ICLR 2026

Characterizing human semantic navigation as trajectories in embedding space

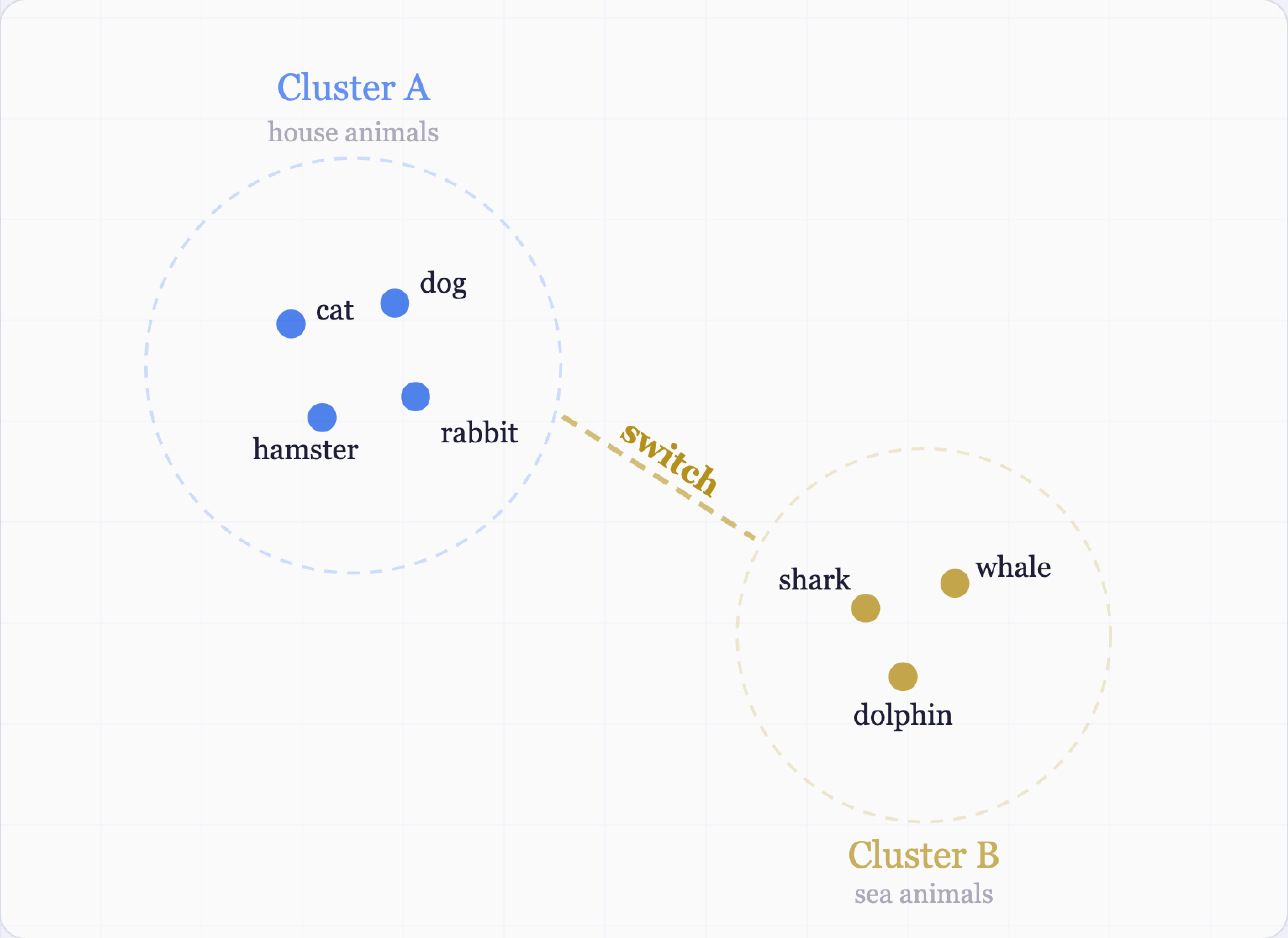
Felipe D. Toro-Hernández[1] · Jesuino Vieira Filho[2] · Rodrigo M. Cabral-Carvalho[1]

All authors contributed equally

[1] Center of Mathematics, Computing and Cognition · Federal University of ABC
[2] Dept. of Computer Science and Operations Research · Université de Montréal

THE PROBLEM

Human semantic navigation: moving through concepts



Navigation through semantic representations is often characterized in terms of *clustering* and *switching*.

Usually, semantic pipelines are labor-intensive, heterogeneous, and hard to compare across studies.

Furthermore, this often misses the step-by-step granularity — the continuous geometry of how meaning unfolds over time.

→ What if we framed semantic retrieval as a **trajectory through geometric space**?

KEY IDEA

Cumulative embeddings capture search history

Each word x_t encodes the full prefix up to step t.

Step 1: embed `"cat"`
Step 2: embed `"cat dog"`
Step 3: embed `"cat dog shark"`
⋮

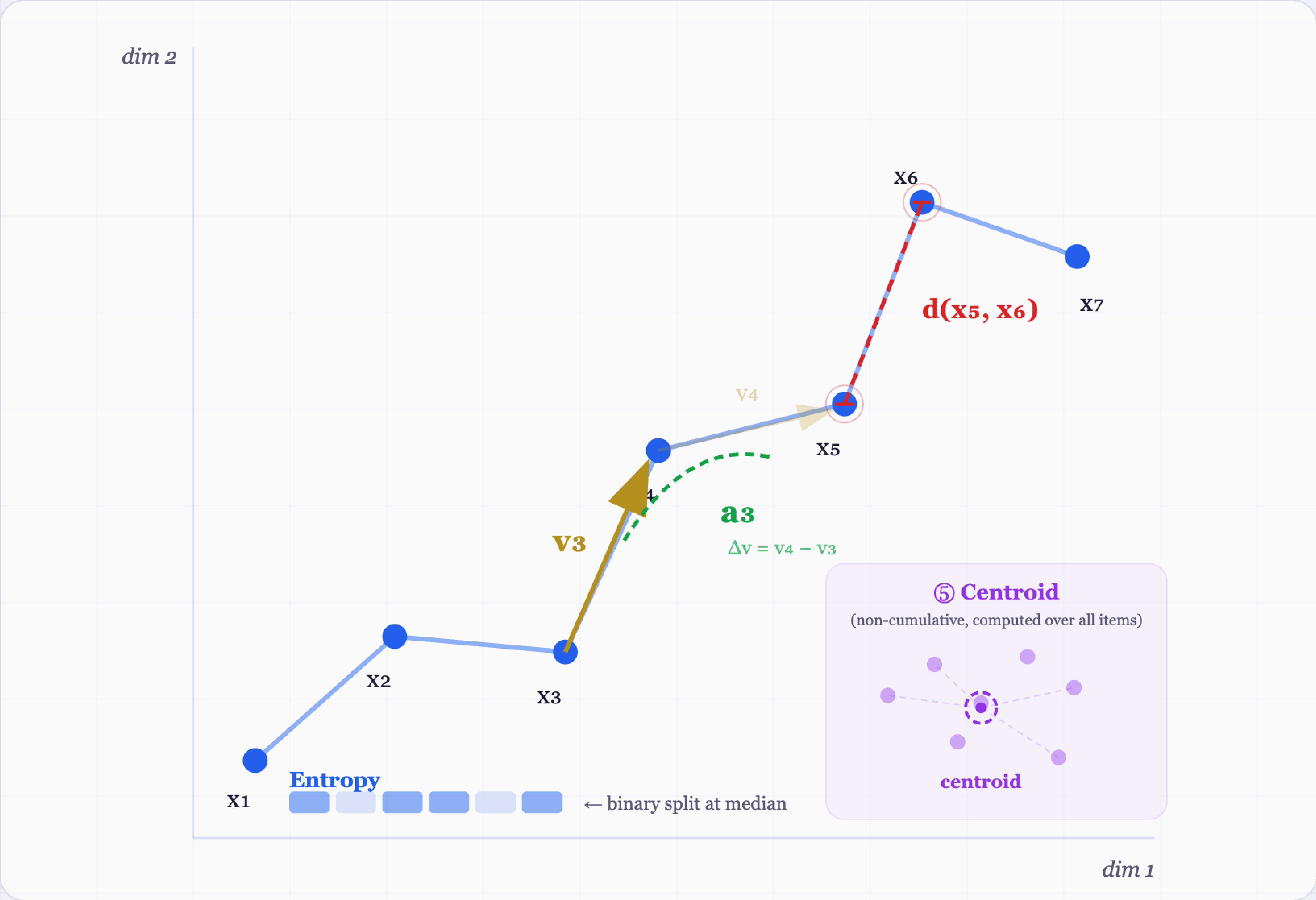
This captures dependencies between successive items — semantic retrieval is inherently cumulative.

Working memory and inhibitory control shape each response based on all previous ones.

Result: a trajectory per participant × concept pair

METHODS

Five physics-inspired trajectory metrics



① Distance to next

Cosine distance between consecutive points.

Semantic jump size.

② Velocity

$v_t = x_{t+1} - x_t$

Direction + magnitude of each step.

③ Acceleration

$a_t = v_{t+1} - v_t$

Low → stable cluster. High → erratic switch.

④ Entropy

Shannon entropy of median-split steps.

Predictability of the search.

⑤ Distance to centroid

Distance to mean position of all items. **Dispersion** of the search.

EXPERIMENTAL SETUP

Four datasets · Four languages · Two tasks

A

Neurodegenerative

ES-CL 🇨🇱 · N=76 · Property Listing Task

🧠 HC · 🔵 PD · 🔴 bvFTD

B

Swear Fluency

EN 🇺🇸 · N=274 · Verbal Fluency Task

🤬 Swear · 🐾 Animals · 🔤 Letters

C

Cross-Linguistic PLT

Property Listing Task · 10 semantic categories

🇮🇹 Italian N=69

🇩🇪 German N=73

Statistics: Generalized Linear Mixed Models (GLMMs) with Tukey HSD post-hoc correction.

Embedding models

- OpenAI text-embedding-3-large
- Google text-embedding-004
- Qwen3-Embedding-0.6B
- fastText (baseline, non-cumul.)

All results reported with OpenAI unless otherwise noted.

Design choices

Causal and bidirectional attention

Cumulative vs. non-cumulative comparison

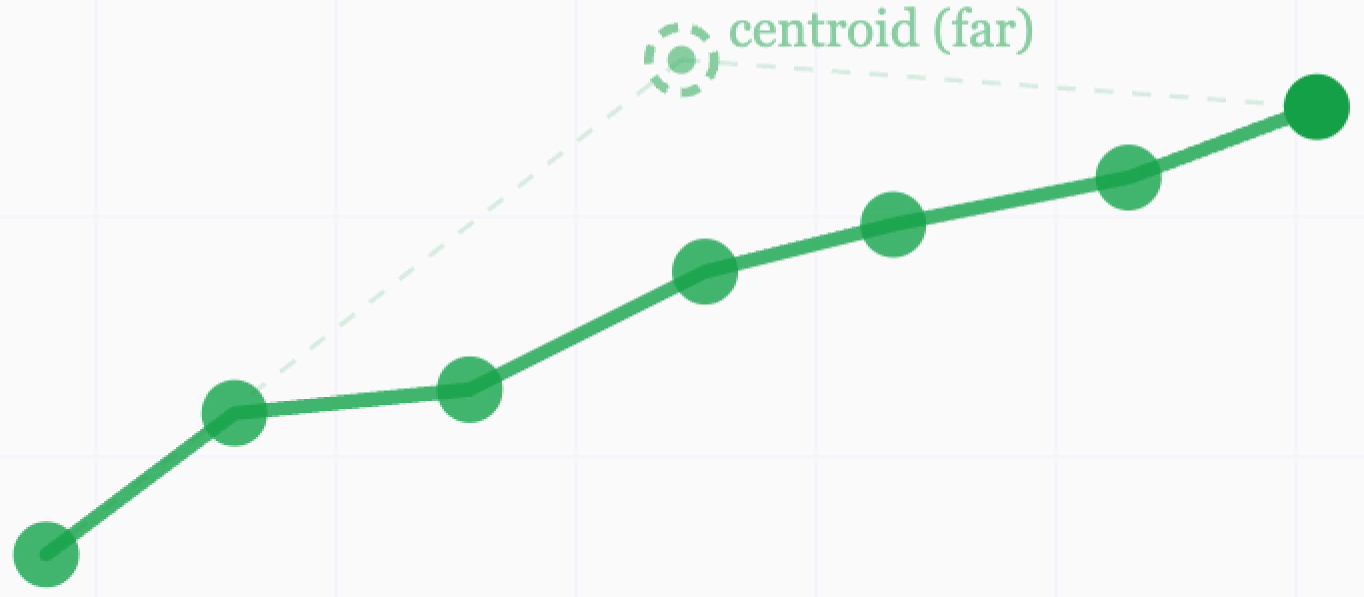
ZCA-whitening tested for anisotropy

RESULTS

Neurodegenerative patients show more erratic, constricted navigation

Healthy Control

Organized · Broad · Predictable

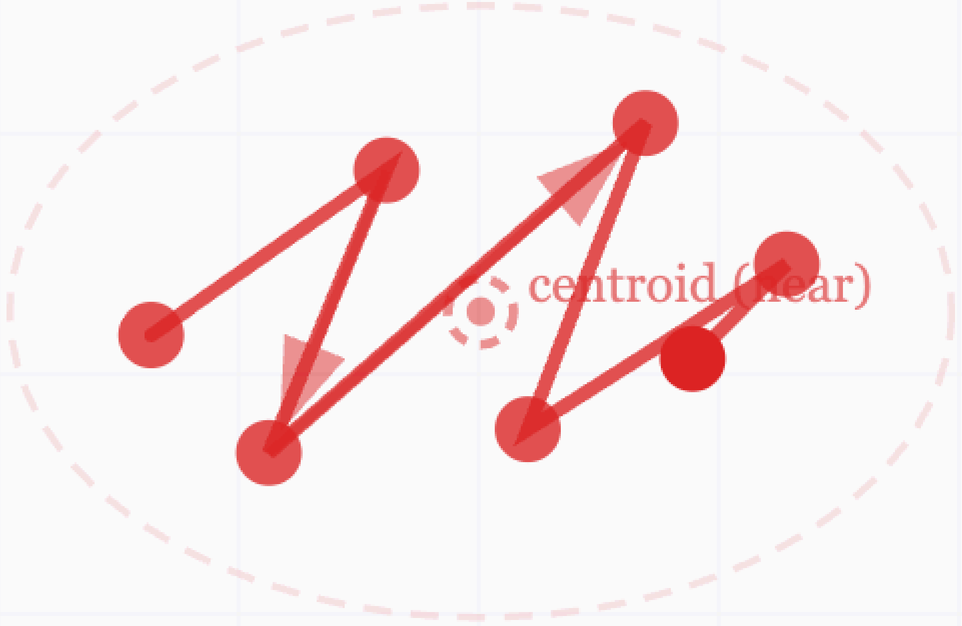


low velocity · low entropy
broad semantic space

↓ distance to next, velocity
↓ acceleration, entropy
↑ centroid distance

PD / bvFTD

Erratic · Constricted · Unpredictable



high velocity · high entropy
constricted semantic space

↑ distance to next, velocity
↑ acceleration, entropy
↓ centroid distance

↑ Higher in patients

Distance to Next — larger semantic jumps

Velocity — erratic movement

Acceleration — abrupt direction changes

Entropy — unpredictable search

↓ Lower in patients

Distance to Centroid — search confined to a tighter neighborhood despite being more volatile

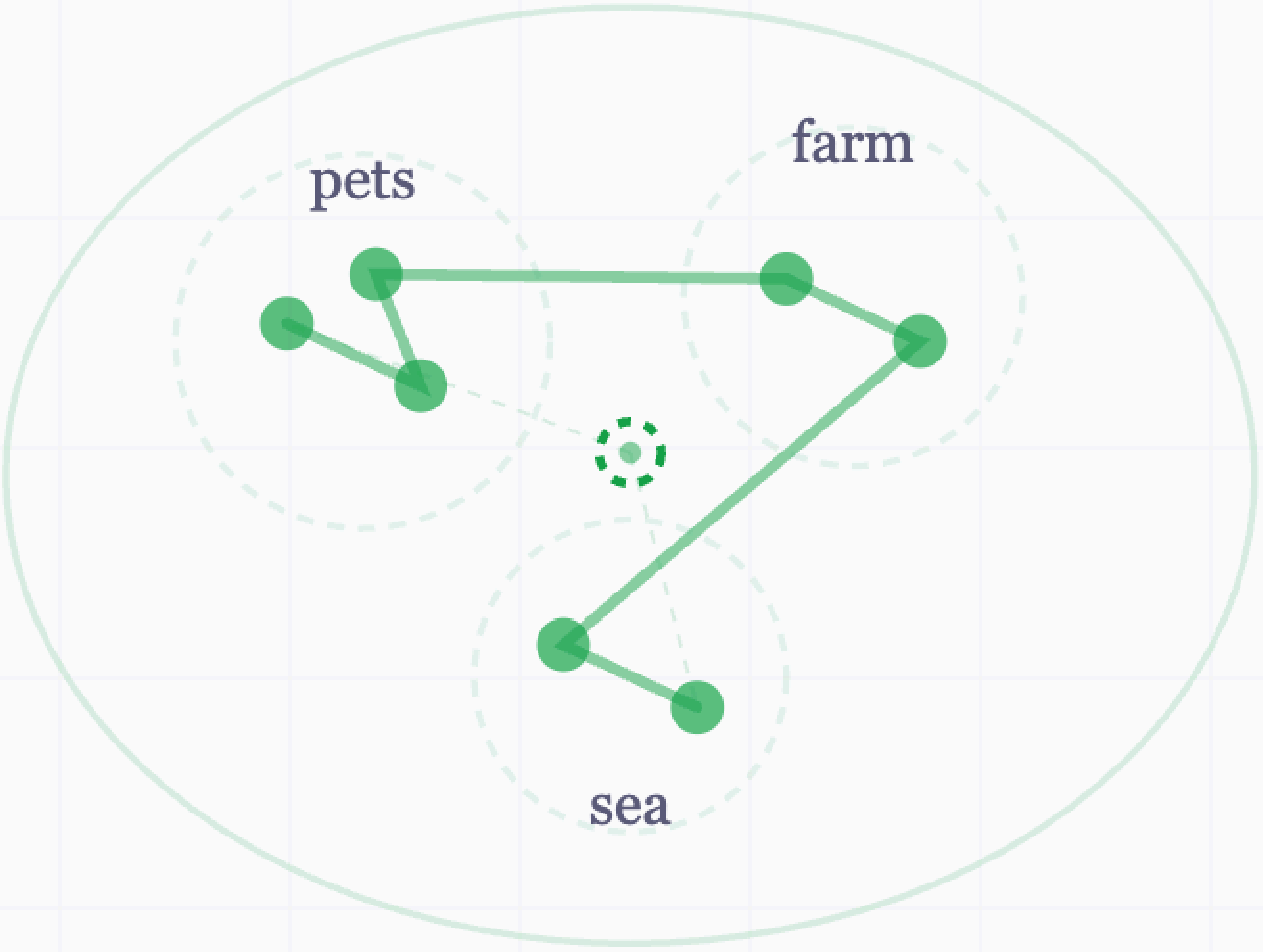
Interpretation: a kinematic signature of executive dysfunction — volatile trajectories within a diminished semantic space.

PD and bvFTD did not differ from each other

RESULTS

Swear words & semantic category structure

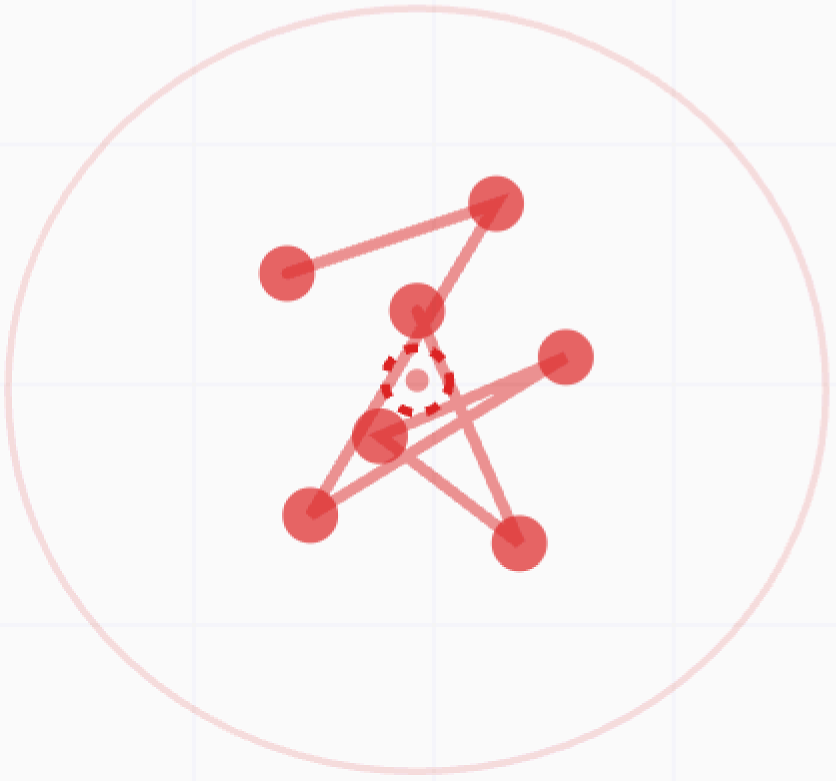
Animals



structured sub-categories
→ organized, broad

↓ velocity, acceleration
↓ entropy
↑ centroid distance
(most organized)

Swear Words



no clear sub-structure
→ compact, erratic

↑ velocity, acceleration
↑ entropy
↓ centroid distance
(most erratic)

Animals

Lowest kinematic values, highest centroid distance. Structured sub-categories (pets, farm, sea...) allow **organized exploitation**.

Swear Words

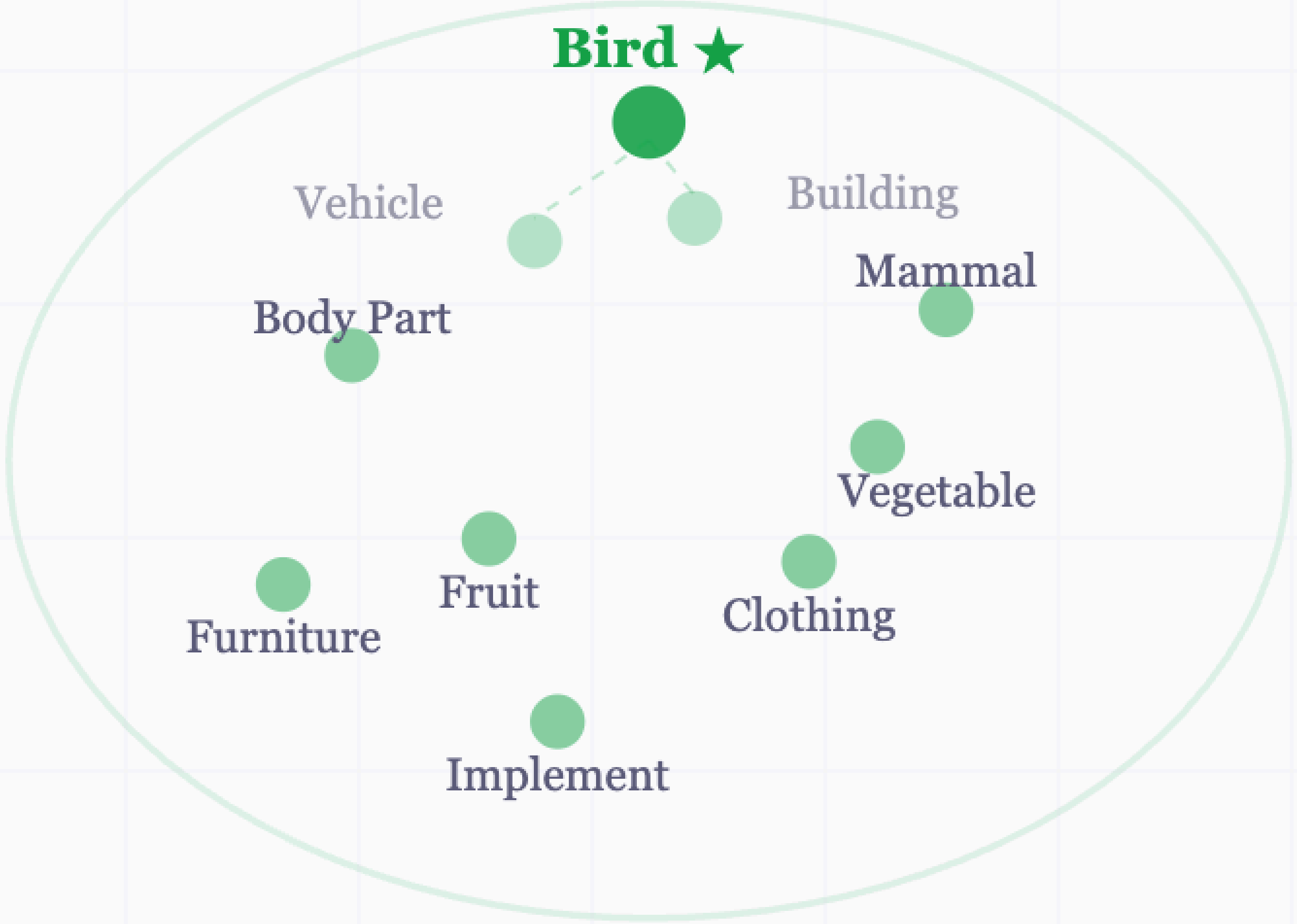
Highest kinematic values across all metrics. Taboo lexicons lack sub-category structure → **high variability in a compact space**.

RESULTS

Cross-linguistic structure: Italian & German

Italian (IT)

N = 69 · 10 categories · Property Listing



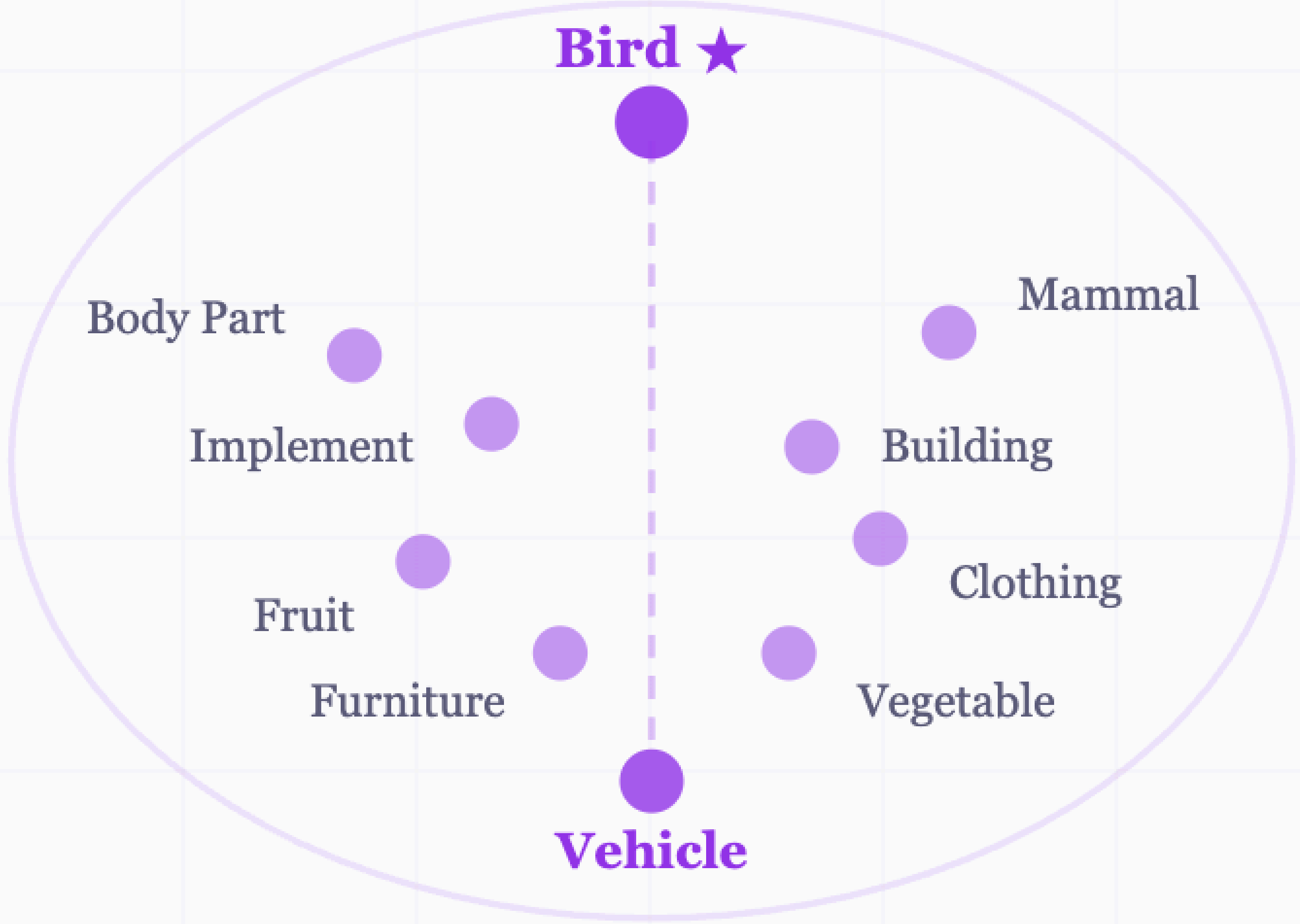
Bird = reference category
Building, Vehicle least separated

Italian findings

63–84 sig. pairs (cumul.) · 70–107 (non-c.)

German (DE)

N = 73 · 10 categories · Property Listing



Bird = reference category
Vehicle most distinct

German findings

80–112 sig. pairs (cumul.) · 111–137 (non-c.)

Separation based primarily on distance-to-next and velocity · positions are illustrative

Italian (IT) · N=69

10 categories · Property Listing Task. Bird is reference. Building and Vehicle **least separated**. Category effects *selective*, not uniform.

German (DE) · N=73

Same 10 categories · Same protocol. Vehicle **most distinct** from Bird. Bird highest velocity & acceleration. Fruit vs. tool-like categories show strongest centroid gaps.

→ Same protocol, but different category effects — *cultural and linguistic structure* shapes how meaning is organized.

KEY FINDING

Cumulative vs. non-cumulative: trajectory length matters

When does each approach win?

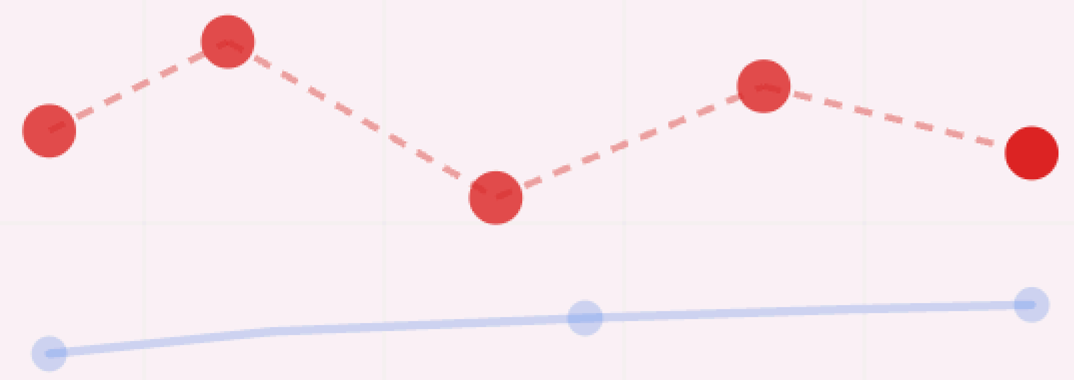
Long Trajectories

Neurodegenerative (≈20 items) · Swear Fluency (≈21)

✓ Cumulative wins — more context

Short Trajectories

Italian (≈5 items) · German (≈5.5)



✓ Non-cumulative wins — less noise

Summary — Significant Differences (Table 2)

	Neuro (long)	Swear (long)	Italian (short)	German (short)
▲ Cumul.	**7–10**	**43–46**	63–84	80–112
▼ Non-c.	4	37–46	**70–107**	**111–137**
	▲ cumul. wins	▲ cumul. wins	▼ non-c. wins	▼ non-c. wins

Number of significant pairwise differences (Tukey HSD) across 4 embedding models. Range shows min–max across models.

Cumulative → long sequences

Longer trajectories provide rich context that cumulative embeddings leverage — **more significant group differences** and higher effect sizes.

Non-cumulative → short sequences

With only ~5 items, there is **too little context** to accumulate. Point-to-point variation retains more discriminative signal.

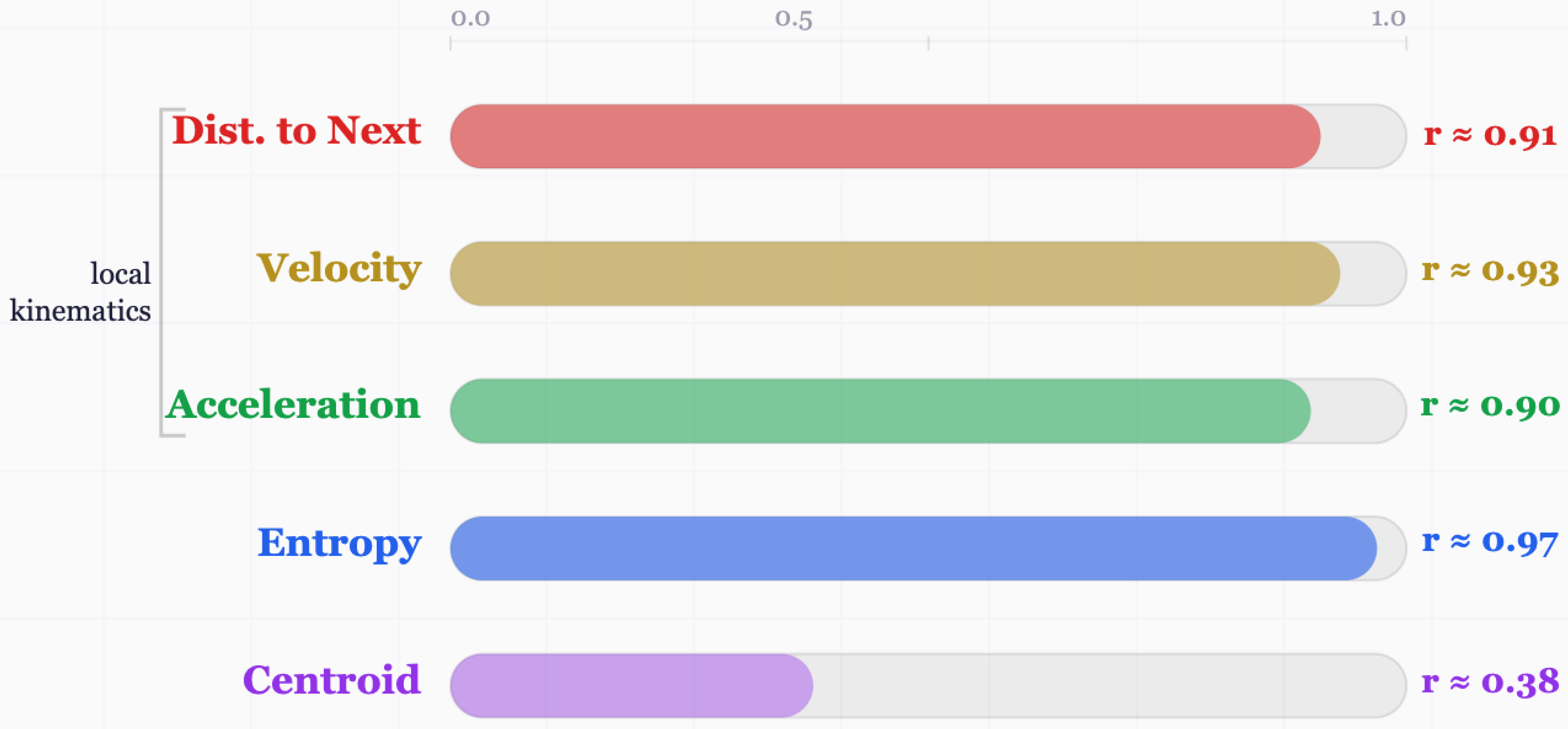
→ Both approaches are *complementary* — trajectory length determines which representation best exposes semantic structure.

ROBUSTNESS

Different models, convergent geometry

Inter-model Pearson r per metric

across the three cumulative transformer encoders (fastText non-cumulative baseline excluded)



Why does centroid diverge?

- Sensitive to embedding anisotropy
- Model-specific "rogue dimensions" distort global centroids
- Potential tool for comparing **how models structure knowledge**

✓ Robust across models

Velocity, Acceleration, Dist-to-Next — high Pearson *r* across encoders. Local dynamics are *encoder-invariant*.

✓ Entropy: near-perfect

Depends on *rank ordering*, not absolute distances. Median binarization absorbs differences.

⚠ Centroid: model-dependent

Sensitive to each model's global geometry. Most pronounced in the neurodegenerative dataset.

→ Models learn *similar local dynamics* despite different architectures (causal vs. bidirectional).

CONCLUSION

Towards a geometric framework for human semantic navigation

① Cumulative trajectory metrics build on traditional clustering/switching and capture fine-grained navigation dynamics.

② Clinical utility — extends existing evidence on semantic navigation signatures associated with neurocognitive dysfunction.

③ Cross-linguistic & cross-category — discriminates semantic categories and reveals language-specific organization.

④ Model-robust — convergent results across three different embedding architectures for local dynamics.

Future: temporal timestamps · more mathematically robust (potentially non-Euclidean) metrics

Paper

UFABC

ICLR International Conference On Learning Representations

Université de Montréal

Code