

Language-Specific Loop Structure in mBERT Attention Graphs

Anna Brežgis

annabrezgis@brandeis.edu

brezgis.com

1 Introduction

Unlike other continuous physical phenomena like temperature or speed, there exists no system of precise measurement for color in the common vernacular. While any layman can easily imagine what 82 degrees Fahrenheit or 25 miles per hour means, to accurately describe color, one must approximate. In other words, language in use employs a discrete category system for color. Consider the affectual differences between *blue*, *cerulean* and *powder blue*. In order to achieve more precision, we either reach for tail-end color words or resort to metaphor. While a color word can express information about the real-world color state of an object, it can also give information about a shared cultural or social analogy, as in *green with envy* or *Picasso's blue period*. It is then natural to suppose that the distributional environment of a color word must be somewhat distinct from that of other color words.

If concepts can be modeled geometrically, such as subspaces in some latent representation, then the question becomes how to access and compare those subspaces. Transformer architectures offer one possible lens. They build nuanced latent representations of conceptual relationships, and the attention mechanism is fundamentally *relational*: each head encodes how a token relates to every other token in context. If a category boundary is salient enough in the language, this might be reflected in the attentional structure surrounding the category's lexical anchors.

This project asks three related questions:

1. Does **distributional evidence** reveal systematic and distinct category boundaries cross-linguistically?
2. Is this structure present in **transformer attention**?
3. Can we detect it with **topological data analysis** (TDA)?

2 Incidentals

- **Task:** Cross-lingual comparative analysis of the topological features of attention signatures via persistent homology
- **Track:** Research track – exploratory methodology
- **Motivation:** My interest in this topic is motivated by a desire to study cultural values and worldviews in language *quantitatively and at scale*.

3 Related Work

The cultural specificity of color is roughly the idea that has motivated a hearty spread of color-categorization studies within psycholinguistics. The most foundational of these is the work by [Berlin and Kay \(1969\)](#), whose cross-linguistic survey of basic color terminology established a now-canonical scalar schema. In their model, languages range from two basic color terms (light/dark – alternatively, white/black) up to eleven (the upper bound of most Indo-European languages, including English), with categories appearing to be added in a near-universal implicational order. This is a landmark in cognitive linguistics and is still the gold standard today. In this paper, it is the main theoretical foundation. What they refer to as *basic color terms*, for clarity I will refer to as *canonical terms*.

Also famously, the blue-shade discrimination task conducted by [Winawer et al. \(2007\)](#) demonstrated an interesting caveat to the [Berlin and Kay \(1969\)](#) schema. Russian goes beyond the eleven-color bound with an obligatory twelfth term, a split between dark blue (*sinij*) and light blue (*goluboy*). When participants were asked to discriminate between shades that approached the *sinij/goluboy* boundary, Russian speakers were consistently faster than their English-speaking coun-

terparts. But when linguistic interference was introduced (participants were asked to hold a string of integers in working memory while completing the task), the Russophone advantage was essentially eliminated, and even negative in some cases. This result led researchers to connect the perceptual advantage to the category distinction native to Russian language. The paper has become something of a "cult classic", referenced in (Boroditsky, 2017)’s popular TED talk, absorbed into common "fun fact" knowledge, and its central claim anchors the present project. If a language obligates specific perceptual boundaries, that distinction should, in principle, be reflected in distributional patterns and therefore in latent model representations. Lillo et al. (2018) confirm that Castilian Spanish, like English, makes do with the standard eleven, providing a useful third anchor for the comparison undertaken here.

3.1 TDA in NLP

Topology as a field is concerned with the global structure of a space, rather than its local or Euclidean properties. Where geometry is defined by distance, angles, and curvature, topology concerns which features of an object survive continuous deformation. Two spaces are considered topologically equivalent if one can be stretched or bent into the other without ripping or gluing. Take the prototypical example of a donut and a coffee cup as the same topological object. Both have a single connected component ($B_0 = 1$) and a single one-dimensional hole ($B_1 = 1$).

Topological data analysis (TDA) as a field has its origins in Carlsson (2009), who proposed that high-dimensional data could be treated as a point cloud or graph upon which persistent homology (more on that soon!) could be computed. This methodology has only recently become tractable at scale. TDA is computationally expensive and benefits from large volumes of data. Additionally, it fits within a broader trend of geometrically-oriented methodologies in machine learning.

Within **natural language processing (NLP)** specifically, Uchendu and Le (2025)’s survey shows TDA as a niche but rapidly growing area, especially within interpretability and AI-detection research. Like many other revelations in NLP, the techniques originated in the wider ML community (specifically image processing and classification) before being adapted to language processing.

Two recent papers applying TDA in NLP an-

chor the methodological design of this project. Draganov and Skiena (2024) inspire the language-specific-features bend. They compute persistent homology on language-specific vector spaces (Fast-Text) and show that the resulting persistence diagrams and topological features can be reliably used to predict language phylogenies. From Kushnareva et al. (2021) I borrow from more directly – the project code is based on their feature extraction pipeline. Their work was to compute persistent homology on transformer (BERT) attention patterns, treating the attention matrix as an undirected graph. The resulting features were shown to distinguish human from machine-generated text reliably, outperforming both count-based and neural methods by over ten percent, a strong indicator of the discriminatory power of global structure. At the time, the frontier of text generation was Open AI’s GPT-2 which dates the paper without diminishing its methodological contribution.

4 Data

*white, black, red, green, yellow, blue,
brown, purple, pink, orange, grey*

Table 1: Berlin and Kay’s eleven basic color terms

The set of canon terms is grounded in the existing literature above: Berlin and Kay (1969)’s eleven basic color terms, Lillo et al. (2018)’s confirmation of the same eleven for Castilian Spanish, and Winawer et al. (2007)’s plus one for Russian (the *sinij/goluboy* split). For each of the Table 1 in each language, I draw KWIC (Key Word In Context) lines from the Leipzig Corpora Collection, specifically the 1M-sentence slices from 2019, 2020, and 2023. I chose Leipzig because it freely provides parallel multilingual corpora. This is an advantage in that content is matched across languages but also a limitation since Leipzig is gathered primarily from news sources, a genre that may not capture the richness or breadth of color terms in actual language use. About 200 KWIC lines per canon term per language, summed over three languages, yields 6,685 attention signatures.

Attention signatures are gathered through processing with mBERT. This is a clear divergence from Kushnareva et al. (2021), who use BERT-base. Because all three languages’ data must be

mapped onto the same representational space, this is a deliberate and indispensable choice.

5 Methodology

I reproduce the feature-extraction pipeline of Kushnareva et al. (2021) with a few small modifications. Several of the original pipeline’s dependencies (numpy, ripser++ and others) had deprecated APIs that needed to be replaced. Finally, I swapped the standard BERT tokenizer for BertTokenizerFast for speed. I process each KWIC line using this pipeline, truncated to 64 tokens.

Two extraction details matter for the comparison. The Russian lemmatizer (pymorphy3) does not strip punctuation, so without stripping it first, a Russian color term touching a comma, period or quote would never be matched, a systematic difference in context between the languages (10.5% of English and 28.6% of Spanish target tokens carry punctuation). And a 32-token cap would truncate 64% of Russian contexts, so the cap is 64 tokens.

The pipeline returns *persistence diagrams* which correspond to the topological features of each (language, color) set. To explain persistent homology briefly: recall the notion of topology as the study of shapes whose properties are invariant under continuous transformation. A persistence diagram captures topological features as a set of (birth, death) pairs as a scale parameter sweeps from zero to infinity. Features that *persist* across many scales are read as structural, while features that appear and disappear quickly are read as noise.

My baseline analysis is hierarchical clustering on language-specific FastText 300-dimensional embeddings, with cosine distance between color-term vectors. This is admittedly a weak comparison in that there are vast differences between static and contextual embeddings, in architecture and in training material. Linear structure does not capture nuanced spatial relationships like hierarchy, loops, or voids. An exact baseline would in fact be difficult to define, given the relative novelty of the methodology and the specific shape of the research question.

5.1 Comparative analysis

The central question is: how *different* or *distinct* are the topological representations of one language’s color space from another’s? To answer this, I need a measure of distance between persistence diagrams.

The two standard measures are Wasserstein and bottleneck distance, both computed pairwise across all diagrams. Wasserstein distance can be intuitively understood as a measure of how much “work” it takes to transform one persistence diagram into another, or more precisely the p -th root of the summation of differences across (birth, death) pairs after optimal matching. Bottleneck distance measures the greatest overall distance of any (birth, death) from one persistence diagram to another, in other words, the most-mismatched pair of the whole set.

Concretely, the p -Wasserstein distance between two persistence diagrams D_1 and D_2 is

$$W_p(D_1, D_2) = \inf_{\gamma: D_1 \rightarrow D_2} \left(\sum_{x \in D_1} \|x - \gamma(x)\|_\infty^p \right)^{1/p},$$

where γ ranges over every possible way of matching points in D_1 to points in D_2 , and where each diagram is first augmented with the diagonal, so any feature without a clean counterpart in the other diagram can pair off with the diagonal cheaply, essentially declaring itself noise. The bottleneck distance is the same equation with the maximum in place of the sum.

I then perform permutation tests at $K = 10,000$ iterations for these distances, asking how likely it is that the feature-label pairings observed in the data would arise by chance under a null distribution in which the language label carries no topological signal. The first is a per-color (or per-domain) permutation test, asking how consistent the topological representations of a specific color are within a language; here, the null distribution itself can be read as a kind of empirical baseline. The second is a per-color-per-language test, asking how consistent the (language, color) representations are within the full set of (language, color) tuples.

5.2 Controls

Two controls ask whether a language effect, if found, is about *color*. **Length.** Even at 64 tokens, Russian contexts are longer than English and Spanish ones, and the size of an attention graph alone changes its topology. I therefore repeat the language test on a subset of contexts matched for length across the three languages (282 of the 680 color contexts in the test sample). **Non-color adjectives.** I run the identical pipeline on Dixon’s core non-color adjectives in the same three languages

(720 contexts; 309 length-matched). If the effect is specific to color, it should be weaker for these.

6 Results

metric	dim	color		adjectives	
		all	matched	all	matched
wasserstein	H_0	79.7	4.5	93.8	4.1
wasserstein	H_1	54.6	34.1	112.2	35.3
bottleneck	H_0	66.7	2.7	71.9	2.6
bottleneck	H_1	68.3	32.8	97.8	35.7

Table 2: Language effect: permutation z ($K = 10,000$) for the head-averaged persistence-diagram distance, on all contexts and on the length-matched subset. Every entry has $p \leq 0.02$; all “all” entries have $p < 10^{-4}$.

metric	dim	within (noise)	cross	$\Delta\%$
wasserstein	H_0	0.8085	0.9479	+17.2%
wasserstein	H_1	0.0828	0.0860	+3.8%
bottleneck	H_0	0.2827	0.2978	+5.4%
bottleneck	H_1	0.0467	0.0478	+2.4%

Table 3: Cross-language distance between diagrams of the same color term, against a within-language noise floor (distance between two random halves of one language’s contexts for that term).

The head-averaged permutation test (Table 2) shows an overwhelmingly large effect size on all contexts ($z = 55\text{--}80$, $p < 10^{-4}$). Read on its own, this is highly statistically significant. Within a given color term (red, blue, green, etc.), there is a clear topological distinction between languages.

The controls change what that distinction means (Figure 1). On length-matched contexts, the H_0 effect (connected components) shrinks to $z = 2.7\text{--}4.5$, small though still significant, while the H_1 effect (loops) holds at $z \approx 33\text{--}34$; head by head, 85 of 144 heads remain significant for H_1 and 41 for H_0 after Benjamini–Hochberg correction (Figure 2). And the non-color adjectives show the same pattern: an even larger effect on all contexts and, once matched for length, effects of the same size as color. The language effect is real, but nothing about it is specific to color.

Tokenization. Matching sentences on word-piece count does not match their tokenization. Even on the length-matched subset, mBERT splits Russian far more finely: 1.8 wordpieces per word against 1.2–1.3 for English and Spanish, 44% continuation pieces against 18–22%, and Russian color terms break into 2.9 pieces against 1.1–1.4. I therefore regress pairwise diagram distance on a

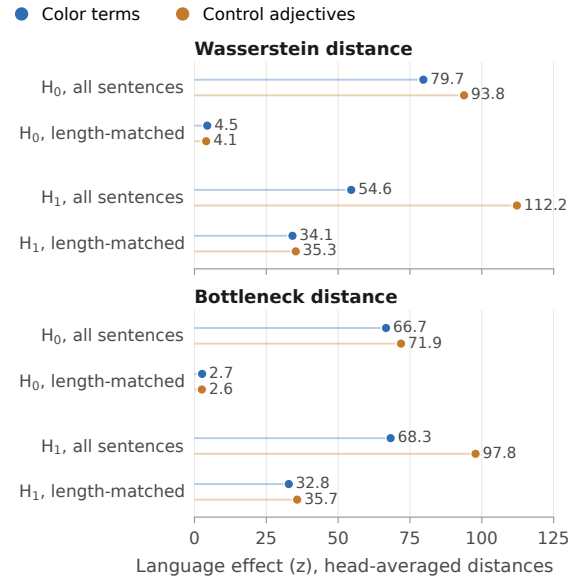


Figure 1: Language effect for color terms and control adjectives. Length-matching removes most of the H_0 effect, the H_1 effect remains, and the control adjectives behave like color terms.

different-language indicator together with a same-concept indicator and the pairwise differences in wordpiece count, fertility, continuation share and target-word pieces, with permutation p -values ($K = 999$) from jointly permuting the rows and columns of the distance matrix. On the full corpus these covariates absorb the language effect (H_0 : standardized β from 0.15 to 0.00, $p = 0.99$). On length-matched contexts the H_0 effect is no longer significant, but the H_1 effect survives ($\beta \approx 0.07\text{--}0.11$, $p = 0.001$), for color and adjectives alike. Finally, a bootstrap over sentences shows no color $\times$ language interaction: the difference in normalized language separation between color and adjectives has a 95% confidence interval spanning zero for every metric and dimension.

But the corresponding slice (Table 3) shows that the actual magnitude of this distinction is small. Diagrams of the same color term from two languages are only 17.2% (Wasserstein, H_0) and 3.8% (H_1) farther apart than two random halves of one language’s sample; for bottleneck distance, 5.4% and 2.4%. The overall result, otherwise stated, is that within a specific color space (again, red, blue, green, etc.) there is a small but statistically significant distinction between languages.

Within the Russian *sinij/goluboy* split specifically, the case the literature most directly predicts, the two blues do not separate at all: the within-Russian permutation test gives $p = 0.57\text{--}0.97$

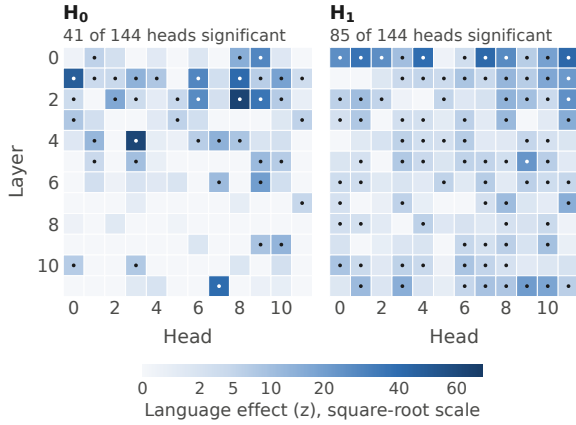


Figure 2: Per-head language effect for color terms on the length-matched subset (Wasserstein distance). Dots mark heads significant after Benjamini-Hochberg correction ($q = 0.05$).

across the four metric and dimension combinations (Figure 3). The translation triples lean the way one would expect if English *blue* lines up with *sinij*: {blue, azul, *sinij*} is somewhat closer than chance ($p = 0.04-0.15$) and {blue, azul, *goluboy*} is not ($p = 0.34-0.98$). But only one of the four tests falls below 0.05, so the asymmetry does not survive correction.

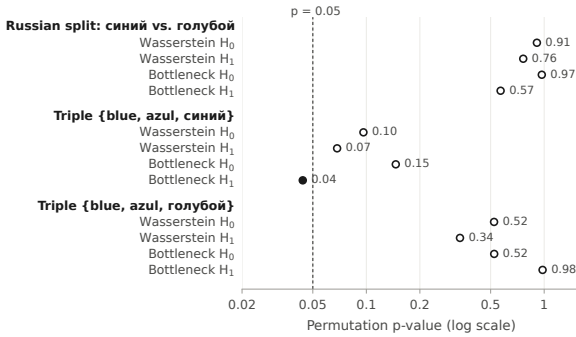


Figure 3: Permutation p -values for the Russian *sinij/goluboy* split and the two blue translation triples. The dashed line marks $p = 0.05$; filled markers fall below it.

6.1 FastText Baseline

The FastText baseline tells a related story. The dendrograms produced by clustering on cosine distance (Figure 4) are interpretable and broadly intuition-matching — some groupings are what one would expect (in English the achromatic *black*, *white*, *gray* and *brown* form their own branch), others less so (*red* sits closest to *blue* in English, as *krasnyj* does to *sinij* in Russian). In English, *black* and *white* merge first; in Russian, *sinij* and *gol-*

uboy fall in different branches. The silhouette plots (Figure 5) tell a quieter story: silhouette scores are modest (best k -means silhouette 0.29 for English at $k = 4$, 0.21 for Spanish at $k = 2$, 0.24 for Russian at $k = 5$; k -means runs on normalized vectors), meaning the clusters themselves are only loosely organized and not cleanly separable.

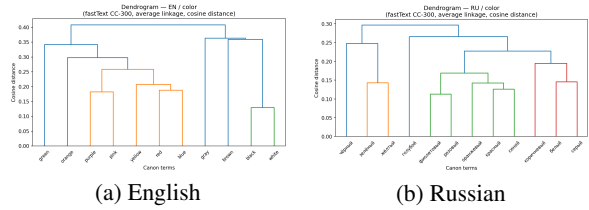


Figure 4: FastText cosine-distance dendrograms over basic color terms.

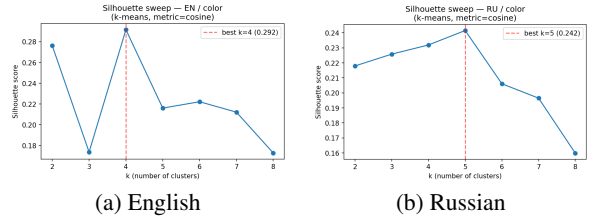


Figure 5: Silhouette plots for the FastText cosine-distance clusters.

7 Analysis and Discussion

The adjective control is the most important result: whatever separates the languages in mBERT’s attention topology, it is not color-specific. A difference that appears just as strongly for *big*, *old* and *good* is about how mBERT processes each language, not about how each language carves up color, and the tokenization analysis shows how much of it is simply how finely mBERT cuts each language into wordpieces. What survives those controls, a language difference in loop structure on matched contexts, is the one open thread: it may reflect syntax or word order, or tokenization effects finer than the summaries used here.

Several explanations are available for the gap between the strong significance of the global test and the small practical effects observed in the *sinij/goluboy* slice. The most salient is that KWIC lines may simply be too noisy a substrate for the kind of highly interpretive, fragile, diffuse, distributed semantic meaning in the domain of color terms. Additionally, two hundred lines per word may also be too few to capture a conceptual bound-

ary cleanly. Scaling up may yield more promising results.

A second concern is whether attention signatures are the correct lens for this problem at all. The idea that “attention modulates cognition and perception” is attractive, and seems natural to extend into the linguistic domain, but it is possible that the representation provided by transformer attention is simply not sufficient or appropriate for picking out something as semantically subtle as a color-category boundary. An approach paralleling Draganov and Skiena (2024), such as computing persistent homology on the last-layer contextual embeddings emitted by the transformer may be more powerful.

TDA for NLP remains a powerful approach and will continue to grow in adoption, especially in the budding area of interpretability research. The semantic domain of color, however, may not be appropriate for its application, and for possibly the reason that has historically made it interesting to cognitive linguists. Its inseparable connection to the physical world and to direct perception may cause more connectivity in multilingual models. Other semantic domains may prove to be more revealing. Emotion terms, culture-specific terms (“key terms” per Wierzbicka (1997)), kinship terms, or profanity. This in turn opens up a more meta-level question: what areas of human language and sociality best distinguish one language group from another?

Code Availability

The project code is available at <https://github.com/brezgis/ph-project>.

AI Use Statement

The code for this project was developed using Claude Code with Claude Opus 4.7 and Claude Opus 5.5. I reviewed generated code manually through GitHub’s pull-request interface before merging. My workflow followed Delfino (2025)’s agent-workflow, which scaffolds agentic code generation around modular task decomposition, gated permissions, and human-in-the-loop review.

References

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

Lera Boroditsky. 2017. How language shapes the way we think. TEDWomen 2017. <https://www.youtube.com/watch?v=RKK7wGAYP6k>.

Gunnar Carlsson. 2009. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

Joe Delfino. 2025. agent-workflow: A workflow for software development with coding agents. GitHub repository. <https://github.com/jdelfino/agent-workflow>.

Ondřej Draganov and Steven Skiena. 2024. The shape of word embeddings: Quantifying non-isometry with topological data analysis. *arXiv preprint arXiv:2404.00500*. <https://arxiv.org/abs/2404.00500>.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649. Association for Computational Linguistics.

Julio Lillo, Fernando González-Perilli, Lilia Prado-León, Anna Melnikova, Leticia Álvaro, José A. Colado, and Humberto Moreira. 2018. Basic color terms (bcts) and categories (bccs) in three dialects of the spanish language: Interaction between cultural and universal factors. *Frontiers in Psychology*, 9:761.

Adaku Uchendu and Thai Le. 2025. Unveiling topological structures from language: A survey of topological data analysis applications in NLP. *arXiv preprint arXiv:2411.10298*. <https://arxiv.org/abs/2411.10298>.

Anna Wierzbicka. 1997. *Understanding Cultures Through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press, New York.

Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.