# CHARACTERIZING HUMAN SEMANTIC NAVIGATION IN CONCEPT PRODUCTION AS TRAJECTORIES IN EMBEDDING SPACE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Semantic representations can be framed as a structured, dynamic knowledge space through which humans navigate to retrieve and manipulate meaning. To investigate how humans traverse this geometry, we introduce a framework that represents concept production as navigation through embedding space. Using different transformer text embedding models, we construct participant-specific semantic trajectories and extract geometric and dynamical metrics—including distance to next, distance to centroid, entropy, velocity, and acceleration. These measures capture both scalar and directional aspects of semantic navigation, providing a computationally grounded view of semantic representation search as movement in a geometric space. We evaluate the framework on four datasets across different languages, spanning different property generation tasks: Neurodegenerative, Swear verbal fluency, Property listing task in Italian, and in German. Across these contexts, our approach distinguishes between clinical groups and concept types, offering a mathematical framework that requires minimal human intervention compared to typical labor-intensive linguistic pre-processing methods. Critically, different embedding models were essentially similar in describing these differences, highlighting similarities between different learned representations despite different training pipelines. By framing semantic navigation as a structured trajectory through embedding space, bridging cognitive modeling with learned representation, thereby establishing a pipeline for quantifying semantic representation dynamics with applications in clinical research, cross-linguistic analysis, and the assessment of artificial cognition. [1]

## 1 INTRODUCTION

Semantic representations are the stored, structured traces of our knowledge about the world (Hills et al., 2015). Retrieving a concept depends on context and draws jointly on experiential details and abstract, shared knowledge—for "dog," this might range from memories of a family pet to generic category knowledge (Xie et al., 2024; Barsalou, 2023). Navigation in semantic representations involves searching within a space that is both dynamic and context-dependent, including features for sensorimotor representations, affective experiences, linguistic encodings, and contextual cues (Hills et al., 2015; Diveica et al., 2025). We adopt the view that semantic retrieval can be understood as navigation through a multidimensional space in which multiple features jointly define each concept. We therefore propose a natural-language–based characterization of human semantic navigation as trajectories in transformer-based embedding space, using the learned representation to quantify how meaning is searched and accessed.

Classical task paradigms in cognitive sciences such as semantic fluency and property listing provide behavioral windows into this search process (Canessa et al., 2024; Canessa & Chaigneau, 2020; Troyer et al., 1997), and formal models have described how people balance exploitation and exploration over time (Hills et al., 2012). Yet these approaches often rely on labor-intensive, heterogeneous pipelines that hinder comparability across studies (Chaigneau et al., 2018). Natural Language Processing (NLP) methods—especially embedding-based analyses—offer scalable alternatives that

[1] Code in: Hidden to maintain anonymous

have already helped differentiate clinical groups and organize conceptual structure (García et al., 2025); for example, word embedding metrics have separated Alzheimer's and Parkinson's patients from controls, and language model (LM) embedding trajectories have characterized psychosis and schizofrenia profiles (Toro-Hernández et al., 2024; Ferrante et al., 2025; He et al., 2024; Nour et al., 2023; Lopes da Cunha et al., 2024; Sanz et al., 2022; Palominos et al., 2024).

Building on this foundation, our framework represents concept production as movement through transformer-based spaces. For each participant, we construct semantic trajectories and extract geometric and dynamical metrics—distance to next, velocity, acceleration, entropy, distance to centroid —that capture both scalar and directional aspects of navigation. This computational approach minimizes manual intervention while preserving rich structure in the data, enabling principled tests of hypotheses about semantic meaning and search in humans and in artificial agents (Xu et al., 2025).

We demonstrate the effectiveness of our approach by applying it to datasets that specifically challenge standard LM embeddings. Our evaluation probes: the clinical utility of embeddings for analyzing natural language in patients with Parkinson's disease (Linz et al., 2017); the semantic consistency of multilingual embeddings across Italian and German (Conneau et al., 2020; Artetxe & Schwenk, 2019); and the atypical geometric properties of swear word embeddings as revealed through a verbal fluency task (Graumas et al., 2019). Critically, our metrics provide novel insights in each of these established research areas by isolating the specific trajectory features that differentiate between different groups and semantic categories. Notably, different embedding models yield essentially similar patterns, suggesting convergent geometry across learned representations despite distinct training pipelines and architectural differences (Valeriani et al., 2023; Doimo et al., 2024; Lee et al., 2025; Wolfram & Schein, 2025). By framing human semantic retrieval as structured trajectories in embedding space, we bridge cognitive modeling with learned representations and establish a pipeline for quantifying semantic dynamics with applications to clinical research and cross-linguistic analysis (Shakeri & Farmanbar, 2023). This approach hold promise for applications, including the classification of brain disorders, the differentiation between concept types, and the testing of core hypotheses about search dynamics in artificial agents, as models that compare human responses to linguistic data with LLMs' generated responses.

## 2 METHODS

### 2.1 DATASETS

To evaluate our metrics, we use four open datasets that vary in language, population, and tasks.

**Neurodegenerative dataset**: Introduced in Toro-Hernández et al. (2024), consists of 62 Chilean Spanish-speaking participants divided into three groups: 20 individuals with Parkinson's disease (PD), 16 with the behavioral variant of frontotemporal dementia (bvFTD), and 26 healthy controls (HC). Participants completed a property listing task (PLT), in which they were asked to generate as many attributes as possible for 10 concrete concepts ("tree," "sun," "clown," "puma," "airplane," "hair," "duck," "house," "shark," and "bed"). Instructions emphasized the inclusion of "physical characteristics, internal parts, appearance, sounds, smells, textures, uses, functions, and typical locations" (Toro-Hernández et al., 2024). Finally, the data in this paper was preprocessed by only extracting content words (nouns, verbs, adjectives, and adverbs).

**Swear fluency dataset**: Introduced by Reiman & Earleywine (2023), includes 274 undergraduate native speakers of U.S. English who performed verbal fluency tasks across domains. In this case, participants were instructed to generate as many items as possible within a given category in one minute (e.g., if the category was "animals," acceptable responses might include "dog," "cat," "lion," or "tiger"). The categories comprised animals, words beginning with F, A, and S, and swear words.

**Italian** and **German datasets**: Drawn from Kremer & Baroni (2011), comprising 69 Italian and 73 German students, respectively. Participants were asked to generate descriptive properties for 50 concrete concepts, divided into 10 categories: : "Bird," "Body Part," "Building," "Clothing," "Fruit," "Furniture," "Implement," "Mammal," "Vegetable," and "Vehicle". The task has a time limit of one minute per item. Participants were encouraged to provide at least four descriptive phrases per concept and were not allowed to return to previously described items once the time expired.

### 2.2 CHARACTERIZING NAVIGATION

Participants generate concept streams—ordered lists of items (e.g., "cat," "dog," ...)—of length $N$. Let item $t$ denote the $t$-th entry. We map each stream to a trajectory in semantic space, $X = (x_1, \ldots, x_N)$, where $x_t$ is the point associated with item $t$. The points are time-indexed ($x_1$ is the first item, $x_2$ the second, etc.). Rather than computing embeddings independently (Linz et al., 2017; Nour et al., 2023), we construct them cumulatively: $x_t$ summarizes items $1{:}t$. For example, if the first two items are "cat" then "dog," $x_2$ encodes "cat dog." This design captures dependencies among successive items, avoids independence assumptions, and yields a distinct trajectory for each participant–concept pair, enabling analysis of navigation dynamics (Figure 1). Because $x_t$ conditions on the full prefix, this approach calls for more complex, sequence-aware embedding representations capable of modeling history-dependent semantics.

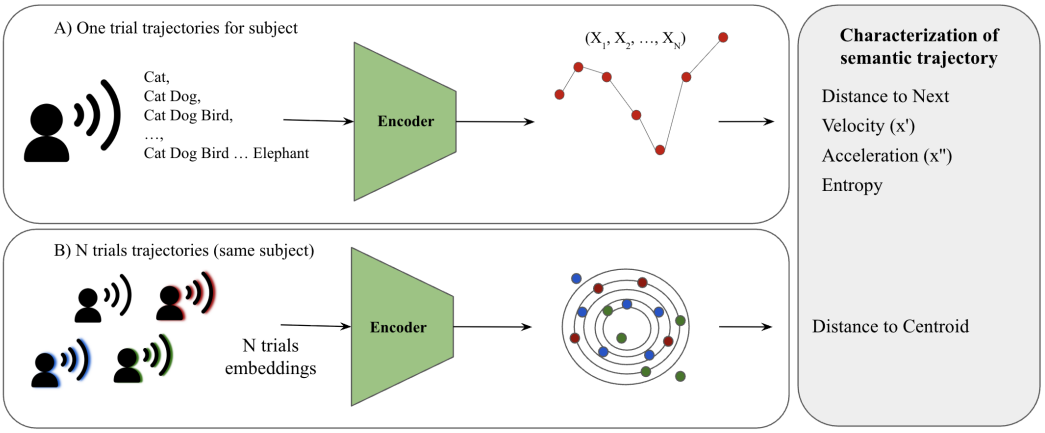


Figure 1: A schematic of the semantic trajectory analysis. (A) In a single trial, a participant generates a cumulative word list. A text encoder then maps each sequential step to a vector embedding, creating a trajectory in semantic space. This path is characterized using dynamical metrics like velocity ($x'$), acceleration ($x''$), and entropy. (B) Across multiple trials for the same subject, the dispersion of the resulting cloud of embeddings trajectories is summarized by measuring the distance of each point to the collective centroid.

Each trajectory is a time-ordered sequence of dense multilingual embeddings. Unless otherwise noted, all results are reported using OpenAI's text-embedding-3-large; results with alternative encoders (Google's text-embedding-004 and Qwen-Embedding-0.6B) (Zhang et al., 2025) are reported in the Appendix A.

#### 2.2.1 DISTANCE TO NEXT

To quantify moment-to-moment change in semantic state, we compute the cosine distance between each pair of successive unit-normalized embeddings, yielding an $N - 1$ length series of step sizes ("semantic jumps") per trajectory. Larger values indicate bigger shifts in meaning from one item to the next. Because trajectories naturally differ in length, we also summarize each series with its mean step size—the average cosine distance across steps—as a length-invariant indicator of average memory-search breadth.

#### 2.2.2 ENTROPY

We also summarize the information contained of the step-distance series with a scale-free approximate Shannon entropy. Distances are split at their within-trajectory median into "high" versus "low," forming a binary sequence whose Shannon entropy is computed and then normalized by the number of valid steps (Pincus et al., 1991). This value is set to zero when all steps fall on a single side of the median and is only estimated when at least three valid steps are available, ensuring stability for short sequences. Given an embedding time series $\{x_t\}_{t=1}^{n}$, let $\theta$ denote its within-trajectory median.

We first binarize the sequence:

$$b_t = \begin{cases} 1, & x_t \geq \theta, \\ 0, & x_t < \theta, \end{cases} \qquad t = 1, \ldots, n. \tag{1}$$

Let $p = \frac{1}{n} \sum_{t=1}^{n} b_t$ be the fraction of ones. The Shannon entropy of the binarized sequence is

$$H = -\,p \log_2 p \;-\; (1-p) \log_2 (1-p), \tag{2}$$

with the convention $H = 0$ when $p \in \{0, 1\}$. This measure can be interpreted as the information richness of fluctuation around a typical step size in a time series (Gao et al., 2008).

#### 2.2.3 Velocity and Acceleration

Beyond scalar cosine distances, we characterize the semantic directional dynamics by computing discrete derivatives of the embeddings themselves, similarly to Nour et al. (2025). It is important to remark that we are assuming an Euclidean dynamics for simplicity, which overlooks the real anisotropic nature of the embedding spaces (Nickel & Kiela, 2017; Ethayarajh, 2019). Velocity is defined as the vector difference between consecutive embeddings, yielding both a direction and a magnitude for each step; the final row has no velocity. Since in the datasets tasks don't have a time stamp, then $\alpha = \Delta t^{-1} = 1$.

$$\mathbf{v}_t = \alpha(\mathbf{x}_{t+1} - \mathbf{x}_t), \quad (t = 1, \ldots, T-1) \tag{3}$$

Acceleration is defined as the difference between successive velocity vectors and quantifies changes in direction or speed from one step to the next; the final two rows have no acceleration. These kinematic quantities retain information about where the trajectory is heading in the high-dimensional space—information that step-wise scalar distances alone cannot convey. By default, derivatives assume a unit time step between items; if timestamps are available, magnitudes can be rescaled accordingly with $\alpha$.

$$\mathbf{a}_t = \alpha(\mathbf{v}_{t+1} - \mathbf{v}_t) = \alpha^2(\mathbf{x}_{t+2} - 2\mathbf{x}_{t+1} + \mathbf{x}_t), \quad (t = 1, \ldots, T-2) \tag{4}$$

#### 2.2.4 Distance to Centroid

To capture how individual properties relate to the overall semantic context, we computed a centroid-based measure. When categorical property labels and their embeddings were available, repeated occurrences of the same property were collapsed to a single instance, ensuring that redundancy did not overweight specific properties. For each unique property, we retained only its first embedding and constructed a centroid vector representing the average position of all unique property embeddings over N trials in each specific concept and unique subject. Each item in the sequence was then assigned the cosine distance between its embedding and this centroid. This measure quantifies how far each produced property lies from the central tendency of the participant's semantic exploration, providing an index of dispersion that complements step-wise trajectory distances.

#### 2.2.5 Embedding model comparison

To compare different models, we will correlate trajectory measures for the same subject—such as distance to the next point, distance to the centroid, entropy, velocity, and acceleration—to determine whether the trajectories are similar across the models (i.e., OpenAI's text-embedding-3-large; Google's text-embedding-004; Qwen-Embedding-0.6B).

#### 2.2.6 Statistical analysis

For each metric, we evaluated group- and concept-level effects using generalized linear mixed models (GLMMs), with each metric as a fixed factor, including participants and concept as random factors to account for repeated measures and individual variability. Models were fitted according to the most appropriate distribution. Based on this procedure, a lognormal distribution was applied to distance to next, entropy, velocity, and acceleration, while a Gaussian distribution was used for the distance-to-centroid metric. Post-hoc pairwise comparisons were adjusted with Tukey's HSD to control the family-wise error rate. Visualizations combine raw distributions (boxplots with jittered points) with model-estimated marginal means and $95\%$ confidence intervals, annotated with

significance levels from the Tukey tests. This approach highlights both the variability in the data and the inferential estimates used for statistical testing. All statistical analyses were conducted in R (v.4.3.1) using the glmmTMB package, selected for its robustness, flexibility, and ease of implementation (Brooks et al., 2017).

# 3 RESULTS

## 3.1 NEURODEGENERATIVE DATASET

There was a significant effect of category across all metrics. For distance to next, healthy controls showed lower values than both bvFTD and PD, while bvFTD and PD did not differ. A similar pattern emerged for velocity and for acceleration: HCs were lower than both patient groups, with bvFTD and PD comparable. For entropy, HC again showed reduced values compared to both groups, with no difference between bvFTD and PD. In contrast, distance-to-centroid showed the opposite pattern: HC exhibited greater distances than both patient groups, which did not differ from each other. These results indicate that semantic navigation in patient groups is characterized by greater spread, higher variability, increased entropy, and more compact clustering relative to controls (see Figure 2).

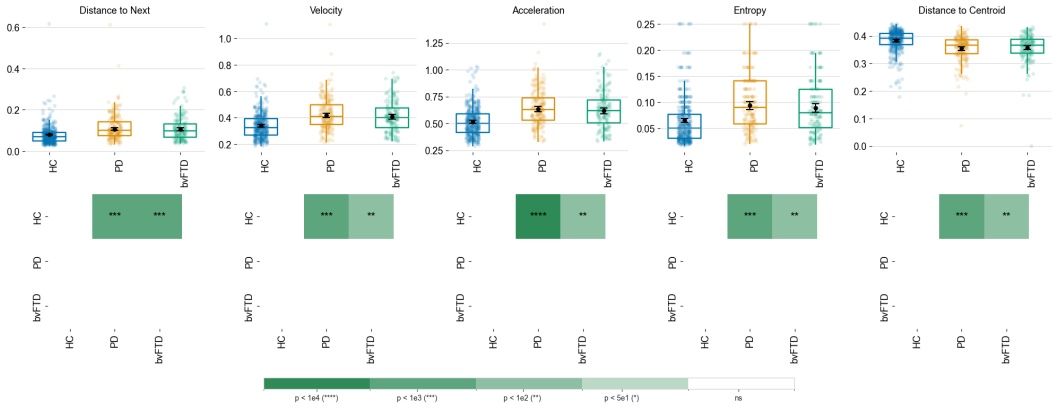


Figure 2: Summary of the metrics for the Neurodegenerative dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across three semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

## 3.2 SWEAR FLUENCY DATASET

For distance to next, all letter categories and swear words were higher than animals, with swear words eliciting the longest distance to next and animals the shortest; letters occupied an intermediate range. Velocity showed the same pattern, with navigation being faster for letters and especially for swear words than for animals. Acceleration mirrored velocity, with letters and, most prominently, swear words exceeding animals. For entropy, animals showed the lowest values; letters were higher, and swear words the highest. Finally, distance-to-centroid reversed the pattern, with animals being farther from the centroid than letters, whereas swear words were markedly closer. Overall, swear words consistently drove the strongest responses across metrics, animals the lowest, and letters clustered in between (see Figure 3).

## 3.3 ITALIAN DATASET

Relative to Bird (reference group), most categories showed shorter distance to next, with Building and Vehicle the least separated from Bird. Velocity followed the same ordering, with Bird exceeding most categories and Building and Vehicle only weakly or not separated. Acceleration mirrored

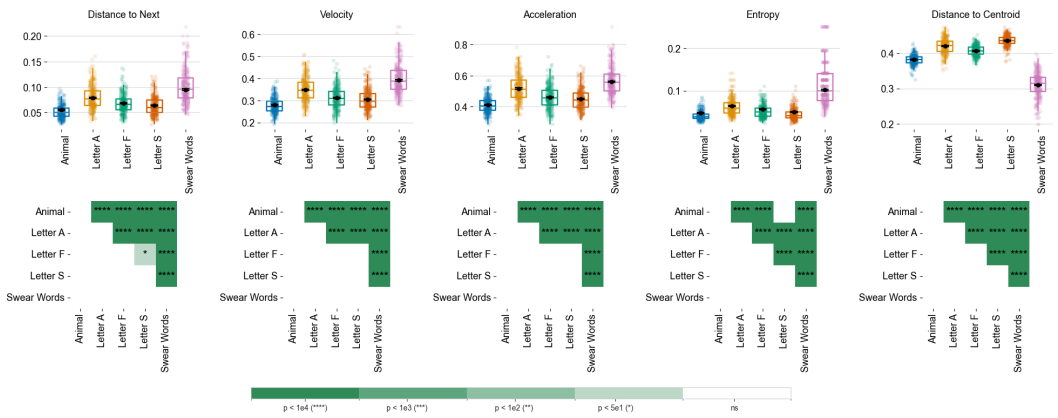


Figure 3: Summary of the metrics for the Swear Fluency dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across five semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

velocity, again showing Bird higher than the bulk of categories. Entropy differences were selective: several categories had lower entropy than Bird, whereas many contrasts were not significant. Distance-to-centroid showed a partially reversed structure: some categories were farther from the centroid than Bird, while others (e.g., Body Part, Clothing, Implement) were closer; several categories showed no difference from Bird (see Figure 4).

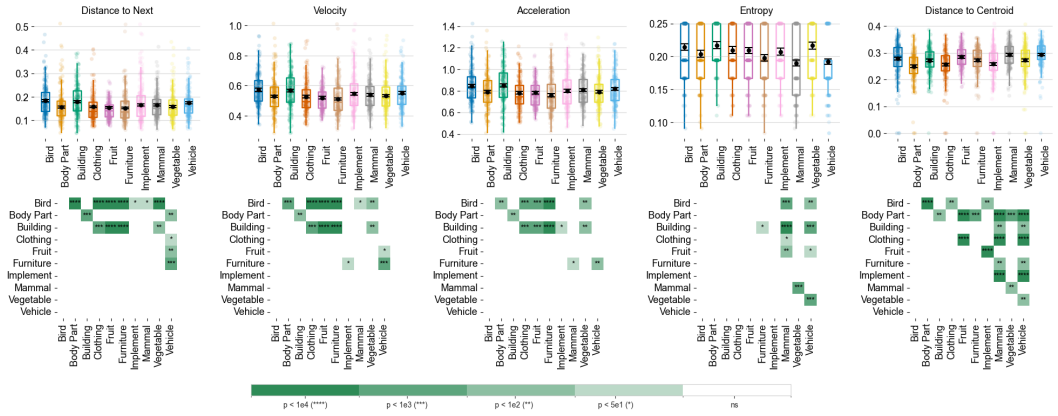


Figure 4: Summary of the metrics for the Italian dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

### 3.4 German Dataset

For distance next, most categories produced shorter values than Bird, with Vehicle and several others most distinct from Bird, and Vegetable showing little to no separation. Velocity showed Bird higher than nearly all categories, with the largest gap against Vehicle. Acceleration followed the same pattern, with Bird exceeding most categories and the clearest separation against Vehicle. Entropy differences were selective: several categories were lower than Bird, while Implement was higher; many others showed no differences. Finally, distance-to-centroid revealed a different structure:

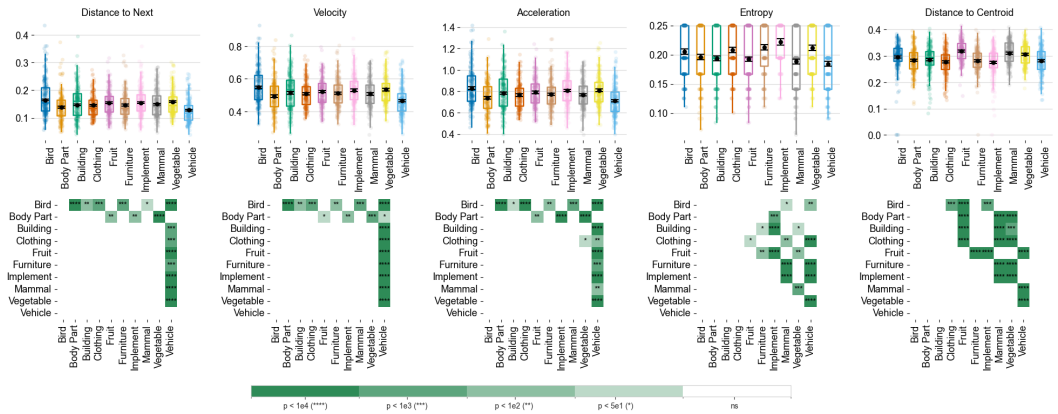


Figure 5: Summary of the metrics for the German dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

some categories (e.g., Fruit, Mammal, Vegetable) were farther from the centroid than Bird, whereas others (e.g., Body Part, Building, Clothing, Furniture, Implement, Vehicle) were closer, with the most pronounced gaps involving Fruit compared to clothing and tool-like categories (see Figure 5).

### 3.5 Model Comparison

To ensure our findings were not dependent on a specific text encoder, we compared the trajectory metrics from three different models. Figure 6 shows the Pearson correlation matrices for five metrics across all four datasets, revealing a generally high degree of robustness. The results exhibit a block-diagonal structure, indicating strong positive correlations for each metric across models. Kinematic measures—velocity, acceleration, and distance-to-next—are highly correlated across models and moderately correlated with one another, consistent with their shared capture of step-wise trajectory dynamics. Two metrics diverge: distance to centroid shows the weakest inter-model correlation, suggesting sensitivity to model-specific embedding geometry, whereas entropy shows near-perfect inter-model correlation because it depends on rank ordering rather than absolute distances; median binarization further stabilizes it when models agree on the relative size of semantic jumps.

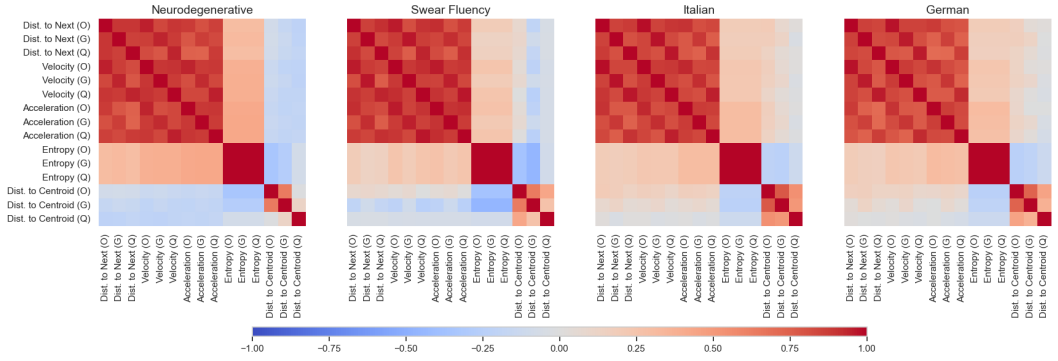


Figure 6: Pearson correlations of metrics across datasets. From left to right: Neurodegenerative, Swear Fluency, Italian, and German. Each heatmap reports pairwise correlations between Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid, computed with three embedding models (O = OpenAI text-embedding-3-large, G = Google text-embedding-004, Q = Qwen-Embedding-0.6B). Color indicates correlation strength (blue = negative, red = positive).

# 4 DISCUSSION

Our results show that each metric indexed a distinct aspect of semantic navigation. Distance to next reliably separated clinical groups from healthy controls in the neurodegenerative dataset and graded the fluency tasks, with the swear-word task yielding the largest effect, followed by the letter task, and then the animal task. This aligns with prior evidence that Parkinson's disease is associated with greater variability in semantic search compared to controls (Toro-Hernández et al., 2024).The kinematic measures of velocity and acceleration reproduced these effects, quantifying the pace of movement and the rate of directional change rather than step size alone. Nonetheless, in the Italian and German datasets, these metrics also uncovered category-specific patterns. While not identical across languages, these patterns consistently revealed an informative structure about how retrieval unfolds within each linguistic context, a finding that complements previous work on cross-lingual similarities (Conneau et al., 2020; Artetxe & Schwenk, 2019).

Entropy captured disorder in search and complemented the dynamic-based metrics. Neurodegenerative groups showed higher entropy than controls, consistent with less predictable, less routine traversal of meaning space, possible due to executive functioning constraints (Birba et al., 2017; Cousins & Grossman, 2017). Entropy was also higher for swear-word fluency than for animal or letter fluency, in line with greater contextual dependence and population variability of taboo lexicons. Cross-language analyses, again revealed category-specific differences without assuming the same pattern in Italian and German, suggesting language-specific modulation of variability. Finally, distance-to-centroid indexed positional centrality and provided information orthogonal to other measures. Healthy controls were farther from the centroid than patient groups in the neurodegenerative dataset, indicating more dispersed searches. Swear-word fluency was more central than animal and letter tasks, consistent with a tighter lexical neighborhood. In both Italian and German datasets, distance-to-centroid differentiated multiple categories, with language-dependent profiles—evidence that centrality exposes structure not fully explained by distance to next, velocity, acceleration, or entropy.

Our findings not only corroborate previous research but also highlight the power of NLP-based approaches to address new scientific questions. Geometrically grounded analyses of language have been useful for capturing complex semantic patterns in neurodegeneration that are often missed by traditional methods (Mota et al., 2023; Zhang et al., 2022). This is evidenced by their successful application in classifying patient groups (Sanz et al., 2022; García et al., 2025), distinguishing cognitive phenotypes (Aresta et al., 2025), predicting disease progression (Šubert et al., 2023), and their extension across languages to conditions like HIV (Gattei et al., 2023) and schizophrenia (Palominos et al., 2024; He et al., 2024; Zhang et al., 2024). Furthermore, our results open new avenues in less-explored domains, such as the semantic navigation underlying the production of swear words. Since swear-word fluency has been linked to substance use (Reiman & Earleywine, 2023) and differential brain activity patterns in schizophrenia (Lee et al., 2019), analyzing its semantic dynamics could provide novel insights into behavioral regulation and inhibitory control. Thus, the continued development of NLP metrics for semantic navigation is crucial for advancing our understanding of human semantic search and its disruptions across diverse clinical and linguistic contexts.

Our results also proved to be discriminative of specific semantic categories. The analysis of category-specific effects in semantic navigation has been crucial for differentiating cognitive profiles in brain pathologies (Shebani et al., 2017). Interestingly, although our approach was effective across contexts and languages, differences in category effects emerged between the Italian and German datasets. This may reflect the flexible nature of lexico-semantic representations, where linguistic structure and cultural conventions shape how meaning is accessed and organized during semantic search (Vigliocco et al., 2009; Barsalou, 2023; Kemmerer, 2023). In this respect, different transformer-based models, trained on distinct corpora, may be expected to capture specific manifestations of semantic structure in divergent ways.

Crucially, our cross-model analyses revealed that trajectory metrics were highly correlated across the three different embedding models, indicating that the observed dynamics are not an artifact of a single encoder, as agrees with previous literature as they generate similar representations (Lee et al., 2025; Wolfram & Schein, 2025). This consistency was particularly strong for metrics capturing the local, step-by-step evolution of the trajectory, such as velocity, acceleration, and entropy. In contrast, the distance-to-centroid metric consistently showed the lowest inter-model correlation,

revealing that while models agree on a trajectory's local dynamics (its shape and variability), they differ significantly in its global positioning. This sensitivity arises because the metric uses a static, global average rather than successive states, making it dependent on the unique high-level geometry of each model's embedding space. It might be a possible tool for comparing how different models structure knowledge. Notably, this dissimilarity between models was most pronounced in the neurodegenerative dataset, potentially reflecting a more complex disruption in semantic navigation.

## 5 Conclusion

In sum, applying five text embedding-based trajectory metrics to fluency and property listing tasks data revealed signatures of semantic navigation: distance-to-next, velocity, and acceleration distinguished neurodegenerative groups from healthy controls in their semantic search; entropy captured irregularity in search (notably higher for swear-word fluency); and distance-to-centroid indexed positional centrality orthogonal to the other measures, exposing category- and language-specific structure. These effects were broadly consistent across three multilingual transformer embedding models, indicating robustness to the choice of encoder for local trajectory dynamics, while lower cross-model agreement for distance-to-centroid highlights model-dependent global geometry. Together, these results shows a geometrically grounded NLP framework for characterizing human semantic retrieval, through across tasks and languages, and open avenues for clinical stratification and cross-model comparisons of how humans and generative LMs traverse meaning space.

## 6 Limitations and Future Work

Although fluency and property-listing tasks are useful across a range of applications, they capture only a partial view of human semantic navigation, in this specific case the tasks didn't contain the time step of the words. This could contribute to more temporal meaningful dynamics. Developing richer speech-based protocols may help probe semantic search and representation more directly, especially when linked to learned representations from language models (LMs). We acknowledge that our assumption of Euclidean dynamics is a simplification that overlooks the anisotropic structure of embedding spaces (Nickel & Kiela, 2017; Ethayarajh, 2019). More mathematically robust—potentially non-Euclidean—metrics are needed to better characterize these trajectories. Furthermore, we used a basic information measure (Shannon entropy); future work should broaden the information-theoretic toolkit for semantic navigation to assess complexity in systems with many interacting variables that jointly shape human semantic representation and retrieval.

Furthermore, future work could apply a similar framework to characterize different LLMs and assess their generative semantic navigation across tasks. The goal is to develop a unified account of trajectories in semantic space that encompasses both humans and generative language models.

## Reproducibility Statement

The code and data required to reproduce the findings of this study are openly available. The source code for all analyses and figure generation is accessible at [hidden link to repository]. The datasets are all from public sources, which are detailed with their respective access links in Appendix A.2.

## References

Simona Aresta, Petronilla Battista, Cinzia Palmirotta, Serena Tagliente, Gianvito Lagravinese, Paola Santacesaria, Allegra Benzini, Davide Mongelli, Brigida Minafra, Christian Lunetta, Adolfo M. García, and Christian Salvatore. Digital phenotyping of parkinson's disease via natural language processing. *NPJ Parkinson's Disease*, 11(1):182, 2025. doi: 10.1038/s41531-025-01050-8.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038/.

Lawrence W. Barsalou. Implications of grounded cognition for conceptual processing across cultures. *Topics in Cognitive Science*, 00:1–8, 2023. ISSN 1756-8765. doi: 10.1111/tops.12661. URL https://doi.org/10.1111/tops.12661.

Agustina Birba, Indira García-Cordero, Giselle Kozono, Agustina Legaz, Agustín Ibáñez, Lucas Sedeño, and Adolfo M. García. Losing ground: Frontostriatal atrophy disrupts language embodiment in parkinson's and huntington's disease. *Neuroscience & Biobehavioral Reviews*, 80: 673–687, 2017. doi: 10.1016/j.neubiorev.2017.07.011.

Mollie Elizabeth Brooks, Kasper Kristensen, Koen Johannes van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker. glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400, 2017. doi: 10.32614/RJ-2017-066.

Enrique Canessa and Sergio E. Chaigneau. Mathematical regularities of data from the property listing task. *Journal of Mathematical Psychology*, 97:102376, 2020. doi: 10.1016/j.jmp.2020.102376.

Enrique Canessa, Sergio E. Chaigneau, and Sebastián Moreno. Describing and understanding the time course of the property listing task. *Cognitive Processing*, 25:61–74, 2024. doi: 10.1007/s10339-023-01282-3.

Sergio E. Chaigneau, Enrique Canessa, Carlos Barra, and Rodrigo Lagos. The role of variability in the property listing task. *Behavior Research Methods*, 50:972–988, 2018. doi: 10.3758/s13428-017-0920-8.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536/.

Katheryn AQ Cousins and Murray Grossman. Evidence of semantic processing impairments in behavioural variant frontotemporal dementia and parkinson's disease. *Cortex*, 93:92–101, 2017. doi: doi:10.1097/WCO.0000000000000498.

Veronica Diveica, Emiko J. Muraki, Richard J. Binney, and Penny M. Pexman. Special issue: The multidimensionality, variability and flexibility of concepts. editorial. *Cortex*, 2025. doi: 10.1016/j.cortex.2025.05.011.

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 18122–18165. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/206018a258033def63607fbdf364bd2d-Paper-Conference.pdf.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019. doi: 10.48550/arXiv.1909.00512. Accepted at EMNLP 2019.

Franco J Ferrante, Daniel Escobar Grisales, María Fernanda López, Pamela Lopes da Cunha, Lucas Federico Sterpin, Jet MJ Vonk, Pedro Chaná Cuevas, Claudio Estienne, Eugenia Hesse, Lucía Amoruso, et al. Cognitive phenotyping of parkinson's disease patients via digital analysis of spoken word properties. *Movement Disorders*, 2025. doi: 10.1002/mds.70005.

Y. Gao, I. Kontoyiannis, and E. Bienenstock. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99, 2008. doi: 10.3390/entropy-e10020071.

Adolfo M. García, Francisco J. Ferrante, Gonzalo Pérez, Julieta Ponferrada, Alejandro Sosa Welford, Nicolás Pelella, Matías Caccia, et al. Toolkit to examine lifelike language v. 2.0: Optimizing speech biomarkers of neurodegeneration. *Dementia and Geriatric Cognitive Disorders*, 54(2):96–108, 2025. doi: 10.1159/000541068.

Carolina A. Gattei, Franco J. Ferrante, Bárbara Sampedro, Lucas Sterpin, Valeria Abusamra, Lorena Abusamra, Paola Andrea Cañataro, and Adolfo M. García. Semantic memory navigation in hiv: Conceptual associations and word selection patterns. *Neuropsychologia*, 189:107934, 2023. doi: 10.1016/j.neuropsychologia.2023.107934.

Leon Graumas, Roy David, and Tommaso Caselli. Twitter-based polarised embeddings for abusive language detection. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–7, 2019. doi: 10.1109/ACIIW.2019.8925049.

Rui He, Claudio Palominos, Han Zhang, Maria Francisca Alonso-Sánchez, Lena Palaniyappan, and Wolfram Hinzen. Navigating the semantic space: Unraveling the structure of meaning in psychosis using different computational language models. *Psychiatry Research*, 333:115752, 2024. doi: 10.1016/j.psychres.2024.115752.

T. T. Hills, P. M. Todd, and M. N. Jones. Foraging in Semantic Fields: How We Search Through Memory. *Topics in Cognitive Science*, 7(3):513–534, 2015. doi: 10.1111/tops.12151. URL https://doi.org/10.1111/tops.12151.

Thomas T. Hills, Michael N. Jones, and Peter M. Todd. Optimal foraging in semantic memory. *Psychological Review*, 119(2):431–440, 2012. doi: 10.1037/a0027373.

David Kemmerer. Grounded cognition entails linguistic relativity: A neglected implication of a major semantic theory. *Topics in Cognitive Science*, 15(4):615–647, 2023. doi: 10.1111/tops.12628.

Gerhard Kremer and Marco Baroni. A set of semantic norms for german and italian. *Behavior Research Methods*, 43:97–109, 2011. doi: 10.3758/s13428-010-0028-x.

Andrew Lee, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *ICLR 2025 Re-Align Workshop*, 2025. Submission 55.

Sang Won Lee, Bumseok Jeong, Jong-Il Park, Gyung Ho Chung, Hyo-Jong Lee, Yin Cui, Woo-Sung Kim, Kang Han Oh, Il Seok Oh, Guang Fan Shen, and Young-Chul Chung. Alteration of semantic networks during swear words processing in schizophrenia. *Psychiatry Investigation*, 16 (1):70–79, 2019. doi: 10.30773/pi.2018.10.09.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra König. Using neural word embeddings in the analysis of the clinical semantic verbal fluency task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short Papers*, pp. 1–7. The Association for Computer Linguistics, 2017. doi: 10.18653/v1/W17-6926. URL https://aclanthology.org/W17-6926/.

Pedro Lopes da Cunha, Facundo Ruiz, Francisco Ferrante, Lucas F. Sterpin, Agustin Ibáñez, Andrea Slachevsky, Adolfo M. García, et al. Automated free speech analysis reveals distinct markers of alzheimer's and frontotemporal dementia. *PLoS ONE*, 19(6):e0304272, 2024. doi: 10.1371/journal.pone.0304272.

Natália Bezerra Mota, Janaina Weissheimer, Ingrid Finger, Marina Ribeiro, Bárbara Malcorra, and Lilian Hübner. Speech as a graph: Developmental perspectives on the organization of spoken language. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2023. doi: 10.1016/j.bpsc.2023.04.004.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

M. M. Nour, D. C. McNamee, Y. Liu, and R. J. Dolan. Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts. *Proceedings of the National Academy of Sciences of the United States of America*, 120(42):e2305290120, 2023. doi: 10.1073/pnas.2305290120. URL https://doi.org/10.1073/pnas.2305290120.

Matthew M. Nour, Daniel C. McNamee, Isaac Fradkin, and Raymond J. Dolan. Charting trajectories of human thought using large language models, 2025. URL https://arxiv.org/abs/2509.14455.

Claudio Palominos, Rui He, Karla Fröhlich, Rieke Roxanne Mülfarth, Svenja Seuffert, Iris E. Sommer, Philipp Homan, Tilo Kircher, Frederike Stein, and Wolfram Hinzen. Approximating the semantic space: word embedding techniques in psychiatric speech analysis. *Schizophrenia*, 10 (114), 2024. doi: 10.1038/s41537-024-00434-6.

Steven M. Pincus, Ian M. Gladstone, and Richard A. Ehrenkranz. A regularity statistic for medical data analysis. *Journal of Clinical Monitoring and Computing*, 7(4):335–345, 1991. doi: 10.1007/BF01619355.

Ann-Kathrin Reiman and Mitch Earleywine. Swear word fluency, verbal fluency, vocabulary, personality, and drug involvement. *Journal of Individual Differences*, 44(1):37–46, 2023. doi: 10.1027/1614-0001/a000379.

Camila Sanz, Facundo Carrillo, Andrea Slachevsky, Gonzalo Forno, Maria Luisa Gorno Tempini, Rodrigo Villagra, Adolfo M. García, et al. Automated text-level semantic markers of alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14(1):e12276, 2022. doi: 10.1002/dad2.12276.

Arezo Shakeri and Mina Farmanbar. Natural language processing in alzheimer's disease research: Systematic review of methods, data, and efficacy. *Frontiers in Digital Health*, 5:1250365, 2023. doi: 10.1002/dad2.70082.

Zubaida Shebani, Karalyn Patterson, Peter J. Nestor, Lara Z. Diaz-de Grenu, Kate Dawson, and Friedemann Pulvermüller. Semantic word category processing in semantic dementia and posterior cortical atrophy. *Cortex*, 93:92–106, 2017. doi: 10.1016/j.cortex.2017.04.016.

Martin Šubert, Michal Novotný, Tereza Tykalová, Jan Hlavnička, Petr Dušek, Evžen Růžička, Dominik Škrabal, Amelie Pelletier, et al. Spoken language alterations can predict phenoconversion in isolated rapid eye movement sleep behavior disorder: A multicenter study. *Annals of Neurology*, 2023. doi: 10.1002/ana.26835.

Felipe Diego Toro-Hernández, Joaquín Migeot, Nicolás Marchant, Daniela Olivares, Franco Ferrante, Raúl González-Gómez, Cecilia González Campo, Sol Fittipaldi, Gonzalo M. Rojas-Costa, Sebastian Moguilner, Andrea Slachevsky, Pedro Chaná Cuevas, Agustín Ibáñez, Sergio Chaigneau, and Adolfo M. García. Neurocognitive correlates of semantic memory navigation in parkinson's disease. *npj Parkinson's Disease*, 10(15), 2024. doi: 10.1038/s41531-024-00727-2.

Angela K. Troyer, Morris Moscovitch, and Gordon Winocur. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, 11 (1):138–146, 1997. doi: 10.1037/0894-4105.11.1.138.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51234–51252. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a0e66093d7168b40246af1cddc025daa-Paper-Conference.pdf.

Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247, 2009. doi: 10.1515/LANGCOG.2009.011.

Christopher Wolfram and Aaron Schein. Layers at similar depths generate similar activations across LLM architectures. In *Proceedings of the Conference on Language Modeling (COLM)*, jul 2025.

Haodong Xie, Rahul Singh Maharjan, Federico Tavella, and Angelo Cangelosi. From concrete to abstract: A multimodal generative approach to abstract concept learning. *arXiv preprint arXiv:2410.02365v1*, 2024. URL https://arxiv.org/abs/2410.02365v1.

Q. Xu, Y. Peng, S. A. Nastase, M. Chodorow, M. Wu, and P. Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*, 9(9):1871–1886, 2025. doi: 10.1038/s41562-025-02203-8. URL https://doi.org/10.1038/s41562-025-02203-8.

Guanyu Zhang, Jinghong Ma, Piu Chan, and Zheng Ye. Graph theoretical analysis of semantic fluency in patients with parkinson's disease. *Computational and Mathematical Methods in Medicine*, 2022:6935263, 2022. doi: 10.1155/2022/6935263.

Han Zhang, Rui He, Claudio Palominos, Ning Hsu, Hintat Cheung, and Wolfram Hinzen. The structure of meaning in schizophrenia: A study of spontaneous speech in chinese. *Psychiatry Research*, 330:116347, 2024. doi: 10.1016/j.psychres.2024.116347.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

# A Appendix

## A.1 Use of Large Language Models Disclosure

We used large language models (LLMs) as a code assistant and text editor to refine implementation details, improve manuscript clarity and grammar, and identify relevant literature. All LLM outputs were thoroughly reviewed and verified by the authors. The conceptual framework and methodology contributions presented in this work are entirely our own.

## A.2 Data Availability

The datasets used in this study are publicly available. The neurodegenerative dataset can be found at https://osf.io/8pufk/, and the swear words dataset is located at https://osf.io/w8drt/. The Italian and German datasets were sourced from the appendices of Kremer & Baroni (2011), available at https://link.springer.com/article/10.3758/s13428-010-0028-x.

### A.3 RESULTS FOR QWEN-EMBEDDING-0.6B

All the previous results were reproduced for Qwen-Embedding-0.6B, which is a small open-source high performance embedding model (Zhang et al., 2025). For Neurodegerative dataset (Fig. 7), swear dataset (Fig. 8), Italian (Fig. 9) and German (Fig. 10) datasets.

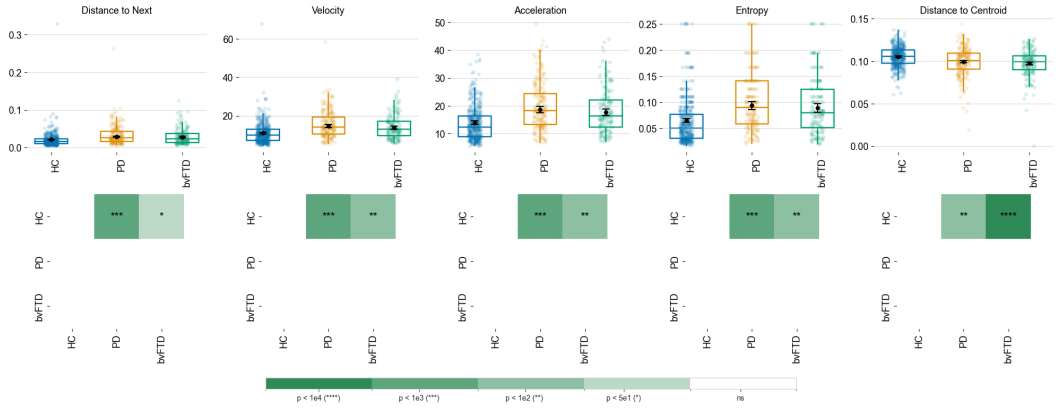


Figure 7: Summary of the metrics for the Neurodegenerative dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across three semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

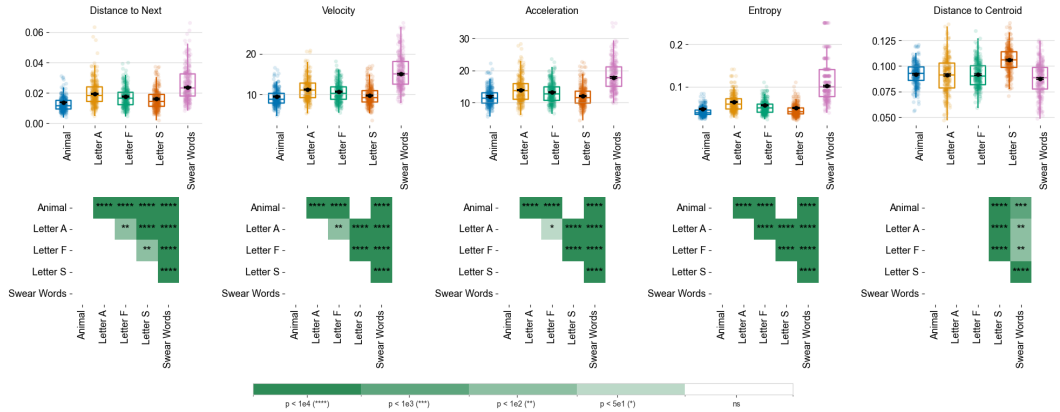


Figure 8: Summary of the metrics for the Swear Fluency dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across five semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

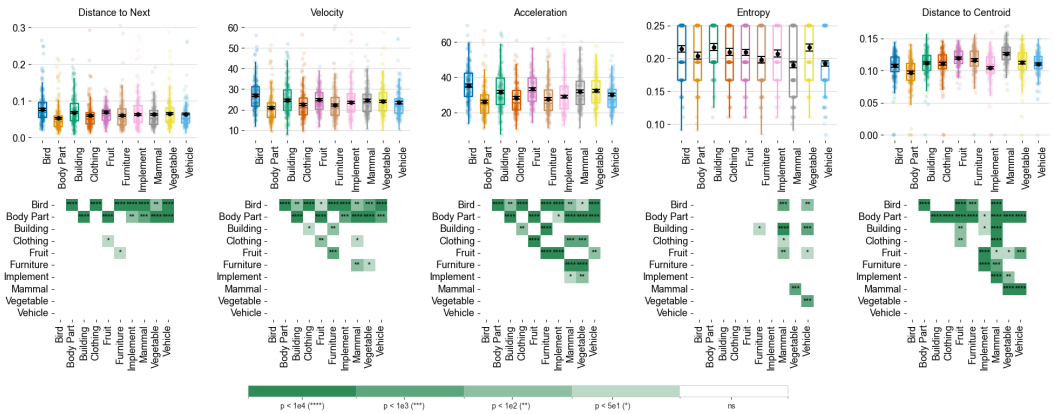


Figure 9: Summary of the metrics for the Italian dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

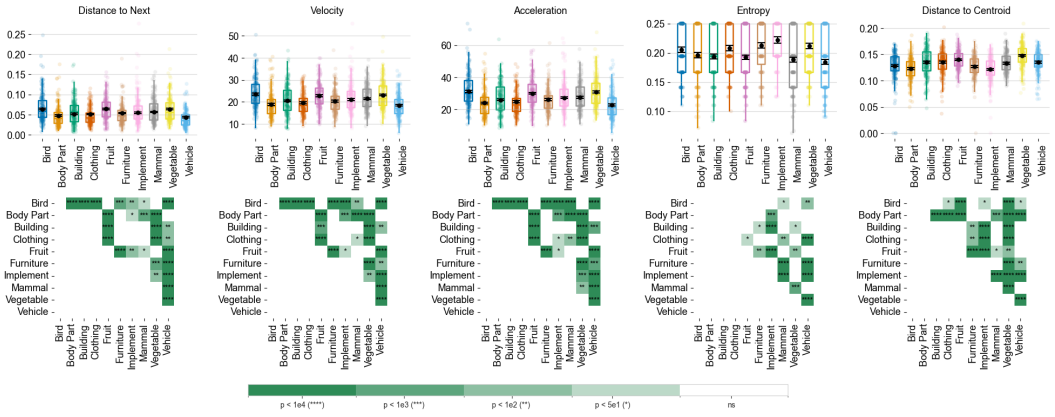


Figure 10: Summary of the metrics for the German dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

### A.4 RESULTS FOR GOOGLE'S TEXT-EMBEDDING-004

All the previous results were reproduced for Google's text-embedding-004. For Neurodegerative dataset (Fig. 11), swear dataset (Fig. 12), Italian (Fig. 13) and German (Fig. 14) datasets.

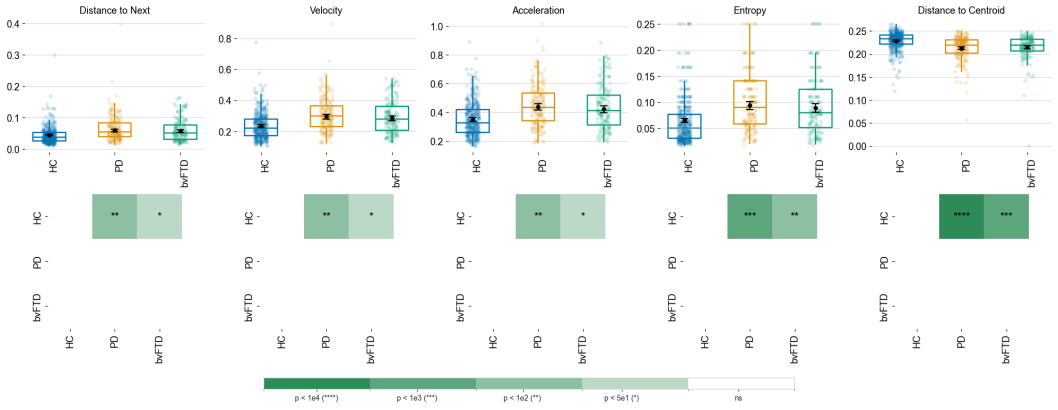


Figure 11: Summary of the metrics for the Neurodegenerative dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across three semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

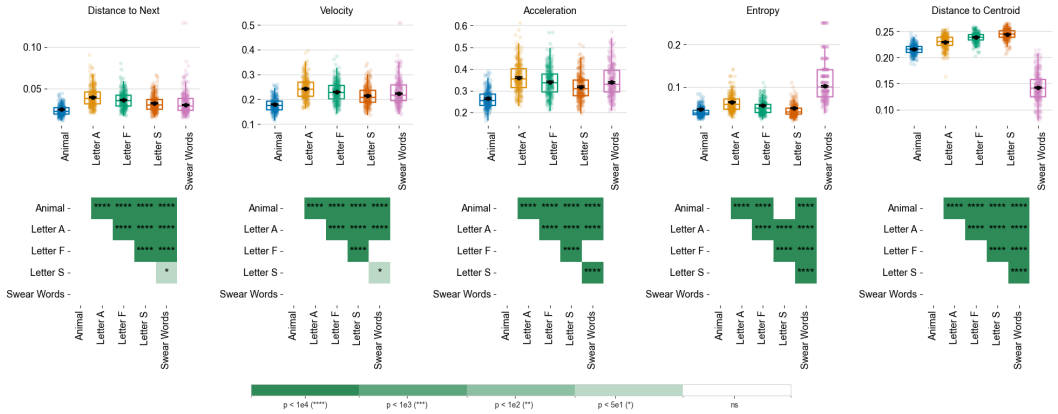


Figure 12: Summary of the metrics for the Swear Fluency dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across five semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

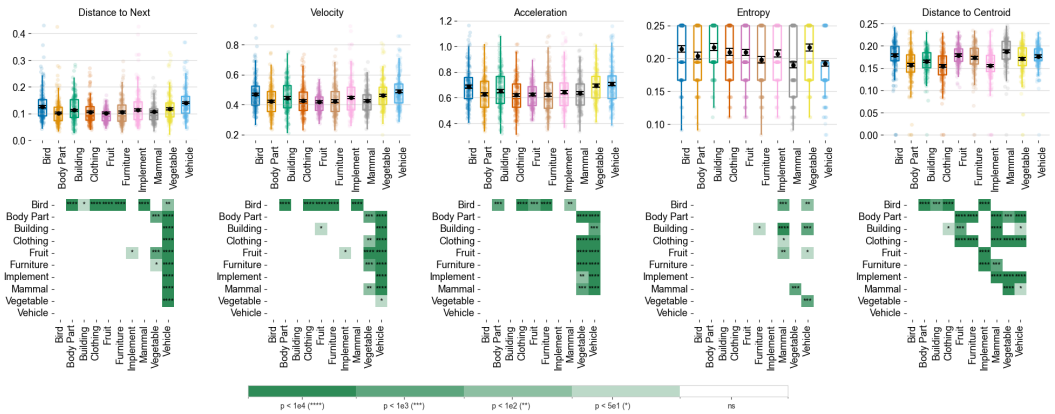


Figure 13: Summary of the metrics for the Italian dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).

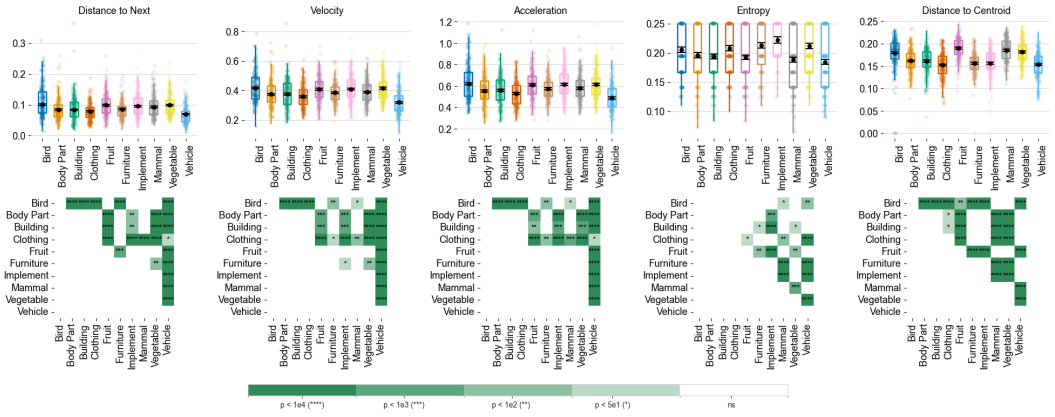


Figure 14: Summary of the metrics for the German dataset. From left to right: Distance to Next, Velocity, Acceleration, Entropy, and Distance to Centroid. Each panel shows distributions across ten semantic categories using boxplots (with individual observations overlaid). Below each panel, a matrix reports pairwise statistical comparisons between categories, with color intensity and asterisks denoting significance levels (see scale).