Topology Matters: How Scale and Alignment Reshape Multilingual Spaces

Multilingual pretrained models enable cross-lingual transfer, yet how their internal spaces encode language identity remains unclear. Large encoders such as XLM-R have been reported to "flatten" linguistic variation [1], while compact instruction-tuned models like mT0 enforce task alignment [2]. A systematic comparison of these families at the level of multilingual topology is still missing.

We propose TopoLingEval, a lightweight framework with three components: (i) *geometric projections*, using PCA to capture global variance and t-SNE to reveal local structure; (ii) *centroid distance analysis*, computing cosine distances across centroids to quantify overlap or separation; and (iii) *typological correlation*, comparing embedding distances against genealogical resources. The framework is model-agnostic and requires only sentence embeddings, making it efficient and reproducible.

Our setup focuses on six diverse languages from TyDiQA [3]: Telugu, Arabic, Bengali, Finnish, Indonesian, and Swahili. For XLM-R, embeddings are obtained by pooling [CLS] tokens; for mT0-small, final decoder hidden states are used. Each language is represented by 200 balanced samples. This design ensures diversity in families, scripts, and resource levels while controlling for size.

Results reveal complementary patterns. XLM-R shows higher global variance in PCA, where some languages (e.g., Finnish, Arabic) spread widely, but collapses locally in t-SNE, with centroid distances averaging only 0.0006. This indicates a "flattened" space where language identity is obscured. By contrast, mT0-small produces tighter, language-specific clusters in t-SNE and substantially larger centroid distances (mean 0.139), suggesting clearer separation.

Table 1: Representative cosine distances between language centroids. Smaller = more overlap.

Language Pair	XLM-R	mT0-small
Arabic–Bengali	0.0006	0.154
Finnish–Swahili	0.0006	0.132
Bengali–Telugu	0.0006	0.136
Mean	0.0006	0.139

These projection patterns are illustrated in Figure 1, where PCA emphasizes global variance for XLM-R, while t-SNE highlights sharper clusters in mT0-small. However, typological correlation is absent in both models (Spearman $\rho \approx 0.01$ for XLM-R, $\rho \approx -0.15$ for mT0-small, both not significant), showing that neither captures genealogical similarity.

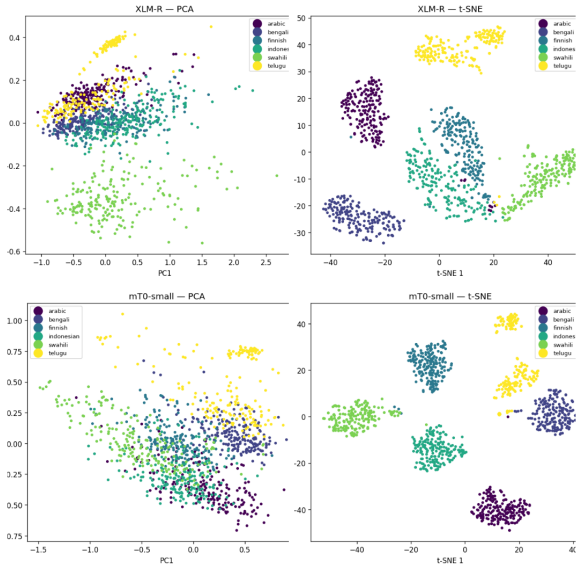

Figure 1: PCA and t-SNE projections: XLM-R (global variance, overlap) vs. mT0-small (local clustering).

These findings highlight a fundamental trade-off. XLM-R promotes overlap but risks erasing language-specific signals. mT0-small enforces separation but without linguistic grounding, which could amplify spurious differences. For low-resource settings, this trade-off is particularly consequential: transfer may fail either by homogenizing minority languages or isolating them in uninformative clusters. Future work should explore hybrid objectives that balance inclusivity with typological sensitivity and adopt evaluation protocols that measure representational fidelity, not just downstream accuracy.

References

[1] T. Pires, E. Schlinger, and D. Garrette. How multilingual is multilingual BERT? In *Proc. ACL*, 2019.

[2] N. Muennighoff, T. Wang, V. Kocijan, D. Simig, L. Sutawika, D. Yogatama, T. Scialom, and T. Schick. Crosslingual generalization through multitask finetuning. In *Proc. ACL*, 2023.

[3] J. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki. TyDi QA: A benchmark for information-seeking QA in typologically diverse languages. In *Proc. ACL*, 2020.