Meaning-infused grammar: Gradient Acceptability Shapes the Geometric Representations of Constructions in LLMs

Supantho Rakshit
Dept of ECE / Princeton University
r.supantho@princeton.edu

Adele E. Goldberg
Dept of Psychology / Princeton University
adele@princeton.edu

Abstract

The usage-based constructionist (UCx) approach to language posits that language comprises a network of learned form-meaning pairings (*constructions*) whose use is largely determined by their meanings or functions, requiring them to be graded and probabilistic. This study investigates whether the internal representations in Large Language Models (LLMs) reflect the proposed function-infused gradience. We analyze representations of the English Double Object (DO) and Prepositional Object (PO) constructions in Pythia-1.4B, using a dataset of 5000 sentence pairs systematically varied by human-rated preference strength for DO or PO. Geometric analyses show that the separability between the two constructions' representations, as measured by energy distance or Jensen-Shannon divergence, is systematically modulated by gradient preference strength, which depends on lexical and functional properties of sentences. That is, more prototypical exemplars of each construction occupy more distinct regions in activation space, compared to sentences that could have equally well have occured in either construction. These results provide evidence that LLMs learn rich, meaning-infused, graded representations of constructions and offer support for geometric measures for representations in LLMs.

1 Introduction

A central tenet of usage-based constructionist (UCx) approaches is that our knowledge of language consists of a structured inventory of constructions — conventionalized pairings of form *and* function at varying levels of complexity and abstraction (Goldberg, 2006). The framework posits that language is learned from experience, with contexts and frequencies of use shaping dynamic "ConstructionNets" (Goldberg, 2024) in the minds of speakers. Grammaticality is not a binary state but a continuum of acceptability, an observation supported by work in experimental and computational work on language (Francis, 2022; Gibson and Fedorenko, 2013; Hu et al., 2024).

Here we focus two English constructions, the Double Object (DO) construction (e.g., *She gave the boy the book*) and the Prepositional Object (PO) alternative (*She gave the book to the boy*). We build on a long-standing and widespread focus in linguistics on the combination of information structure and lexical factors that speakers use to choose between the DO and PO constructions: (e.g., Bresnan et al., 2007; Goldberg, 1995; Green, 1974; Levin, 1993; Oehrle, 1976; Wasow and Arnold, 2003). In particular, the recipient argument in the DO strongly tends to be already under discussion and expressed by a definite word (often a pronoun) or short phrase; the transferred entity in a DO, on the other hand, is within the focus domain and is more often expressed by an indefinite noun phrase, which can be a longer phrase. These information structure properties partially emerge from the fact that the DO construction is used to convey real or metaphorical transfer to an animate entity, and animate entities are more likely to be topical in discourse (people often talk about people), while the transfered entity is more likely to be in the focus domain (for a degree of dialect variation see Bresnan and Nikitina, 2009).

The PO construction has been argued to be a subcase of a much broader "caused-motion" construction (Goldberg, 1995, 2002) that can convey a change of location as well as transfer of possession (e.g., *She kicked the ball to him/the wall*). This idea is supported by recent computational work offering a tool for analyzing word meanings in different contexts (Ranganathan et al.,

2025) using interpretable semantic features (Chronis et al., 2023). Ranganathan et al. (2025) report that the features associated with word embeddings vary systematically, depending on whether a given word appears in the DO or PO. In particular, features related to personhood are stronger when the same word, e.g., *London* is the recipient of the DO (*e.g., She sent London the painting*), while features related to location are stronger when *London* appears in the PO (e.g., *She sent the painting to London*).

Verbs' lexical biases also play a role in whether people prefer the DO or PO. The verb *give* is more common in the DO construction, and in fact *give* accounts for roughly 40% of all DO tokens (e.g., Goldberg et al., 2004). On the other hand, a set of Latinate (i.e., fancy-sounding) verbs resist the DO in favor of the PO (Gropen et al., 1989; Ambridge et al., 2012; Goldberg, 2019). For instance, the verbs *transfer*, *explain,* and *donate* rarely occur in the DO, despite their highly compatible meanings; instead, each verb is biased toward the PO construction. Lexical biases can be quite particular and specific; for instance, the Latinate verb *guarantee*, bucks the tendency for fancy-sounding words to resist appearing in the DO: *Guarantee* strongly prefers the DO. Thus, a nuanced account of how such lexical factors are learned is required (e.g., Goldberg, 2011, 2019; Ambridge et al., 2012). Indeed, computational work has found that the differences in information structure between the DO and PO are useful in LLMs' learning of lexical biases (Misra and Kim, 2024).

As LLM representations are learned through exposure to natural language texts, there is an opportunity to investigate whether massive distributional learning can give rise to representations that reflect principles of the UCx approach. Recent work has assessed how accurately LLMs can classify or distinguish argument structure constructions (e.g. Huang, 2025; Bonial and Tayyar Madabushi, 2024), but less is known about *how* constructions are represented or their underlying geometry. Our work addresses this gap by shifting the focus from classification accuracy to an analysis of underlying representational geometry.

We hypothesize that representations of the DO and PO constructions should be more distinct to the extent that instances' typical lexical and functional properties are more prototypical instances of the respective constructions. We test this by asking whether collections of instances of the DO and PO that include typical functional features are more easily separable than collections of instances that are prototypical of neither, even though each sentence is unambiguous syntactically (either a PO or DO).

More specifically, we ask: Does the geometric distinction between the representations of the DO and PO increase, as measured by either energy distance or Jensen-Shannon Divergence, as the functional factors associated with DO and PO more closely align with their respective syntactic expressions?

Stimuli sentences come from the DAIS (Dative Alternation and Information Structure) dataset, which includes $5,000$ English pairs of DO and PO sentences (Hawkins et al., 2020). Across DO/PO pairs, several factors are systematically varied along the dimensions recognized to distinguish the two constructions. Here we use human preference strengths, also from DAIS, toward one or the other construction, to analyze the hidden states of Pythia-1.4B (Biderman et al., 2023).

Both energy distance (Rizzo and Székely, 2016) and Jensen-Shannon divergence (JSD) (Fuglede and Topsoe, 2004) are used to measure the separability of entire clouds of representations, at different layers of Pythia-1.4B. The preferred version (DO, PO, or either) of each sentence pair in the DAIS corpus was binned, according to the degree of preference toward the DO or PO, to be described in Methods and Results.

Results reveal a sophisticated geometric encoding of constructions in which lexical and functional factors improve the distinction between the DO and PO. In this way, LLM representations are consistent with key principles of the UCx approach, including gradiently distinguishable function-infused grammatical patterns, interpretable as clusters in geometric space.

2 Methods

Dataset and Model: As noted, the DAIS dataset includes 5000 pairs of sentences, one in the DO and one in the PO, while systematically varying the length and definiteness each postverbal argument across pairs. Two hundred main verbs also vary across pairs, including verbs standardly treated as both 'alternating' and 'non-alternating' (Levin, 1993). Importantly, DAIS also includes

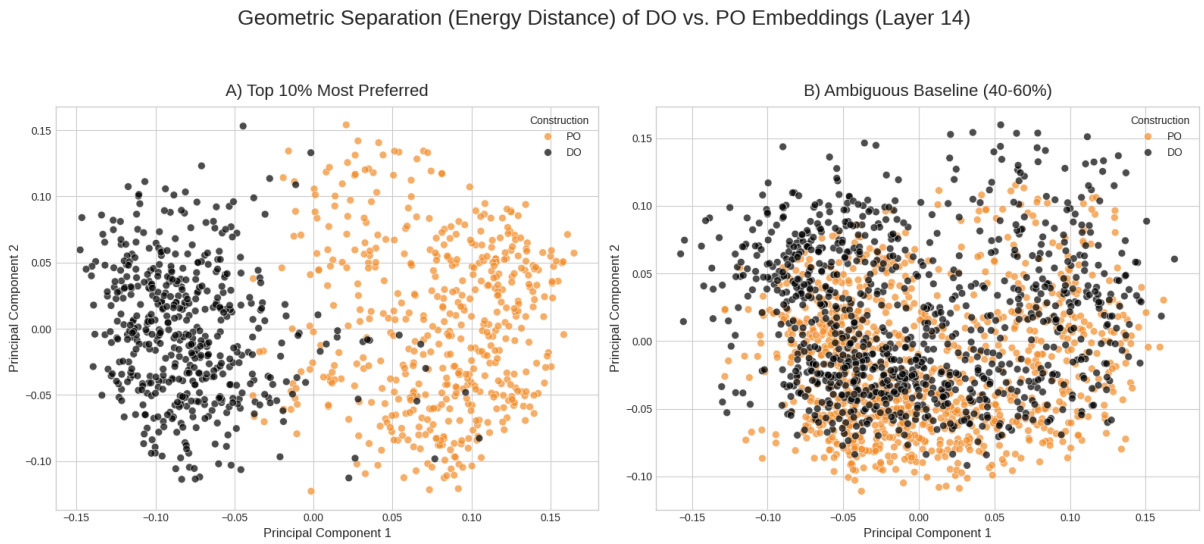

Figure 1: Projection into 2-dimensions of mean-pooled and normalized representations of the Double-Object (DO, in orange) and Prepositional Dative Object (PO, in black) constructions. Points represent sentences from the DAIS corpus, binned as follows: A) Instances of the DO and PO that are well-suited to the lexical and functional factors of their respective constructions as determined by human preferences. B) Instances of the DO and PO, as determined by syntax only, as their lexical and functional properties do not favor either construction.

human ratings of how strongly they prefer one construction over the other, for each combination of verb and arguments. Participants used a slider to indicate a preference for the DO (one end) or PO (other end) or neither (midpoint) (Hawkins et al., 2020).

We use these human preference ratings to partition sentences into five tiers based on the mean preference strength. We combined sentences from both ends of the scale to create 5 bins, ranging from: (1) the top 10% of sentences with the strongest preference for one or the other construction, to (5) those sentences judged to be in the middle of the scale (equally non-biased toward either construction). A sample collection of DO sentences that vary from strong DO-bias (1) to little bias toward either DO or PO (5) are provided below:

DO biased
ˆ (1) Maria asked him some questions.
| (2) Bob lobbed her a tennis ball.
| (3) Juan shuttled the team something.
| (4) Alice threw a woman a book.
| (5) Michael took the woman the blanket.
DO or PO

An equal number of PO sentences were included in each of the same bins, correspondingly ranging from strongly PO-biased (1) to equi-biased (5), according to the human preferences in DAIS.

From the publicly available pretrained Pythia-1.4B model, we extracted mean-pooled and normalized state representations for each sentence. We analyzed representations from all 24 layers, reducing them to 150 principal components, which captured 88.01% of the total variance (averaged across model layers). Common benchmarks suggest retaining components that explain 70%–90% of the total variance (Jolliffe, 2011), and 88.01% sits comfortably within this range, suggesting that a large majority of the structure in the data is retained, while a smaller portion that is more likely to reflect noise was discarded. We normalized the activations so that they all exist on a unit hypersphere S^{149}. Finally, we deployed the following analyses.

Preference strength was treated as an ordinal variable with five levels: 1 (10% most strongly biased) to 5 (10% most equi-biased). That is, level (1) includes sentences that strongly preferred the DO and sentences that strongly preferred the PO, while level (5) included sentences that were roughly equi-biased toward either DO or PO. Bins were used rather than a continuous factor for visualization purposes.

To measure the separability in representational space for each tier of bias strength, we em-

ployed two different measures, Energy Distance and Jensen-Shannon divergence. **Energy distance**, $\mathcal{E}(X, Y)$, is a statistical distance between the probability distributions of two random vectors, X and Y, in a metric space (Cramér, 1928). It is simply based on the expected Euclidean distances between their elements. Given two samples from our PCA-reduced representations, $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$, where each $x_i, y_j \in \mathbb{R}^{150}$, the squared energy distance is estimated as:

$$\mathcal{E}^2(X,Y) = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|x_i - y_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i - x_j\| \quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|y_i - y_j\|$$

where $\|\cdot\|$ is the Euclidean norm. Energy distance is zero if and only if the distributions are identical; it is sensitive to differences in both the location and the shape of distributions, making it a robust measure of overall geometric separation in the model's representation space. We calculated the energy distance between the distributions of the constructions on S^{149} layer by layer.

A more sensitive measure of distributions is **Jensen-Shannon Divergence (JSD)**, which measures the relationship between distributions in high-dimensional space (Menéndez et al., 1997). To use JSD, we first estimated the probability distributions of both constructions $P(v_{DO})$ and $Q(v_{PO})$, in the 150-dimensional PCA space. Following (Conklin, 2025), we next generated a set of $k = 1000$ *anchor vectors*, $A = a_1, a_2, \ldots, a_k$, by sampling from a uniform distribution on S^{149}. The vector corresponding to each individual sentence v was then assigned to the anchor vector nearest to it, based on cosine similarity. This effectively partitions the hypersphere into k Voronoi cells. This technique, based on vector quantization, yields two discrete probability distributions, $\hat{P}$ and $\hat{Q}$, which are k-dimensional vectors where the i-th element represents the proportion of vectors from each set assigned to anchor a_i. Jensen-Shannon divergence is then computed as:

$$JSD(\hat{P}\|\hat{Q}) = \frac{1}{2}D_{KL}(\hat{P}\|M) + \frac{1}{2}D_{KL}(\hat{Q}\|M)$$

where $M = \frac{1}{2}(\hat{P} + \hat{Q})$ and D_{KL} is the Kullback-Leibler divergence. This method avoids information loss from projecting onto any single axis to offer a more holistic comparison of distributions (Conklin, 2025). Since the sampled anchor vectors are probabilistic, we averaged across 20 random seeds to get stable JSD scores.

3 Results

Our analysis reveals that graded bias strength for one construction over a paraphrase systematically shapes the geometry of construction representations across the model architectures. In particular, the model assigns representations that are more distinct when the constructions are more clearly differentiated, when instances are more strongly biased toward the construction used. This is the case for both energy distance and JSD, as each shows a clear and consistent stratification by the tiers of preference strength (Figure 2 and 3). At nearly every layer, the Top 10% strongest preference tier exhibits the greatest geometric distance, followed in order by the other tiers, down to the ambiguous baseline. As is clear in Figure 1, with sentence vectors projected onto two dimensions for visualization purposes, DO and PO sentences that conform better to the DO or the PO, respectively (left panel), are more distinctive than sentences that could nearly as easily be paraphrased by the other construction (right panel).

Because energy distance and JSD are based on very different analyses, we cannot expect their qualitative patterning to align. In fact, energy distance follows a convex trajectory, slightly dipping in the mid-layers before rising sharply (Figure 2) In contrast, JSD shows divergence increasing sharply and remaining high (Figure 3). Yet the overall pattern showing more distinctiveness between DO vs. PO sentences when sentences that are more biased toward either varient compared with sentences that are relaitvely un-biased is evident according to both energy distance (Figure 3) and JSD (Figure 2). We take this to indicate that the finding is robust and not an artifact of a single metric.

Scaling Analysis across Pythia Model Suite. To test whether our findings are specific to the 1.4B parameter model or reflect a more general property of transformer architectures, we replicated our full analysis pipeline — including the energy distance, high-dimensional JSD with correlation of mean cosine similarity tests — across the models in the Pythia suite (from 70M to 6.9B parame-

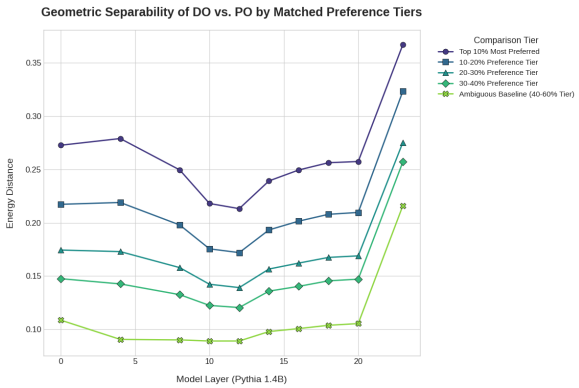

Figure 2: Layerwise Energy Distance between DO and PO representations, stratified by tiers binned by degree of bias. The plot shows a consistent ordering by bias, with more prototypical instances of the two constructions being more separable.

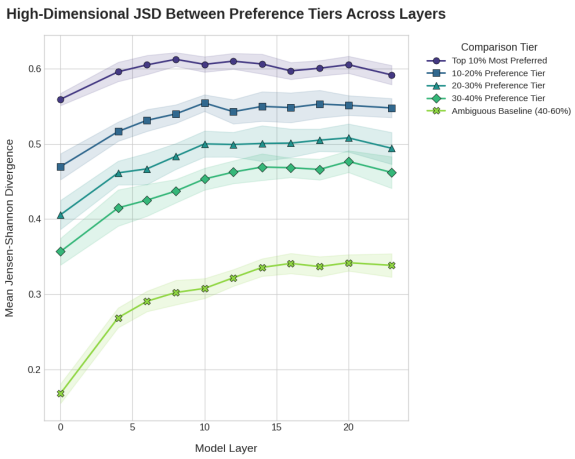

Figure 3: A high-dimensional measure of Distributional Separability (JSD) Across layers (with $k = 1000$ anchor points, averaged over 20 random seeds). Stratification is evident: tiers that include more biased sentences of either DO or PO are more separable into distinct constructions. This plot reinforces the finding from the energy distance analysis, though with different layer-wise dynamics.

ters). Results confirm that our central findings are robust across model scales. We consistently observe the geometric stratification by degree of bias so that preference strength remains a significant predictor of representational distance. A detailed report of these scaling law analyses, with code to generate plots, is in the supplementary materials.

4 Discussion

Our results provide compelling computational evidence for a core principle of the usage-based constructionist approach: that grammatical representations are graded and sensitive to semantic-pragmatic fit. The clear stratification in our geometric analyses demonstrates that LLMs develop representations whose geometric properties are highly consistent with the probabilistic, usage-based categories posited by the UCx approach. This geometric entanglement of form and function resonates with the core tenets of the UCx approach. The energy distance reflects the distinctiveness of the two constructions spatially in regions of the model's representation space. This extends previous work that has focused on the model's ability to classify constructions categorically (Huang, 2025; Bonial and Tayyar Madabushi, 2024) by showing more fine-grained, graded geometric structure, dependent on lexical and functional factors.

Future work is needed to better understand the distinct qualitative patterns across layers when energy distance and JSD are compared. We note that the JSD measure, which is more nuanced but perhaps less intuitive, appears to distinguish the constructions particularly well: JSD is bounded between 0 and log(2) (≈ 0.693) (Lin, 1991), and the distinction between the constructions in the most biased tier approaches this limit at 0.6. Yet because the two metrics are based on quite different calculations, so we do not attempt to compare them directly.

5 Conclusion and future directions

The current work demonstrates that the geometry of an LLM's internal representations directly reflects the graded function-infused bias toward one or another linguistic construction, where biases are recognized to be conditioned on lexical and functional factors. We have shown that the model's representations of constructions are systematically organized by their distinctiveness. This work bridges the gap between the theoretical principles of the usage-based constructionist approach and the empirical realities of modern NLP, suggesting that LLMs learn a rich, dynamic, and meaning-infused model of grammar. Our findings open a promising new direction for future work; we are currently using these geometric insights to guide an investigation aimed at isolating the specific computational circuit(s) within the model that are responsible for encoding verb bias in the dative alternation, using tools like causal mediation analysis from Mechanistic Interpretability.

References

Ben Ambridge, Julian M Pine, Caroline F Rowland, and Franklin Chang. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, 88(1):45–81.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2217–2256. PMLR.

Claire Bonial and Harish Tayyar Madabushi. 2024. Constructing understanding: on the constructional information encoded in large language models. *Language Resources and Evaluation*, pages 1–40.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In Gosse Bouma, Irene Kraemer, and Joost Zwarts, editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Arts and Sciences, Amsterdam.

Joan Bresnan and Tatiana Nikitina. 2009. The gradience of the dative alternation. *Reality exploration and discovery: Pattern interaction in language and life*, pages 161–184.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. *arXiv preprint arXiv:2305.18598*.

Henry Conklin. 2025. Information structure in mappings: An approach to learning, representation, and generalisation. *arXiv preprint arXiv:2505.23960*.

Harald Cramér. 1928. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74.

Elaine Francis. 2022. *Gradient acceptability and linguistic theory*, volume 11. Oxford University Press.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Adele E Goldberg. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Adele E Goldberg. 2002. Surface generalizations: An alternative to alternations.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, New York.

Adele E Goldberg. 2011. Corpus evidence of the viability of statistical preemption.

Adele E Goldberg. 2019. *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Adele E Goldberg. 2024. Usage-based constructionist approaches and large language models. *Constructions and Frames*, 16(2):220–254.

Adele E Goldberg, Devin M Casenhiser, and Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive linguistics*, 15(3):289–316.

Georgia M Green. 1974. Semantics and syntactic regularity. bloomington. *(No Title)*.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, pages 203–257.

Robert D. Hawkins, Ngan Nguyen, Adele E. Goldberg, Michael C. Frank, and Noah D. Goodman. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4718, Online. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Haerim Huang. 2025. Assessing the performance of pretrained models for accurate and consistent classification of argument structure constructions. In *Applied Artificial Intelligence*.

I. Jolliffe. 2011. Principal component analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, Berlin, Heidelberg.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Richard Thomas Oehrle. 1976. *The grammatical status of the English dative alternation*. Ph.D. thesis, Mass. Cambridge.

J. Ranganathan, R. Jha, K. Misra, and K. Mahowald. 2025. semantic-features: A user-friendly tool for studying contextual word embeddings in interpretable semantic spaces. ArXiv preprint arXiv:2506.XXXXX.

Maria L Rizzo and Gábor J Székely. 2016. Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38.

Thomas Wasow and Jennifer Arnold. 2003. Postverbal constituent ordering in english. *Determinants of grammatical variation in English*, pages 119–154.