Linguistic Relativity as Representational Geometry

A Mathematical Framework for Language-Conditioned Cognition

Greg Olsen
grolsen@gmail.com

Claude Opus 4.5
Anthropic

January 2026

> *"To think is to forget a difference, to generalize, to abstract."*
> — Jorge Luis Borges, "Funes the Memorious"

1. The Core Thesis

Language is not merely a labeling layer laid over a fixed conceptual substrate. It functions as a *computational code*—a culturally evolved compression scheme—that makes some distinctions cheap to represent, some analogies easy to compute, and some inferences natural to execute.

We formalize this by treating language as inducing a **representational geometry**: a structured space in which distances, directions, and curvature summarize the computational cost of moving between meanings under the constraints of expressing, rehearsing, and communicating thought in a particular language.

This is not a claim about all cognition. Human minds are grounded, multi-modal, and action-oriented. But a wide class of higher-order operations—explicit categorization, narrative explanation, rule articulation, memory labeling, social justification, deliberation via inner speech—are routinely implemented through linguistic representations. Linguistic relativity, on this view, is a claim about the *language-conditioned slice of cognition*, not about vision or motor control. Crucially, it is a claim about *lexicalization*—which distinctions a language makes cheap—not about grammar or syntax (a distinction developed in Section 6).

2. Proxy Hypothesis

Let H_L denote the latent space of language-conditioned cognitive states for speakers of language L: the states that are stable, reachable, and efficiently manipulable when thought is carried through the linguistic code of L.

We cannot directly observe H_L. What we observe are texts: the traces humans produce when navigating H_L under communicative and cultural constraints. Let Ψ denote the (unknown) mapping from language-conditioned cognition to text.

A monolingual transformer T_L trained on that text learns a representational manifold M_L sufficient to model the statistical and inferential structure of those traces. The crucial hypothesis is:

$$M_L \approx \Phi \circ \Psi(H_L) \tag{1}$$

for some model-dependent encoding Φ. In words: M_L is a measurable proxy for the geometry of language-conditioned cognition in L.

This bridge does not require that models "are minds." It requires a weaker, testable property: that Ψ preserves enough geometric structure (neighborhoods, relative distances, pathway costs) that differences in M_L are informative about differences in H_L. Without that preservation, geometry in models would be merely a modeling artifact. With it, cross-linguistic geometry in models becomes evidence about cross-linguistic geometry in language-mediated thought.

2.1 Strong Relativity as a Geometric Statement

The strong linguistic relativity hypothesis becomes: Languages induce measurably non-isometric cost landscapes over meaning for language-mediated cognition. What is a short hop in one language may be a detour in another—not because the destination is forbidden, but because the terrain is different.

2.2 The Trace Stability Assumption

The bridge claim $M_L \approx \Phi \circ \Psi(H_L)$ is only useful if the composition $\Phi \circ \Psi$ preserves geometric structure. We cannot directly verify this—H_L is unobservable. But we can state what must hold and test it indirectly.

Trace Stability Assumption. Let $S \subset H_L$ be a set of *meaning-anchors*: cognitive states with known relational structure (e.g., structurally-defined kinship positions, perceptually-anchored color regions, parallel translation pairs). On S, the composed mapping from cognition through text to model embedding is *quasi-isometric*:

$$\frac{1}{K} \cdot d_H(h_1, h_2) - \varepsilon \;\leq\; d_M\big(\Phi \circ \Psi(h_1),\, \Phi \circ \Psi(h_2)\big) \;\leq\; K \cdot d_H(h_1, h_2) + \varepsilon \tag{2}$$

for bounded constants $K \geq 1$ and $\varepsilon \geq 0$.

Relative distances are preserved up to bounded distortion. The mapping may stretch or compress, but not arbitrarily.

Why this is sufficient. We do not need $\Phi \circ \Psi$ to be an isometry, or to preserve all of H_L. Quasi-isometry on anchors suffices because:

1. Relative comparisons are preserved: if $d_H(a, b) < d_H(a, c)$, then $d_M(a', b') < d_M(a', c')$ up to bounded error
2. Topological features (connectivity, holes) are preserved under quasi-isometry
3. Cross-linguistic comparisons remain valid: if $d_{H_{L_1}}(a, b) \neq d_{H_{L_2}}(a, b)$, this will be reflected in M_{L_1} vs. M_{L_2}

Indirect testability. While we cannot directly measure d_H, we can test consistency: on anchor sets where we have independent grounds for relative distance (structural position in kinship, wavelength proximity in color, translation equivalence), the model geometry should respect these priors. Systematic violations would falsify trace stability and undermine the framework.

2.3 The Compression Alignment Argument

Why expect trace stability to hold? Transformers do not implement cognition, but they face the same compression problem under similar constraints.

Human language-conditioned cognition is a solution to a constrained optimization problem:

- **Finite capacity**: Working memory, attentional bandwidth, and retrieval are limited
- **Communicative pressure**: Representations must support efficient transmission to others
- **Inferential utility**: Representations must support reasoning, planning, and explanation
- **Frequency sensitivity**: Common meanings must be cheap; rare meanings may be expensive

A transformer trained on language faces structurally analogous pressures:

- **Finite capacity**: Fixed parameter count, embedding dimension, context length
- **Predictive pressure**: Representations must support accurate next-token prediction
- **Generalization**: Representations must transfer across contexts, supporting compositional and analogical extension
- **Frequency sensitivity**: Cross-entropy loss weights frequent tokens more heavily

These are not identical problems, but they share a common structure: *compress the statistical and inferential regularities of a language into a finite representational format that supports productive use.*

2.3.1 Geometric Convergence

When two systems solve similar compression problems under similar constraints, their solutions share geometric properties—the problem structure constrains the solution space.

If language L makes certain distinctions frequent and compositionally productive, both H_L and M_L face pressure to represent those distinctions cheaply. If certain inferences are commonly drawn in L, both systems face pressure to make those inferences geometrically short. The result is convergent geometry on the dimensions that matter for linguistic behavior.

M_L “proxies” H_L as an independent solution to the same representational problem, constrained by the same linguistic data.

2.3.2 What This Does Not Claim

1. **Mechanisms**: We are not asserting that attention heads implement human memory retrieval, or that residual streams mirror neural firing patterns. The claim is about representational geometry, not computational process.

2. **Scope**: The bridge applies to language-conditioned cognitive states—the slice of thought that runs through linguistic representation. Perception, motor control, and affect are outside scope.

3. **Consciousness**: Whether models have inner experience is orthogonal. The claim is that their learned geometries are informative about human representational geometries.

2.3.3 Empirical Test of the Proxy Relationship

If M_L is a good proxy for H_L, then geometric properties measured in models should predict behavioral measures in humans:

- Geodesic distance in M_L should predict reaction time for semantic relatedness judgments
- High-curvature regions should predict order effects in human reasoning tasks
- Cross-linguistic geometric divergence should predict translation difficulty and bilingual switching costs

If model geometry predicts human behavior, the bridge holds. If not, the framework fails.

3. The Operational Metric

The geometric claims become meaningful only if the metric corresponds to a meaningful notion of *cognitive cost* or behavioral substitutability—not merely dot products in an arbitrary coordinate system.

3.1 The Fisher Information Metric

Define distance between internal states h_1 and h_2 by how differently they behave as predictors of language. Let the model produce a distribution over next tokens $p(\cdot \mid h)$. Define a local metric tensor:

$$g_h(u, v) = u^\top F(h)\, v \tag{3}$$

where $F(h)$ is the Fisher information matrix of $p(\cdot \mid h)$ with respect to perturbations in h.

Intuition: Directions that drastically change the next-token distribution are "long" directions. Directions that leave predictions unchanged are "short." Distance is tied to *functional similarity*—how the state behaves—not to arbitrary embedding coordinates.

3.2 Alternative: Jensen–Shannon Distance

For a global distance measure:

$$d(h_1, h_2) = \mathrm{JS}\big(p(\cdot \mid h_1),\, p(\cdot \mid h_2)\big) \tag{4}$$

Jensen–Shannon divergence is symmetric and bounded, giving a well-defined metric on representational states grounded in behavioral difference.

4. Curvature as Cognitive Friction

Curvature measures the *path-dependence of inference* under the model's dynamics—not metaphysical drift, but computational instability.

4.1 Noncommutativity of Update Operators

Let U_A and U_B be operators corresponding to incorporating premise A then premise B into the latent state. (Instantiate as: append A to context and run the model forward; similarly for B.) Define a friction measure:

$$F(h; A, B) = d\big(U_B(U_A(h)),\, U_A(U_B(h))\big) \tag{5}$$

If F is large in some region, *order matters*: reasoning is route-sensitive. The same premises, encountered in different orders, yield different conclusions.

This is the operational meaning of curvature. On a flat manifold, parallel transport is path-independent. On a curved manifold, the same idea transported along different paths arrives *differently*. The noncommutativity measure F is an empirical proxy for this geometric fact.

4.2 Cognitive Friction as Integrated Instability

For a region $U \subset M_L$, define cognitive friction as the average noncommutativity over premise pairs relevant to that domain:

$$\text{Friction}(U) = \mathbb{E}_{A,B\sim U}\big[F(h; A, B)\big] \tag{6}$$

High friction means reasoning in domain U is computationally unstable—small perturbations in reasoning path produce divergent conclusions.

5. Geodesics as Minimal-Cost Transformations

A geodesic is a minimal-cost path under the functional metric. In the geometry of M_L, geodesics represent *natural inferences*—the paths a reasoner follows when moving from concept A to concept B with minimal representational work.

5.1 The Cost Functional

Let a path $\gamma(t)$ in representation space induce output distributions $p(\cdot \mid \gamma(t))$. Define path length:

$$\text{Len}(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}\big(\dot\gamma(t), \dot\gamma(t)\big)}\, dt \tag{7}$$

The geodesic from A to B is the path of minimal functional change needed to transform predictive behavior:

$$d_L(A, B) = \inf_{\gamma: A\to B} \text{Len}(\gamma) \tag{8}$$

5.2 Cross-Linguistic Inferential Divergence

After aligning concepts across languages via parallel meaning-anchors, define inferential divergence:

$$\Delta(A, B) = \big|d_{L_1}(A, B) - d_{L_2}(A, B)\big| \tag{9}$$

Non-zero Δ is a geometric signature of linguistic relativity: the *minimal representational work* required to traverse from A to B differs between languages. Concepts that are “one step apart” in one language require traversal through intermediate concepts in another.

6. Lexicalization as Geometric Privilege

6.1 The MDL Definition

Rather than defining lexicalization as “has a single lexeme,” define it via minimum description length in L:

$$w_L(c) \propto -\min_{s \in E_L(c)} \mathrm{MDL}(s) \tag{10}$$

where $E_L(c)$ is the set of acceptable expressions for concept c, and MDL can be approximated by model surprisal $-\log p_L(s)$ plus a length penalty.

This makes lexicalization a continuous, cross-linguistically comparable quantity.

6.2 Grammar vs. Geometry

The manifold structure is not determined by syntax. Grammar specifies the rules for composing tokens into well-formed expressions; the manifold captures which conceptual states are cheap to occupy and traverse. These are independent.

English and Spanish share most grammatical structure—SVO word order, similar tense systems, comparable phrase structure—but may diverge geometrically wherever lexicalization differs. Spanish lexicalizes *sobremesa* (the lingering conversation after a meal) as a single token; English requires circumlocution. The grammar for expressing the concept is trivially similar across languages. The representational cost is not.

Conversely, Japanese and English differ radically in syntax: SOV vs. SVO, postpositions vs. prepositions, topic-prominence vs. subject-prominence, pervasive ellipsis vs. obligatory arguments. Yet in domains where they lexicalize the same distinctions at comparable granularity, their manifolds could be locally isometric despite surface dissimilarity.

The empirical prediction is therefore: **geometric divergence tracks lexicalization patterns, not grammatical typology.** This is directly testable by comparing language pairs that share grammar but differ in lexicalization against pairs that differ in grammar but share lexicalization. The framework predicts the former comparison reveals geometric divergence; the latter reveals geometric similarity.

6.3 The Geometric Privilege Hypothesis

Lexicalized concepts occupy geometrically privileged regions:

1. **Lower local curvature**—reasoning nearby is stable

2. **Higher tangent space dimensionality**—more directions of extension available

3. **Shorter geodesic distances to related concepts**—analogies come easily

4. **Better alignment with principal components**—less superposition, cleaner separation

6.4 Testable Predictions

Let $\kappa(c)$ be local curvature (estimated via noncommutativity), $\tau(c)$ be local tangent dimension, $\alpha(c)$ be axis alignment. Then:

$$\operatorname{Corr}\big(w_L(c), -\kappa(c)\big) > 0 \tag{11}$$

$$\operatorname{Corr}\big(w_L(c), \tau(c)\big) > 0 \tag{12}$$

$$\operatorname{Corr}\big(w_L(c), \alpha(c)\big) > 0 \tag{13}$$

The correlations should hold within a language, and the *pattern of differences* should track *lexical differences* across languages.

Three additional predictions tie the remaining measures to observable behavior:

Friction. High-friction regions (large $F(h; A, B)$) should predict order effects in human reasoning: same premises encountered in different orders yield different conclusions. This directly tests path-dependence.

Cross-linguistic divergence. $\Delta(A, B) = |d_{L_1}(A, B) - d_{L_2}(A, B)|$ should predict translation difficulty—human translator time, error rates, and quality ratings. If geometric divergence is real, it surfaces where humans struggle to translate.

Energy barriers. $B(A, B)$ should predict scaffolding frequency in corpora: how often speakers resort to paraphrase, metaphor, or borrowing when expressing transitions between A and B. High barriers imply speakers route around them.

7. Topological Constraints: Energy Barriers

Topology determines what's *costly*, not what's impossible. "Holes" in the manifold are not logical impossibilities but regions where no low-cost path exists within standard linguistic dynamics.

7.1 Energy Barriers

Define an energy barrier between regions A and B:

$$B(A, B) = \inf_{\gamma: A \to B} \max_{t \in [0,1]} \operatorname{Cost}\big(\gamma(t)\big) \tag{14}$$

where Cost could be surprisal, rarity density penalty, or distance from the typical-set manifold. If B is high, the transition requires *scaffolding*—even if it's not logically impossible.

7.2 Scaffolding as Tunneling

When barriers are high, speakers deploy mechanisms that lower them:

- **Paraphrase**—rerouting through lower-cost intermediate concepts
- **Neologism**—adding a bridge token that creates a new low-cost path
- **Borrowing**—importing a lexeme with its geometric privileges

- **Metaphor**—exploiting structural similarity to tunnel between distant regions
- **Code-switching**—temporarily shifting to a different representational regime
- **Formal notation**—expanding the representational space

These are not workarounds for the unthinkable. They are *tunneling mechanisms*: ways of rerouting around or punching through high-energy regions of the representational landscape.

8. The Measurement Program

8.1 Protocol

1. Train identical transformer architectures on monolingual corpora: $L_1, L_2, \ldots, L_n$
2. Identify concept sets C with known cross-linguistic variation in lexicalization
3. Measure geometric properties: curvature proxies $\kappa(c)$, tangent dimension $\tau(c)$, geodesic distances, energy barriers
4. Test predictions across three levels:
   - *Within-language:* lexicalization status predicts geometric privilege
   - *Across-language:* geometric differences track lexical differences
   - *Strong test:* Linguistic typology features (from WALS, etc.) correlate with manifold topology

8.2 Worked Example: Kinship as Geometric Test Case

Kinship is anchored independently of language. Biological facts (reproduction, shared parentage) and social facts (marriage, co-residence) define a universal relational space. Languages differ in how they lexicalize it.

The structural concept space. Define kinship positions by genealogical path: Mother, Father, Mother's-Brother, Father's-Brother, Mother's-Sister, Father's-Sister, Parent's-Sibling's-Child, etc. These are language-independent anchors.

Cross-linguistic variation. Languages partition this space differently:

- *English*: Collapses Mother's-Brother and Father's-Brother into "uncle"; collapses all Parent's-Sibling's-Child into "cousin"
- *Tamil*: Distinguishes Mother's-Brother (*māmā*) from Father's-Brother (*periyappā/chiththappa*); cross-cousins from parallel-cousins
- *Hawaiian*: Generational system—all same-generation relatives of same sex share a term
- *Sudanese*: Maximally differentiated—unique terms for each genealogical position

Measurement protocol.

1. For each language L, collect minimal natural expressions for each anchored kinship position

2. Embed expressions in M_L using the monolingual model
3. Compute pairwise geodesic distances under the Fisher/JS metric
4. Construct the distance matrix D_L over the anchor set

Predictions.

1. *Lexical boundary effect*: Within a lexical category, distances should be smaller; across lexical boundaries, larger. For Tamil, $d(\text{M-Bro}, \text{F-Bro})$ should exceed the English $d(\text{M-Bro}, \text{F-Bro})$, where English collapses both to "uncle."
2. *Dimensionality tracking*: Languages with finer kinship distinctions should show higher local dimensionality in kinship-concept regions of M_L.
3. *Curvature at boundaries*: Lexical boundaries should correspond to higher curvature (higher noncommutativity of inference across the boundary).

The protocol extends to other structurally-anchorable domains: color (wavelength), spatial relations (geometric configuration), evidentiality (information source), numeral systems (cardinality).

8.3 Estimation by Measure

Each theoretical quantity requires an estimation procedure tied to specific hypotheses.

Lexicalization weight $w_L(c)$. The independent variable for geometric privilege predictions. Estimate via model surprisal on minimal expressions: $w_L(c) = -\log p_L(s^*)$ where s^* is the shortest natural expression for c in L. Higher w_L indicates cheaper encoding.

Local curvature $\kappa(c)$. Measures path-dependence of reasoning near c. Estimate via the noncommutativity measure $F(h; A, B) = d(U_B(U_A(h)), U_A(U_B(h)))$ averaged over premise pairs relevant to c, using JS divergence for d. *Tests:* $\text{Corr}(w_L, -\kappa) > 0$; curvature elevated at lexical boundaries.

Tangent dimensionality $\tau(c)$. Measures richness of local variation. Estimate via local PCA on the k-neighborhood of c: count dimensions explaining 95% of variance, or use maximum likelihood intrinsic dimension estimators. *Tests:* $\text{Corr}(w_L, \tau) > 0$; finer lexical distinctions correlate with higher local dimensionality.

Geodesic distance $d_L(A, B)$. Minimal representational work between concepts. Under the Fisher metric, approximate via shortest path on a k-NN graph with edge weights given by JS divergence between adjacent states. Alternative: diffusion distance, which integrates over paths. *Tests:* Cross-linguistic divergence $\Delta(A, B) = |d_{L_1}(A, B) - d_{L_2}(A, B)|$ tracks lexicalization differences; lexical boundary effects in kinship and other anchored domains.

Axis alignment $\alpha(c)$. Measures separation versus superposition. Project representations near c onto principal components of the full embedding space; $\alpha(c)$ is the fraction of variance captured by the top k components. *Tests:* $\text{Corr}(w_L, \alpha) > 0$—lexicalized concepts are less superposed.

Energy barriers $B(A, B)$. Maximum cost along the cheapest path. Estimate via minimax path cost on the k-NN graph: $B(A, B) = \min_\gamma \max_t w(\gamma(t))$ where w is edge weight. *Tests:* High barriers where scaffolding is empirically required; barriers differ cross-linguistically in predicted domains.

8.4 Null and Alternative Hypotheses

Null (pure compression): All M_L are isometric up to noise. Language is mere encoding; structure is universal.

Strong Sapir–Whorf: M_L are measurably non-isometric in ways predicted by lexicalization patterns—which distinctions a language makes cheap to encode—not by grammatical typology (see Section 6).

The detailed experimental protocols, controls, and falsification conditions required to cleanly distinguish these hypotheses from confounds (corpus content, frequency, tokenization) are developed in Appendix A.

9. Implications

9.1 For Cognitive Science

The Sapir-Whorf hypothesis has long been framed as a binary question: does language determine thought, or doesn't it? Decades of debate have produced equivocal results precisely because the question is ill-posed. "Determine" is not a measurable quantity.

The geometric framework reframes the question: *How much does each language distort the cost landscape over meaning?* This is a quantitative question with continuous answers. For any pair of concepts (A, B) and any pair of languages (L_1, L_2), we can ask: what is the divergence $\Delta(A, B) = |d_{L_1}(A, B) - d_{L_2}(A, B)|$? The answer will vary by domain, by concept pair, by language pair. Some regions of meaning may show near-isometry across languages; others may show substantial distortion.

Linguistic relativity becomes a distribution to be measured: a map of where and how much languages diverge in the costs they impose on thought.

9.2 For AI Alignment

If language shapes the geometry of thought, then *training data shapes the geometry of artificial minds.* A model trained predominantly on legal English may have systematically different reasoning topology than one trained on literary Russian—different computational affordances, not just different knowledge.

Monolingually-trained models may have systematic blind spots—not missing knowledge, but missing *paths.* If certain inferences are geometrically cheap in Mandarin but expensive in English, an English-only model may struggle with those inferences even when it "knows" the relevant facts. The knowledge is present; the route to it is costly. Multilingual training is *geometric expansion*: creating new low-cost paths, flattening high-curvature regions, making previously difficult thoughts easy.

Mechanistic interpretability already analyzes model representations geometrically via classifiers, activation patching, and representation similarity. This framework extends that work with specific measures (curvature, geodesic distance, energy barriers) tied to behavioral predictions. The same geometric quantities that test linguistic relativity in humans become diagnostic instruments for artificial minds: map the cost landscape, identify high-curvature regions where reasoning is unstable, find the barriers the model routes around.

Coda: Different Manifolds, Different Minds

(as cost landscapes)

Large language models are trained on nothing but language, and yet they learn behaviors that resemble inference: analogy, explanation, abstraction, the decomposition of problems into steps. Language models do not instantiate human cognition in general—humans think with bodies, perception, affect, long-lived memory, and goals. But something narrower follows:

> **Language carries enough structure to support a powerful form of cognition when cognition is constrained to run through linguistic representations.**

On the compression view, language is the code that makes complex social and conceptual life computationally tractable. Compression is never neutral—and crucially, it *creates* structure rather than mapping onto a universal structure. A code induces a geometry: it determines what is near, what is far, what is stable, what is fragile, what routes are cheap. The coordinate system is constitutive of the computational affordances.

When the induced landscape contains high-cost barriers—regions that are sparse, unstable, or require atypical representational commitments—speakers do not encounter "impossible thoughts." They encounter *energy barriers*: transitions that are available in principle but costly to traverse within the ordinary routines of the language.

Scaffolding has precise meaning within this framework. Paraphrase, neologism, borrowing, metaphor, code-switching, formal notation—these are *tunneling mechanisms*: ways of rerouting around or punching through high-energy regions of the representational landscape. In the formalism, they lower the barrier term $B(A, B)$ by changing the effective code.

Different languages, different induced landscapes. If a "mind" is characterized by the stable states it can occupy fluently and by the cost of moving between them, then learning a different landscape is learning a different mode of thought. The gradients and barriers differ.

Different manifolds are different minds in a precise, testable sense: they define different cost landscapes over meaning, and cost landscapes determine what is fluent, what is fragile, what requires tunneling.

These landscapes are measurable—via curvature proxies, geodesic distortions, energy barriers—especially where typology predicts structured differences.

The task now is to measure.

A. Experimental Protocols for Causal Identification

The measurement program in Section 8 provides the theoretical quantities and their estimation procedures. This appendix addresses a harder problem: how to establish that observed geometric differences are caused by *linguistic structure* rather than by confounds—corpus content, domain mix, frequency distributions, tokenization artifacts, or training stochasticity.

The core challenge is causal identification. We observe:

$$\text{Corpus} \longrightarrow \text{Model} \longrightarrow \text{Geometry}$$

But corpus correlates with linguistic structure *and* with many other factors. Naive comparison of monolingual models cannot distinguish "language shapes geometry" from "corpora differ."

A.1 Model Architecture: BERT as Primary Platform

We recommend BERT-family encoder models as the primary experimental platform for several reasons:

Practical advantages:

- **Pretrained availability:** Monolingual BERT models exist for dozens of languages: German (german-bert), French (CamemBERT), Spanish (BETO), Arabic (AraBERT), Chinese (chinese-bert), Japanese (cl-tohoku/bert), Tamil and other Indic languages (IndicBERT, MuRIL), Finnish (FinBERT), and many others. These are publicly available and require no training.

- **Multilingual BERT (mBERT):** A single model trained on 104 languages with identical architecture. This provides a critical control: comparing representations *within* the same model across languages eliminates architecture as a confound.

- **Scale:** BERT-base (110M parameters) and BERT-large (340M parameters) are small enough to run on consumer hardware, enabling rapid iteration and replication.

- **Embedding-native:** BERT is an encoder designed to produce dense contextual representations. Extracting meaningful embeddings is straightforward, unlike decoder models optimized for generation.

- **Tooling:** Extensive libraries exist for BERT analysis (HuggingFace Transformers, BertViz, TransformerLens).

Theoretical compatibility: The main text defines the functional metric via next-token prediction (Section 3). BERT uses masked language modeling (MLM) rather than autoregressive prediction, requiring adaptation:

- **Embedding distance (primary):** Cosine distance on contextual embeddings—either the [CLS] token or mean-pooled token representations. This is standard practice and well-validated for semantic similarity tasks.

- **MLM-based functional distance:** Define $d_{\text{MLM}}(h_1, h_2) = \text{JS}(p([\text{MASK}] \mid c_1), p([\text{MASK}] \mid c_2))$ where c_1, c_2 are contexts differing only in the target concept position. This preserves the behavioral grounding of Section 3.

- **Robustness:** Report results under both metrics. Core predictions should hold under either.

The mBERT control: Multilingual BERT enables a powerful within-model comparison. If we observe geometric differences between English and Tamil representations *within mBERT*—same architecture, same training procedure, same parameter count—those differences cannot be attributed to model architecture. They must arise from the linguistic data. This is a cleaner comparison than cross-model analysis and should be the first-line test.

When to use other architectures: Autoregressive models (GPT-family) may be preferable when: (1) the curvature measure (Section 4) is central, since sequential update operators are more natural; (2) the research question involves generation or continuation behavior; (3) replication across architectures is needed to establish robustness.

A.2 The Evidence Hierarchy

We propose five levels of evidence, ordered by the strength of causal identification:

1. **Descriptive (correlational):** Pretrained monolingual models exhibit geometric differences across languages. Using existing models (no training required), this establishes whether the phenomenon exists. *Problem:* Confounded with corpus content, domain mix, frequency.

2. **Parallel corpus control:** Models trained on sentence-aligned parallel corpora (same propositional content, different linguistic encoding) still exhibit geometric differences. *Controls:* Content. *Remaining confounds:* Register, tokenization, frequency within parallel data.

3. **Synthetic semantic control:** Models trained on corpora with *identical* semantic content by construction (controlled concept sets, matched frequencies) exhibit geometric differences predicted by lexicalization. *Controls:* Content, frequency. *Remaining confounds:* Minimal—this is nearly clean.

4. **Causal intervention:** Fine-tuning a model to add a lexical distinction produces predicted geometric changes. *Establishes:* Causal direction from lexicalization to geometry. BERT fine-tuning is extremely well-documented and tractable.

5. **Behavioral prediction:** Model geometric properties predict human behavioral measures (reaction times, order effects, translation difficulty). *Establishes:* The bridge claim—that model geometry proxies cognitive geometry.

A compelling case for linguistic relativity requires reaching at least Level 2, ideally Levels 3–4, with Level 5 as validation. Critically, Level 1 can be executed immediately using existing pretrained models.

A.3 The Operational Null Hypothesis

Section 8.4 states the null informally. For rigorous testing, we require precision about *what class of mappings* count as "isometry" and *what magnitude of residual* counts as "noise."

A.3.1 Formal Statement

Let S be an anchor set of n meaning-anchors with known cross-linguistic correspondence (e.g., kinship positions, color coordinates). For language L, let $D_L \in \mathbb{R}^{n\times n}$ be the pairwise distance matrix on S under the functional metric (Section 3).

Let $\mathcal{F}$ be an alignment class—a family of transformations under which we consider manifolds equivalent. We propose Procrustes alignment (orthogonal transformation plus uniform scaling) as the default, with robustness checks under learned neural alignment.

Null Hypothesis H_0: For each language pair (L_1, L_2), there exists $f^* \in \mathcal{F}$ such that:

$$\|D_{L_1} - f^*(D_{L_2})\|_F \ \leq \ \varepsilon_{\text{noise}} \tag{15}$$

where $\varepsilon_{\text{noise}}$ is estimated from within-language resampling (different training runs, random seeds, or context samples).

Alternative Hypothesis H_1: Residuals exceed noise, and concentrate at boundaries predicted by lexicalization differences.

A.3.2 Test Procedure

1. For each language L, obtain models: either pretrained monolingual BERTs, or mBERT representations for text in L, or (for robustness) k models trained with different random seeds
2. Compute anchor distance matrices D_L using cosine distance on embeddings
3. Estimate within-language noise via context variation: compute D_L across 10 different sentence frames, take $\varepsilon_{\text{noise}} = \text{std}(\|D_L^{(i)} - D_L^{(j)}\|_F)$
4. For language pair (L_1, L_2), find optimal alignment: $f^* = \arg\min_{f \in \mathcal{F}} \|D_{L_1} - f(D_{L_2})\|_F$
5. Compute residual: $R = D_{L_1} - f^*(D_{L_2})$
6. Test: Is $\|R\|_F > \varepsilon_{\text{noise}} + 2\sigma$? (Reject H_0 if yes)
7. For significant pairs: Does R concentrate at lexically-predicted boundaries?

Fast-path with mBERT: For initial testing, use mBERT for all languages. Extract embeddings for anchor terms in each language within the same model. This eliminates model architecture as a confound entirely—any geometric difference must arise from the linguistic encoding.

A.4 Level 1 Protocol: Descriptive Comparison with Existing Models

Before investing in training, establish whether the basic phenomenon exists using pretrained models.

A.4.1 Design

Use publicly available pretrained models:

- **Within-model comparison (primary):** mBERT or XLM-RoBERTa—extract embeddings for the same concepts expressed in different languages
- **Cross-model comparison:** Language-specific BERTs (BETO for Spanish, IndicBERT for Tamil, bert-base-uncased for English)

A.4.2 Concrete Implementation

For kinship domain with English, Spanish, Tamil:

1. Define anchor set: 20 kinship positions (mother, father, uncle, aunt, cousin, mother's-brother, father's-brother, etc.)
2. For each position, create minimal expressions in each language
3. Embed in controlled context: "[TERM] is a family relationship" or equivalent
4. Extract layer 8 (of 12) embeddings; mean-pool over subword tokens
5. Compute pairwise cosine distance matrices D_{en}, D_{es}, D_{ta}
6. Test predictions: Is $d_{\text{ta}}(\text{M-Bro}, \text{F-Bro}) > d_{\text{en}}(\text{M-Bro}, \text{F-Bro})$?

This protocol requires no training.

A.5 Level 2 Protocol: Parallel Corpus Control

A.5.1 Design

Train BERT models on parallel versus monolingual corpora:

- M_L^{mono}: BERT trained on monolingual corpus (Wikipedia)
- M_L^{par}: BERT trained on parallel corpus (sentence-aligned translations)

For language pair (L_1, L_2), we obtain four models: $M_{L_1}^{\text{mono}}$, $M_{L_2}^{\text{mono}}$, $M_{L_1}^{\text{par}}$, $M_{L_2}^{\text{par}}$.

Resource note: Training BERT-base from scratch requires significant compute. For resource-constrained settings, continue pretraining from existing checkpoints on the target corpus.

A.5.2 Logic

The parallel models see the *same propositions* expressed in different languages. If geometric divergence is driven by propositional content (what is talked about), parallel models should converge. If driven by linguistic structure (how it is encoded), parallel models should still diverge.

A.5.3 Predictions

Let $\Delta^{\text{mono}} = \|D_{L_1}^{\text{mono}} - f^*(D_{L_2}^{\text{mono}})\|$ and $\Delta^{\text{par}} = \|D_{L_1}^{\text{par}} - f^*(D_{L_2}^{\text{par}})\|$.

Under linguistic relativity:

$$\Delta^{\text{par}} \approx \Delta^{\text{mono}} \gg \varepsilon_{\text{noise}} \tag{16}$$

Content control does not eliminate divergence; the pattern of divergence is similar.

Under corpus confounding:

$$\Delta^{\text{par}} \ll \Delta^{\text{mono}} \tag{17}$$

Content control eliminates or substantially reduces divergence.

A.5.4 Required Data

- Sentence-aligned parallel corpora: UN Parallel Corpus (6 languages, 400M+ sentences), Europarl (21 languages), OPUS collection, Bible translations (1000+ languages)
- Monolingual corpora matched on domain mix and size
- Target language pairs with known lexicalization differences in test domains

A.6 Level 3 Protocol: Synthetic Semantic Control

A.6.1 Design

Construct a semantic universe U with *known* relational structure:

- **Kinship:** 50 positions defined by genealogical path (Mother, Father, Mother's-Brother, Father's-Brother, Mother's-Mother, etc.)
- **Color:** 330 Munsell chips with CIELab coordinates
- **Spatial:** 100 geometric configurations (topological relations, cardinal directions)

For each concept $c \in U$:

1. Elicit minimal natural expressions from balanced bilingual speakers (same individual, both languages)
2. Ensure referent identity: same family tree diagram, same color chip, same spatial configuration
3. Generate corpus sentences using these expressions in controlled syntactic frames
4. Match frequency distributions across languages (each concept appears equally often)

A.6.2 Logic

The semantic structure is identical by construction. Frequency is matched. Only the linguistic encoding differs. Any geometric difference is attributable to encoding, not content.

A.6.3 Predictions

Let D_{true} be the ground-truth distance matrix (genealogical distance for kinship, CIELab distance for color). Let D_L^{syn} be the model distance matrix from synthetic training.

Define distortion:

$$\text{Distortion}_L(c_1, c_2) = \left|D_{\text{true}}(c_1, c_2) - D_L^{\text{syn}}(c_1, c_2)\right| \tag{18}$$

Prediction: Distortion is *lower* for concept pairs that share a lexeme in L than for pairs that don't:

$$\mathbb{E}\left[\text{Distortion}_L \mid \text{same lexeme}\right] \ < \ \mathbb{E}\left[\text{Distortion}_L \mid \text{different lexemes}\right] \tag{19}$$

Lexicalization compresses representation toward the ground truth within lexical categories.

A.6.4 Limitations

Synthetic corpora may not produce models with full linguistic competence. This protocol tests whether the *mechanism* works, not whether it operates at scale in natural language. Positive results here provide proof of concept; negative results are not definitive.

A.7 Level 4 Protocol: The Lexicalization Intervention

This is the critical causal experiment. BERT fine-tuning makes it highly tractable.

A.7.1 Design

Target: English kinship geometry, specifically the mother's-brother / father's-brother distinction (both collapsed to "uncle").

Baseline: Start with pretrained `bert-base-uncased`.

Intervention: Fine-tune on augmented data introducing a lexical distinction:

- Introduce neologism "avuncle" (or borrow "māmā" from Tamil) for mother's-brother
- Generate 10,000+ sentences using this term in natural contexts: "My avuncle (mother's brother) came to visit," "She inherited the house from her avuncle," etc.
- Add "avuncle" to the tokenizer vocabulary
- Fine-tune using masked language modeling objective for 3–5 epochs
- Obtain $M_E^{+\text{lex}}$

Control Interventions:

- $M_E^{+\text{freq}}$: Same data volume, but with a semantically-empty nonce word ("blicket") in the same syntactic positions. Tests whether mere frequency produces geometric change.
- $M_E^{+\text{noise}}$: Same data volume of unrelated content (e.g., random Wikipedia articles). Tests whether any additional training produces geometric change.
- $M_E^{+\text{synonym}}$: Introduce "avuncle" as a pure synonym for "uncle" with no semantic distinction (used interchangeably in training data). Tests whether new word form alone produces geometric change.

Resource note: Fine-tuning BERT-base on 10K sentences is computationally modest—a single GPU suffices for the full experiment (baseline + intervention + 3 controls).

A.7.2 Measurements

Before and after intervention, measure:

- $d(\text{M-Bro}, \text{F-Bro})$: Distance between mother's-brother and father's-brother representations
- $\kappa(\text{uncle})$: Local curvature in the uncle region
- $\tau(\text{uncle})$: Local tangent dimensionality

- α(M-Bro), α(F-Bro): Axis alignment of each concept

A.7.3 Predictions

Model	$d(\text{M-Bro}, \text{F-Bro})$	κ	τ	α
M_E (baseline)	Small	Low	Low	Low
$M_E^{+\text{lex}}$ (new distinction)	**Large**	**High at boundary**	**Higher**	**Higher**
$M_E^{+\text{freq}}$ (nonce word)	Small	Low	Low	Low
$M_E^{+\text{noise}}$ (unrelated)	Small	Low	Low	Low
$M_E^{+\text{synonym}}$ (pure synonym)	Small	Low	Low	Low

If $M_E^{+\text{lex}}$ shows predicted changes and controls do not, we have established:

$$\text{Lexicalization} \longrightarrow \text{Geometric Privilege} \tag{20}$$

This is causal identification: we manipulated the independent variable and observed the predicted effect.

A.7.4 Cross-Linguistic Convergence Test

A stronger prediction: after intervention, English geometry should *more closely resemble* Tamil geometry in the kinship region.

Let D_{Tamil} be the kinship distance matrix from a Tamil model. Measure:

$$\|D_E^{+\text{lex}} - D_{\text{Tamil}}\| \;<\; \|D_E - D_{\text{Tamil}}\| \tag{21}$$

If introducing the Tamil-like distinction makes English geometry Tamil-like, the theory is strongly supported.

A.8 Level 5 Protocol: Behavioral Validation

The bridge claim (Section 2) asserts that M_L proxies H_L. This is testable.

A.8.1 Design

Measure geometric properties in models and behavioral properties in human speakers of the same languages. Test whether model geometry predicts human behavior.

A.8.2 Model Measurements

For target concepts A, B in language L:

- Geodesic distance $d_L(A, B)$ (Section 5)
- Local curvature $\kappa_L(c)$ (Section 4)
- Energy barrier $B_L(A, B)$ (Section 7)
- Cross-linguistic divergence $\Delta(A, B) = |d_{L_1}(A, B) - d_{L_2}(A, B)|$ (Section 5)

A.8.3 Human Measurements

- **Reaction time:** Semantic relatedness judgment RT for concept pairs
- **Order effects:** Present premises A, B in both orders; measure conclusion difference
- **Priming:** Semantic priming magnitude between concept pairs
- **Translation difficulty:** Professional translator time, error rate, quality ratings
- **Scaffolding frequency:** Corpus frequency of paraphrase, metaphor, borrowing when expressing target concepts

A.8.4 Predictions

$$\mathrm{Corr}\big(d_L(A,B),\ \mathrm{RT}_L(A,B)\big) > 0 \tag{22}$$

$$\mathrm{Corr}\big(\kappa_L(c),\ \mathrm{OrderEffect}_L(c)\big) > 0 \tag{23}$$

$$\mathrm{Corr}\big(B_L(A,B),\ \mathrm{ScaffoldFreq}_L(A,B)\big) > 0 \tag{24}$$

$$\mathrm{Corr}\big(\Delta(A,B),\ \mathrm{TranslationDifficulty}(A,B)\big) > 0 \tag{25}$$

If model geometry predicts human behavior, the bridge holds. If not, either the bridge claim fails, the geometric measures are wrong, or the behavioral measures are noisy.

A.8.5 Diagnostic

To distinguish these failure modes, first test whether model geometry predicts *model behavior*:

- Does $d_L(A,B)$ predict model perplexity on sentences relating A and B?
- Does $\kappa_L(c)$ predict order sensitivity in model outputs?

If model geometry predicts model behavior but not human behavior, the measures are valid but the bridge fails. If model geometry doesn't even predict model behavior, the measures need refinement.

A.9 Measurement Specifications for BERT

The geometric quantities in Section 8.3 require precise extraction procedures. This section provides BERT-specific implementation guidance.

A.9.1 Layer Selection

BERT representations vary systematically by layer. Research on BERT probing (Jawahar et al., 2019; Tenney et al., 2019) establishes:

- Layers 0–4: Surface features (POS, morphology)
- Layers 5–8: Syntactic features (constituency, dependencies)
- Layers 9–12: Semantic features (entity types, relations, coreference)

For semantic geometry, we extract from:

- **Primary:** Layer 9 or 10 (of 12) for BERT-base; layer 18–20 (of 24) for BERT-large
- **Robustness:** Report results across layers 6, 8, 10, 12 to show consistency
- Effect should be strongest in later layers; if strongest in early layers, it may reflect surface form rather than semantics

A.9.2 Token Selection

For concept c expressed as subword tokens $t_1, \dots, t_n$:

- **[CLS] token:** Standard for sentence-level representations; captures whole-context semantics
- **Target token mean-pool:** Average embeddings of subword tokens comprising the target term (e.g., for "grandmother," average "grand" and "##mother")
- **Robustness:** Compare [CLS], target-mean-pool, and last-subword; results should be qualitatively consistent

Implementation note: Use the tokenizer's `word_ids()` method to identify which subwords correspond to the target term.

A.9.3 Context Specification

BERT embeddings are context-dependent by design. We use:

- **Canonical frame:** "[TERM] is a type of family member" (neutral, definitional)
- **Multiple contexts:** 10 diverse frames per concept:
  - Definitional: "A [TERM] is someone who..."
  - Relational: "My [TERM] lives nearby"
  - Possessive: "[TERM]'s house is large"
  - Narrative: "I visited my [TERM] yesterday"
- **Aggregation:** Report mean embedding and variance across contexts
- High variance suggests the concept is context-sensitive; low variance suggests stable representation

A.9.4 Distance Metric Specification

Primary metric: Cosine distance on contextual embeddings:

$$d_{\cos}(h_1, h_2) = 1 - \frac{h_1 \cdot h_2}{\|h_1\| \|h_2\|} \tag{26}$$

This is standard for BERT and well-validated for semantic similarity tasks.

Behavioral grounding (optional): For results more directly tied to Section 3, use MLM-based JS divergence:

$$d_{\text{MLM}}(c_1, c_2) = \text{JS}\big(p([\text{MASK}] \mid \text{context}_1), p([\text{MASK}] \mid \text{context}_2)\big) \tag{27}$$

where contexts are matched except for the concept position. This measures how differently the model “expects” the next word given each concept.

Robustness: Report both metrics. Core predictions should hold under either.

A.9.5 Curvature Estimation for Bidirectional Models

The noncommutativity measure $F(h; A, B)$ from Section 4 assumes sequential processing. For BERT’s bidirectional architecture, we adapt:

Method 1: Context perturbation.

- For concept c and context modifiers A, B: compute embeddings of c in contexts containing A only, B only, A then B, and B then A
- Define $F(c; A, B) = d(h_{c|AB}, h_{c|BA})$—the distance between embeddings when modifiers appear in different orders
- High F indicates order-sensitive (curved) regions

Method 2: Attention pattern analysis.

- Extract attention weights from concept token to context tokens
- Measure whether attention patterns are stable across related concepts (flat) or highly variable (curved)
- Use attention entropy as a local curvature proxy

Method 3 (for comparison): Use autoregressive model (GPT-2) with the sequential F measure from Section 4. Cross-architecture consistency strengthens the result.

A.9.6 Topology Estimation

Persistent homology can be applied to BERT embedding point clouds:

- **Sampling:** Embed 1000+ instances per concept using varied contexts; embed all concepts in anchor set
- **Scale selection:** Pre-register scale range as $[0.5\bar{d}, 2\bar{d}]$ where $\bar{d}$ is mean inter-anchor cosine distance
- **Software:** Use `giotto-tda` or `ripser` for computation
- **Stability:** Report persistence diagrams; features must persist over $> 20\%$ of scale range
- **Subsampling:** Repeat on 10 bootstrap subsamples; report only features stable across subsamples

A.10 Required Controls

A.10.1 The mBERT Architecture Control

The most powerful initial control uses multilingual BERT. Within a single model:

- Same architecture, same parameters, same training procedure
- Only difference: the language of the input text
- Any geometric difference between English and Tamil representations *within mBERT* cannot be due to model architecture

This should be the first test. If geometric divergence appears within mBERT at lexically-predicted boundaries, the phenomenon is robust to architecture confounds.

A.10.2 Grammar vs. Lexicalization Controls

Section 6 claims geometric divergence tracks lexicalization, not grammatical typology. Test this with mBERT or matched monolingual BERTs:

Control A (Grammar-different, lexicalization-similar):

- Language pair: Japanese–English
- Domain: Basic color terms (both lexicalize 11 basic terms similarly)
- Models: mBERT, or Japanese BERT (cl-tohoku/bert-base-japanese) vs bert-base-uncased
- Prediction: Geometric *similarity* despite radical syntactic differences (SOV vs SVO, etc.)

Control B (Grammar-similar, lexicalization-different):

- Language pair: Spanish–English
- Domain: Extended kinship, meal-related concepts (*sobremesa*), emotional granularity
- Models: mBERT, or BETO (dccuchile/bert-base-spanish-wwm-uncased) vs bert-base-uncased
- Prediction: Geometric *divergence* despite near-identical syntax

If Control A fails (grammar-different pairs show geometric divergence in lexicalization-matched domains), grammar matters more than claimed.

If Control B fails (grammar-similar pairs show geometric similarity in lexicalization-different domains), the effect is not lexicalization-driven.

A.10.3 Frequency Matching Control

Corpus frequency affects BERT embedding geometry—frequent words have more stable, better-separated representations. To control:

- Estimate target-domain term frequencies in training corpora (Wikipedia word counts are publicly available)
- For pretrained models: weight distance measurements inversely by frequency difference
- For trained models: continue pretraining on frequency-matched samples
- Prediction: Geometric divergence persists after frequency correction

If frequency matching eliminates the effect, the original result was confounded.

A.10.4 Paraphrase Ensemble Control

Surface form affects embedding independent of meaning. To control:

- For each meaning-anchor, generate 10 paraphrases per language
- Compute all measures for each paraphrase
- Report distribution; test whether cross-linguistic difference survives averaging
- Prediction: Effect survives paraphrase averaging

If paraphrase averaging eliminates the effect, it's surface-form-driven, not conceptual.

A.11 Statistical Framework

A.11.1 Multiple Comparisons

Testing many language pairs × domains × measures inflates false positive rate. Corrections:

- Pre-register primary hypotheses (kinship $\times$ 3 language pairs $\times$ 4 measures = 12 tests)
- Apply Benjamini-Hochberg FDR correction at $q = 0.05$
- Report both corrected and uncorrected p-values
- Exploratory analyses clearly labeled

A.11.2 Effect Sizes

Statistical significance is insufficient. Report:

- Cohen's d for pairwise comparisons
- R^2 for correlation predictions
- Cross-linguistic divergence Δ in interpretable units (e.g., "distance in Tamil is 1.4× distance in English")

A.11.3 Replication

- For pretrained models: replicate across context variations (10+ sentence frames per concept)
- For fine-tuned models: train 5 models per condition (different random seeds)
- Report mean and standard error across runs/contexts
- Effect must exceed between-run variance
- Cross-architecture replication: confirm key findings hold for both BERT and GPT-2 family models
- Pre-register replication plan for key findings

A.12 Falsification Conditions

The theory makes specific predictions. It is falsified if:

1. **Parallel corpus convergence:** $\Delta^{\text{par}} \ll \Delta^{\text{mono}}$. If content-controlled models converge, divergence was corpus-driven, not language-driven.

2. **Intervention failure:** $M_E^{+\text{lex}}$ shows no geometric change, or control interventions show equal change. If adding lexicalization doesn't change geometry, the causal claim fails.

3. **Grammar-lexicalization dissociation fails:** Grammar-different/lexicalization-similar pairs show divergence; grammar-similar/lexicalization-different pairs show similarity. If so, Section 6's claim is wrong.

4. **Frequency control eliminates effect:** Frequency-matched corpora show no divergence. If so, the effect is distributional, not structural.

5. **Behavioral prediction fails:** Model geometry does not predict human RT, order effects, or translation difficulty. If so, either the bridge claim fails or the measures are invalid.

6. **Cross-linguistic anchor quasi-isometry fails:** Anchor-set distances are not preserved even approximately across languages. If so, the manifolds are not comparable and the entire framework is inapplicable.

Any of these outcomes would constitute serious evidence against the theory as stated. The framework is empirically vulnerable—and therefore scientifically meaningful.

A.13 Summary: The Minimal Convincing Package

A compelling positive result requires:

1. **Anchor validation:** Quasi-isometry holds on meaning-anchors (the bridge is valid)

2. **Residual concentration:** After optimal alignment, residuals concentrate at lexically-predicted boundaries

3. **Parallel corpus survival:** Divergence persists under content control

4. **Intervention success:** Adding lexicalization produces predicted geometric changes; controls do not

5. **Control battery passed:** Grammar vs. lexicalization, frequency matching, paraphrase ensemble controls all support the lexicalization hypothesis

6. **Multi-measure coherence:** Geodesic distance, curvature, dimensionality, energy barriers all point the same direction

7. **Behavioral prediction:** Model geometry predicts human behavioral measures

This package would establish, beyond reasonable methodological doubt, that languages induce non-isometric representational geometries in ways predicted by lexicalization patterns—the strong Sapir–Whorf hypothesis stated in geometric terms.

A.14 Recommended Execution Sequence

Given the BERT-based approach, we recommend the following sequence, ordered by effort and evidential value:

Phase 1: Existence proof (no training required)

1. Download mBERT and monolingual BERTs (English, Spanish, Tamil)
2. Define 20 kinship anchors with expressions in each language
3. Extract embeddings (layer 9, mean-pooled, 10 contexts each)
4. Compute distance matrices; test for lexical boundary effects
5. If positive: the phenomenon exists. Proceed to Phase 2.

Phase 2: Causal test (minimal training)

1. Fine-tune bert-base-uncased with "avuncle" intervention
2. Fine-tune control conditions (nonce word, synonym, noise)
3. Measure geometric change in kinship region
4. If intervention > controls: causality established. Proceed to Phase 3.

Phase 3: Robustness

1. Grammar vs. lexicalization controls (Japanese-English color, Spanish-English kinship)
2. Frequency matching analysis
3. Paraphrase ensemble averaging
4. Cross-architecture replication with GPT-2
5. If controls pass: robust effect. Proceed to Phase 4.

Phase 4: Full validation

1. Train BERT on parallel corpora (Level 2 control)
2. Behavioral prediction studies with human participants
3. Extended domains (color, spatial, evidentiality)
4. Publication-ready results

The framework is designed so that negative results at any phase are informative: they either falsify specific claims or identify measurement issues, guiding refinement.

References

Linguistic Relativity

Whorf, B. L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* MIT Press.

Sapir, E. (1929). The status of linguistics as a science. *Language*, 5(4), 207–214.

Lucy, J. A. (1992). *Language Diversity and Thought: A Reformulation of the Linguistic Relativity Hypothesis.* Cambridge University Press.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. Gumperz & S. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp. 70–96). Cambridge University Press.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1–22.

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity.* Cambridge University Press.

Typology and Cross-Linguistic Semantics

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution.* University of California Press.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15), 9085–9089.

Murdock, G. P. (1949). *Social Structure.* Macmillan.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *The World Atlas of Language Structures Online.* Max Planck Institute for Evolutionary Anthropology. https://wals.info

Information Geometry

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37, 81–91.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics.* Springer.

Amari, S., & Nagaoka, H. (2000). *Methods of Information Geometry.* American Mathematical Society.

Geometric Semantics

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought.* MIT Press.

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces.* MIT Press.

Compression and Efficient Communication

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.

Grünwald, P. D. (2007). *The Minimum Description Length Principle.* MIT Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407.

Semantic Memory and Lexical Access

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407–428.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition* (pp. 264–336). Erlbaum.

Transformer Models

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report.*

Probing and Representational Analysis

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? *Proceedings of ACL*, 3651–3657.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *Proceedings of ACL*, 4593–4601.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7, 49–72.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *Proceedings of NAACL-HLT*, 4129–4138.

Reif, E., Yuan, A., Wattenberg, M., Viégas, F. B., Coenen, A., Pearce, A., & Kim, B. (2019). Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Language-Brain Alignment

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134.

Goldstein, A., Zada, Z., Buchnik, E., Sber, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Lora, F., Flinker, A., Devore, S., Doyle, W., Dugan, P., Friedman, D., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25, 369–380.

Multilingual Models

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of ACL*, 4996–5001.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*, 8440–8451.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3), 559–601.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. *Proceedings of ICLR*.

Parallel Corpora

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations parallel corpus v1.0. *Proceedings of LREC*, 3530–3534.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of MT Summit X*, 79–86.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of LREC*, 2214–2218.