{spc}} = f_{\text{spc}}([CLS])$.

The outputs are concatenated into h, so that $h = h_{\text{inv}} \parallel h_{\text{spc}} \in \mathbb{R}^d$, where each encoder maps the d-dimensional "[CLS]" embedding to a $d/2$-dimensional vector (i.e., 768→384 each), so that the concatenated representation matches the original dimensionality. The joint feature h is used in place of the original `[CLS]` embedding for downstream tasks.

Our choice of an adversarial objective is motivated by two considerations. First, adversarial alignment is the predominant mechanism for enforcing invariance in both cross-lingual transfer (Wang et al., 2023; Huang et al., 2021) and domain generalization. (Ganin et al., 2016; Ngo and Nguyen, 2024). Using the same mechanism allows us to make controlled comparisons with this prior work. Second, our research goal is not to propose a new alignment strategy, but to test whether complementing invariant representations with explicitly modeled variety-specific features improves generalization. The adversarial loss on f_{inv} ensures that variety-specific signals are excluded from the invariant branch, allowing us to directly evaluate the contribution of the variety-specific encoder f_{spc}.

To train each encoder to extract variety-invariant and variety-specific features, each encoder is paired with its own discriminator (D_{inv} and D_{spc}) that performs classification on what variety the input belongs to. A gradient-reversal layer G_λ is inserted in front of D_{inv} to selectively update parameters to fool D_{inv} (Ganin and Lempitsky, 2015).

$$G_\lambda(z) = z, \qquad \frac{\partial G_\lambda}{\partial \mathbf{z}} = -\lambda \mathbf{I}, \tag{4}$$

where λ is a hyperparameter. Contrastingly, f_{spc} learns to help D_{spc} by producing easily distinguishable features. The discriminators yield two loss terms :

$$\begin{aligned} L_{\text{inv}} &= L_{CE}(D_{\text{inv}}(G_\lambda(h_{\text{inv}})), y_{\text{var}}), \\ L_{\text{spc}} &= L_{CE}(D_{\text{spc}}(h_{\text{spc}}), y_{\text{var}}), \end{aligned} \tag{5}$$

where L_{CE} denotes Cross Entropy Loss.

Lastly, the task loss is employed for the fine-tuning objective and to ground the representation extraction in right directions.

$$L_{\text{task}} = L_{\text{task}}(f_{\text{task}}(h), y_{\text{task}}). \tag{6}$$

Finally, the objective function is defined as $L_{\text{total}} = L_{\text{inv}} + L_{\text{spc}} + L_{\text{task}}$.

4 Experiments

In the following section, we provide experiments designed to evaluate the effectiveness of (i) source selection method TOPPing, (ii) framework VAÇAÍ-Bowl. We further investigate (iii) *why linguistic dissimilarities contribute to transfer, and to what extent?* For our experiments, we utilize structured prediction tasks to proxy MLM performance on downstream tasks.

4.1 Experimental Setup

Benchmark. DialectBench provides datasets and benchmarks for low-resource varieties with annotations that group varieties into language clusters, allowing direct visualization of performance gaps within the same cluster. For our experiment, we select 10 target low-resource varieties from DialectBench that have no training dataset available for the variety. For source varieties, we utilize the rest of DialectBench and sample high-resourced varieties representative of distinctive language clusters from Universal Dependencies (Nivre et al., 2017; de Marneffe et al., 2021). These source variety sets are used for TOPPing variety selection. We evaluate on dependency parsing (DEP) and part-of-speech tagging (POS) task, which both are structured prediction tasks. We evaluate DEP using Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS), which measure correctness of predicted head and head+relation label, respectively. For POS tagging, we report token-level F1. Please refer to Appendix C for detailed results of downstream tasks across varieties.

Source Selection Baselines. In Figure 1, we illustrate how automated source language selection using TOPPing can be a simple yet effective method for source language selection especially for unseen varieties. This selection is applicable to diverse cross-lingual transfer scenarios, not limited to a specific method. As a widely-used baseline, we implement **LangRank** where the source varieties are selected based on pre-annotated linguistic features and dataset-dependent features (Lin et al., 2019). For a fair comparison, we use the full set of available LangRank features to obtain its best performance. Please refer to Appendix B for detailed information on selected source varieties for each target variety.

Language Generalization Baselines. To compare the VAÇAÍ-Bowl framework, we implement two baselines : (1) **MLM** baseline illustrates performance when the model is simply finetuned on source languages. (2) **Alignment** baseline where the model leverages adversarial training to learn only variety-invariant features, adopted from DG and robust training is implemented (Huang et al., 2021).

Implementation Details. We utilize mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020) as MLMs for all tasks. We use the two models in their base size, where mBERT has 110M and XLM-R has 125M number of parameters. Specifically, we utilize publicly available models—`bert-base-multilingual-cased` for mBERT and `xlm-roberta-base` for XLM-R—downloaded from Huggingface.[2] We set the learning rate as 2e-4, batch size as 64 for mBERT. We set the learning rate as 5e-5, batch size as 64 for XLM-R. Overall, λ for gradient-reversal layer is searched in [0.1, 0.5, 1.0] for the following experiments. Parameters are optimized with Adam optimizer (Kingma and Ba, 2015). We finetune each model for 10 epochs and halt at step size of 1000 to not exceed the finetuning steps of zero-shot cross-lingual steps. In accordance with the benchmark evaluation procedure, results are reported from the checkpoint obtaining the highest UAS for the DEP task. Please refer to Appendix D for detailed information on parameter search regarding lambda.

4.2 Quantitative Results

In accordance with the previous discussions, a model that can learn both the variety-invariant and variety-specific features should show higher generalization performance regardless of source varieties. Also, this performance should be boosted with model-agnostic source selection TOPPing that preserves noticeable differences in source varieties. At the same time, TOPPing is also expected to perform comparatively to LangRank which prioritizes linguistic similarities. For scores across LAS evaluation metric, please refer to Appendix C.1.

Source Selection. Table 1 and Table 2 presents results on DEP task, using mBERT and XLM-R as backbones, respectively. Comparing the scores reported using each LangRank and TOPPing, it is noticeable that in Table 1, evaluations made using TOPPing outperforms LangRank across all methods in 9 out of 10 varieties. Also, simply finetuning

[2] https://huggingface.co/models

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source is eng										
mBERT◇	38.14	13.51	8.95	26.12	26.89	50.22	36.77	19.41	36.77	33.21
mBERT	39.13	13.03	12.91	30.03	29.79	49.86	42.61	20.81	42.49	32.01
source selected using LangRank (Lin et al., 2019)										
mBERT	43.90	22.13	10.47	33.97	32.51	59.38	46.45	27.79	51.12	36.14
+Alignment	49.58	25.98	16.40	36.33	33.37	59.68	50.49	29.90	52.60	34.67
+VAÇAÍ-Bowl (OURS)	**51.00**	27.05	17.21	36.90	35.32	63.02	50.92	32.62	52.23	36.29
source selected using TOPPing (OURS)										
mBERT	44.55	34.10	15.18	40.83	36.80	63.99	52.54	35.63	57.22	37.21
+Alignment	45.30	31.56	16.53	40.72	36.52	62.96	52.04	38.80	54.84	35.99
+VAÇAÍ-Bowl (OURS)	46.34	**36.39**	**19.00**	**42.29**	**38.19**	**64.29**	**54.90**	**39.67**	**57.74**	**37.67**

Table 1: Quantitative results on UAS scores using mBERT as backbone on dependency parsing task evaluated across low-resource varieties from Faisal et al. (2024). ◇ refers to value reported in original paper. Underlined refers to best performing on target under controlled source variety. **Bold** refers to best performing on the target variety.

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source is eng										
XLM-R◇	43.50	11.15	4.23	30.91	32.14	43.78	34.70	28.28	34.70	28.75
XLM-R	52.58	13.61	4.38	31.50	30.60	53.79	42.92	30.20	43.60	24.50
source selected using LangRank (Lin et al., 2019)										
XLM-R	55.58	28.44	10.95	41.51	33.27	60.37	47.36	41.25	43.75	33.08
+Alignment	57.41	28.44	12.47	40.49	35.80	59.51	47.61	42.16	46.95	34.30
+VAÇAÍ-Bowl (OURS)	**58.55**	31.23	12.51	42.41	36.13	59.82	48.38	42.53	47.47	34.56
source selected using TOPPing (OURS)										
XLM-R	55.07	30.57	11.27	40.50	**38.04**	**63.80**	48.73	40.76	52.23	35.68
+Alignment	56.54	28.67	8.47	39.60	35.85	63.13	50.81	40.20	56.62	34.76
+VAÇAÍ-Bowl (OURS)	57.50	**31.97**	**13.98**	**44.66**	37.76	63.44	**51.65**	40.99	**57.74**	**36.60**

Table 2: Quantitative results on UAS scores using XLM-R as backbone on dependency parsing task evaluated across low-resource varieties from Faisal et al. (2024). ◇ refers to value reported in original paper. Underlined refers to best performing on target under controlled source variety. **Bold** refers to best performing on the target variety.

the mBERT model on TOPPing itself beats the best score obtained using LangRank for 8 out of 10 varieties. In Table 3, this phenomenon is more apparent across additional evaluation metrics and task. TOPPing achieves higher performance enhancement for low-resource varieties overall, for both mBERT and XLM-R.

In cases where LangRank enhances transferability and generalization, Gheg Albanian (aln) and Umbrian (xum), it is notable that the target varieties all fall into Indo-European family.

Task	Metric	LangRank	TOPPing
DEP	UAS	7.48	**10.27**
	LAS	7.82	**10.01**
POS	F1	7.36	**9.34**

Table 3: Comparative results of source selection methods for each task. Each value denotes the *average of absolute improvement in score (in points)* for both mBERT and XLM-R across all methods, compared to the English fine-tuned baseline.

Language Generalization. Table 4 illustrates that our proposed architecture, VAÇAÍ-Bowl, improved model performance regardless of source selection across different evaluation schemes. Specifically, in Table 1 VAÇAÍ-Bowl achieves the highest performance on 9 out of 10 target varieties across both source selection methods for mBERT. Paying close attention to other baselines, it is notable that the Alignment method, which attempts to enforce alignment by pulling diverse variety embeddings together, fails to surpass the performance on fine-tuned MLM baseline for certain varieties. Specifically, for varieties {*gug, koi, kpv, lij, nds, gsw, xum*} in Table 1 and {*gug, gun, koi, kpv, lij, sma*} in Table 2. We refer to this phenomenon as **alignment-induced failures**. When this occurs, VAÇAÍ-Bowl overcomes the failures of alignment by utilizing variety-specific attributes in all cases. Especially in Table 1, for 6 out of 7 alignment-induced fails ob-

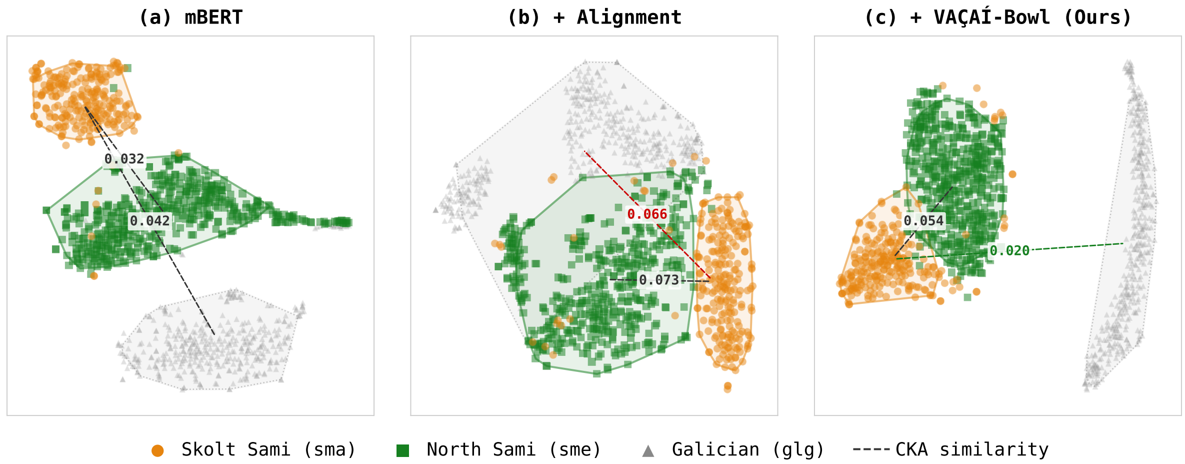

Figure 3: t-SNE visualization of sentence embeddings for Skolt Sami (sma), North Sami (sme), and Galician (glg) from final representations. (a) Pre-trained mBERT separates all three varieties with low inter-variety similarity. (b) Alignment indiscriminately collapses all representations, including the unrelated Galician. (c) VAÇAÍ-Bowl selectively increases similarity between related varieties while pushing the unrelated Galician further apart.

Task	Metric	MLM	Alignment	VAÇAÍ-Bowl
DEP	UAS	7.84	8.58	**10.21**
	LAS	8.50	8.67	**9.58**
POS	F1	8.50	7.73	**8.83**

Table 4: Comparisons of methods for each task. Each value denotes the *average of absolute improvement in score (in points)* for both mBERT and XLM-R across all methods, compared to the English fine-tuned baseline.

served using TOPPing, VAÇAÍ-Bowl outperforms all methods on the target variety. Also, VAÇAÍ-Bowl performs consistently better than Alignment method across all varieties.

Overall, results presented across low-resource varieties across backbones, tasks, and source selection methods show promising results of VAÇAÍ-Bowl. It can improve model performance up to 27% (the case of *gug* with LangRank as source selection) of simply finetuning the MLM. We additionally conduct ablation experiments on each component of the loss function, confirming that both $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{spc}}$ contribute to generalization performance (Appendix E).

4.3 Qualitative Results

To examine how each training objective reshapes the representation space, we compute pairwise CKA similarity (Kornblith et al., 2019) between sentence embeddings of three varieties: Skolt Sami (sma) and North Sami (sme) (related Uralic varieties), and Galician (an unrelated Romance variety). Higher CKA indicates greater representational similarity. Figure 3 visualizes scores for pre-trained mBERT, the Alignment baseline, and VAÇAÍ-Bowl on corresponding embedding spaces via t-SNE.

Alignment uniformly increases CKA across all pairs, including the unrelated Galician, whose similarity to Skolt Sami rises from 0.042 to 0.066. This confirms that alignment collapses representations indiscriminately, regardless of linguistic relatedness.

VAÇAÍ-Bowl behaves selectively. It moderately increases similarity between the related Sami varieties (0.032→0.054), reflecting shared Uralic structure captured by the invariant encoder. Crucially, it simultaneously *decreases* similarity of Skolt Sami (sma) to the unrelated Galician (0.042→0.020), indicating that the variety-specific encoder actively separates linguistically distant varieties. This selective behavior of bringing related varieties closer while pushing unrelated ones apart is precisely the property that alignment-only methods lack, and explains why VAÇAÍ-Bowl recovers from alignment-induced failures observed in Tables 1–2.

Figure 4 further analyzes how each similarity metric in TOPPing affects source selection for variety *sma*. LangRank selects sources (*est, sme*) that correlate with phylogenetic similarity, yet TOPPing (*glg, sme*) yields stronger performance—suggesting lexical overlap contributes more than phylogenetic proximity alone. The effect is also model-dependent: XLM-R remains robust across diverse source pairs, likely due to its multi-script pre-training, while mBERT benefits more from careful source selection.

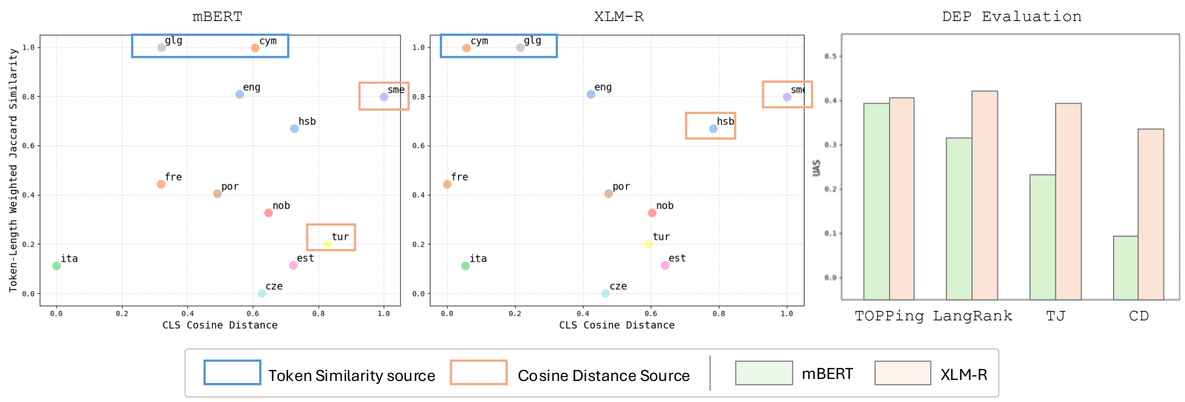

Figure 4: Analysis on source selection method. Two plots on the left illustrate TOPPing source selection scheme on Skolt Sami (sma). The x-axis is closeness of Cosine Distance of "[CLS]" tokens (CD), and the y-axis is token-length weighted Jaccard Similarity (TJ). We experiment VAÇAÍ-Bowl on two more source selections in rightmost plot; Taking two sources with highest TJ and two sources with highest CD. Note that TOPPing selects North Sami (sme), Galician (glg) and LangRank selects Estonian (est), North Sami (sme).

4.4 Analysis

Q. When is TOPPing most effective? TOPPing is most effective for low-resource or under-documented varieties, where typological or lexical metadata are sparse and pre-annotated descriptions are unavailable. Experimental results show that it remains robust even without such annotations, outperforming methods like LangRank under these conditions. In contrast, LangRank occasionally performs better for Indo-European family varieties where dense typological and lexical metadata (e.g., from URIEL and WALS) provide rich feature vectors and reliable genealogical signals for ranking candidate sources. However, for under-documented varieties, every candidate appears equally (dis)similar to LangRank, leading to poor source selection. Therefore, TOPPing serves as an effective source selection for varieties with minimal annotation and sparse resources.

Q. When is VAÇAÍ-Bowl most effective? VAÇAÍ-Bowl is most effective for low-resource varieties, where alignment-based methods tend to fail. For instance, with mBERT backbone (see Table 1), for Umbrian (xum), alignment alone led to a a performance drop in UAS of 36.14 to 34.67(-1.47) under LangRank and from 37.21 to 35.99(-1.22) under TOPPing when alignment was applied alone. In contrast, incorporating variety-specific features through VAÇAÍ-Bowl raised performance to 36.29($+1.62$) and 37.67($+1.77$), respectively. The framework enables the model to overcome alignment-induced failures, showing that enforcing uniform representations across distinct linguistic domains is insufficient. Instead, effective generalization and zero-training transfer require preserving the diversity inherent in each variety.

Overall, contrary to prevailing assumptions in NLP, certain cross-lingual transfer scenarios benefit more from dissimilar language pairs than from closely related ones. These findings underscore the importance of preserving variety-specific information so that models can better generalize to unseen and low-resource varieties, with potential applicability beyond reported cases.

5 Conclusion

This paper introduces a Language Generalization pipeline that tackles the twin challenges of selecting helpful high-resource varieties and learning representations that preserve, rather than erase, distinctiveness of varieties. Our source-selection strategy, TOPPing, is designed to select two neighboring varieties based on distinct criteria—(i) lexical overlap and (ii) proxy for phylogenetic similarity (using pre-trained MLM). Unlike prior works that utilize pre-annotated information for each variety, our method does not require prior knowledge regarding the target data, ensuring scalability. Coupled with the lightweight VAÇAÍ-Bowl dual-encoder, one branch aligning varieties and the other amplifying variety-specific cues, our framework delivers consistent gains on dependency parsing. Experiments across ten low-resource varieties, TOPPing with VAÇAÍ-Bowl lifts zero-shot UAS by an average of 50.63% and 58.6%, using mBERT and XLM-R, respectively. This beats alignment-centric baseline and even rescues cases where full alignment hurts ("alignment-induced fails"). Beyond parsing, the approach is model-agnostic, computation-friendly, and immediately applicable to other tasks.

Limitations

The methods presented in this research proved to be effective in handling under-represented varieties that pre-trained MLMs cannot easily generalize to. Although we suggest an end-to-end pipeline that does not require any human annotated work on either source or target varieties, the method still requires unique selection of source varieties for each training. To counterpart this computational complexity, our method trains with only addition of MLP encoders, discriminators, and the task-specific head. This approach significantly reduces both the model size and training overhead compared to methods that require full fine-tuning of MLMs. Yet, it should be recognized that the ultimate goal of Language Generalization is to leverage only a limited set of language varieties to develop a model capable of robust generalization across all varieties, regardless of their resource availability.

Ethical Considerations

We study a method to enhance zero-shot cross-lingual transfer to very low-resourced varieties, which are often not provided with sufficient data for training or evaluation. While we aim to develop language technologies targeting under-represented language communities, we still lack such coverage, limiting our research to datasets that are publicly available. Nevertheless, our approach provides a valuable step toward addressing the gap, by not simply aligning low-resource varieties with high-resource ones, but instead encouraging the model to recognize and preserve the linguistic differences that define them. All resources used in this research are publicly available, and no personal or sensitive information was collected or utilized. We do not anticipate any potential harm arising from this study.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under multiple grants funded by the Korea government (MSIT), including the ICT Creative Consilience Program (IITP-2026-RS-2020-II201821), ITRC(Information Technology Research Center)(RS-2021-II212052), and AI Graduate School Support Program (Sungkyunkwan University)(RS-2019-II190421).
This work was also supported by the Institute of Information & Communications Technology Planning&Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00437337, Development of technology for linking and utilizing security information and event management with privacy-preserving internal security data).

References

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Ayu Purwarianti, and Alham Fikri Aji. 2024. LinguAlchemy: Fusing typological and geographical elements for unseen language generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3912–3928, Miami, Florida, USA. Association for Computational Linguistics.

Niyati Bafna, Emily Chang, Nathaniel R. Robinson, David R. Mortensen, Kenton Murray, David Yarowsky, and Hale Sirin. 2025. Dialup! modeling the language continuum by adapting models to dialects and dialects to models. *Preprint*, arXiv:2501.16581.

Niyati Bafna, Kenton Murray, and David Yarowsky. 2024. Evaluating large language models along dimensions of language variation: A systematik invesdigatiom uv cross-lingual generalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18742–18762, Miami, Florida, USA. Association for Computational Linguistics.

Steve Bakos, David Guzmán, Riddhi More, Kelly Chutong Li, Félix Gaschi, and En-Shiun Annie Lee. 2025. AlignFreeze: Navigating the impact of realignment on the layers of multilingual models across diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 562–586, Albuquerque, New Mexico. Association for Computational Linguistics.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'11, page 2178–2186, Red Hook, NY, USA. Curran Associates Inc.

Verena Blaschke, Felicia Körner, and Barbara Plank. 2025. Add noise, tasks, or layers? MaiNLP at the VarDial 2025 shared task on Norwegian dialectal slot and intent detection. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 182–199, Abu Dhabi, UAE. Association for Computational Linguistics.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

J.K. Chambers and P. Trudgill. 1998. *Dialectology*. Cambridge Textbooks in Linguistics. Cambridge University Press.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023a. Enhancing cross-lingual learning: Optimal transfer language selection with linguistic similarity. *Science Talks*, 6:100226.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023b. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Preprint*, arXiv:1505.07818.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haeji Jung, Changdae Oh, Jooeon Kang, Jimin Sohn, Kyungwoo Song, Jinkyu Kim, and David R Mortensen. 2024. Mitigating the linguistic gap with phonemic representations for robust cross-lingual transfer. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 200–211, Miami, Florida, USA. Association for Computational Linguistics.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018a. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018b. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ying Li, Jianjian Liu, Zhengtao Yu, Shengxiang Gao, Yuxin Huang, and Cunli Mao. 2024. Representation alignment and adversarial networks for cross-lingual dependency parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7687–7697, Miami, Florida, USA. Association for Computational Linguistics.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. One language, many gaps: Evaluating dialect fairness and robustness of large language models in reasoning tasks. *Preprint*, arXiv:2410.11005.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. 2024. Exploring alignment in shared cross-lingual spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6326–6348, Bangkok, Thailand. Association for Computational Linguistics.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–10–I–18. JMLR.org.

Nghia Trung Ngo and Thien Huu Nguyen. 2024. Zero-shot cross-lingual transfer learning with multiple source and target languages for information extraction: Language selection and adversarial training. *Preprint*, arXiv:2411.08785.

Duke Nguyen, Aditya Joshi, and Flora Salim. 2025. Harnessing test-time adaptation for nlu tasks involving dialects of english. *Preprint*, arXiv:2503.12858.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vipul Rathore, Rajdeep Dhingra, Parag Singla, and Mausam. 2023. ZGUL: Zero-shot generalization to unseen languages using multi-source ensembling of language adapters. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6969–6987, Singapore. Association for Computational Linguistics.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Aarohi Srivastava and David Chiang. 2025. We‘re calling an intervention: Exploring fundamental hurdles in adapting language models to nonstandard text. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, pages 45–56, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Saedeh Tahery, Sahar Kianian, and Saeed Farzi. 2024. Cross-lingual NLU: Mitigating language-specific impact in embeddings leveraging adversarial learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4158–4163, Torino, Italia. ELRA and ICCL.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online. Association for Computational Linguistics.

Fei Wang, Kuan-Hao Huang, Kai-Wei Chang, and Muhao Chen. 2023. Self-augmentation improves zero-shot cross-lingual transfer. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.

Siyin Wang, Jie Zhou, Qin Chen, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Domain generalization via causal adjustment for cross-domain sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5286–5298, Torino, Italia. ELRA and ICCL.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9749–9764, Singapore. Association for Computational Linguistics.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2023. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Appendix

A Language Codes

In this section, we provide the ISO 639-3 and Universal Dependency dataset code for the varieties used in this paper.

Variety	ISO 639-3	UD-code
gheg	aln	UD_Gheg-GPS
paraguay mbya guarani	gug	UD_Mbya_Guarani-Thomas
brazil mbya guarani	gun	UD_Mbya_Guarani-Dooley
permyak komi	koi	UD_Komi_Permyak-UH
zyrian komi	kpv	UD_Komi_Zyrian-IKDP
ligurian	lij	UD_Ligurian-GLT
central alemanic	gsw	UD_Swiss_German-UZH
skolt saami	sma	UD_Skolt_Sami-Giellagas
low saxon	nds	UD_Low_Saxon-LSDC
umbrian	xum	UD_Umbrian-IKUVINA

Table 5: Target Varieties in Section 4 with their ISO 639-3 and Universal Dependencies code

Variety	ISO 639-3	UD-code
italian	ita	UD_Italian-MarkIT
norwegian	nor	UD_Norwegian-Bokmaal
north sami	sme	UD_North_Sami-Giella
portuguese	por	UD_Portuguese-Bosque
spanish	spa	UD_Spanish-AnCora
finnish	fin	UD_Finnish-TDT
estonian	est	UD_Estonian-EDT
catalan	cat	UD_Catalan-AnCora
indonesian	ind	UD_Indonesian-CSUI
galician	glg	UD_Galician-CTG
galician	glg	UD_Galician-TreeGal
turkish	tur	UD_Turkish-Penn
turkish	tur	UD_Turkish-IMST
serbian	srp	UD_Serbian-SET
croatian	hrv	UD_Croatian-SET
czech	ces	UD_Czech-CAC
slovak	slk	UD_Slovak-SNK
russian	rus	UD_Russian-SynTagRus
old church slavonic	chu	UD_Old_Church_Slavonic
belarusian	bel	UD_Belarusian-HSE
ukrainian	ukr	UD_Ukrainian-IU
upper sorbian	hsb	UD_Upper_Sorbian-UFAL
bulgarian	blg	UD_Bulgarian-BTB
irish	gle	UD_Irish-IDT
welsh	cym	UD_Welsh-CCG

Table 6: Target Varieties in Section 4 with their ISO 639-3 and Universal Dependencies code

B Selected Source Varieties

In this section, we list the selected source varieties in Section 4. The LangRank selected source varieties are same for both experiments.

cluster	target variety	LangRank		TOPPing	
		source	UAS	source	UAS
albanian	gheg	italian finnish	51.00	turkish german italian	46.34
gallo-italian	ligurian	catalan spanish	63.02	portuguese italian	64.29
high german	central alemannic	norwegian bokmal italian	52.23	german turkish german	57.74
komi	komi-zyrian	indonesian turkish	35.32	bulgarian russian	38.19
	komi-permyak	estonian portuguese	36.90	bulgarian turkish	42.29
saami	skolt saami	estonian north saami	32.62	north saami galician	39.67
sabellic	umbrian	estonian turkish	35.07	north african arabic italian	37.67
tupi-guarani	paraguay mbya guarani	spanish hindi	27.05	turkish italian	36.39
	brazil mbya guarani	finnish turkish	17.21	english upper sorbian	19.00
west low german	low saxon	english italian	50.92	turkish german english	54.90

Table 7: Selected two source varieties for Dependency Parsing and its scores with VAÇAÍ-Bowl using mBERT as backbone. Abbreviations (spk) and (wrt) each refer to mode of dataset, spoken and written, respectively.

cluster	target variety	LangRank		TOPPing	
		source	UAS	source	UAS
albanian	gheg	italian finnish	58.55	norwegian bokmaal italian	57.50
gallo-italian	ligurian	catalan spanish	59.82	portuguese italian	63.44
high german	central alemannic	norwegian bokmaal italian	47.47	german turkish german	57.74
komi	komi-zyrian	indonesian turkish	35.32	bulgarian russian	37.76
	komi-permyak	estonian portuguese	42.41	bulgarian turkish	44.66
saami	skolt saami	north saami north saami	42.53	north saami galician	40.99
sabellic	umbrian	estonian turkish	34.56	north african arabic italian	32.77
tupi-guarani	paraguay mbya guarani	spanish hindi	31.23	turkish italian	31.97
	brazil mbya guarani	finnish turkish	11.27	english italian	13.98
west low german	low saxon	english italian	48.38	german turkish german	51.65

Table 8: Selected two source varieties for Dependency Parsing and its scores with VAÇAÍ-Bowl using XLM-R as backbone. Abbreviations (spk) and (wrt) each refer to mode of dataset, spoken and written, respectively.

C Detailed Experimental Results

C.1 Dependency Parsing Results

We report UAS scores in Table 1 initially for direct comparison to the benchmark DialectBench - that provides only UAS scores. However, to better support the results, we also report the labeled attachment scores (LAS) in this section.

In Table 9 and Table 10, LAS scores generally follow the trends of UAS scores reported in the main table. The combination of TOPPing and VAÇAÍ-Bowl enhances the model's ability to capture dependencies in 8 out of 10 low-resource varieties. For the remaining varieties, the following observations hold: For *aln*, VAÇAÍ-Bowl still yields improvements in LAS across both source variety selection methods. For *nds*, the LAS score of TOPPing + VAÇAÍ-Bowl (34.89) remains comparable to the best-performing method (34.96), with UAS scores of 54.90 and 52.54, respectively—indicating a minimal trade-off.

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source selected using LangRank (Lin et al., 2019)										
mBERT	20.43	7.54	1.42	15.30	13.17	36.89	31.44	13.35	34.00	4.59
+Alignment	22.00	7.95	1.35	18.12	14.22	36.90	34.89	14.06	33.18	3.69
+VAÇAÍ-Bowl (OURS)	22.00	7.05	1.35	18.00	15.85	40.21	35.25	15.12	35.42	4.59
source selected using TOPPing (OURS)										
mBERT	20.53	10.41	2.32	18.11	20.33	42.29	34.96	17.19	37.05	5.20
+Alignment	18.21	10.74	2.21	18.45	20.10	43.19	32.61	16.06	34.15	4.90
+VAÇAÍ-Bowl (OURS)	20.79	12.13	2.44	18.79	20.33	43.79	34.89	18.21	37.65	5.20

Table 9: Quantitative results on LAS scores using mBERT as backbone on dependency parsing task evaluated across selected low-resource varieties from DialectBench.

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source selected using LangRank (Lin et al., 2019)										
XLM-R	28.87	5.34	1.16	21.04	14.80	38.10	29.01	21.12	27.60	2.60
+Alignment	29.73	6.23	0.81	20.02	16.47	37.68	28.98	19.49	29.67	3.52
+VAÇAÍ-Bowl (OURS)	30.8	7.46	3.10	22.95	15.89	37.43	29.37	19.53	29.39	4.75
source selected using TOPPing (OURS)										
mBERT	27.93	9.51	2.38	20.36	20.10	43.29	29.82	18.06	35.42	5.51
+Alignment	28.92	9.59	1.08	21.82	20.38	43.83	30.70	19.27	36.24	3.98
+VAÇAÍ-Bowl (OURS)	29.47	10.41	1.10	22.05	21.75	43.84	31.30	20.12	36.86	5.21

Table 10: Quantitative results on LAS scores using XLM-R as backbone on dependency parsing task evaluated across selected low-resource varieties from DialectBench.

C.2 Part-of-Speech Tagging Results

In this section, we report the specific evaluated results on POS tagging downstream task for mBERT and XLM-R also represented in Table 4.

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source selected using LangRank (Lin et al., 2019)										
mBERT	45.76	32.07	12.08	41.63	45.63	62.61	66.54	39.93	58.64	22.61
+Alignment	47.79	32.46	11.69	41.48	46.47	60.56	68.25	37.38	57.16	22.14
+VAÇAÍ-Bowl (OURS)	45.51	32.17	14.54	42.71	45.55	60.48	68.36	36.38	56.89	23.81
source selected using TOPPing (OURS)										
mBERT	43.80	33.37	11.19	48.68	47.71	68.32	65.48	42.03	59.26	26.56
+Alignment	45.57	33.34	9.51	49.40	48.53	68.70	63.93	44.06	61.22	28.26
+VAÇAÍ-Bowl (OURS)	50.50	33.53	10.89	49.34	48.21	69.30	66.41	43.99	61.25	25.93

Table 11: Quantitative results on F1 scores using mBERT as backbone on part-of-speech tagging task.

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source selected using LangRank (Lin et al., 2019)										
XLM-R	56.34	31.54	5.01	54.66	50.87	62.57	60.20	53.83	48.74	25.28
+Alignment	57.17	33.37	6.17	55.25	50.48	61.15	61.18	53.11	47.40	24.64
+VAÇAÍ-Bowl (OURS)	57.97	32.94	6.24	55.97	51.71	63.54	60.35	53.83	50.46	24.97
source selected using TOPPing (OURS)										
XLM-R	56.91	33.11	7.33	53.87	53.73	67.98	61.95	53.89	60.16	31.07
+Alignment	57.81	34.08	4.95	55.36	53.22	68.70	61.34	54.54	58.35	29.22
+VAÇAÍ-Bowl (OURS)	58.35	34.88	5.47	54.02	55.16	69.97	60.81	55.17	60.21	29.23

Table 12: Quantitative results on F1 scores using XLM-R as backbone on part-of-speech tagging task.

D Parameter Search for Lambda of Gradient-Reversal Layer

lambda	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
0.1	46.09	35.00	15.03	36.56	40.61	64.72	52.13	37.18	56.40	36.14
0.5	45.41	35.25	15.17	35.70	40.27	63.74	52.36	40.12	54.91	37.37
1.0	46.34	36.39	19.00	42.29	38.19	64.29	54.90	39.67	57.74	37.67

Table 13: Ablation study on VAÇAÍ-Bowl performance with mBERT backbone based on different lambda values.

For the gradient-reversal layer used to adversarially train invariant feature encoder in Section 3.2, we provide an ablation study on its affects on performance.

E Ablation on Loss Components

	$\mathcal{L}_{\text{inv}}$	$\mathcal{L}_{\text{spc}}$	$\mathcal{L}_{\text{task}}$	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
w/o both	–	–	✓	45.99	34.02	15.60	37.35	36.19	62.69	51.76	37.82	54.84	37.21
w/o $\mathcal{L}_{\text{spc}}$	✓	–	✓	46.49	29.19	15.35	40.27	36.08	63.17	53.05	35.56	56.25	26.27
w/o $\mathcal{L}_{\text{inv}}$	–	✓	✓	**46.74**	20.82	15.44	40.33	36.95	63.23	53.20	38.24	54.69	27.18
Full	✓	✓	✓	46.34	**36.39**	**19.00**	**42.29**	**38.19**	**64.29**	**54.90**	**39.67**	**57.74**	**37.67**

Table 14: Ablation study on loss components of VAÇAÍ-Bowl using mBERT with TOPPing source selection (UAS). Removing $\mathcal{L}_{\text{inv}}$ disables adversarial invariance training; removing $\mathcal{L}_{\text{spc}}$ removes the variety-specific objective. The full objective consistently outperforms all ablated variants, confirming that both branches contribute to generalization.

To isolate the contribution of each loss term in VAÇAÍ-Bowl, we conduct an ablation study by selectively removing $\mathcal{L}_{\text{inv}}$ (adversarial invariance) and $\mathcal{L}_{\text{spc}}$ (variety-specific discrimination) from the full objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{spc}} + \mathcal{L}_{\text{task}}$. All experiments use mBERT with TOPPing source selection.

Table 14 reports UAS scores across all ten target varieties. Removing $\mathcal{L}_{\text{spc}}$ leads to notable degradation on varieties such as gug and sma, where variety-specific cues are most beneficial. Removing $\mathcal{L}_{\text{inv}}$ similarly degrades performance, indicating that enforcing invariance on one branch is necessary for the other to capture complementary variety-specific structure. Training with only $\mathcal{L}_{\text{task}}$ (no disentanglement) yields the weakest average, validating the dual-encoder design. The full objective achieves the highest average UAS (43.65), confirming that both branches are essential for effective generalization.