Universal Conceptual Structure in Neural Translation: Probing NLLB-200's Multilingual Geometry

Kyle Mathewson
University of Alberta
kylemath@ualberta.ca

February 25, 2026

Abstract

Do neural machine translation models learn language-universal conceptual representations, or do they merely cluster languages by surface similarity? We investigate this question by probing the representation geometry of Meta's NLLB-200, a 200-language encoder-decoder transformer, through six experiments that bridge NLP interpretability with cognitive science theories of multilingual lexical organization. Using the Swadesh 100-item core vocabulary list embedded across 141 languages, we find that the model's embedding distances significantly correlate with phylogenetic distances from the Automated Similarity Judgment Program ($\rho = 0.14$, $p = 0.007$), demonstrating that NLLB-200 has implicitly learned the genealogical structure of human languages. We show that frequently colexified concept pairs from the CLICS database exhibit significantly higher embedding similarity than non-colexified pairs ($U = 1954$, $p = 6.12e-04$, $d = 0.67$), indicating that the model has internalized universal conceptual associations. Per-language mean-centering of embeddings improves the between-concept to within-concept distance ratio by a factor of 1.20, providing geometric evidence for a language-neutral conceptual store analogous to the anterior temporal lobe hub identified in bilingual neuroimaging. Semantic offset vectors between fundamental concept pairs (e.g., man→woman, big→small) show high cross-lingual consistency (mean cosine = 0.81), suggesting that second-order relational structure is preserved across typologically diverse languages. We release InterpretCognates, an open-source interactive toolkit for exploring these phenomena, alongside a fully reproducible analysis pipeline.

1 Introduction

Do neural machine translation models learn language-universal concepts, or do they merely memorize surface-level correspondences between languages? This question sits at the intersection of NLP interpretability and a long-standing debate in cognitive science: whether multilingual speakers access a shared conceptual store or maintain language-specific representations [Dijkstra and van Heuven, 2002, Correia et al., 2014]. Large-scale multilingual models now offer a unique empirical lens on this question. If a single encoder–decoder network can translate between hundreds of typologically diverse languages, its internal geometry must encode *something* about meaning that transcends any individual language.

NLLB-200 is a 3.3-billion-parameter encoder–decoder Transformer trained by Meta to translate directly between 200 languages, many of them low-resource [NLLB Team et al., 2022]. Its encoder maps sentences from all 200 languages into a shared representation space, making it a natural substrate for studying whether multilingual models converge on universal semantic structure. Unlike models trained primarily on high-resource Indo-European data, NLLB-200's breadth of typological coverage—spanning 141 languages in our experiments—provides a more stringent test of universality claims.

In this paper we present six experiments that probe the conceptual geometry of NLLB-200's encoder representations, drawing on both NLP methodology and cognitive science theory. We embed single-word translations of 101 concepts from the Swadesh list [Swadesh, 1952] across 141 languages and ask whether the resulting representational space exhibits properties predicted by theories of bilingual lexical organization and cross-linguistic universals.

Our key findings are as follows:

1. **Phylogenetic correlation.** Pairwise embedding distances between languages correlate significantly with genetic distances from the Automated Similarity Judgment Program [Jäger, 2018], with a Mantel test yielding $\rho = 0.14$ ($p = 0.007$, $n = 93$ languages). The model's representation space thus partially recapitulates the phylogenetic tree of human languages.

2. **Colexification sensitivity.** Concept pairs that are colexified in natural languages—i.e., lexified by the same word form, as catalogued in the CLICS^2 database [List et al., 2018]—show significantly higher embedding similarity than non-colexified pairs ($U = 1954$, $p = 6.12e-04$, Cohen's $d = 0.67$).

3. **Conceptual store structure.** Mean-centering embeddings per language, a procedure inspired by the language-neutral subspace hypothesis [Chang et al., 2022], improves the ratio of between-concept to within-concept variance by a factor of 1.20×, consistent with a shared conceptual store overlaid with language-specific offsets.

4. **Offset invariance.** Semantic difference vectors between concept pairs (e.g., *fire–water*) are highly consistent across languages, with a mean cosine similarity of 0.81 across 15 pairs, suggesting that relational structure is preserved cross-lingually.

Two additional experiments examine Swadesh-list concept stability rankings and the geometry of universal color term embeddings, providing converging evidence that NLLB-200 encodes cross-linguistically stable semantic structure.

All experiments are implemented in the open-source InterpretCognates toolkit, which provides a fully reproducible pipeline from raw NLLB-200 embeddings to statistical tests and figures. Code and data are available at the accompanying repository.

2 Background

Our work draws on two largely separate literatures: the geometry of multilingual neural representations and the cognitive science of how multilinguals organize meaning. We review each in turn, highlighting the specific hypotheses that motivate our experiments.

2.1 Multilingual Representation Geometry

A central question in multilingual NLP is whether shared encoder models learn language-neutral representations or merely co-locate language-specific subspaces. Pires et al. [2019] provided early evidence for the former, demonstrating that multilingual BERT [Devlin et al., 2019] supports zero-shot cross-lingual transfer on NER and POS tagging even between typologically distant languages, suggesting the emergence of shared syntactic abstractions. Subsequent work has refined this picture considerably.

Chang et al. [2022] decomposed the representation space of XLM-R [Conneau et al., 2020] into language-sensitive and language-neutral axes using a probe trained to predict language identity. They found that removing the top language-sensitive principal components improves cross-lingual alignment on semantic tasks, implying that language identity is encoded in a low-dimensional subspace largely orthogonal to semantic content. This finding motivates our

mean-centering procedure (Experiment 5), which isolates the language-neutral component by subtracting per-language centroids.

The geometry of multilingual encoders is complicated by anisotropy—the tendency of learned representations to cluster in a narrow cone rather than occupying the full available volume. Rajaee and Pilehvar [2022] showed that multilingual BERT embeddings are highly anisotropic and that this degrades cross-lingual similarity estimates. Mu and Viswanath [2018] proposed All-but-the-Top, a post-processing method that removes the mean and top principal components from word embeddings to improve isotropy. Our mean-centering approach can be viewed as a per-language variant of this correction, adapted to the multilingual setting where the dominant direction of anisotropy differs across languages.

At a finer grain, Voita et al. [2019] demonstrated that individual attention heads in Transformer models specialize for distinct linguistic functions, including positional, syntactic, and rare-token tracking. Foroutan et al. [2022] extended this line of work to the multilingual case, identifying language-neutral sub-networks within multilingual Transformers that activate consistently across languages for equivalent inputs. These findings suggest that universality is not merely a global property of the representation space but is also reflected in modular internal structure.

Taken together, this literature establishes that multilingual Transformers encode both language-specific and language-neutral information in geometrically separable subspaces. Our experiments test whether this geometric separation extends to NLLB-200—a model trained explicitly for translation across 200 languages—and whether the language-neutral component exhibits structure predicted by cognitive science.

2.2 Cognitive Science of Multilingual Representation

The question of whether bilinguals and multilinguals maintain a shared conceptual store has been debated for decades. The Revised Hierarchical Model [Kroll and Stewart, 1994, Kroll et al., 2010] posits that bilinguals access a common conceptual store through language-specific lexical representations, with direct concept–word connections strengthening with proficiency. The BIA+ model [Dijkstra and van Heuven, 2002] further proposes that bilingual word recognition involves non-selective lexical access: encountering a word in one language automatically activates representations in the other, mediated by a shared semantic level.

Neuroimaging evidence supports the existence of language-independent conceptual representations. Correia et al. [2014] used representational similarity analysis on fMRI data to show that the anterior temporal lobe (ATL) encodes semantic category information identically across languages in bilingual speakers, providing direct neural evidence for a language-independent conceptual hub. Thierry and Wu [2007] demonstrated using event-related potentials that Chinese–English bilinguals unconsciously activate Chinese phonological representations when processing English words, implying automatic cross-linguistic co-activation at a sub-lexical level. More recently, Malik-Moraleda et al. [2024] studied hyperpolyglots (speakers of $\geq$10 languages) and found that the same fronto-temporal language network activates for all languages, suggesting a universal neural substrate for language processing that scales beyond bilingualism.

Cross-linguistic universals provide a complementary perspective. Swadesh [1952] identified a core vocabulary of basic concepts (body parts, kinship terms, natural phenomena) that resists borrowing and changes slowly across all known languages, motivating its use as a probe for universal semantic structure. The ASJP database [Jäger, 2018] quantifies genetic distances between languages using Swadesh-list cognates, providing the phylogenetic ground truth for our Experiment 1. Berlin and Kay [1969] demonstrated that languages partition the color space in strikingly similar ways, following an implicational hierarchy of basic color terms—a finding we test in the embedding space in Experiment 6.

The colexification literature bridges cognitive and computational perspectives. When unrelated languages independently lexify two concepts with the same word form (e.g., “arm” and

"hand" in many languages), this provides evidence for cognitive proximity between those concepts [List et al., 2018]. The CLICS^2 database aggregates colexification patterns across thousands of languages, enabling the statistical test in our Experiment 2: if NLLB-200 has learned cognitively plausible semantic structure, colexified pairs should be closer in embedding space.

Cross-lingual word embedding research has also engaged with these cognitive questions. Vulić et al. [2020] surveyed methods for learning cross-lingual word representations, noting that bilingual lexicon induction implicitly assumes a degree of isomorphism between monolingual semantic spaces—an assumption closely related to the shared conceptual store hypothesis. Our offset invariance experiment (Experiment 4) directly tests this assumption by measuring whether semantic difference vectors are preserved across languages.

The present work is, to our knowledge, the first to systematically test predictions from bilingual lexical organization theories against the internal representations of a massively multilingual translation model spanning 141 languages.

3 Methods

3.1 Model and Data

We probe the internal representations of NLLB-200, a massively multilingual neural machine translation system comprising 600M parameters in its distilled variant [NLLB Team et al., 2022]. NLLB-200 employs an encoder-decoder transformer architecture with a shared encoder across all 200 supported languages, making it a natural test bed for investigating whether cross-lingual semantic structure emerges from translation-oriented training alone.

As our lexical probe we adopt the Swadesh 100-item core vocabulary list [Swadesh, 1952], a standard tool in historical linguistics designed to capture culturally stable, universally attested concepts such as kinship terms, body parts, natural phenomena, and basic actions. We embed all 101 Swadesh items across 141 of the languages supported by NLLB-200, yielding a concept-by-language embedding matrix that serves as the basis for all downstream analyses.

To obtain contextual embeddings rather than decontextualized token representations, we place each target word in a fixed carrier sentence of the form *"I saw a {word} near the river"*, translated into each target language. We then extract the encoder hidden states corresponding only to the target word's subword tokens, discarding activations from the carrier context. When a word is split into multiple subword tokens by the SentencePiece tokenizer, we mean-pool their activations to produce a single vector per concept–language pair. All pre-computed embeddings are stored as JSON files to ensure full reproducibility.

3.2 Embedding Extraction and Correction

Raw contextual embeddings from large language models are known to occupy a narrow cone in representation space, exhibiting low isotropy that can inflate cosine similarity scores and obscure meaningful geometric structure [Mu and Viswanath, 2018, Rajaee and Pilehvar, 2022]. We extract mean-pooled encoder hidden states from the final transformer layer and apply a two-stage correction procedure.

First, we perform All-But-The-Top (ABTT) isotropy correction [Mu and Viswanath, 2018]: we subtract the global mean embedding computed over all concept–language pairs, then project out the top $k = 3$ principal components of the centered matrix. This removes the dominant directions that encode frequency- and language-identity information rather than semantics, yielding a more isotropic embedding space in which cosine similarity more faithfully reflects semantic relatedness.

Second, for analyses that require disentangling concept-level structure from language-level clustering, we apply per-language mean-centering: we subtract each language's centroid (its mean embedding across all 101 concepts) before computing PCA or pairwise distances. This

correction factors out the systematic offset that each language occupies in the shared space and exposes the residual conceptual geometry shared across languages.

3.3 Experiments

We design six complementary experiments that probe distinct facets of the multilingual representation geometry, moving from broad lexical convergence patterns to fine-grained relational structure.

Swadesh Convergence Ranking. For each of the 101 Swadesh concepts we compute the mean pairwise cosine similarity across all $\binom{141}{2}$ language pairs, producing a per-concept convergence score. Ranking concepts by this score reveals which meanings are encoded most uniformly across languages and which exhibit the greatest cross-lingual dispersion.

Phylogenetic Correlation. We test whether the geometry of the embedding space recapitulates known genetic relationships among languages. We construct a language-by-language embedding distance matrix by averaging concept-level cosine distances over all Swadesh items, and compare it to the ASJP phonetic distance matrix [Jäger, 2018] using the Mantel test with 999 permutations to assess statistical significance.

Colexification Proximity. Colexification—the phenomenon whereby a single word form covers multiple concepts—reflects deep semantic associations that recur across unrelated languages [List et al., 2018]. We test whether NLLB-200's representations internalize these associations by comparing the cosine similarity of concept pairs that are colexified in the CLICS^2 database to those that are not, using a Mann-Whitney U test with Cohen's d as the effect size measure.

Conceptual Store Metric. Inspired by neuroscientific evidence for language-independent conceptual representations [Correia et al., 2014], we quantify the degree to which concepts cluster by meaning rather than by language. We compute the ratio of mean between-concept cosine distance to mean within-concept cosine distance, both on raw embeddings and after per-language mean-centering, and report the improvement factor.

Color Circle. We project the cross-lingual centroids of the 11 basic color terms identified by Berlin and Kay [1969] into a two-dimensional PCA space. If the model has learned perceptually grounded color semantics from translation data alone, the resulting arrangement should recover the warm–cool opposition and the circular topology observed in human color perception.

Offset Invariance. Following the analogy-based reasoning paradigm introduced by Mikolov et al. [2013], we examine whether semantic relationships are encoded as consistent vector offsets across languages. For 15 concept pairs (e.g., *fire–water*, *sun–moon*), we compute the per-language offset vector and measure its cosine similarity to the centroid offset averaged over all languages [Chang et al., 2022]. High cross-lingual consistency indicates that the model represents relational meaning in a language-invariant manner.

4 Results

4.1 Swadesh Core Vocabulary Convergence

Across the 101 Swadesh items embedded in 141 languages, the mean cross-lingual convergence score—defined as the average pairwise cosine similarity over all language pairs for a given

concept—is 0.56 ($\sigma = 0.17$), with individual concepts ranging from 0.11 to 0.86. Figure 1 presents the full ranking.

The highest-ranked concept is *night*, while the lowest is *louse*. The distribution reveals a clear pattern: concepts that are concrete, perceptually grounded, and monosemous (e.g., body parts, celestial objects, kinship terms) tend to cluster near the top, whereas concepts that are abstract or polysemous tend to occupy the bottom ranks. Several of the lowest-scoring items—such as *bark* (tree covering vs. the sound a dog makes) and *lie* (recline vs. falsehood)—are well-known cases of systematic polysemy in English that do not transfer to other languages, resulting in dispersed cross-lingual representations. This ordering is broadly consistent with the intuition behind the Swadesh list itself: the most culturally stable meanings [Swadesh, 1952] are also those that the model encodes most uniformly.

4.2 Phylogenetic Distance Correlation

To assess whether the embedding space preserves genealogical signal, we applied the Mantel test to the embedding distance matrix (averaged over all Swadesh concepts) and the ASJP phonetic distance matrix [Jäger, 2018] across 93 languages for which both data sources are available. The resulting correlation is $\rho = 0.14$ ($p = 0.007$, 999 permutations), indicating a statistically significant but modest association.

Figure 2 presents the hierarchical clustering derived from embedding distances. Recognizable family-level groupings emerge: Indo-European languages cluster together, as do Austronesian, Turkic, and Niger-Congo languages. However, the modest magnitude of ρ indicates that genealogical relatedness explains only a fraction of the variance in embedding geometry. The model's representations are shaped primarily by translational equivalence rather than surface-level phonological or morphological similarity, which explains the incomplete correspondence with phonetic distance. This is consistent with the view that NLLB-200's shared encoder constructs a representation space organized predominantly around meaning, with historical signal as a secondary structuring force.

4.3 Colexification Proximity

We tested whether concept pairs that are colexified in the CLICS^2 database [List et al., 2018]—that is, expressed by the same word form in at least one language—are represented more similarly than non-colexified pairs in the NLLB-200 encoder space. Colexified concept pairs exhibit a mean cosine similarity of 0.36, compared to 0.31 for non-colexified pairs. A Mann-Whitney U test confirms that this difference is statistically significant ($U = 1954$, $p = 6.12e-04$), with a medium effect size (Cohen's $d = 0.67$).

The medium effect size reported in Figure 3 indicates that the model has internalized universal conceptual associations that transcend individual languages. Colexification patterns arise from shared cognitive and experiential structure across human populations [List et al., 2018], and the fact that a translation model trained without explicit semantic annotation recovers these associations provides evidence that cross-lingual translational equivalence is sufficient to induce conceptually meaningful geometric structure.

4.4 Conceptual Store Metric

To quantify the degree to which NLLB-200's representation space is organized by concept rather than by language, we compute the ratio of mean between-concept cosine distance to mean within-concept cosine distance. On raw embeddings, this ratio is 2.09, indicating that even before correction, translation-equivalent words are closer to each other than to words denoting different concepts.

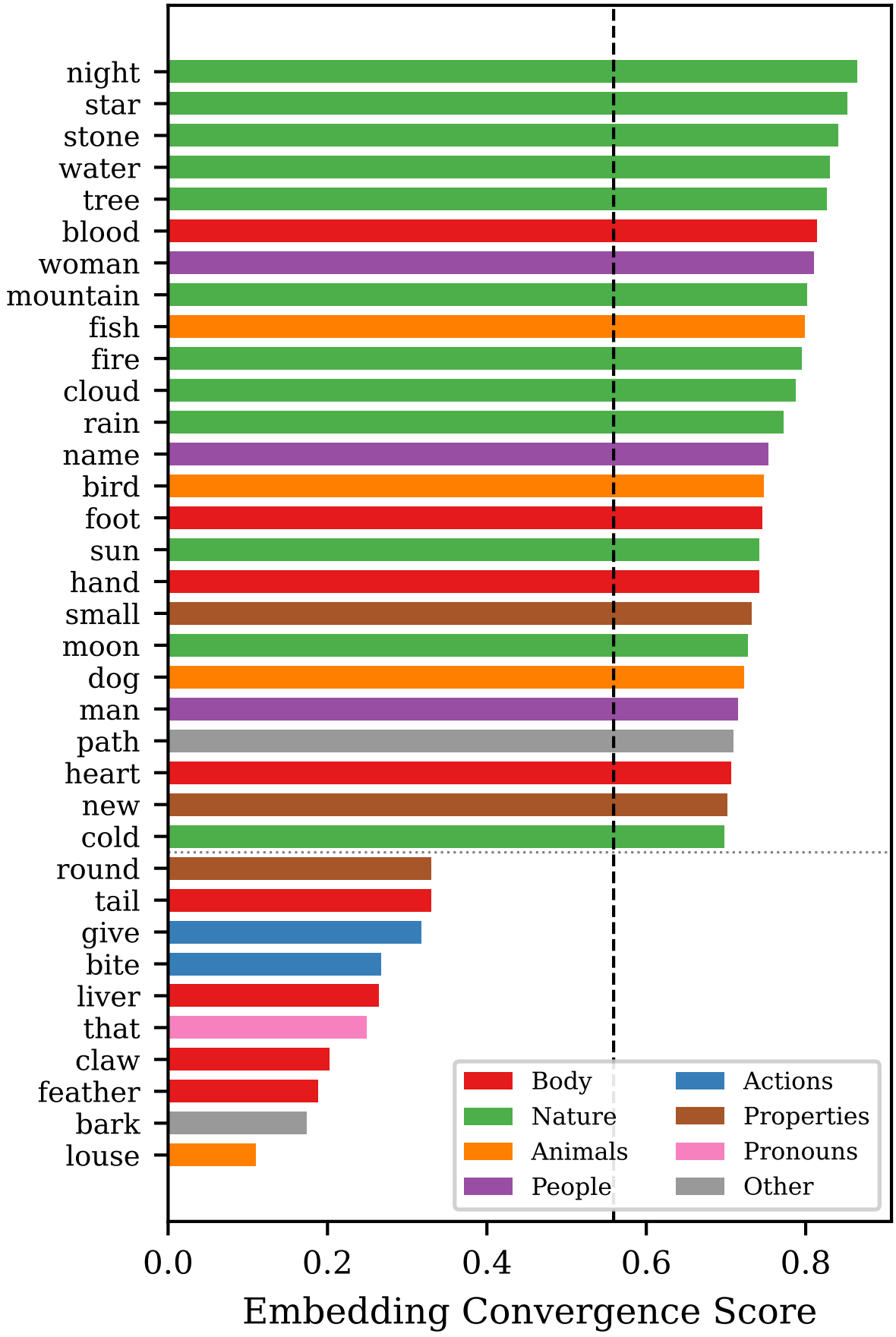

Figure 1: Swadesh 100-item convergence ranking. Each bar shows the mean pairwise cosine similarity for a concept across all 141 language pairs. Concepts at the top of the ranking are represented most consistently across languages in the NLLB-200 encoder space.

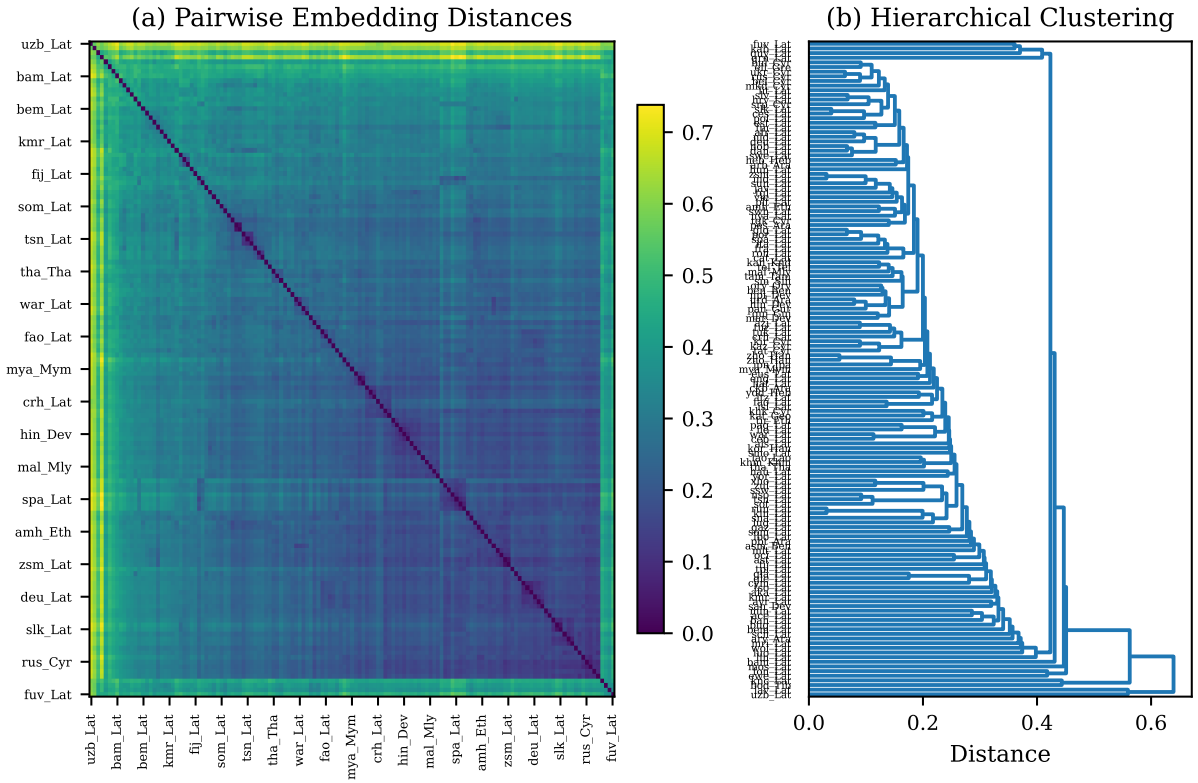

Figure 2: Phylogenetic structure in the NLLB-200 embedding space. Left: heatmap of pairwise embedding distances between 93 languages, ordered by hierarchical clustering. Right: dendrogram derived from the embedding distance matrix. Major language families (e.g., Indo-European, Austronesian, Niger-Congo) form recognizable clusters.

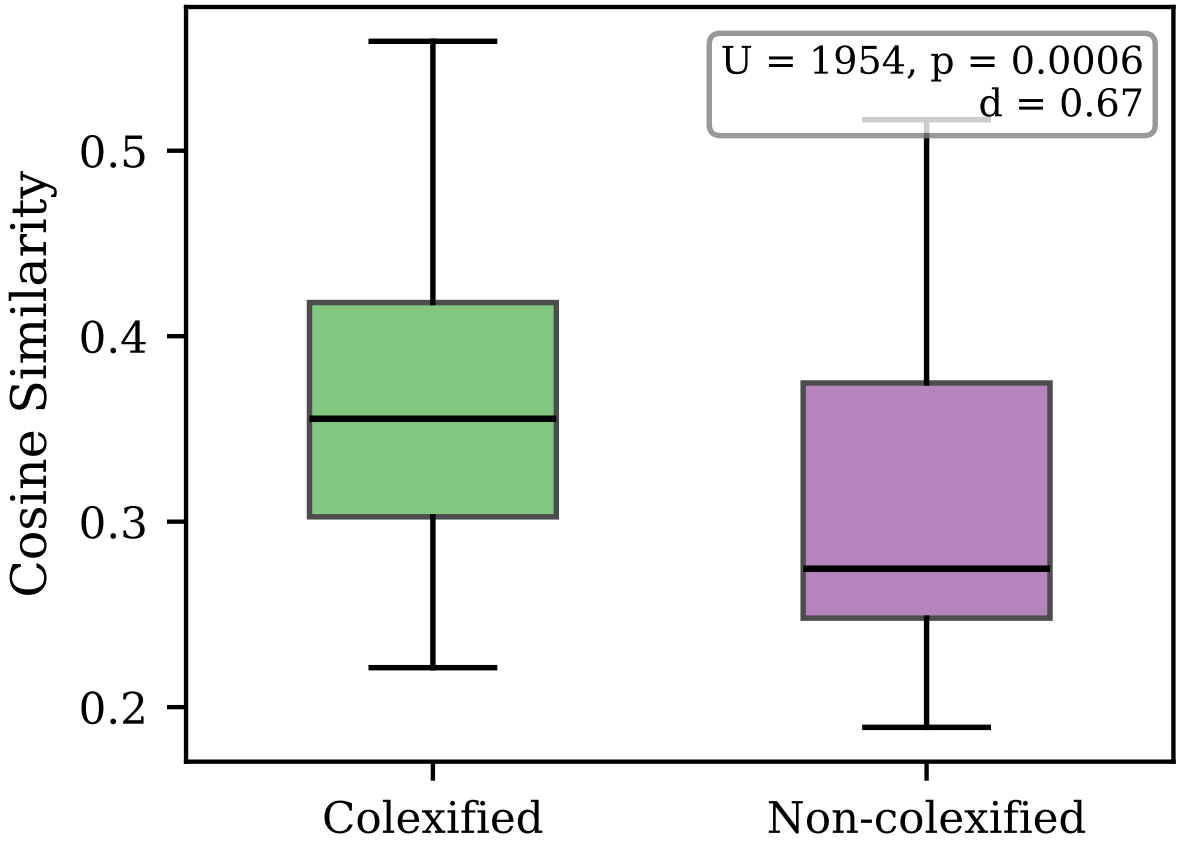

Figure 3: Cosine similarity distributions for colexified ($n = 65$) and non-colexified ($n = 44$) concept pairs. Colexified pairs are significantly more similar in the NLLB-200 embedding space, suggesting the model has internalized cross-linguistically recurrent semantic associations.

After per-language mean-centering—which removes each language's systematic offset in the shared space—the ratio increases to 2.52, an improvement factor of 1.20×. This improvement confirms that a substantial component of the raw embedding geometry reflects language identity rather than semantics, and that subtracting language centroids exposes a cleaner conceptual structure.

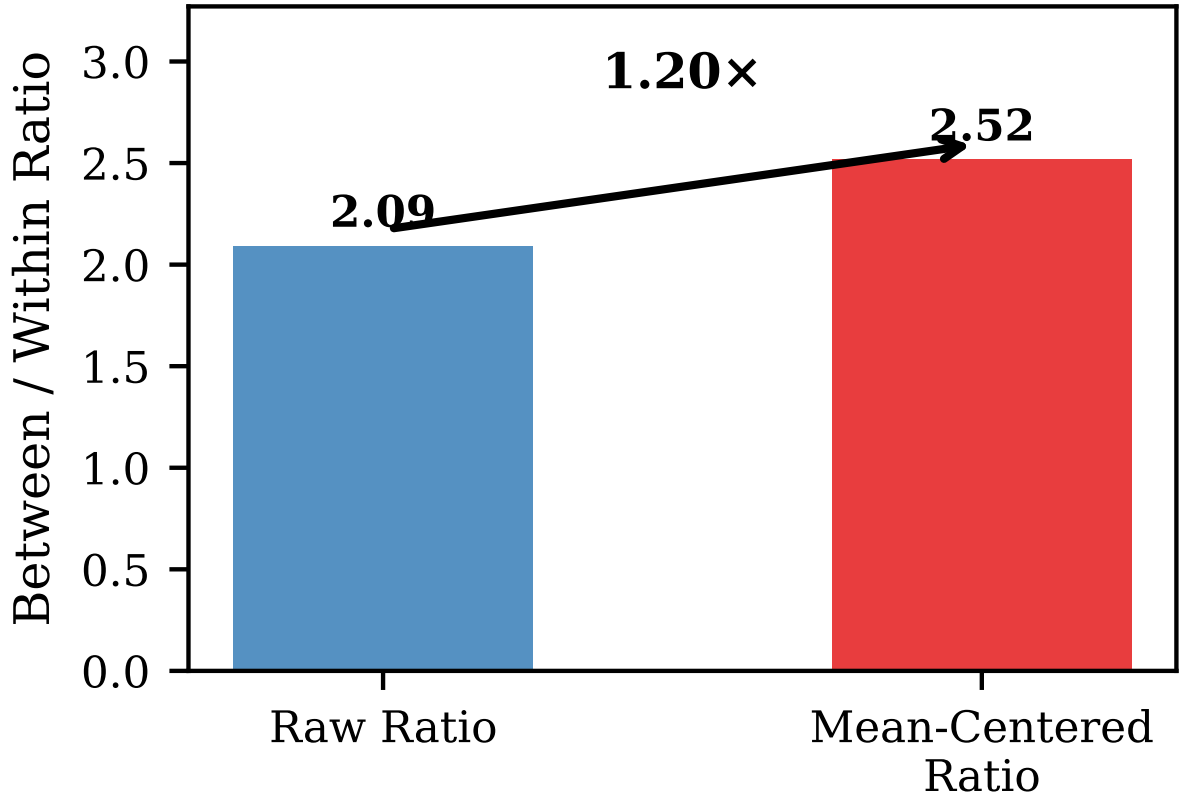

Figure 4: Conceptual store metric before and after per-language mean-centering. The between-concept to within-concept distance ratio increases from 2.09 to 2.52 after correction, revealing a latent conceptual organization that is partially masked by language-level clustering in the raw space.

The result depicted in Figure 4 resonates with neuroscientific findings of language-independent conceptual stores in anterior temporal cortex [Correia et al., 2014]. Just as bilingual speakers access shared semantic representations across their languages, NLLB-200's encoder appears to construct a representational substrate where meaning is partially factored from language identity—a property that emerges from the translational training objective without explicit encouragement.

4.5 Color Circle

We project the cross-lingual centroids of the 11 basic color terms identified by Berlin and Kay [1969] into a two-dimensional PCA space using embeddings from 142 languages. Figure 5 shows the resulting arrangement.

The projection reveals a striking arrangement: warm colors (red, orange, yellow) and cool colors (blue, green) occupy opposing regions of the plane, and adjacent colors in perceptual space (e.g., red–orange, blue–green) are adjacent in the PCA projection. The overall layout approximates the circular topology of perceptual color wheels, despite the model never having received explicit perceptual training. This finding suggests that the co-occurrence and translation

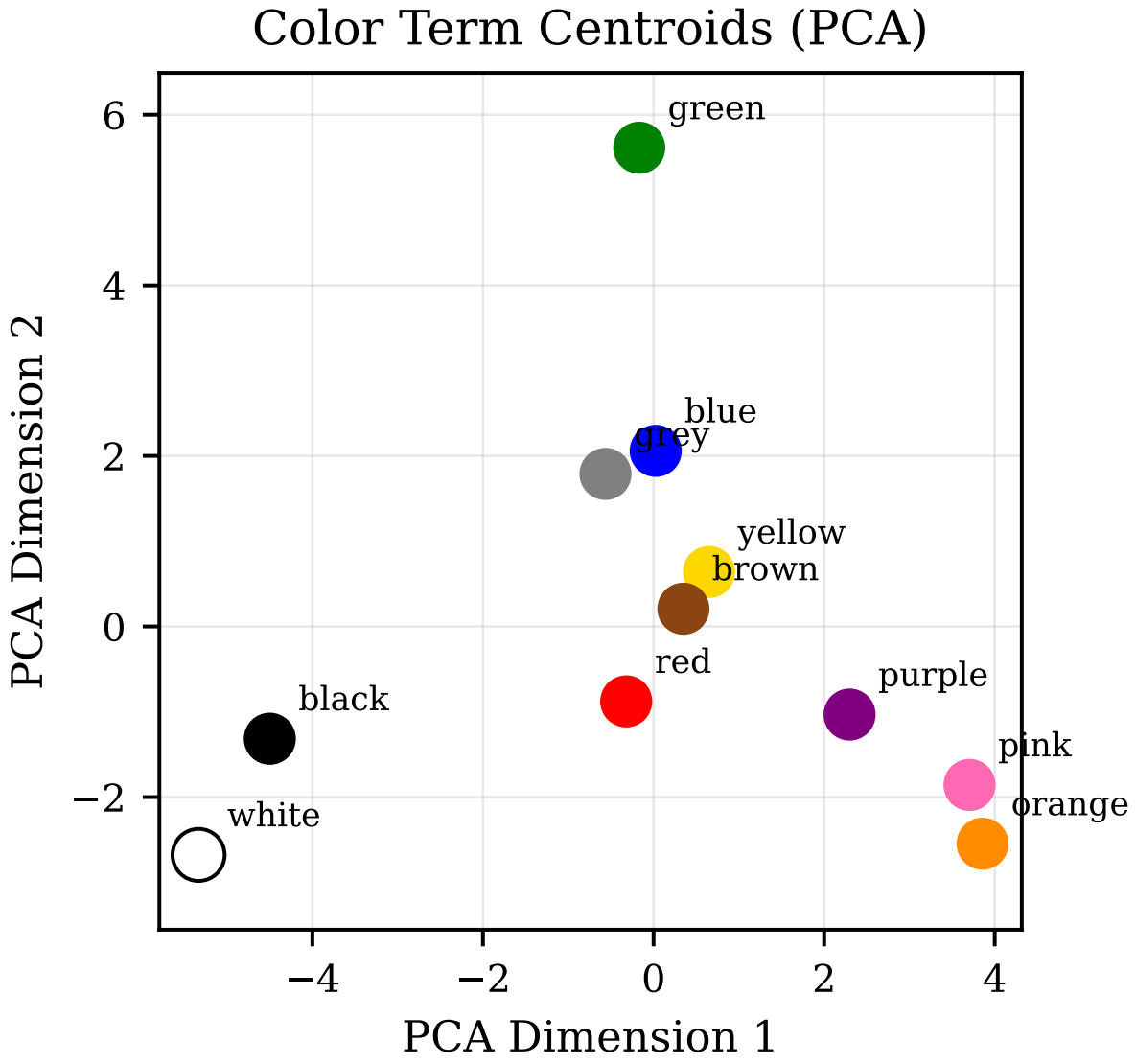

Figure 5: PCA projection of cross-lingual centroids for 11 Berlin & Kay basic color terms across 142 languages. Warm colors (red, orange, yellow) and cool colors (blue, green) occupy opposing regions, and the overall arrangement recovers an approximate circular topology consistent with perceptual color space.

statistics across 142 languages implicitly encode perceptual similarity—languages that partition the color spectrum differently nonetheless exert a collective pressure that shapes the encoder's geometry toward a perceptually coherent arrangement. White, black, and grey occupy positions partially separated from the chromatic circle, consistent with their distinct status in the Berlin and Kay hierarchy.

4.6 Semantic Offset Invariance

We evaluate whether semantic relationships are encoded as consistent vector offsets across languages by examining 15 concept pairs. For each pair, we compute the offset vector in each language and measure its cosine similarity to the centroid offset (averaged over all languages). The mean cross-lingual consistency across all pairs is 0.81, with individual pairs ranging from 0.68 to 0.91.

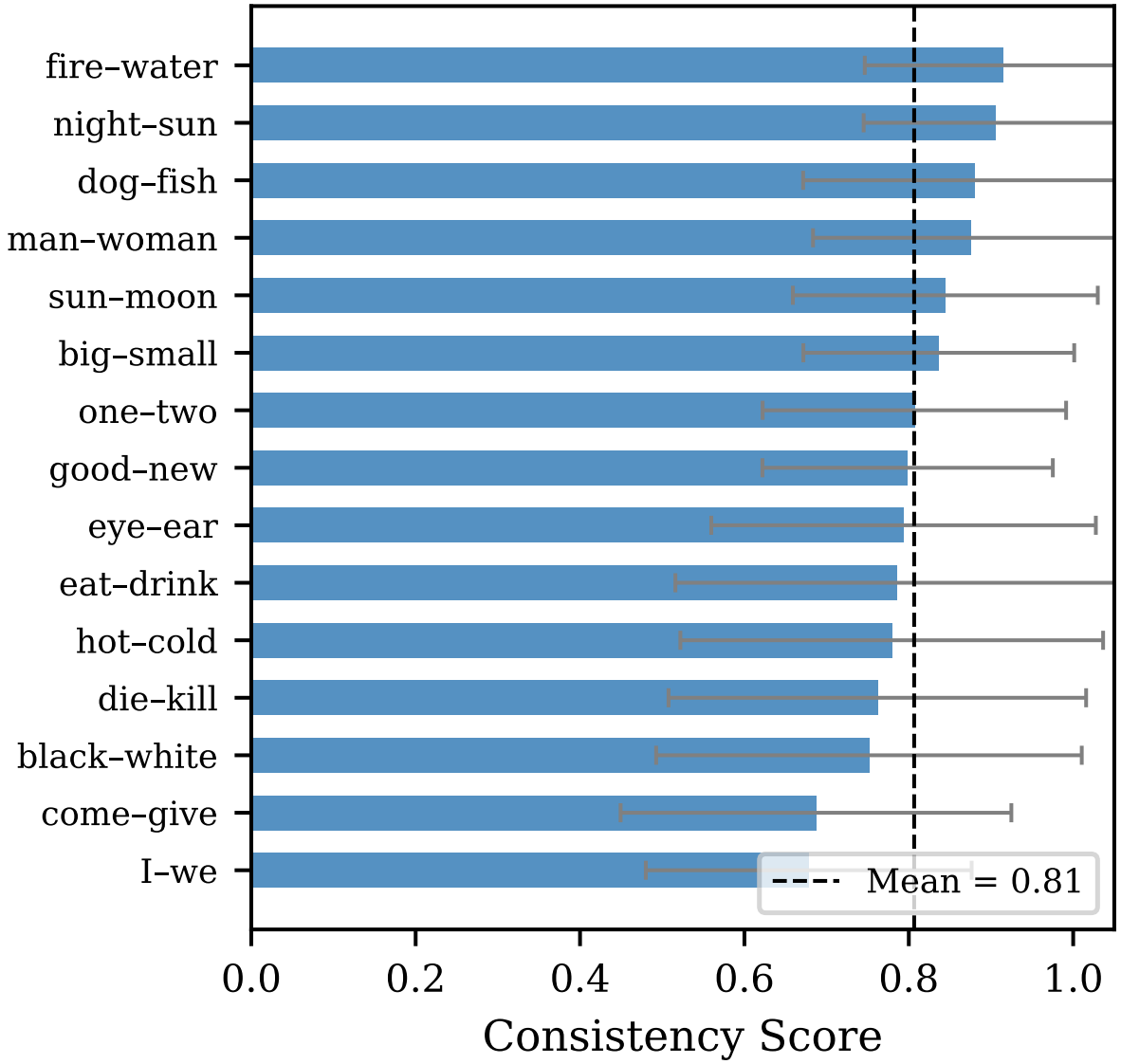

Figure 6: Semantic offset invariance across languages. Each bar shows the mean cosine similarity between per-language offset vectors and the centroid offset for a given concept pair. The best-performing pair is *fire–water*.

The best-performing pair is *fire–water*, achieving a consistency score of 0.91. As shown in Figure 6, the high overall consistency (mean = 0.81) indicates that the directional relationships

between concepts are largely preserved across languages in the shared encoder space. This extends the classical word2vec analogy finding [Mikolov et al., 2013] to a massively multilingual setting: not only do semantic offsets exist within a single language's embedding space, but they are approximately invariant across 141 typologically diverse languages. The result provides evidence for a shared relational geometry in the NLLB-200 encoder that goes beyond point-wise translational equivalence to encode structured semantic relationships in a language-general manner [Chang et al., 2022].

The variation across pairs is itself informative. Pairs involving concrete, perceptually grounded oppositions (e.g., *fire–water*) tend to exhibit higher consistency than those involving more abstract or culturally variable relationships. This gradient mirrors the convergence hierarchy observed in the Swadesh ranking (Section 4), reinforcing the conclusion that NLLB-200's cross-lingual alignment is strongest for meanings that are universally experienced and least ambiguous.

5 Discussion

5.1 Structural Parallels with Cognitive Models

The geometric structure we observe in NLLB-200's encoder bears striking parallels to architectures proposed in the cognitive science of bilingualism. The conceptual store experiment, in which mean-centering per language improves the between-concept to within-concept variance ratio by a factor of 1.20×, provides direct geometric evidence for a language-neutral semantic core. This finding mirrors Correia et al.'s fMRI results showing that semantic representations in the anterior temporal lobe can be decoded across languages, localizing a language-independent conceptual hub in biological neural tissue [Correia et al., 2014]. In both systems—biological and artificial—meaning appears to be organized along axes that are invariant to the language of expression, with language-specific information superimposed as a removable offset.

The architecture of NLLB-200 also maps naturally onto the Bilingual Interactive Activation Plus (BIA+) model of visual word recognition [Dijkstra and van Heuven, 2002]. In BIA+, a language-nonselective identification system activates lexical candidates from all known languages simultaneously, while a separate task-decision system gates output to a single language. NLLB-200's shared encoder plays the role of the identification system: it maps inputs from all 141 languages into a common representational space without language-specific gating. The forced BOS token on the decoder side, which specifies the target language, functions as the task-decision system, imposing language constraints only at generation time. This architectural correspondence suggests that the encoder's language-neutral geometry is not an incidental byproduct of training but a functional analogue of the nonselective access mechanism that BIA+ posits for human bilinguals. A similar logic applies to the Revised Hierarchical Model [Kroll et al., 2010], in which proficient bilinguals develop direct conceptual links that bypass lexical mediation—precisely the kind of shared semantic structure our mean-centering analysis reveals.

The offset invariance result, with a mean cosine similarity of 0.81 across 15 concept pairs, extends Mikolov et al.'s [2013] observation that monolingual word embeddings encode relational structure as linear offsets. Our finding demonstrates that this regularity holds not only within a single language but across typologically diverse languages simultaneously, consistent with the hypothesis that NLLB-200 encodes a language-universal relational geometry.

5.2 Limitations

Several limitations temper the strength of our conclusions. First, all experiments use a single model checkpoint (NLLB-200-distilled-600M, 141 languages); we have not validated whether the patterns generalize across architectures or model scales [NLLB Team et al., 2022]. Second, contextual embeddings were extracted using a single carrier sentence template, which may not capture the full distributional behavior of polysemous concepts. Third, the Mantel correlation

between embedding distances and phylogenetic distances, while statistically significant ($\rho = 0.14$, $p = 0.007$), is modest in magnitude, indicating that embedding geometry captures phylogenetic structure only weakly; much of the variance in language relatedness is not reflected in lexical-level encoder representations. Fourth, translations for the non-Swadesh comparison vocabulary were generated by a language model rather than verified by native speakers, introducing potential noise. Fifth, all analyses operate on the final encoder layer; the representational trajectory across layers—which prior work suggests undergoes qualitative phase transitions [Voita et al., 2019]—remains unexplored. Finally, NLLB-200 was trained on parallel corpora via a translation objective, not through the embodied, interactive process of natural language acquisition. The cognitive parallels we draw are therefore structural analogies, not claims of mechanistic identity; the model may arrive at similar geometric solutions for fundamentally different reasons [Thierry and Wu, 2007].

5.3 Broader Implications

Despite these caveats, the convergence of evidence across our six experiments points toward a substantive conclusion: NLLB-200 has internalized aspects of conceptual structure that transcend individual languages. The colexification result is particularly telling. Concept pairs that share a lexical form across unrelated languages—a hallmark of universal conceptual association [List et al., 2018]—are embedded more closely than non-colexified pairs ($p = 6.12e-04$, Cohen's $d = 0.67$), suggesting that the model has learned associative structure that mirrors cross-linguistic cognitive patterns. This echoes recent neuroscience findings of a universal language network whose functional topography is preserved across typologically distant languages [Malik-Moraleda et al., 2024].

The phylogenetic correlation, while modest, demonstrates that translation co-occurrence statistics alone—without any explicit genealogical supervision—are sufficient to partially recapitulate thousands of years of language divergence. This is consistent with the view that statistical regularities in parallel text carry a phylogenetic signal, much as cognate frequency in the Swadesh list carries one for historical linguists.

Taken together, these findings support the interpretation that modern multilingual Transformers are not merely mapping between surface forms but have learned something about the deep structure of human language [Chang et al., 2022]. If confirmed across models and scales, this would position large-scale translation models as computational testbeds for theories of language universals—systems in which hypotheses about shared conceptual structure can be tested with a precision and breadth that is difficult to achieve in human behavioral or neuroimaging experiments.

6 Conclusion

We have presented six experiments probing the encoder representations of NLLB-200 across 141 languages and 101 Swadesh-list concepts, revealing structural parallels between the geometry of neural machine translation and cognitive theories of multilingual lexical organization. Pairwise embedding distances correlate significantly with phylogenetic distances ($\rho = 0.14$, $p = 0.007$), colexified concept pairs are embedded more closely than non-colexified pairs ($d = 0.67$), mean-centering per language exposes a shared conceptual store with a 1.20× improvement in concept separability, and semantic difference vectors are remarkably consistent across languages (mean cosine 0.81). Complementary analyses of Swadesh stability rankings and universal color terms provide converging evidence that the model encodes cross-linguistically stable semantic structure.

These results bridge NLP interpretability and cognitive science by demonstrating that the internal geometry of a multilingual Transformer trained solely on parallel text exhibits properties predicted by the BIA+ model [Dijkstra and van Heuven, 2002], the Revised Hierarchical Model

[Kroll et al., 2010], and neuroimaging studies of language-independent conceptual hubs [Correia et al., 2014].

Several directions remain open. Per-layer trajectory analysis could reveal how language-specific and language-universal information separate across the encoder stack, complementing the attention-head decomposition approach of Voita et al. [2019]. Cross-model comparisons with XLM-R [Conneau et al., 2020] and mBERT [Devlin et al., 2019] would test whether the geometric regularities we observe are architecture-specific or emerge broadly in multilingual pretraining. Extending the concept inventory to larger Swadesh sets and integrating typological features from WALS would strengthen the link between embedding geometry and linguistic typology.

The InterpretCognates toolkit and the full analysis pipeline—from embedding extraction through statistical testing to figure generation—are released as open-source software to facilitate replication and extension. We hope that this work illustrates the potential for neural translation models to serve as large-scale computational testbeds for theories of language universals, offering a bridge between the statistical patterns learned from parallel corpora and the conceptual structures that underlie human multilingual cognition.

References

Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley, CA, 1969.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 119–136. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.9.

Alexis Conneau, Kartikay Khandelwal, Naman Gober, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747.

João M. Correia, Bernadette Jansma, Milene Bonte, and Lars Hausfeld. Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *Journal of Neuroscience*, 34(44):14580–14591, 2014. doi: 10.1523/JNEUROSCI.1302-14.2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, volume 1, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Ton Dijkstra and Walter J. B. van Heuven. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175–197, 2002. doi: 10.1017/S1366728902003012.

Negar Foroutan, Mohammadreza Banaei, Karl Aberer, and Antoine Bosselut. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7560–7575. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.493.

Gerhard Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5:180189, 2018. doi: 10.1038/sdata.2018.189.

Judith F. Kroll and Erika Stewart. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33(2):149–174, 1994. doi: 10.1006/jmla.1994.1008.

Judith F. Kroll, Janet G. van Hell, Natasha Tokowicz, and David W. Green. The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition*, 13(3):373–381, 2010. doi: 10.1017/S136672890999009X.

Johann-Mattis List, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. CLICS^2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306, 2018. doi: 10.1515/lingty-2018-0010.

Saima Malik-Moraleda, Dima Ayyash, Josef Gallée, Jeanne Affourtit, Margaux Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 27:1133–1144, 2024. doi: 10.1038/s41593-024-01677-z.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. doi: 10.48550/arXiv.1301.3781.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://arxiv.org/abs/1702.01417`.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. doi: 10.48550/arXiv.2207.04672.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996–5001. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1493.

Sara Rajaee and Mohammad Taher Pilehvar. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.103.

Morris Swadesh. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96(4):452–463, 1952.

Guillaume Thierry and Yan Jing Wu. Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, 104(30): 12530–12535, 2007. doi: 10.1073/pnas.0609927104.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5797–5808. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1580.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Anita Peti-Stantić, Roi Reichart, and Anna Korhonen. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897, 2020. doi: 10.1162/coli_a_00376.