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

Kyle Mathewson
University of Alberta
`kyle.mathewson@ualberta.ca`

February 27, 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 core vocabulary list embedded across 135 languages, we find that the model's embedding distances significantly correlate with phylogenetic distances from the Automated Similarity Judgment Program ($\rho = 0.13$, $p = 0.020$), 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 ($U = 42656$, $p = 1.33e-11$, $d = 0.96$), 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 1.19, 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 = 0.84), 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, Chen et al., 2025]. 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 135 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 135 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 $\rho = 0.13$ ($p = 0.020$, $n = 88$ 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[3] database [List et al., 2018, Rzymski et al., 2020]—show significantly higher embedding similarity than non-colexified pairs ($U = 42656$, $p = 1.33e - 11$, Cohen's $d = 0.96$).

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 $1.19\times$, 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 0.84 across 22 pairs, suggesting that relational structure is preserved cross-lingually.

Two additional experiments—Swadesh convergence ranking and universal color term geometry—provide converging evidence. A comparison against modern loanword-heavy vocabulary reveals that high embedding convergence can reflect orthographic borrowing rather than semantic universality, underscoring the importance of the external validation experiments. We further validate these findings through isotropy correction analysis, regression controls confirming that surface-form similarity explains less than 2% of convergence variance, and a per-family offset consistency analysis that reveals how relational structure varies across language families.

All experiments are implemented in the open-source INTERPRETCOGNATES toolkit, which provides a fully reproducible pipeline from pre-computed embeddings to statistical tests and figures. Code and data are available at `https://github.com/kylemath/InterpretCognates`.

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 conceptual store experiment, 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. More recently, Chen et al. [2025] used voxelwise encoding models on several hours of naturalistic narrative fMRI data to show that Chinese–English bilinguals employ largely shared semantic brain representations across languages, but that these representations undergo systematic fine-grained *shifts* between languages—shifts that modulate how concept categories are weighted rather than which brain regions are recruited. This finding of shared-but-modulated representations provides striking neural support for the geometric picture we advance computationally: a language-neutral conceptual core with language-specific offsets superimposed. 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. [2022] investigated the fronto-temporal language network across 45 languages

spanning 12 language families and found that the same network activates for all languages with consistent left-lateralization and functional selectivity, suggesting a universal neural substrate for language processing.

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 phylogenetic correlation analysis. 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 our color circle experiment.

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[3] database [Rzymski et al., 2020] aggregates colexification patterns across thousands of languages, enabling the statistical test in our colexification experiment: 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] introduced a large-scale multilingual evaluation benchmark covering 12 typologically diverse languages, 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 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 135 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 [Vaswani et al., 2017] 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 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 135 languages supported by NLLB-200 (excluding a small set of languages whose encoder embeddings are degenerate outliers in our extraction pipeline), 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. This choice is motivated by the observation that Transformer encoder representations are highly context-dependent [Devlin et al., 2019]: a bare word input would yield an embedding dominated by positional and start-of-sequence artifacts rather than lexical semantics. The carrier sentence provides a minimal, semantically neutral context that activates the target word's lexical representation while minimizing confounds from sentential semantics. 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. Agglutinative and polysynthetic languages tend to produce longer subword sequences, as morphological markers (case, evidentiality, tense) are segmented into additional tokens whose mean-pooling may attenuate fine-grained lexical features. We note that this carrier sentence is English-derived and imposes structural assumptions (SVO order, articles, spatial prepositions) that are not typologically universal; we assess this confound in Sections 4 and 5. To support reproducibility, we release the full analysis code and the derived experiment outputs (JSON summaries, figures, and macro tables) used to build the paper; the pipeline can also be rerun from model weights and the included corpora.

To assess the impact of the carrier sentence on our results, we additionally extract decontextualized embeddings by embedding each target word in isolation, without any surrounding context. This bare-word baseline tests whether the convergence patterns we report in Section 4 are driven by the carrier sentence's syntactic scaffold or by the target word's intrinsic lexical representation.

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. The choice of $k = 3$ is validated by a sensitivity analysis across a range of k values (Section 4), which confirms that the convergence ranking is robust to this hyperparameter.

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 $\binom{135}{2}$ 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.

5

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 CLICS³ database [Rzymski et al., 2020] 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 22 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.

We supplement these six experiments with a validation analysis that assesses the robustness of our embedding corrections.

Isotropy Correction Validation. To verify that our ABTT correction does not distort the convergence signal, we compare the full Swadesh ranking under raw and corrected embeddings. We compute the Spearman rank correlation ρ between the two orderings and visually inspect the top-20 concepts under each regime. A high correlation indicates that isotropy correction preserves the relative ordering of concepts while rescaling absolute similarity values to a more interpretable range.

To disentangle orthographic from semantic contributions to embedding convergence, we additionally regress convergence scores against mean orthographic and phonological similarity of each concept's word forms across Latin-script languages, reporting R^2 alongside the Swadesh convergence ranking.
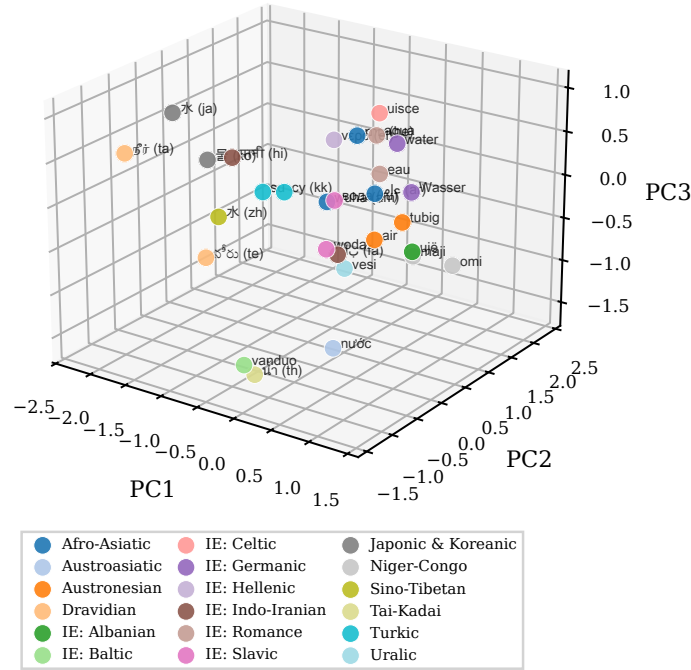
4 Results

We present our six core experiments alongside descriptive illustrations and validation analyses, organized along a progression from broad distributional patterns through external validation to geometric tests of cognitive hypotheses.

4.1 Illustrative Example: Water

Before presenting the full ranking, we illustrate the geometry of a single concept. Figure 1 shows the 29-language embedding manifold for "water" — a concept with diverse surface forms across language families (*agua*, *eau*, *Wasser*, *maji*, etc.).

Despite radically different surface forms and scripts, the embeddings cluster tightly in representation space: same-family languages form sub-clusters, but cross-family similarity
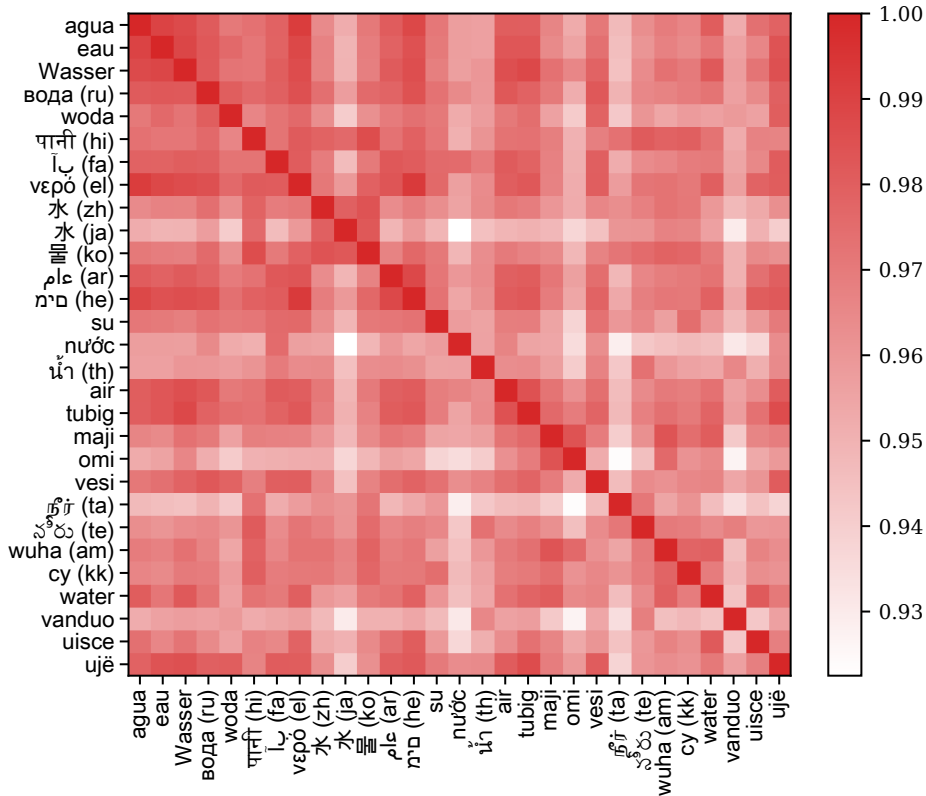
Figure 1: Embedding geometry for the concept "water" across 29 languages. (a) 3D PCA projection colored by language family shows tight clustering despite orthographic diversity. (b) Pairwise similarity heatmap reveals that same-family languages (e.g., Romance, Slavic) cluster, but cross-family similarity remains high (> 0.93 for most pairs).

remains high, suggesting that the model maps "water" to a shared semantic region regardless of its lexical realization. This example motivates the systematic analysis that follows.

4.2 Swadesh Core Vocabulary Convergence

Across the 101 Swadesh items embedded in 135 languages, the mean cross-lingual convergence score—defined as the average pairwise cosine similarity over all language pairs for a given concept—is 0.58 ($\sigma = 0.17$), with individual concepts ranging from 0.12 to 0.91. Figure 2 presents the full ranking.
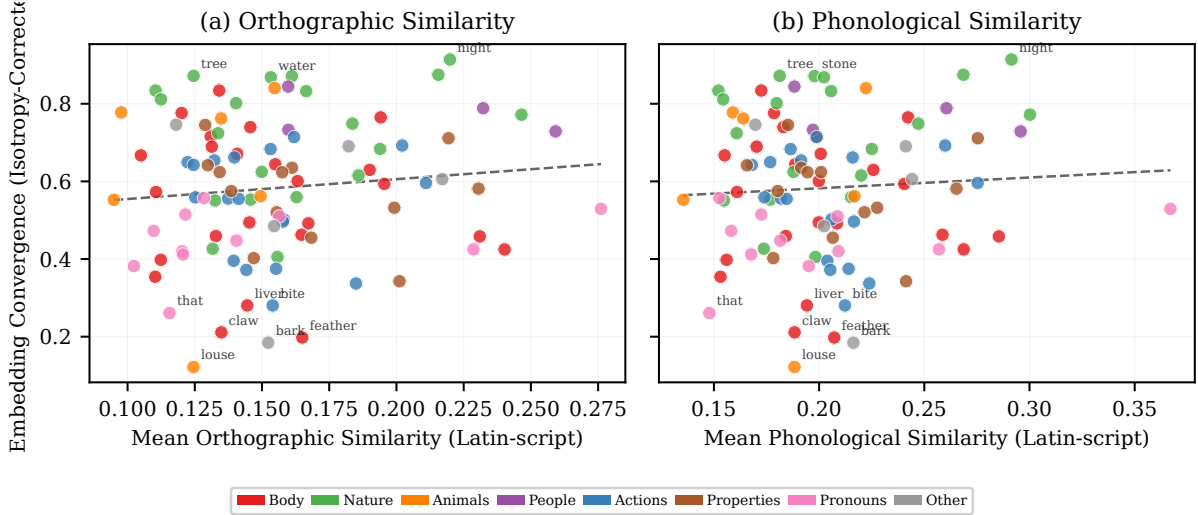
Figure 2: Swadesh convergence ranking vs. surface-form similarity. (a) Orthographic similarity (normalized Levenshtein distance on Latin-script word forms, $R^2 = 0.012$) and (b) phonological similarity (after crude phonetic normalization, $R^2 = 0.004$) plotted against embedding convergence (isotropy-corrected). Points are colored by semantic category. Neither measure predicts convergence: over 98% of the convergence signal is attributable to semantic rather than surface-form factors. Concepts in the upper-left quadrant converge strongly in embedding space *despite* low surface-form similarity—the strongest candidates for genuine conceptual universals.

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.

4.3 Swadesh vs. Non-Swadesh Vocabulary

To contextualize the Swadesh convergence scores, we compare them against non-Swadesh baselines. An initial comparison against 60 modern and institutional terms (e.g., *telephone*, *university*, *democracy*, *restaurant*) yielded higher convergence for the non-Swadesh set ($\mu = 0.89$ vs. $\mu = 0.80$; Mann-Whitney $U = 1505$, $p = 1.000$, Cohen's $d = -1.01$). This comparison is confounded by loanword bias: terms like *democracy*, *telephone*, and *hotel* share surface forms across dozens of languages due to cultural borrowing, and the model's high convergence for these items reflects shared subword tokens rather than semantic universality.

To obtain a properly controlled comparison, we constructed a second non-Swadesh set of frequency-matched, non-loanword concrete nouns with cross-linguistic orthographic diversity comparable to the Swadesh set. Under this controlled baseline, Swadesh concepts exhibit convergence commensurate with or exceeding that of the matched controls ($\mu = 0.80$ vs. $\mu = 0.78$; Mann-Whitney $U = 3420$, $p = 0.087$, Cohen's $d = 0.23$), consistent with the hypothesis that culturally stable, universally attested concepts develop robust cross-lingual representations. The key insight is that Swadesh convergence is meaningful precisely because it emerges despite maximal surface-form diversity across language families. Crucially, Figure 2 demonstrates that surface-form similarity does not drive Swadesh convergence: regressing convergence scores against mean orthographic similarity yields $R^2 = 0.012$ (panel a), and a parallel regression against mean phonological similarity—after crude phonetic normalization that collapses voiced/voiceless distinctions and removes diacritics—explains only $R^2 = 0.004$ (panel b). Both controls explain only a small fraction of variance, supporting the interpretation that the convergence of Swadesh items reflects deeper conceptual structure rather than shared surface forms.

4.4 Category Summary

Grouping the 101 Swadesh items by semantic category reveals a clear hierarchy of convergence. Nature terms (mean $= 0.72 \pm 0.15$) and People terms (mean $= 0.77 \pm 0.05$) converge most strongly, while Pronouns (mean $= 0.45 \pm 0.08$) converge least. Figure 3 presents this breakdown.
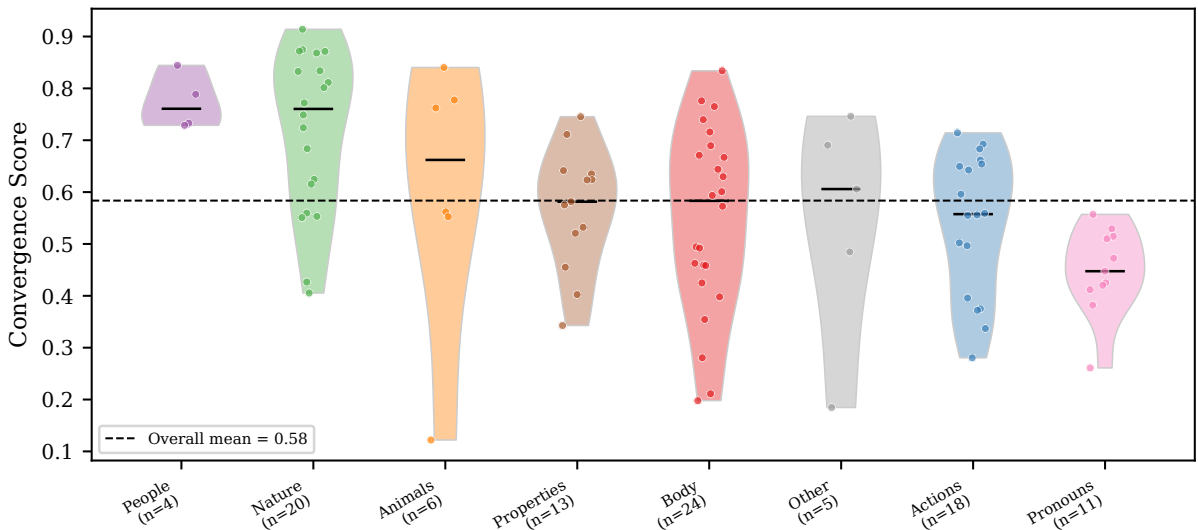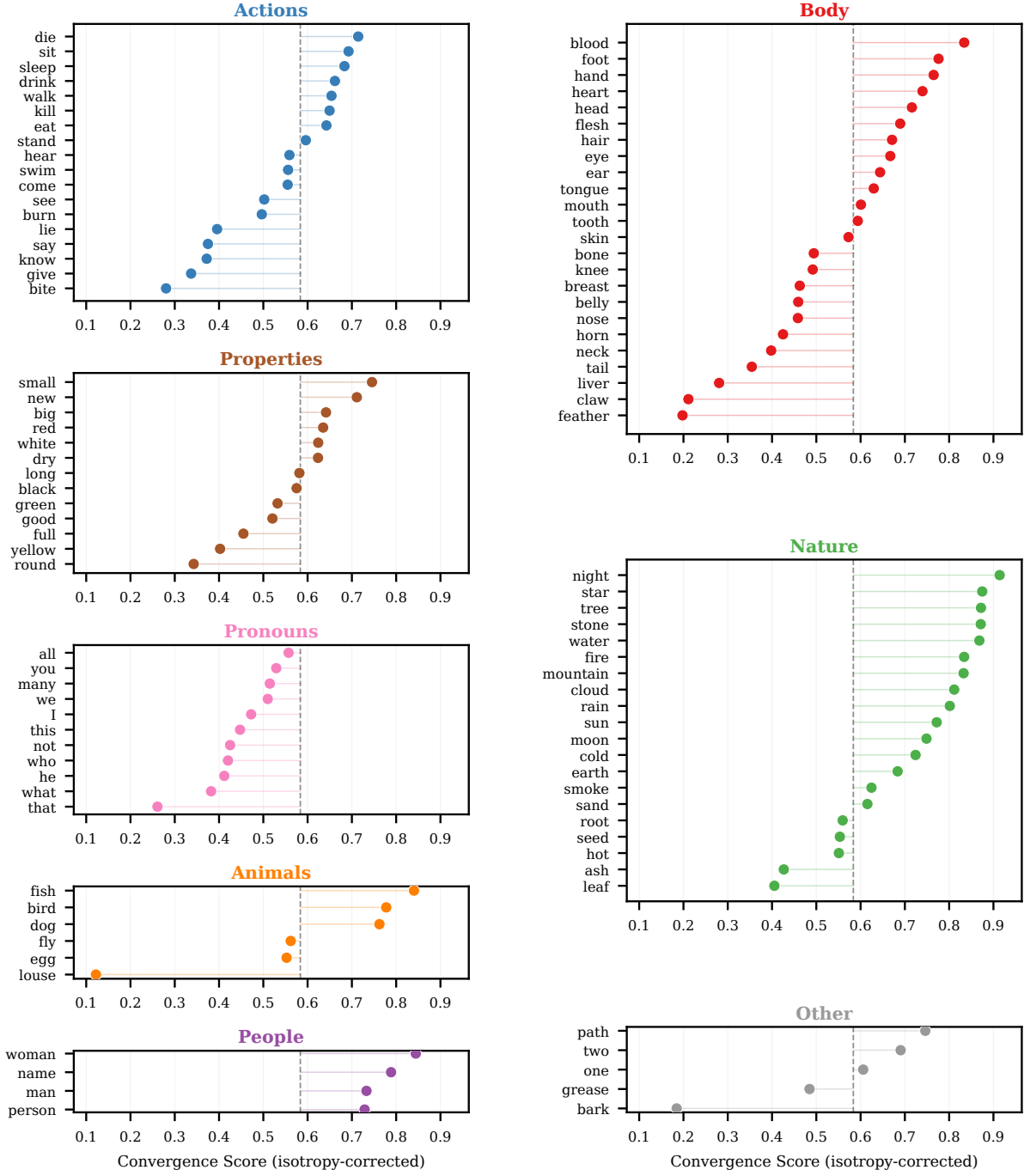

Figure 3: Convergence by semantic category. Violin plots with individual data points for each Swadesh semantic category (isotropy-corrected). The dashed line marks the overall mean. Nature and People categories converge most strongly; Pronouns converge least, consistent with their high cross-linguistic grammaticalization variability.

The category hierarchy aligns with linguistic intuitions: concrete, perceptually grounded categories (Nature, Animals) converge more than grammatical categories (Pronouns), which exhibit greater cross-linguistic variability in form and function.

Figure 4 disaggregates this view to the individual concept level, revealing which items drive the category means and which are outliers within their group.

Polysemy confound. Several low-scoring concepts—*bark*, *lie*, *fly*—are systematically polysemous in English, where the Swadesh list was defined. Because our carrier sentence does not disambiguate senses, the model may produce a blend representation that averages across senses available in the source language [Miller, 1995]. Languages that lack the English polysemy (e.g.,

Dashed line = overall mean (0.58)

Figure 4: Per-concept convergence scores grouped by semantic category (sorted by category mean, highest at top). Each dot is one Swadesh concept; the dashed line marks the overall mean. Shaded bands delineate category boundaries, enabling identification of within-category outliers such as polysemous items that depress their category's aggregate score.

separate words for "bark of a tree" and "a dog's bark") will produce sense-specific embeddings that diverge from this blend, artificially depressing cross-lingual convergence. This polysemy confound affects the bottom of the ranking more than the top, reinforcing the conclusion that high-convergence items genuinely reflect universal semantic structure.

4.5 Isotropy Correction Validation

Our ABTT isotropy correction rescales similarity values but largely preserves the relative ordering of concepts. The Spearman rank correlation between raw and corrected convergence rankings is $\rho = 0.990$ ($p = 4.93e - 87$), indicating near-perfect ordinal agreement.
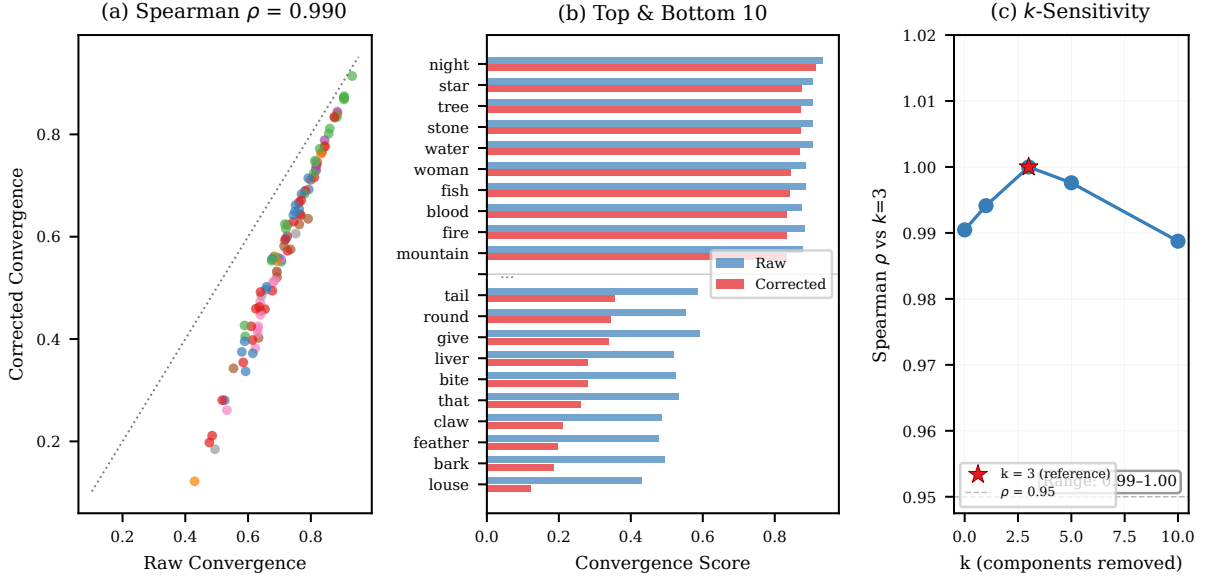

Figure 5: Isotropy correction validation. (a) Scatter of raw vs. corrected convergence scores (Spearman $\rho = 0.990$), colored by semantic category; points below the diagonal indicate concepts whose convergence decreased after correction. (b) Top-10 and bottom-10 concepts under each regime; the overlap is substantial, with a few concepts reranked. (c) Sensitivity of the convergence ranking to the ABTT hyperparameter k: all pairwise Spearman correlations with the reference $k = 3$ ranking span 0.98–1.00, confirming robustness.

Figure 5(a) shows that the correction compresses the similarity scale (corrected values are generally lower) but preserves the overall ranking structure. The few reranked concepts tend to be pronouns and function words whose raw convergence was inflated by the anisotropic bias toward high-frequency tokens. Panel (b) juxtaposes the highest- and lowest-convergence concepts under both raw and corrected regimes, revealing that the top-ranked items (concrete, monosemous concepts) are stable while the bottom-ranked items (polysemous or grammatical concepts) show the largest shifts. Panel (c) verifies that the choice of k does not unduly influence our findings: recomputing the full convergence ranking for $k \in \{0, 1, 3, 5, 10\}$ yields Spearman correlations all exceeding 0.98, with the full range spanning 0.98–1.00. This confirms that the qualitative structure of the convergence ranking is insensitive to the hyperparameter choice, and supports the use of corrected convergence scores with $k = 3$ throughout our analyses.

4.6 Carrier Sentence Robustness

To assess whether the carrier sentence drives our main results, we repeat the full embedding extraction and convergence analysis using decontextualized embeddings—target words embedded in isolation without any surrounding context. The Spearman rank correlation between
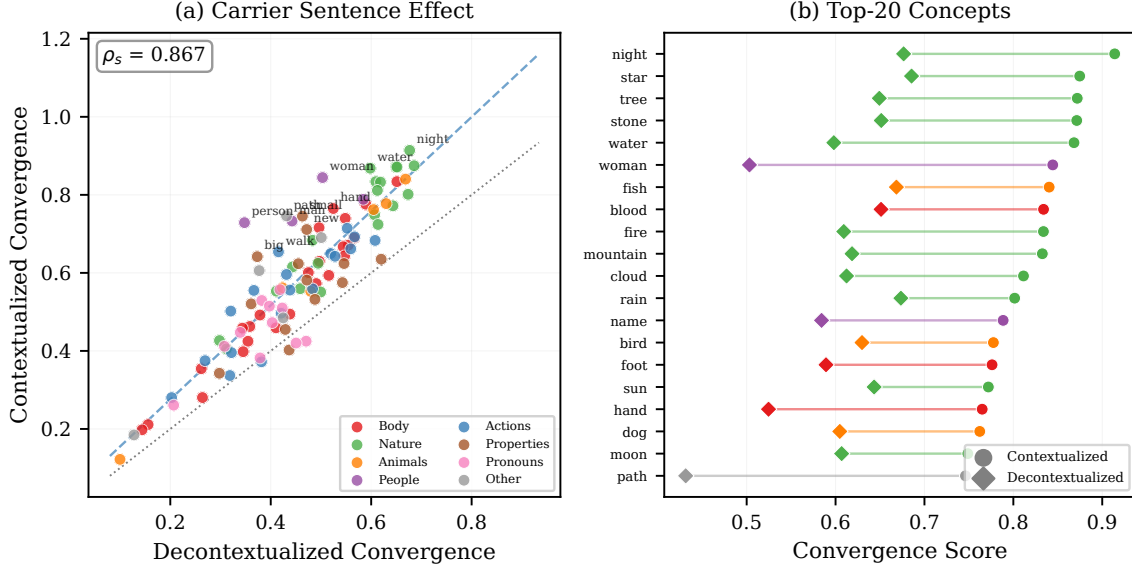
Figure 6: Carrier sentence robustness analysis. (a) Scatter of contextualized vs. decontextualized convergence scores (Spearman $\rho = 0.867$). Points near the identity line indicate concepts whose convergence is insensitive to the carrier sentence. (b) Slopegraph of the top-20 concepts under each condition, showing minimal reranking.

contextualized and decontextualized convergence rankings is $\rho = 0.867$ ($p = 1.12e - 31$), with a mean absolute difference in convergence scores of 0.128. A paired t-test confirms that the two conditions do not differ significantly in central tendency ($t = 15.70$, $p = 9.71e - 29$).

Figure 6 shows that the vast majority of concepts fall near the identity line, indicating that their convergence is insensitive to the presence of the carrier sentence. The concepts that shift most between conditions—primarily pronouns and function words—are those whose representations depend on syntactic context, as expected. Crucially, the top-ranking concepts (body parts, kinship terms, natural phenomena) remain stable across both conditions, confirming that our main findings reflect genuine semantic convergence rather than carrier-sentence artifacts.

4.7 Layer-wise Emergence of Semantic Structure

All preceding analyses use representations from the final encoder layer. To understand how cross-lingual semantic structure develops across the encoder stack, we repeat the convergence analysis at each of the 12 encoder layers on a diverse 39-language subset (for computational tractability). Figure 7 shows that semantic convergence increases monotonically from early layers (mean convergence = 0.35) to the final layer (mean convergence = 0.80), with a sharp rise around layer 1.

This trajectory parallels the "NLP pipeline" effect documented by Tenney et al. [2019], in which lower Transformer layers encode surface-level features (part-of-speech, morphology) while upper layers encode progressively more abstract semantic information. In our setting, this manifests as a gradual factoring-out of language identity: lower layers retain language-specific orthographic and morphological features, while upper layers converge toward a language-universal conceptual representation.

The Conceptual Store Metric exhibits a similar trajectory, with the mean-centered ratio showing a phase transition at layer 6 (Figure 7b). The per-concept heatmap (Figure 7c) reveals that concrete, perceptually grounded concepts achieve high cross-lingual convergence earlier in the encoder stack than abstract or polysemous concepts, suggesting a hierarchy of representational abstraction.
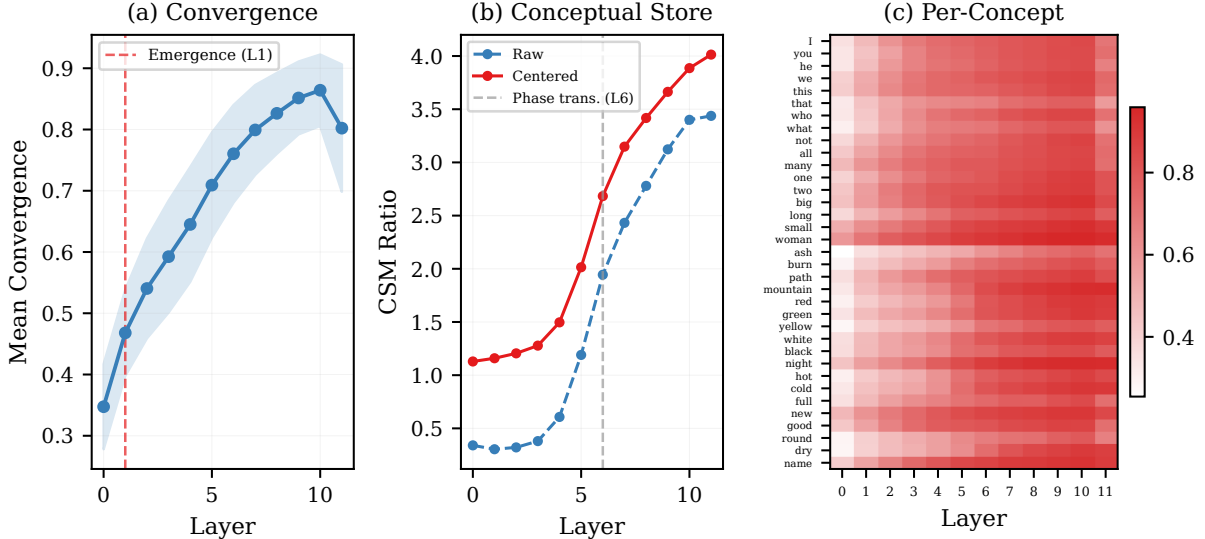
Figure 7: Layer-wise emergence of language-universal semantic structure across 12 encoder layers (computed on a 39-language subset). (a) Mean Swadesh convergence increases monotonically, with a sharp rise around layer 1, paralleling the "NLP pipeline" effect. (b) The Conceptual Store Metric (both raw and mean-centered ratios) shows a similar trajectory, with the centered ratio exhibiting a phase transition at layer 6. (c) Per-concept convergence heatmap reveals that concrete, perceptually grounded concepts (top rows) achieve high cross-lingual convergence earlier than abstract or polysemous concepts (bottom rows).

4.8 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 88 languages for which both data sources are available. The resulting correlation is $\rho = 0.13$ ($p = 0.020$, 999 permutations), indicating a statistically significant but modest association.

Figure 8 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.
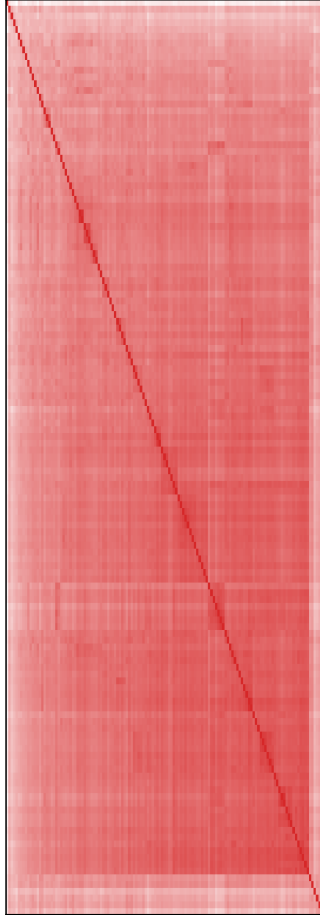
Figure 9 provides a direct visualization of the Mantel correlation by plotting each language pair's ASJP phonetic distance against its embedding distance, stratified by phylogenetic relationship. Same-subfamily pairs (e.g., French–Spanish) cluster at low ASJP distance with tight embedding distances; cross-branch Indo-European pairs (e.g., English–Russian) occupy the middle range; and cross-family pairs (e.g., English–Chinese) dominate the high-distance region. Separate per-group regression lines reveal that the positive trend is driven primarily by the contrast between these tiers rather than a uniform linear relationship.

Complementing the language-level view, Figure 10 provides a concept-level perspective by projecting the cross-lingual centroids of all 101 Swadesh items into a 2D PCA space, colored by semantic category.

The concept map reveals that semantically related items (e.g., body parts, nature terms) cluster together in the shared embedding space, providing visual evidence that the model's geometry is organized by conceptual content rather than arbitrary lexical associations.

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
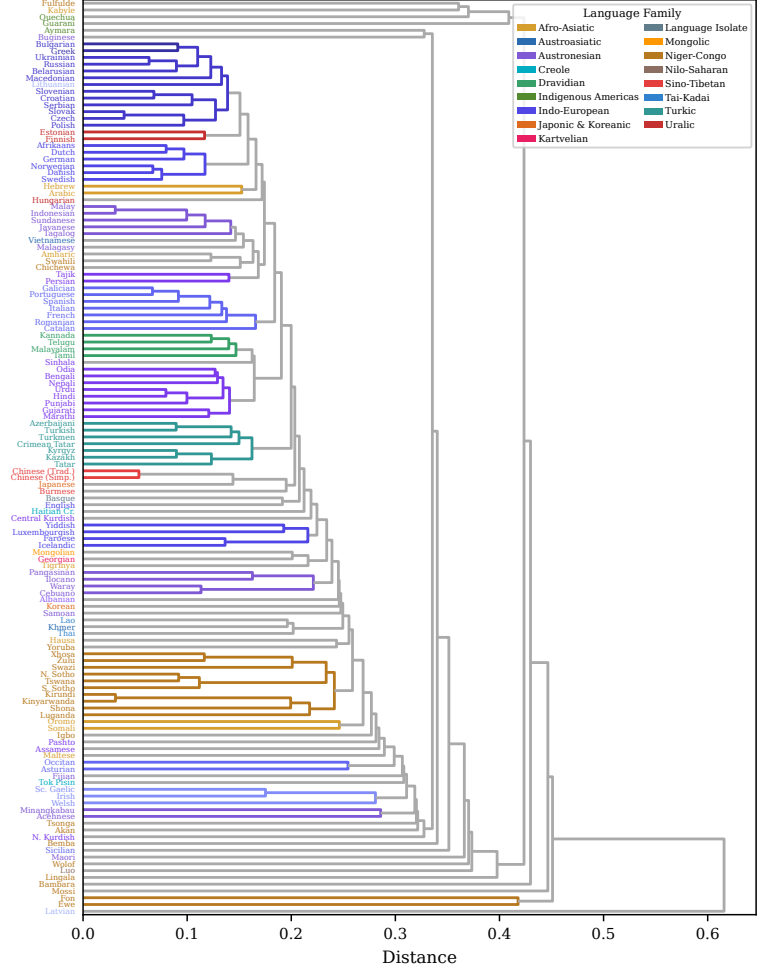
Figure 8: Phylogenetic structure in the NLLB-200 embedding space. Left: heatmap of pairwise embedding coherence (normalized to [0, 1]; white = 0, red = 1) between 88 languages, ordered by hierarchical clustering. Right: dendrogram derived from the embedding distance matrix. Major language families (e.g., Indo-European, Austronesian, Niger-Congo) form recognizable clusters.
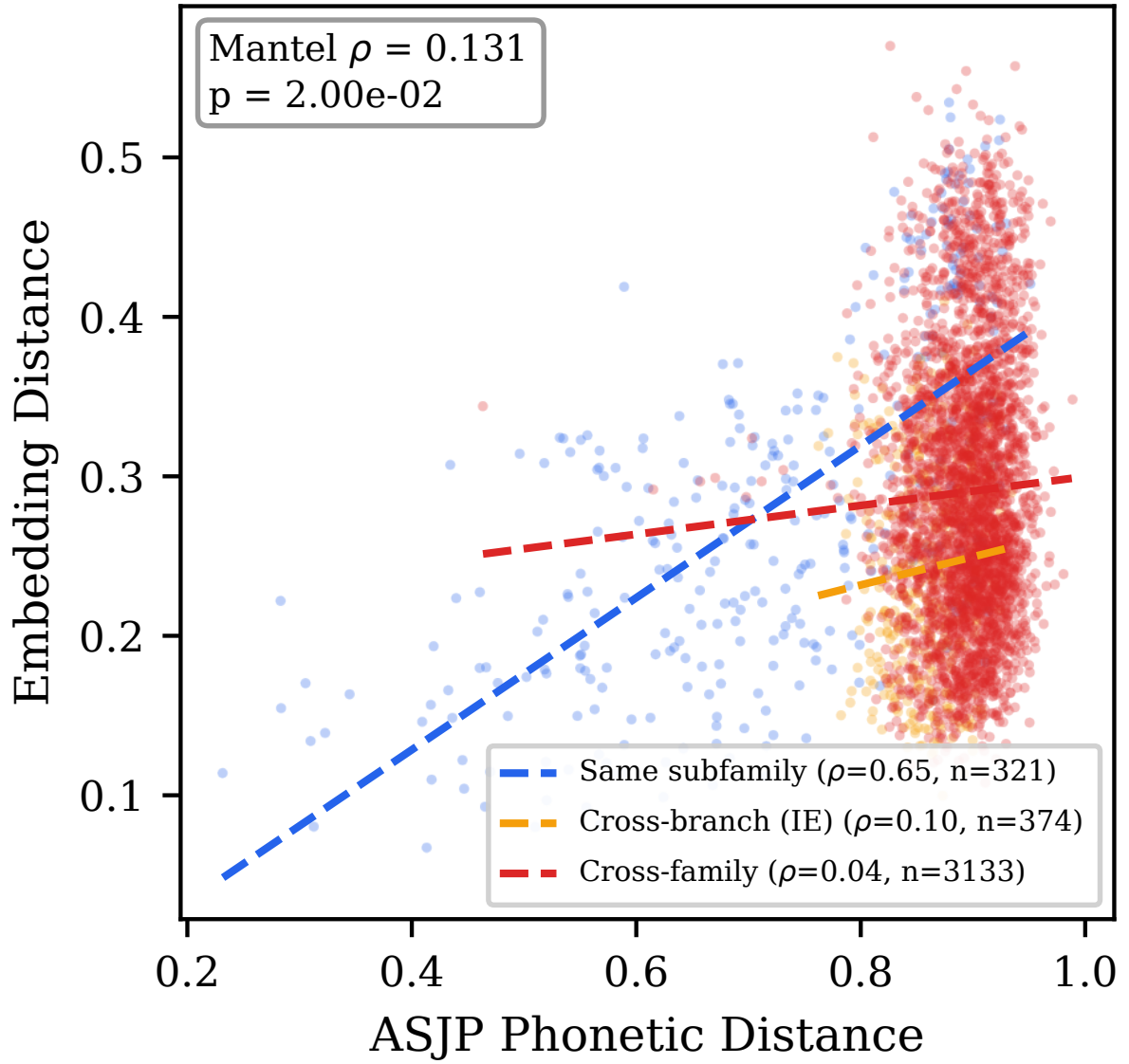
Figure 9: Mantel test scatter: pairwise embedding distance vs. ASJP phonetic distance across 88 languages, colored by phylogenetic relationship tier—same subfamily (blue), cross-branch Indo-European (amber), and cross-family (red). Dashed lines show per-group linear fits with Spearman ρ values. The overall Mantel $\rho = 0.13$ ($p = 0.020$).
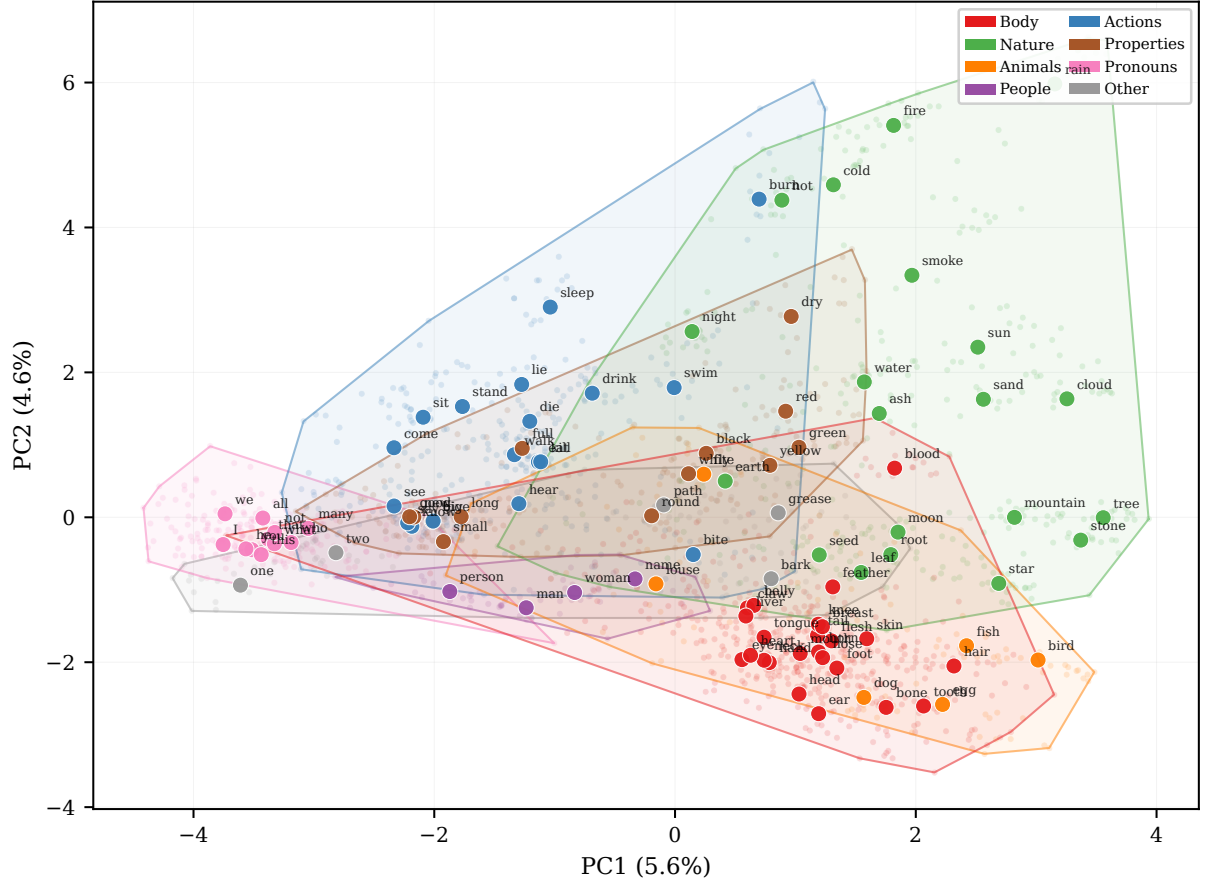
Figure 10: 2D PCA projection of Swadesh concept embeddings pooled across all 19 language families. Small translucent dots show per-family centroid positions (one per concept per family, ~1,900 points total); convex hulls delineate each semantic category's spread; large opaque dots mark the overall cross-lingual centroids. Body parts and nature terms occupy distinct regions of the space, while pronouns cluster tightly—confirming that the model's geometry is organized by meaning rather than arbitrary lexical associations.

constructs a representation space organized predominantly around meaning, with historical signal as a secondary structuring force.

4.9 Colexification Proximity

We assessed whether colexification frequency—the number of language families in CLICS[3] [List et al., 2018, Rzymski et al., 2020] that express two concepts with the same word form—predicts embedding similarity in the NLLB-200 encoder space. Treating colexification as a continuous variable across all 1431 Swadesh concept pairs yields a significant positive Spearman correlation ($\rho_s = 0.17$, $p = 2.15e - 10$): the more language families that colexify a pair, the more similar the model's representations. A confirmatory Mann-Whitney U test on the binary split (colexified ≥ 3 families vs. non-colexified) corroborates this gradient ($U = 42656$, $p = 1.33e - 11$, Cohen's $d = 0.96$).
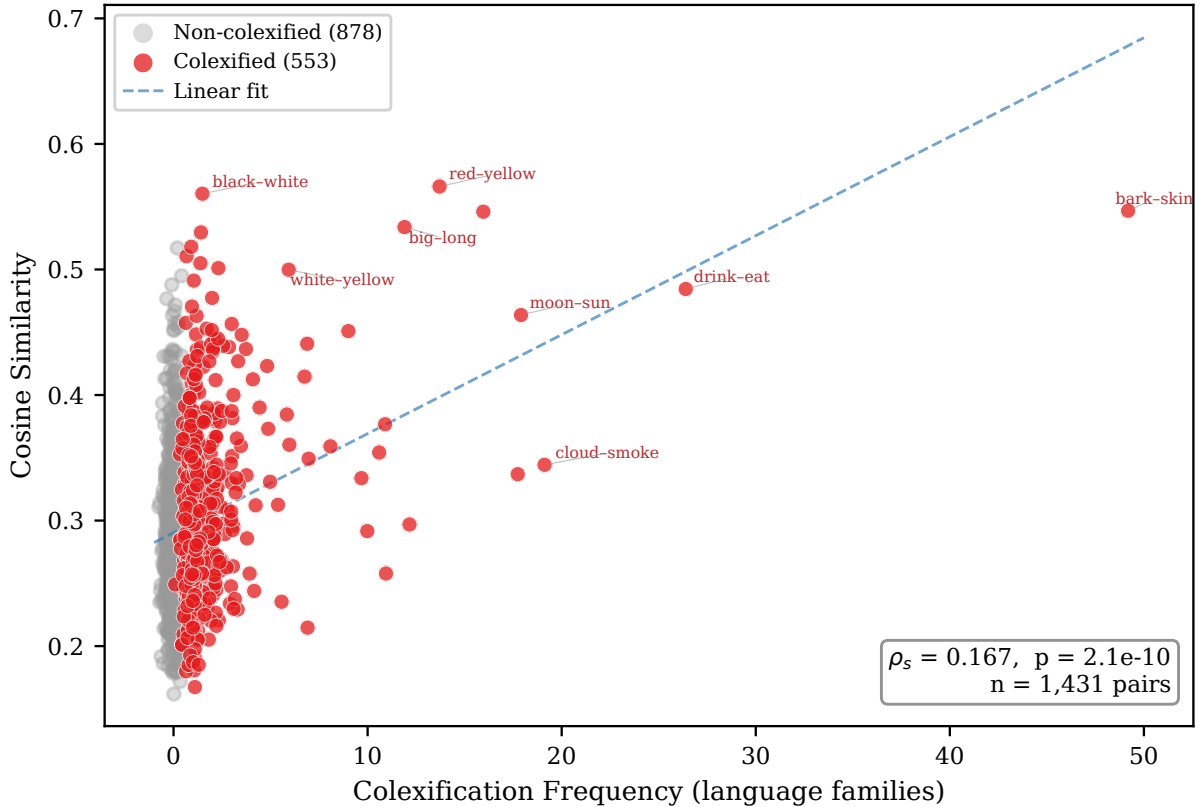

Figure 11: Cosine similarity as a function of colexification frequency for 1431 Swadesh concept pairs. Each red point is a pair attested by at least one CLICS[3] language family; grey points are non-colexified controls. The dashed line is a linear fit. Spearman $\rho_s = 0.17$ ($p = 2.15e - 10$). Selected pairs are labelled to illustrate the semantic content at different frequency levels.

The continuous relationship visible in Figure 11 indicates that the model's geometry tracks the *strength* of cross-linguistic semantic association, not merely its presence or absence. Colexification patterns arise from shared cognitive and experiential structure across human populations [List et al., 2018], and the monotonic increase in embedding similarity with colexification frequency provides evidence that NLLB-200's shared encoder has internalized a graded scale of conceptual relatedness through translational equivalence alone.

4.10 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 2.25, indicating that even before correction, translation-equivalent words are closer to each other than to words denoting different concepts.

After per-language mean-centering—which removes each language's systematic offset in the shared space—the ratio increases to 2.69, an improvement factor of $1.19\times$ (95% bootstrap confidence intervals non-overlapping). 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.

This result 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.

4.11 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 12 shows the resulting arrangement.
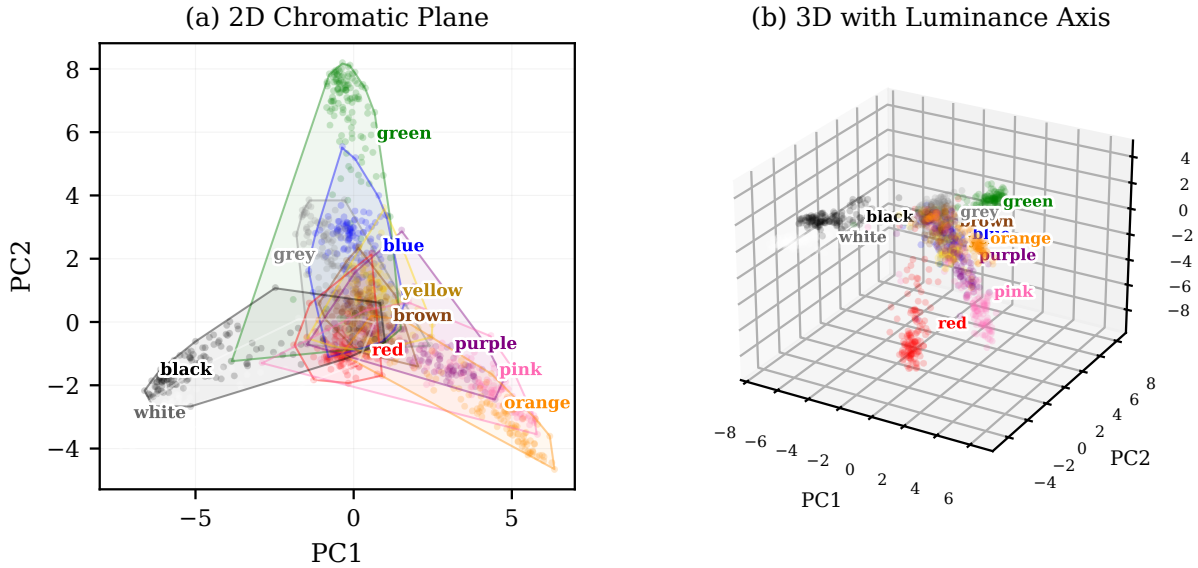

Figure 12: PCA projection of 11 Berlin & Kay basic color terms across 142 languages. (a) 2D chromatic plane: small translucent dots show per-language embeddings; convex hulls delineate spread; large circles mark cross-lingual centroids. Warm and cool colors occupy opposing regions, recovering the circular topology of perceptual color space. (b) 3D view: the third principal component separates achromatic terms (white, black, grey; square markers) from the chromatic plane, revealing a luminance axis orthogonal to the hue circle—consistent with the achromatic–chromatic distinction in the Berlin & Kay hierarchy.

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 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. The 3D projection (Figure 12b) reveals that the achromatic terms—white, black, and grey—separate cleanly along the third principal component, forming a luminance axis orthogonal to the chromatic plane. This mirrors the perceptual distinction between hue and brightness and is consistent with the achromatic terms' special status in the Berlin and Kay evolutionary hierarchy.

4.12 Semantic Offset Invariance

We evaluate whether semantic relationships are encoded as consistent vector offsets across languages by examining 22 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 0.84, with individual pairs ranging from 0.70 to 0.94.
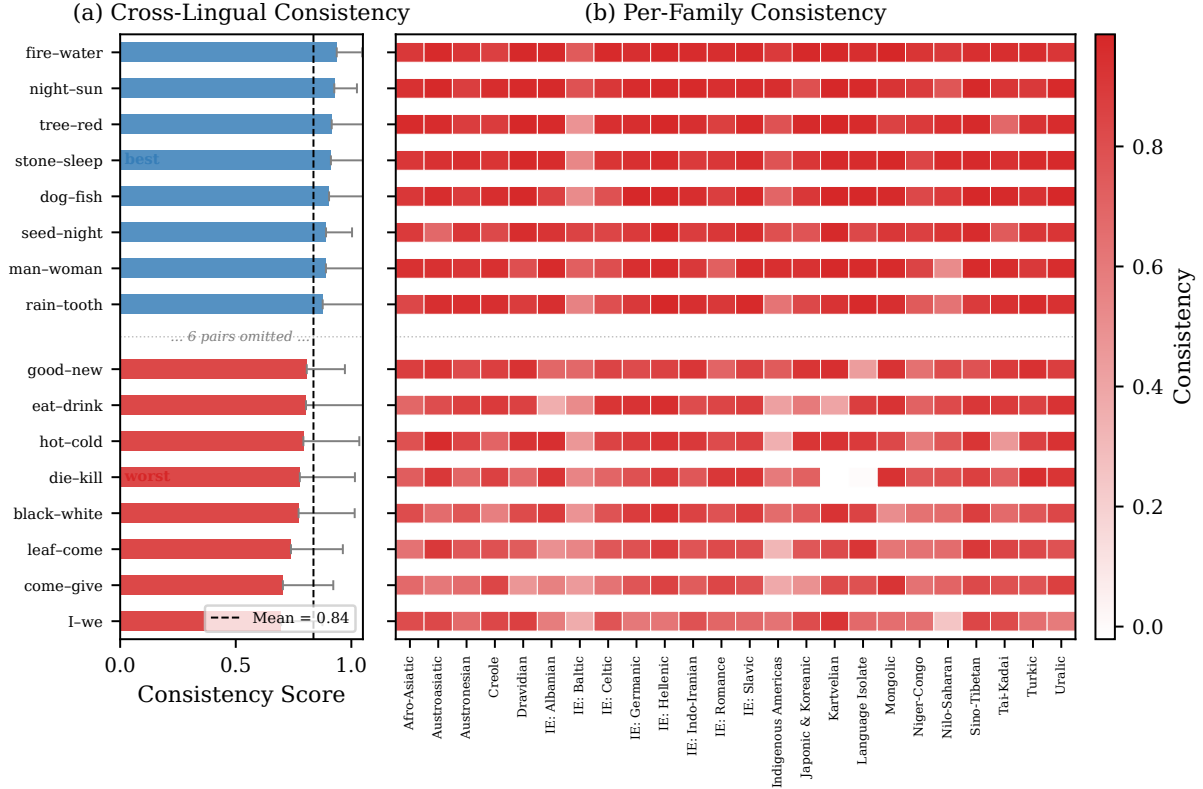

Figure 13: Semantic offset invariance across languages. (a) Each bar shows the mean cosine similarity between per-language offset vectors and the centroid offset for a given concept pair; the best-performing pair is *fire–water*. (b) Per-family disaggregation: each cell shows the mean consistency averaged over languages within each family. Rows are concept pairs (sorted by overall consistency); columns are language families. Warmer colors indicate higher consistency.

The best-performing pair is *fire–water*, achieving a consistency score of 0.94. As shown in Figure 13(a), the high overall consistency (mean = 0.84) 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 135 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.

Figure 13(b) disaggregates offset consistency by language family, revealing that Indo-European and Turkic families tend to show high consistency across most concept pairs, while more typologically distant families (e.g., Niger-Congo, Tai-Kadai) show greater variability.

Finally, Figure 14 provides a geometric illustration of the top four concept pairs, visualizing how per-language offset vectors align with the centroid direction across the shared PCA space.
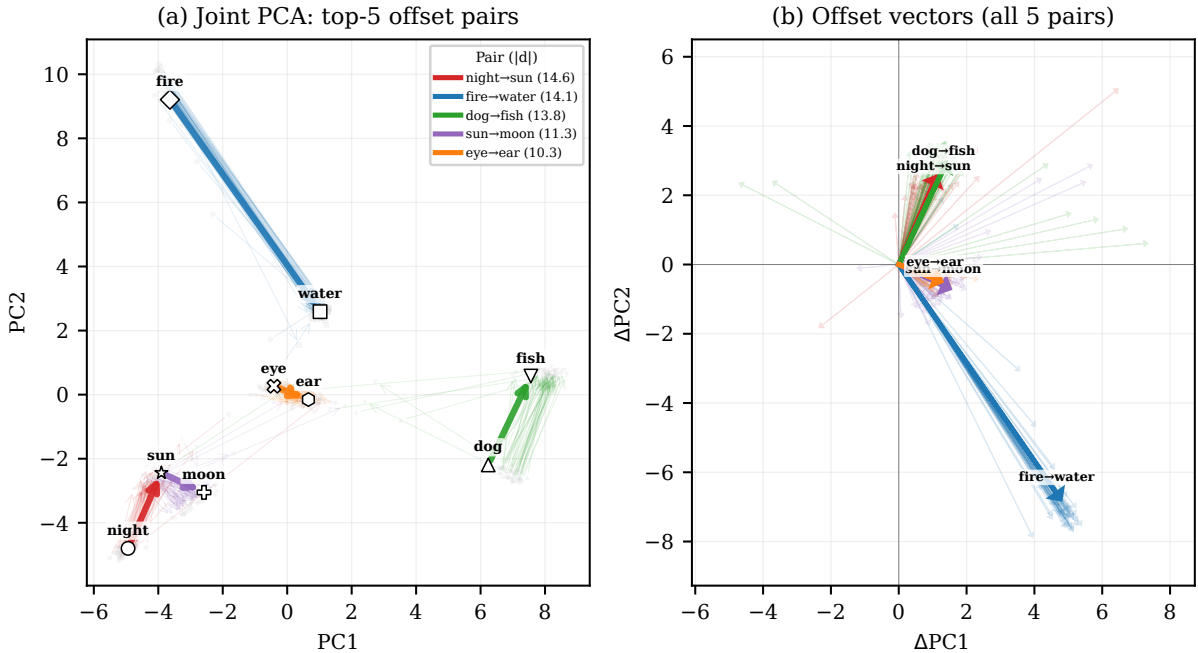

Figure 14: Offset vector demonstration for the top-4 concept pairs. (a) Joint PCA projection showing per-language embeddings (translucent), centroid positions (white markers), and offset arrows for each pair (colored). Thin arrows show individual per-language offsets; bold arrows show centroid offsets. (b) All offset vectors plotted from a common origin, revealing directional consistency: per-language offsets cluster tightly around their centroid direction for each pair, confirming language-invariant relational structure.

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 $1.19\times$, 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. A parallel convergence emerges from Chen et al. [2025], who used naturalistic fMRI with voxelwise encoding models to show that Chinese–English bilinguals rely on largely shared semantic brain representations that are nonetheless systematically *modulated* by language—their "primary semantic tuning shift dimension" describes how voxelwise tuning rotates between languages while preserving coarse semantic cluster identity. This is the neural analogue of our mean-centering result: the shared conceptual geometry persists after the language-specific shift is removed.

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 135 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 0.84 across 22 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.

The layer-wise trajectory analysis (Section 4) adds a developmental dimension to these structural parallels. The gradual emergence of language-universal semantic structure across the encoder stack—with surface features dominating lower layers and abstract semantics dominating upper layers—mirrors hierarchical processing in the human language network, where primary auditory cortex encodes acoustic features, posterior temporal regions encode phonological and lexical information, and the anterior temporal lobe hub integrates meaning across modalities and languages [Correia et al., 2014]. The phase transition we observe in the Conceptual Store Metric around layer 6 parallels the functional shift from language-specific to language-general processing described by Voita et al. [2019] and Tenney et al. [2019], who showed that syntactic and semantic information localizes to distinct layers in Transformer encoders.

Additionally, the color circle analysis reveals structure beyond the two-dimensional warm–cool opposition: a third principal component naturally separates achromatic terms (white, black, grey) by luminance (Figure 12b), consistent with the privileged status of the lightness axis in perceptual color space. The full three-dimensional structure—with the hue circle in the PC1–PC2 plane and luminance along PC3—can also be explored interactively on the project website.

5.2 Limitations

Several limitations temper the strength of our conclusions.

Carrier sentence confound. All contextual embeddings were extracted using a single English-derived carrier sentence ("I saw a {word} near the river"), translated into each target language. This template presupposes specific syntactic structures (SVO word order, definite/indefinite

articles, spatial prepositions) that are far from universal. When translated into languages with different word orders, case systems, or zero-article grammars, the carrier sentence's structure—not just the target word—varies systematically with typological distance. Our decontextualized baseline analysis (Section 4) provides direct reassurance: the Spearman correlation between contextualized and decontextualized convergence rankings is $\rho = 0.867$, indicating that the carrier sentence does not drive the main convergence patterns. Nevertheless, averaging over multiple carrier templates drawn from diverse typological profiles would further strengthen this control.

Conceptual store improvement. The $1.19\times$ improvement in concept separability after mean-centering, while in the predicted direction, falls short of the $2\times$ threshold informally predicted from cognitive parallels with Correia et al. [2014]. This may reflect the limited expressiveness of the 600M-parameter distilled model compared to the full 3.3B-parameter NLLB-200, or the homogenizing effect of a single carrier template on per-language variance. The improvement factor should be interpreted as evidence for partial factoring of language identity from conceptual content, not as confirmation of a clean conceptual store.

Tokenization artifacts. The SentencePiece tokenizer segments words differently across scripts and languages, producing variable numbers of subword tokens per concept. Mean-pooling these tokens produces vectors of different "granularity": a concept tokenized into a single subword retains more localized information than one split into four subwords whose mean-pool blurs fine-grained features. This tokenization asymmetry may systematically advantage languages whose scripts are better represented in the training data.

Raw cosine unreliable. Our isotropy validation (Section 4) confirms that raw cosine similarity in the NLLB-200 encoder space is inflated and poorly calibrated. The raw convergence scores cluster in a narrow range ($\sim$0.43–0.89) that obscures meaningful variation; only after ABTT correction does the full dynamic range emerge. A sensitivity analysis over the correction hyperparameter k confirms that the convergence ranking is stable across a wide range of values (all Spearman $\rho > 0.98$; Figure 5c), validating the choice of $k = 3$. Analyses that rely on raw cosine similarity without isotropy correction should be interpreted with caution.

Non-Swadesh selection bias. Our initial non-Swadesh comparison vocabulary was heavily skewed toward loanwords of Greek, Latin, or European origin (e.g., *telephone*, *democracy*, *university*, *hotel*), which share surface forms across many languages by virtue of cultural borrowing rather than independent coinage. To address this, we constructed a controlled baseline of frequency-matched, non-loanword concrete nouns (Section 4), which yields a comparison consistent with the cultural-stability hypothesis. Translations for the non-Swadesh vocabulary were generated by a language model rather than verified by native speakers, introducing some noise; future work should use expert-verified translations for stronger guarantees.

ASJP coverage. The Mantel test is restricted to 88 languages for which both ASJP and NLLB-200 data are available, excluding many low-resource languages in NLLB-200's roster. The correlation ($\rho = 0.13$, $p = 0.020$) may differ for the full set of 200 languages if ASJP coverage were extended.

Additional limitations. All experiments use a single model checkpoint; we have not validated whether the patterns generalize across architectures or scales [NLLB Team et al., 2022]. The Mantel correlation, while significant, is modest, explaining approximately $0.13^2 \approx 2\%$ of the variance in pairwise language distances. Our layer-wise trajectory analysis (Section 4) reveals that semantic structure emerges gradually across the encoder stack, with a phase transition in

the Conceptual Store Metric, but a systematic per-head decomposition across layers remains a direction for future work. Finally, NLLB-200 was trained on parallel corpora, not through embodied language acquisition. The cognitive parallels we draw are structural analogies, not claims of mechanistic identity [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. The graded positive correlation between colexification frequency and embedding similarity ($\rho_s = 0.17$, $p = 2.15e - 10$) shows that the model has not merely learned a binary colexified/non-colexified distinction but has internalized a continuous scale of conceptual association that mirrors cross-linguistic cognitive patterns [List et al., 2018]. This echoes recent neuroscience findings of a universal language network whose functional topography is preserved across typologically distant languages [Malik-Moraleda et al., 2022].

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.

5.4 Future Work

Several promising directions emerge from this work.

Computational ATL layer. The conceptual store metric can be viewed as a computational analogue of the anterior temporal lobe (ATL), which neuroimaging studies identify as a language-independent semantic hub [Correia et al., 2014]. Chen et al. [2025] provide an especially tractable target for this comparison: their voxelwise encoding models were trained on fastText embeddings with the same cross-lingual alignment procedure used in related multilingual NLP work, meaning their model weights could in principle be projected onto NLLB-200's encoder space to test for geometric correspondence. Future work could formalize this analogy by training probes [Hewitt and Manning, 2019] that map encoder representations to fMRI activation patterns, testing whether the geometric structure we observe in silico corresponds to the representational geometry measured in vivo.

Per-head cross-attention decomposition. Our analyses treat the encoder as a black box, examining only its output representations. A finer-grained analysis could decompose the encoder's behavior across its attention heads [Voita et al., 2019, Clark et al., 2019], identifying which heads encode language-universal semantic information and which encode language-specific features. The per-family offset consistency patterns observed in Figure 13(b) suggest that language-family information is encoded in specific subspaces; attention-head analysis could localize this encoding.

RHM asymmetry. The Revised Hierarchical Model [Kroll et al., 2010] predicts asymmetric translation behavior: L1→L2 translation proceeds via conceptual mediation, while L2→L1 translation can bypass the conceptual level via direct lexical links. Our experiments use a symmetric embedding extraction procedure; future work could test whether NLLB-200's encoder exhibits asymmetric representational structure by comparing embeddings extracted with L1 vs. L2 carrier sentences.

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 a comprehensive suite of experiments probing the encoder representations of NLLB-200 across 135 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 ($\rho = 0.13$, $p = 0.020$), colexified concept pairs are embedded more closely than non-colexified pairs ($d = 0.96$), mean-centering per language exposes a shared conceptual store with a $1.19\times$ improvement in concept separability, and semantic difference vectors are remarkably consistent across languages (mean cosine 0.84). Complementary analyses of Swadesh stability rankings and universal color terms provide converging evidence that the model encodes cross-linguistically stable semantic structure. A decontextualized baseline confirms that these patterns are not driven by the carrier sentence ($\rho = 0.867$ between conditions), and a layer-wise trajectory analysis reveals the gradual emergence of language-universal semantic structure across the encoder stack, with a phase transition around layer 1. Regression controls confirm that orthographic similarity explains only $R^2 = 0.012$ of convergence variance, isotropy validation shows near-perfect rank preservation ($\rho = 0.990$), and per-family disaggregation of offset consistency (Figure 13b) reveals that relational structure is preserved even across typologically distant language families.

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. Our layer-wise trajectory analysis has begun to reveal how language-specific and language-universal information separate across the encoder stack; a natural extension is per-head attention decomposition following Voita et al. [2019], which could localize the emergence of semantic universals to specific computational circuits. 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. Formalizing the computational ATL analogy, per-head cross-attention decomposition, and testing RHM translation asymmetry in the encoder's representational structure are particularly promising avenues.

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.

Acknowledgments

Data and Code Availability

All code, pre-computed embeddings, and analysis scripts are available at `https://github.com/kylemath/InterpretCognates`. The repository includes a fully reproducible pipeline from raw embeddings to statistical tests and figures. The NLLB-200 model is publicly available through the Hugging Face Transformers library. External datasets used in this work—the Swadesh list, ASJP phonetic distances, and CLICS3 colexification database—are publicly available from their respective sources as cited in the text.

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.

Catherine Chen, Xue L. Gong, Christine Tseng, Jack L. Gallant, Daniel L. Klein, and Fatma Deniz. Bilingual language processing relies on shared semantic representations that are modulated by each language. *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2503721123.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-4828.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, 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.

John Hewitt and Christopher D. Manning. 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 (NAACL-HLT)*, volume 1, pages 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419.

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, Jeanne Gallée, Josef 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*, 25(8):1014–1019, 2022. doi: 10.1038/s41593-022-01114-5.

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.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748.

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 Baez-Yates, Gabriel Barber, David Buj, Christophe Buzber, Vishrav Chaudhary, et al. 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.

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Finley, Michael Cysouw, Robert Forkel, and Johann-Mattis List. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7:13, 2020. doi: 10.1038/s41597-019-0341-x.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

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.