Harnessing Multilingual Geometry for Low-Resource Language Representation

Fabrice Niyigaba*, **Syed Ali Haider***, **Jiho Lee**, **Soroush Vosoughi**
Dartmouth College

Abstract

Extremely low-resource languages represent one of the most urgent challenges in natural language processing. For many of these languages, available data consists of only a few hundred sentences or even small word lists, making conventional supervised approaches largely impractical. Recent work suggests that multilingual representations may share a universal geometric structure across languages. If such geometry exists, it means that representations learned in high-resource languages contain reusable structure that can be aligned to bootstrap models for languages with little or no labeled data. In this work, we investigate this hypothesis through a controlled study of cross-lingual embedding geometry using two morphologically distinct languages: English and Turkish. We study the existence of this shared geometry and analyze how much of it can be harnessed to bootstrap downstream task performance. Our findings, though nuanced, suggest that there may exist a linear transformation that can partially align representations across languages and slightly improve downstream performance. This opens an intriguing area of research: how representations learned in high-resource languages might be used to breathe life into languages with extremely little or no labeled data.

1 Introduction

Low-resource languages have recently gained renewed attention in natural language processing. While many languages lack large annotated datasets, sustained community effort can gradually build resources and improve coverage. However, an even more difficult class of languages exists: *extremely low-resource languages*. For these languages, available corpora may consist of only a few hundred sentences, and in some cases the number of fluent speakers is itself rapidly declining (Yang et al., 2025; Matsuura et al., 2020). In such settings, conventional supervised approaches are largely infeasible (Yang et al., 2025). This motivates an intriguing question: **how can we develop NLP systems for languages where reliable ground-truth data may never exist?**

One promising direction comes from recent work suggesting that multilingual models encode languages within a partially shared geometric structure. Rather than learning completely independent representations for each language, models appear to organize linguistic information into overlapping subspaces that capture universal aspects of syntax, semantics, and morphology. If such a shared geometry truly exists, it offers a promising possibility for endangered and extremely low-resource languages: their representations may not need to be learned from scratch, but only *aligned* with those of high resourced languages.

However, prior attempts to study this phenomenon largely rely on pretrained multilingual encoders such as mBERT or XLM (Jha et al., 2025; Cao et al., 2020). While powerful, these models introduce significant confounds. Their training data is fixed and massive, spanning dozens of languages, which makes it difficult to isolate the contribution of individual languages or to systematically vary the amount of available data. As a result, it remains unclear whether the observed geometric structure reflects a genuine linguistic property or an artifact of large-scale pretraining.

In this work, we propose a different approach. Rather than analyzing pretrained models, we construct a controlled experimental setting in which embedding spaces are trained from scratch and the amount of data available for each language can be precisely controlled. Within this sandbox, we investigate whether simple linear transformations can align embedding spaces across languages and enable cross-lingual transfer. Our findings suggest that such transformations can improve performance

*Equal contribution.

on downstream tasks, supporting the hypothesis that cross-lingual embedding spaces share recoverable structure. While preliminary, this result offers an intriguing implication: if universal geometric structure is real, even languages with extremely limited data may benefit from models trained on better-resourced neighbors.

2 Related Work

Representation alignment. Linear mappings between monolingual vector spaces capture translation correspondences (Mikolov et al., 2013). Contextual embeddings implicitly encode hierarchical syntactic and semantic properties within their geometry (Conneau et al., 2018; Liu et al., 2019; Hewitt and Manning, 2019; Tenney et al., 2019). More closely related to our work, Schuster et al. (2019) align these embeddings via linear transformations for zero-shot parsing. We extend this formulation to test whether pairwise geometric alignment can be recovered without parameter updates or target-language supervision in extremely low-resource settings.

Multilingual structural universals. Multilingual models induce structural universals with approximately shared syntactic subspaces (Chi et al., 2020; Devlin et al., 2019; Nivre et al., 2020), though language-specific components drive cross-lingual misalignment (Chang et al., 2022). The Platonic Representation Hypothesis posits neural networks naturally converge toward a shared statistical reality (Huh et al., 2024), empirically showing that unsupervised methods can translate embeddings across disparate architectures without paired data (Jha et al., 2025). We investigate whether this shared structure is accessible via simple linear transformations between a high-resource anchor and a simulated low-resource language.

Geometric manipulation. Pre-trained representations can be geometrically steered to alter linguistic properties without updating encoder weights (De Raedt et al., 2021). This lightweight manipulation suggests that cross-lingual alignment gaps, typically mitigated through full-model fine-tuning (Cao et al., 2020), might be resolved without parameter changes. We address this gap by simulating an extreme low-resource scenario to test whether a linear transformation learned from a high-resource anchor can successfully recover shared geometric structure.

3 Data

Our experiments combine three complementary multilingual resources providing parallel text, lexical supervision, and syntactic annotations. Parallel sentences are drawn from **OPUS-100**, a large-scale multilingual corpus containing aligned text across 100 languages. These sentence pairs are used to train comparable distributional representations across languages. To obtain word-level supervision for alignment, we use bilingual dictionaries from **MUSE**, which provide high-quality translation pairs used to construct aligned embedding sets. For downstream evaluation we use **Universal Dependencies (UD)** treebanks, which provide consistent cross-linguistic part-of-speech (POS) annotations. POS labels allow us to evaluate whether cross-lingual mappings preserve syntactic structure when transferring representations between languages.

4 Methodology

We formalize the task as learning a linear transformation between embedding spaces of two languages. Let language A and language B have embedding matrices

$$X \in R^{n \times d}, \quad Y \in R^{m \times d},$$

where n and m denote the number of embeddings and d is the embedding dimensionality. Given a set of aligned translation pairs extracted from bilingual dictionaries, we construct paired matrices $X_s, Y_s \in R^{k \times d}$ containing k aligned embeddings. Our objective is to learn a transformation $W \in R^{d \times d}$ such that

$$X_s W \approx Y_s.$$

This formulation assumes that cross-lingual embedding spaces differ primarily by a linear transformation.

To obtain the underlying representations, we train distributional word embeddings independently for each language using **FastText**. FastText extends Word2Vec by representing each word as a composition of character n-grams, allowing the model to produce embeddings for out-of-vocabulary tokens. This property is particularly beneficial in low-resource settings where vocabulary coverage is limited.

We estimate the mapping W using two alignment methods. First, we use the **Orthogonal Procrustes** to test whether the two embedding spaces

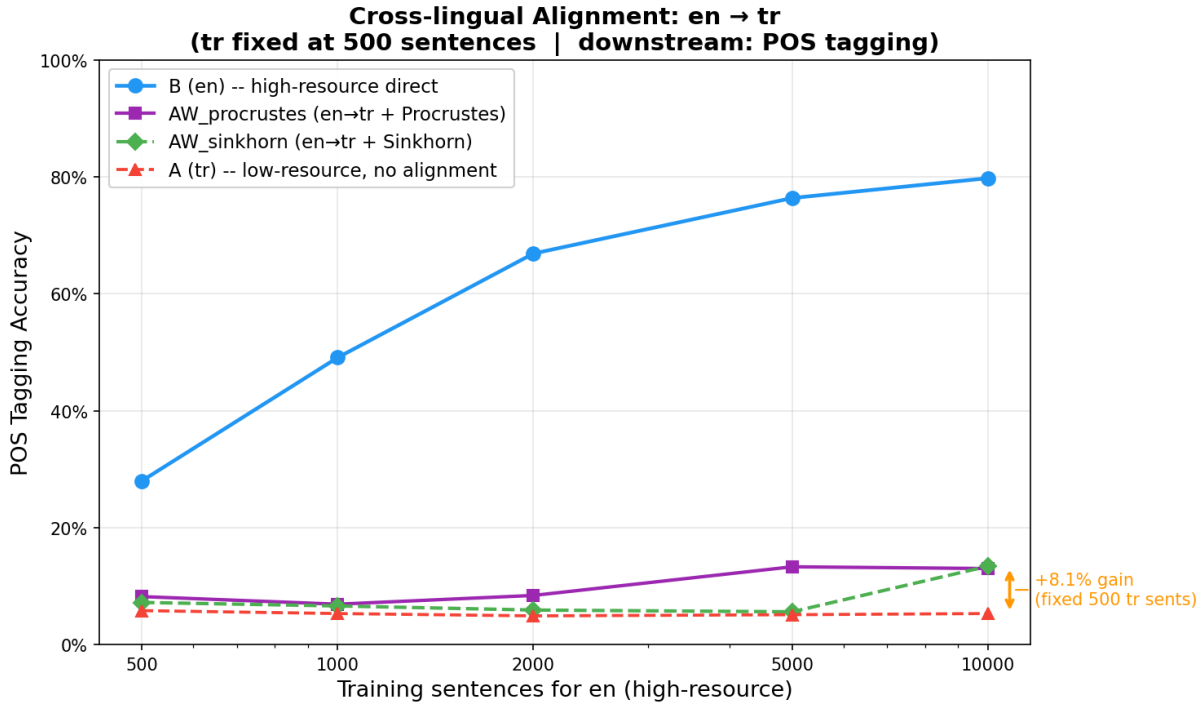

Figure 1: Cross-lingual POS tagging accuracy from English to Turkish under varying English training sizes.

are approximately isomorphic under a rigid transformation. The objective is

$$W^* = \arg\min_W \|X_s W - Y_s\|_F^2 \quad \text{s.t.} \quad W^\top W = I.$$

The orthogonality constraint restricts the transformation to rotations and reflections, preventing scaling or distortion of the embedding geometry. The optimal solution admits a closed-form expression via singular value decomposition (SVD): if $U\Sigma V^\top = \text{SVD}(X_s^\top Y_s)$, then $W^* = UV^\top$.

Second, when direct correspondences between embeddings are unavailable or sparse, we estimate alignment using entropy-regularized optimal transport solved with the **Sinkhorn algorithm**. We compute pairwise distances between embeddings to form a cost matrix C, and solve

$$T^* = \arg\min_{T \in \Pi} \langle T, C \rangle - \epsilon H(T),$$

where T is a transport matrix representing how probability mass moves between embedding distributions, Π denotes the set of valid transport plans, and $H(T)$ is an entropy regularizer. The resulting transport matrix provides a soft correspondence between embedding spaces, enabling alignment without relying on explicit bilingual supervision.

Our experimental design models controlled high-resource and low-resource settings. We fix language B as a hypothetical low-resource language and vary the amount of training data available for language A. This allows us to analyze how the quality of the learned mapping changes as cross-lingual supervision increases. This setup differs from prior work that aligns contextual embeddings from pretrained multilingual models. Training embeddings directly from parallel text allows us to control the amount of training data and simulate arbitrary resource regimes.

To evaluate the usefulness of the learned transformation, we perform cross-lingual POS tagging transfer. A logistic regression classifier is trained on embeddings from language A. The learned mapping W is then applied to embeddings from language B to produce transformed representations

$$Y' = YW.$$

The classifier trained on X is evaluated directly on Y'. Performance on the POS tagging task measures whether the transformation preserves syntactic information across languages.

5 Results

We conduct experiments using English as the high-resource language and Turkish as the low-resource language. We select this pair because the two languages differ substantially in morphology and typological structure. English is largely analytic,

whereas Turkish is highly agglutinative with rich morphological composition. This difference makes cross-lingual alignment non-trivial and provides a challenging setting for evaluating geometric transfer. In all experiments we fix the Turkish training corpus to 500 sentences and vary the English training corpus from 500 to 10,000 sentences. Figure 1 shows the downstream POS tagging accuracy under these settings.

The blue curve (B) represents the performance of the POS classifier trained and evaluated on English embeddings. As the number of English training sentences increases, the classifier accuracy improves consistently. This behavior follows well-known scaling trends in representation learning, where additional training data leads to improved downstream performance.

The red curve (A) represents the zero-shot transfer baseline, where the POS classifier trained on English is applied directly to Turkish embeddings without any alignment. Performance in this setting is extremely low (approximately 8%), indicating that the embedding spaces of the two languages are not directly compatible. We attribute this to the substantial structural and morphological differences between English and Turkish, which prevent a classifier trained on English representations from generalizing to Turkish without transformation.

Applying geometric alignment significantly improves transfer. Both Procrustes and Sinkhorn mappings increase POS accuracy relative to the unaligned baseline. The strongest improvement is observed with the Procrustes transformation, which achieves a gain of +8.1% over the unaligned Turkish embeddings when English is trained on 5,000 sentences. This result is notable because the improvement arises solely from a linear geometric transformation learned using only 500 sentences of the hypothetical low-resource language.

However, we also observe a plateau effect. Increasing English training data from 5,000 to 10,000 sentences does not produce further gains in cross-lingual accuracy. While Procrustes consistently outperforms Sinkhorn in our setting, the saturation suggests that only a limited portion of the cross-lingual structure can be recovered through simple linear alignment. This creates open questions about how much universal geometry truly exists across languages and how effectively it can be harnessed through linear transformations alone.

6 Conclusion

In this paper, we investigated the hypothesis that multilingual representations share a recoverable geometric structure across languages. Using English and Turkish as a case study; two languages with substantially different morphology and syntactic organization. We examined whether simple linear transformations can align their embedding spaces and enable cross-lingual transfer. Our results show that a transformation W learned through geometric alignment improves downstream POS tagging accuracy on Turkish embeddings, despite being trained with only 500 Turkish sentences.

This finding has important implications for multilingual NLP. First, it suggests that representations learned in one language may encode reusable abstract linguistic structure that can be transferred across typologically different languages. Second, it indicates that cross-lingual generalization may depend less on scale than on the geometry of representation spaces. Finally, our results highlight alignment itself as a form of weak supervision: even without annotated data in the target language, geometric transformations can partially recover useful structure for downstream tasks.

These observations are particularly relevant for extremely low-resource and endangered languages where available corpora may consist of only a few hundred sentences or even word lists. In such settings, even modest gains obtained through geometric alignment can have disproportionate impact. While our results remain preliminary, they provide encouraging evidence that the latent structure of multilingual representations may offer a viable path toward building NLP systems for languages with little or no labeled data.

7 Limitations

While our results are encouraging, several limitations must be acknowledged. First, the generalizability of our findings remains uncertain. Although English and Turkish differ substantially in morphology and syntactic structure, experiments on a single language pair cannot establish whether the observed geometric alignment holds across a broader set of typologically diverse languages. Future work should extend this analysis to additional language families and linguistic typologies.

Second, our experiments rely on FastText embeddings, an extension of the Word2Vec framework that incorporates subword information. While this

choice is well suited for low-resource settings, it remains unclear whether the geometric alignment observed here would transfer to contextual embedding models such as BERT, XLM, or other transformer-based architectures. Evaluating whether similar alignment properties hold in contextual representation spaces is an important direction for future research.

Finally, our results suggest the presence of diminishing returns as the amount of training data for the high-resource language increases. In particular, we observe a plateau when increasing English training data from 5,000 to 10,000 sentences under our best-performing alignment method. This may indicate limits to how much cross-lingual structure can be recovered through simple linear transformations. A deeper investigation of alternative alignment techniques and more expressive transformations may be necessary to fully understand the structure and limitations of multilingual embedding spaces.

References

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Maarten De Raedt, Fréderic Godin, Pieter Buteneers, Chris Develder, and Thomas Demeester. 2021. A simple geometric method for cross-lingual linguistic transformations with pre-trained autoencoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10108–10114, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. PMLR.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. 2025. Harnessing the universal geometry of embeddings. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*. Curran Associates, Inc.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2622–2628, Marseille, France. European Language Resources Association.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *Preprint*, arXiv:1309.4168.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. 2025. NüshuRescue: Reviving the endangered nüshu language with AI. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034, Abu Dhabi, UAE. Association for Computational Linguistics.