Harnessing Linguistic Dissimilarity for Zero-Shot Language Generalization on Low-Resource Varieties

Anonymous ACL submission

Abstract

Despite the fact that much communication takes place in them, low-resource language varieties used in specific regions or by specific groups remain neglected in the development of Multilingual Language Models. A great deal of cross-lingual researches focus on inter-lingual language transfer which strives to align allied varieties and suppress linguistic differences between them. 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 dependency parsing task as proxy for performance on other downstream tasks. Together, the methods outperform all baselines across 10 low-resource varieties.

1 Introduction

Multilingual Language Models (MLMs) have led to great strides in Natural Language Processing (NLP), enabling language technologies in more than one hundred languages (Pires et al., 2019; Conneau et al., 2020). Developers of these models have usually treated languages as discrete entities, despite decades of research in linguistics showing that languages lay on a continuum of similarity (Lin et al., 2019). Crucially, a large portion (perhaps the majority) of real-world communication is carried out in language varieties that are not represented among the quasi-standardized set of roughly one hundred languages that comprise MLM training data. When confronted with such variants, existing models fail (Faisal et al., 2024).

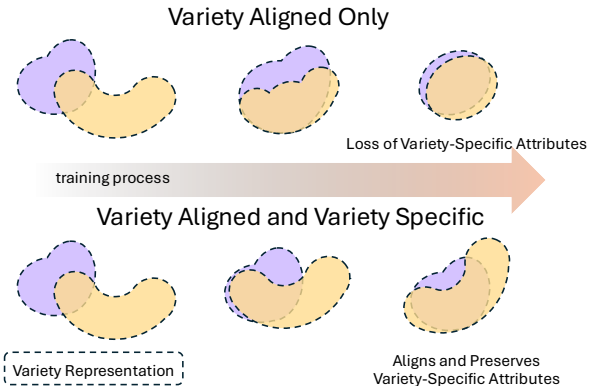

Figure 1: Visualization on training to align two different varieties in the embedding space, comparative to aligning and at the same time preserving variety-specific attributes.

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).

Past modeling approaches to language variation, including the large volume of research in interlingual transfer, tended to focus on similarity-based alignment between language varieties. For example, Yang et al. (2022) proposes instance-level regularization to minimize representational gaps, thereby improving transferability. While such feature-level alignment can be helpful to some extent, it disregards linguistic variances developed in real-world. In this paper, instead, we propose

a variety-aware Language Generalization framework that effectively generalizes to low-resource varieties, even when the variety is unseen during MLM pre-training. 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 zero-shot 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 dependency parsing (DEP), which we believe to be an informative proxy for performance on other downstream tasks.

Our key contributions are as follows:

- This paper suggests 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).

Given this limitation, several works have attempted to address the gap. Inspired by findings in cross-lingual transfer (Snæbjarnarson et al., 2023; Bafna et al., 2024), recent approaches focus on developing distance metrics that rank varieties by similarity to identify the most similar variety best suited to support low-resource varieties. Specifically, Bafna et al. (2025) propose two approaches: one that trains models on artificially generated variants, and another that adapts input data at inference time to closely resemble the high-resource standard. Similarly, Nguyen et al. (2025) define the dialectal gap between language variants and apply test-time adaptation techniques to improve model performance on nonstandard inputs.

2.2 Zero-shot Cross-lingual Transfer

Prevailing methods to improve cross-lingual transferability for language varieties without explicit training suggest various methods to align language representations. Wang et al. (2023) introduce a self-augmentation approach that substitutes tokens in English dataset with tokens from different variety and then perturbing to distill token-level alignment. X-MIXUP reduces cross-lingual representation discrepancy of parallel sentences, so that the target representation is explicitly pulled toward the source (Yang et al., 2022). Wu and Monz (2023) re-parameterises the embedding table with a graph network that forces meaning-similar words to converge on the same coordinate region. Huang et al. (2021) adopts robust training strategies, such as randomized smoothing, to enhance cross-lingual transfer in zero-shot settings. All these methods focus on aligning the target variety to a source variety; while treating variety-specific information as something to be eliminated or suppressed rather than be a potential 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. However, these rely on predefined languages, making it inapplicable to low-resource varieties.

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). This has been a long-standing

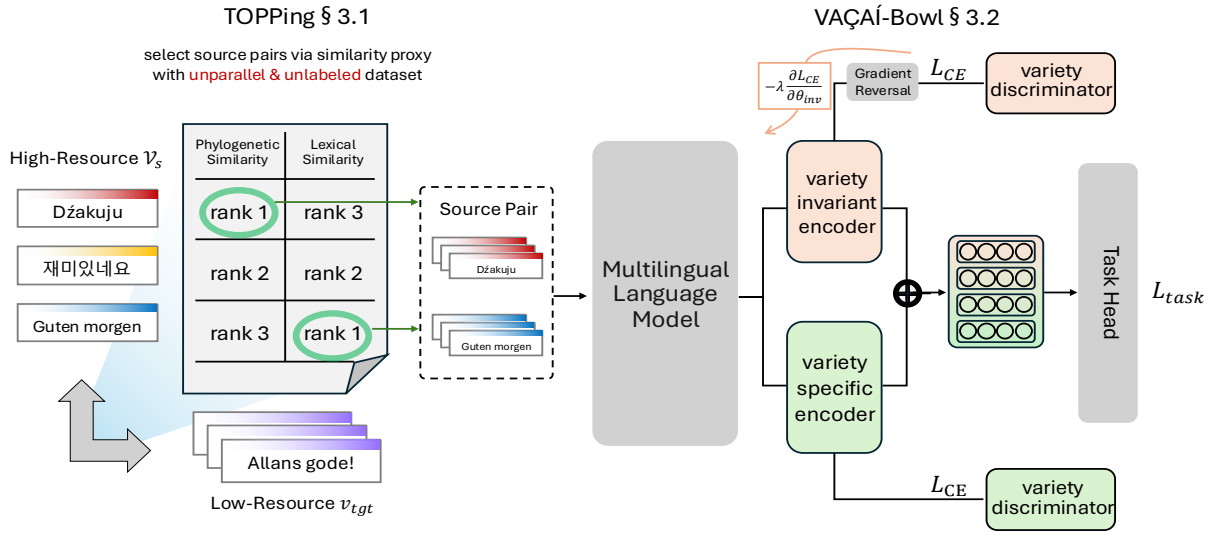

Figure 2: Overall framework of this paper. Using TOPPing, with just unparallel and unlabeled datasets, we can select source varieties with not only shared but also distinctive features to capture relationship between varieties. From the obtained source variety pair, VA learns the semantic differences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

problem in machine learning, where models trained with a limited set of training data fail to successfully perform in domain variations. A very basic approach for DG is to align the different domains together, aiming to learn domain-invariant features that are robust across domains (Muandet et al., 2013; Li et al., 2018a,b), which is also adopted in NLP tasks (Wang et al., 2024, 2021; Li et al., 2024).

In multilingual NLP, Jung et al. (2024) analyze multilingual modeling from a domain-level perspective, treating each language as a separate domain. Building on this view, we explore the possibility to frame zero-shot transfer to low-resource varieties as a DG problem, where the goal is to generalize to unseen linguistic domains without explicit training. While most prior works focus on learning domain-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 in a zero-shot manner, and a method to carefully select exploitable high-resource source variety pair for 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. Although works like LangRank provide a principled method for selecting source varieties for training, it relies on Lang2Vec and URIEL, collections of pre-annotated information such as geometric distance, phylogenetic similarity based on Glottolog, World Atlas of Language Structures, and Syntactic Structures of World Languages (Lin et al., 2019; Littell et al., 2017). This fine-grained approach is applicable for only predefined set of varieties, leaving otu low-resource varieties that are not present in URIEL. LangRank is thus inherently biased towards high-resourced varieties, with 31% of presented languages missing annotations (Toossi et al., 2024).

To mitigate the constraints of low-resource varieties, we suggest **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 dis-

tances 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 between two varieties with 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 frozen 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 proxy the lexical distance between two varieties using token overlap (Blaschke et al., 2025). Here, we introduce **token-length weighted Jaccard Similarity** as lexical overlap calculation tailored for 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 :

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

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

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

The pair $\langle v_{\text{sim}}, v_{\text{overlap}} \rangle$, selected by independent ranking, leaves room for diversity in source pair selection. TOPPing can therefore offer a strong yet inexpensive cues without requiring labeled or parallel dataset, suitable for low-resource scenarios.

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 L**OW**-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. Specifically, we argue that a model must learn not only to align but also to distinguish varieties in order to fully grasp the linguistic characteristics on an unseen variety.

In order to model variety-invariant and variety-specific features, we use a frozen 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.

$$h_{\text{inv}} = f_{\text{inv}}([CLS]) \tag{4}$$

- Variety-specific encoder f_{spc} is trained normally to emphasize variety-specific features.

$$h_{\text{spc}} = f_{\text{spc}}([CLS]) \tag{5}$$

The outputs are concatenated into h.

$$h = h_{\text{inv}} \,\|\, h_{\text{spc}} \in \mathbb{R}^{2d}, \tag{6}$$

where both encoders output d-dimensional vectors. The joint feature h is used in place of original "[CLS]" embedding for downstream tasks.

To train each encoders to successfully 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

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source is eng										
mBERT$^{\diamond}$	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 selected low-resource varieties from DialectBench. $^{\diamond}$ refers to value reported in original paper. Underlined scores refer to best performing on target under controlled source variety. **Bold** scores refer to best performing on the target variety.

G_λ is inserted in front of D_{inv} to selectively update parameters to fool D_{inv} (Ganin and Lempitsky, 2015).

$$G_\lambda(z) = z, \qquad \frac{\partial G_\lambda}{\partial \mathbf{z}} = -\lambda \mathbf{I}, \tag{7}$$

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

$$\begin{aligned} L_{inv} &= L_{CE}(D_{\text{inv}}(G_\lambda(h_{\text{inv}})), y_{\text{var}}), \\ L_{spc} &= L_{CE}(D_{\text{spc}}(h_{\text{spc}}), y_{\text{var}}), \end{aligned} \tag{8}$$

where L_{CE} denotes Cross Entropy Loss.

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

$$L_{\text{task}} = L_{\text{task}}(f_{\text{task}}(h), y_{\text{task}}). \tag{9}$$

Finally, the objective function is defined as follows :

$$L_{total} = L_{inv} + L_{spc} + L_{task}. \tag{10}$$

4 Experiments

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 target low-resource varieties that face following constraints : First, there is no training dataset available for the variety. Second, the zero-shot performance does not meet 50% of criteria. 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 task, which is a structured prediction task.

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. We implement a **LangRank** baseline where the source varieties are selected based on pre-annotated linguistic features and dataset-dependent features (Lin et al., 2019). Please refer to Appendix B for detailed information on selected source languages.

Language Generalization Baselines. To compare the VAÇAÍ-Bowl framework, we implement two baselines : (1) **MLM** baseline illustrates performance when the model is simply finetuned

Methods	Varieties									
	aln	gug	gun	koi	kpv	lij	nds	sma	gsw	xum
source is eng										
XLM-R$^{\diamond}$	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 selected low-resource varieties from DialectBench. $^{\diamond}$ refers to value reported in original paper. Underlined scores refer to best performing on target under controlled source variety. **Bold** scores refer to best performing on the target variety.

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 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 following experiments. Parameters are optimized using Adam optimizer. 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. Please refer to Appendix C for detailed information on parameter search.

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.

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 the mBERT model on TOPPing itself beats the best score obtained using LangRank for 8 out of 10 varieties. These results show that TOPPing is an effective source selection method even without pre-annotated descriptions of varieties. In cases where LangRank enhances transferability and generalization, *aln* and *xum*, it is notable that the target varieties all fall into Indo-European family. This advantage is largely attributable to the dense typological and lexical metadata available for Indo-European languages in resources such as URIEL and WALS, which furnish LangRank with informative feature vectors and reliable genealogical signals for ranking candidate sources. Thus, for targets that are partially represented with characteristics of Indo-European family, LangRank can discriminate between closely related sources. However, for under-documented varieties every candidate looks equally (dis)similar, which leads to unsuitable source selection.

Language Generalization. Our proposed architecture, 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

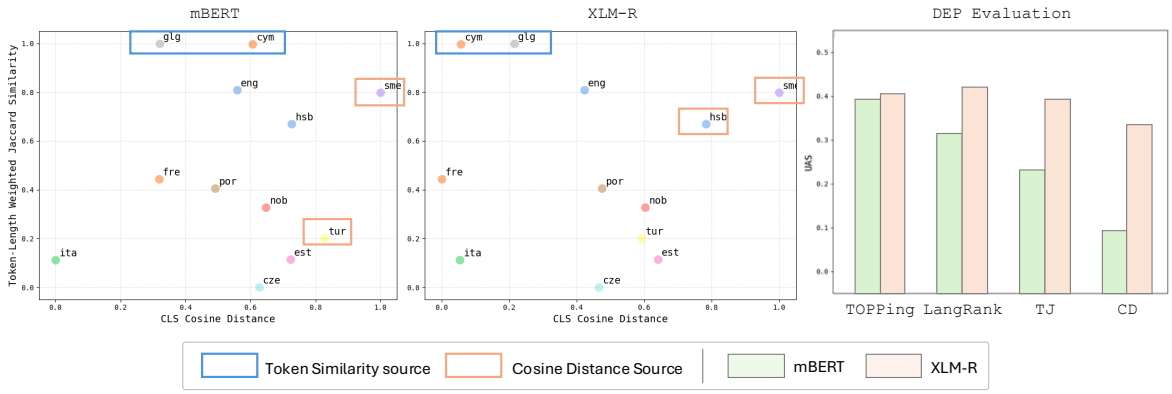

Figure 3: Analysis on source selection method. Two plots on the right illustrates TOPPing source selection scheme on variety *sma*. The x-axis is closeness of Cosine Distance of "[CLS]" tokens (CD), and the y-axis is token-length weighted Jaccard Similarity (TJ). We experiment VAÇAÍ-Bowl on two more source selections ; Token Similarity source which takes two sources with highest TJ and Cosine Distance source which takes two sources with highest CD. Note that TOPPing selects *glg, sme* and LangRank selects *est, sme*.

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, sna*} in Table 2. We refer to this phenomenon as **alignment-induced fails**. When this occurs, VAÇAÍ-Bowl overomes the fails of Alignment by utilizing variety-specific attributes under for all cases. Especially in in Table 1, for 6 out of 7 alignment-induced fails observed using TOPPing, VAÇAÍ-Bowl outperforms all methods on the target variety. This illustrates promising results from utilizing variety-specific cues. Also, VAÇAÍ-Bowl performs consistently better than Alignment method across all varieties. This suggests that merely forcing alignment across distinct linguistic domains is insufficient. Rather, effective generalization and zero-shot performance requires the model to adapt to and preserve the diversity inherent in different language varieties.

4.3 Qualitative Results

In Figure 3, we can observe how different similarity metrics used to obtain TOPPing in Section 3.1 affect performances in VAÇAÍ-Bowl. LangRank utilizes *est, sme* as source varieties, which seem largely correlated with the [CLS] Cosine Distance we use to proxy phylogenetic similarity. TOPPing's performance hints that phylogenetic similarity may not contribute to model performance as much, considering it utilizes *glg, sme* as source varieties. Additionally, as can be observed in the case of CD, performance is also model-dependent. XLM-R tends to stay robust on diverse source varieties unline mBERT. Beyond the observed performance gains, incorporating variety-specific knowledge offers valuable linguistic insights. 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 respect, rather than erase, distinctiveness of varieties. With TOPPing, we automatically choose a linguistically overlapping and complementary source pair for any unseen variety, requiring no labels or parallel data. 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 (MLM layers stay frozen), 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 freezes the MLM and train only the 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.

References

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 investigatiom 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.

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.

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.

Hasti Toossi, Guo Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz, and En-Shiun Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the URIEL knowledge base. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 233–241, Mexico City, Mexico. 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.

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	nds	UD_Swiss_German-UZH
skolt saami	sma	UD_Skolt_Sami-Giellagas
low saxon	gsw	UD_Low_Saxon-LSDC
umbrian	xum	UD_Umbrian-IKUVINA

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

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.

B Selected Source Varieties

In this section, we list the selected source varieties used to train VAÇAÍ-Bowl in Section 4. Table 5 lists TOPPing selected source varieties using mBERT as embedding backbone, which was used to produce results for Table 1. Table 6 lists TOPPing selected source varieties using XLM-R as embedding backbone, used to evaluate results on 2. The LangRank selected source varieties are same for both experiments, as LangRank does not take consideration of the MLM used in training.

C Parameter Search for Lambda of Gradient-Reversal Layer

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.

D Number of Model Parameters

We use mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020)in their base size, where mBERT has 110M and XLM-R has 125M number of parameters.

Variety	ISO 639-3	UD-code
gheg	ita	UD_Italian-MarkIT
paraguay mbya guarani	nor	UD_Norwegian-Bokmaal
brazil mbya guarani	sme	UD_North_Sami-Giella
permyak komi	zho	UD_Chinese-GSDSimp
zyrian komi	por	UD_Portuguese-Bosque
ligurian	spa	UD_Spanish-AnCora
central alemanic	fin	UD_Finnish-TDT
skolt saami	est	UD_Estonian-EDT
low saxon	cat	UD_Catalan-AnCora
umbrian	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 4: Source Varieties in Section 4 with their ISO 639-3 and Universal Dependencies code

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 5: 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 estonian	north saami north saami	42.53	north saami galician	40.99
sabellic	umbrian	estonian turkish	-	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 6: 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.

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