
Response to Reviewer #1

> Zheng and colleagues re-analyze a prior fMRI dataset (from what I understand, anyway, this was a re-analysis) to test the neural localization of temporal association vs semantic similarity relations. In the study, participants were exposed to a sequential transition structure among object images, creating temporally closer and farther relations among them. These objects were also separately rated for semantic similarity. Temporal and semantic distances were used to predict neural adaptation among image pairs. nearby but non-overlapping peaks in the medial temporal lobe were found to encode temporal and semantic distances.

We thank the reviewer for their helpful feedback. We have addressed their comments below point by point.

> 1 intro: some points of clarification needed
>
> 1.1- the idea of a 'cognitive map' seems to be a central part of the stated hypothesis and conclusion. However, what defines or even characterizes a map, specifically, is not stated here. As far as I can tell, what is investigated is simply distances in temporal relational or semantic space. Is this correct? E.g, - "a map of semantic relationships" -- does this just mean, semantic distances, given what was measured? Historically the term cognitive map has additional meaning beyond distances or relational links, no? Not sure why this concept needs to apply here.

We would like to thank the reviewer for highlighting the importance of clarifying our use of the term “cognitive map”. To address these concerns:

Cognitive map definition: Cognitive maps are conceptualized as domain-general internal models of our environment representing relational knowledge that helps us understand the world around us (Eichenbaum & Cohen, 2014). These maps allow for generalization and inference.

Characteristics: Cognitive maps can be high-dimensional, with each dimension possessing an inherent metric. The exact metric may differ between dimensions—e.g. embodying Euclidean distances or resembling a city-block metric. The key feature uniting these metrics is the proximity of similar stimuli within this cognitive space (Gärdenfors & Zenker, 2015), symmetry and conformity to geometric norms, most notably betweenness and equidistance (Bellmund et al., 2018; Gärdenfors, 2004). These features enable straightforward computations of distances between any pairs of states without the need for expensive step-by-step simulations as well as generalization, because a property of two stimuli x and y can be inferred to be shared by any stimulus z falling between x and y (Bellmund et al., 2018).

Our findings: Our observed repetition suppression signal in the hippocampus scales with semantic distance and a measure of the graph structure. This representation aligns with the defining features of a cognitive map: Relationships can be quantified in terms of a metric, this metric is symmetric (see also point 2.3 below) and it adheres to geometric norms. To address your specific query on "a map of semantic relationships": While we indeed measure semantic distances, the organization, relational structure, and generalizable attributes of this knowledge make it more than mere distances; it resembles important properties of a cognitive map.

Nonetheless, we now use the term “map-like” throughout the manuscript as a more nuanced way to describe our findings. We believe that this more clearly reflects that the representation possesses some important, but not necessarily all, characteristics of a cognitive map.

We also included a paragraph to clarify our definition of a cognitive map and its application to our findings more clearly. We believe that these clarifications and modifications enhance the clarity of our message. We now say:

“The hippocampal-entorhinal system builds rich models of the world, called cognitive maps, that account for the relationships between locations, events, and experiences (e.g., Behrens et al., 2018; Eichenbaum & Cohen, 2014; Moser, Kropff, & Moser, 2008; O’Keefe & Nadel, 1978; Tolman, 1948). These maps capture the similarity between symmetric, high-dimensional relationships in a cognitive space, satisfying geometric constraints such as betweenness and equidistance (Bellmund et al., 2018; Gärdenfors, 2004). Abstracting and organizing relational information in this way facilitates flexible behavior, enabling generalization and inference.” (Introduction, page 2)

“Specifically, we observed that repetition suppression of signals in the hippocampus scales with semantic distance. This representation aligns with the defining features of a cognitive map: Relationships can be quantified in terms of a metric, this metric is symmetric and it adheres to geometric norms (Bellmund et al., 2018; Gärdenfors, 2004; Gärdenfors & Zenker, 2015).” (Discussion, page 20)

“Importantly, while both map-like structures localized to the hippocampal formation, the semantic map was located in more posterior regions of the hippocampal formation than the transition structure and thus anatomically distinct.” (Abstract, page 1)

“Notably, although both map-like structures were represented in the hippocampal-entorhinal system, the semantic map was localized in more posterior regions than the transition structure.” (Introduction, page 4)

“For example, when participants acquire new knowledge about the relationships between objects by being exposed to experimentally generated object sequences, the hippocampal formation extracts the associated transition structure and stores it as map-like structural representations (Garvert et al. 2017).” (Discussion, page 20)

References:

Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function?. Neuron, 83(4), 764-770. http://dx.doi.org/10.1016/j.neuron.2014.07.032

Gärdenfors, P., & Zenker, F. (Ed.).(2015). Applications of Conceptual Spaces: the Case for Geometric Knowledge Representation. Cham: Springer Verlag. https://doi.org/10.1007/978-3-319-15021-5

Gärdenfors, P. (2004). Conceptual spaces: The geometry of thought. MIT press. https://doi.org/10.7551/mitpress/2076.001.0001

1.2 - p 3 first para, several unclear phrases:

- "were already embedded in semantic structures" -- what does that mean?

We have revised the sentence:

“In Garvert et al. (2017), participants acquired new relational knowledge about everyday objects which were already linked by semantic connections. Here, participants were exposed to object sequences following a pseudo-random walk along a graph.” (Introduction, page 3)

-"Here, participants were trained on object sequences" - trained to do what?

Thank you for highlighting that this sentence can be misunderstood. Participants were exposed to object sequences following a pseudo-random walk along a graph structure while learning to associate a random stimulus orientation with a button press. On the day of testing, they were exposed again to the object sequences and asked to report on 10% of the trials whether a gray patch had been presented on the preceding object.

We agree with the reviewer that some more clarification is needed here. We now say:

"Here, participants were exposed to object sequences following a pseudo-random walk along a graph." (Introduction, page 3)

"On the day of training (day 1), participants were exposed to object sequences in an implicit learning task. The object transitions followed a pseudo-random walk along a graph (Figure 1A) which was unknown to the participants. This means that each object could only be followed by an immediate neighbor in the graph structure. Participants performed a behavioral cover task, in which they learned to associate a random stimulus orientation with a specific button press…In the scanning session (day 2)...In 10% of the fMRI trials, participants performed an unrelated cover task, reporting whether a gray patch had been present on the preceding object (Figure 1B)." (Methods, page 5)

- "matched the stimuli used in Garvert et al. (2017) with photographs of the same objects" - why not the exact stimuli? (I see this is explained later, but here it is confusing without any context)

Thank you, we have revised the text:

"...Specifically, we constructed a model of object similarity that isolates the semantic relationships reflecting high-level conceptual knowledge acquired from experience from the low-level perceptual attributes of specific objects (Rosch & Lloyd, 1978; Tversky, 1977). To this end, we matched the stimuli used in Garvert et al. (2017) with photographs of the same objects and asked a separate participant population to assess their similarity using a triplet odd-one-out task (Hebart, Zheng, Pereira, & Baker, 2020)." (Introduction, page 3)

-"that precisely localized to" -- odd phrase, "was precisely localized in"?

Thank you for pointing this out, we have revised this accordingly.

2.1 Methods: many details missing or not clearly described

2.1 some details of fMRI study unclear

- It should be stated clearly if this the same fMRI dataset published in Garvert et al 2017. In this manuscript many details of fMRI experiment are missing; they should be summarized to make sure everything makes sense. E.g., reference are made to the patch task but the task was never introduced or described. Why was a different task used in the fMRI than training,and generally, motivation for it? Properties of trial durations, numbers, etc would be helpful to report.

We had stated in the abstract that this paper is the result of a re-analysis of the Garvert et al. (2017) data. We now added very explicit statements about this also to the introduction and the methods sections to make this clear:

“Here, we ask whether prior semantic knowledge about objects would be simultaneously mapped in the same hippocampal system which also represents knowledge about transition structure. We reanalyzed the functional magnetic resonance imaging (fMRI) data from Garvert et al. (2017). Specifically, we constructed a model of object similarity that …" (Introduction, page 3)

“We reanalyzed the data from the fMRI study by Garvert et al. (2017), where 23 human participants (15 male, 8 female, meanage = 23.5, SDage = 3.7, age range 18-31) were tested.…” (Methods, page 4)

Thank you for pointing out that our manuscript would benefit from a more explicit description of the tasks, we agree that this is important background information that should not be omitted. We have now added more information about the original study (incl. description of the patch task, explanation of a different task used in the MRI than training, properties of trial duration and numbers). We hope this now provides sufficient information about the task for the readers.

“On the day of training (day 1), participants were exposed to object sequences in an implicit learning task. The object transitions followed a pseudo-random walk along a graph (Figure 1A) which was unknown to the participants. This means that each object could only be followed by an immediate neighbor in the graph structure. Participants performed a behavioral cover task, in which they learned to associate a random stimulus orientation with a specific button press. For example, the left-facing motorcycle was linked to button F, while the right-facing motorcycle corresponded to button J. The graph structure was the same for all participants. The link distance between any pair of objects in the graph is defined as the minimum number of links between this pair of objects (e.g., in the example displayed in Figure 1A, the link distance between the rabbit and the leaf is two), which ranges from one to four. For each participant, a subset of 12 objects was selected from a total of 31 objects used in the study, and randomly assigned to the 12 nodes on the graph. The objects covered a wide range of semantic categories (e.g., furniture, plants, body parts, animals; see figure 2B, top rows for the full set of objects used). Only one object within a semantic category was assigned to a participant (e.g. either banana or strawberry, but not both) and each participant was assigned a unique set of objects. Participants were trained for 12 blocks, with 132 transitions in each block.

In the scanning session (day 2), 7 out of the 12 training objects were used and presented in randomized order to reduce the total number of stimulus–stimulus transitions and thereby increase statistical power for the fMRI adaptation analysis. The transitions no longer followed the graph structure, but were pseudo-randomized in such a way that each possible stimulus-stimulus transition occurred exactly ten times per block (no stimulus repetitions). To reduce the motor responses in the scanner, a different behavioral cover task was employed that was orthogonal to the imaging analysis of interest: In 10% of the fMRI trials, participants performed an unrelated cover task, reporting whether a gray patch had been present on the preceding object (Figure 1B). This means that participants were not required to pay active attention to the object identity. The fMRI session consisted of 3 blocks, with 420 transitions per block. Stimuli were presented for 1 s, with a jittered inter-trial interval generated from a truncated Poisson distribution with a mean of 2 s." (Methods, pages 4-5)

2.2 Points of clarification needed

- "trained on object sequences whose transitions followed a psuedo-random walk along a graph" -- I can make sense of this but I am not sure all readers would. Perhaps, e.g., images of objects appeared one by one in sequence, and the transitions between objects were governed by a probability such that..." etc. I think it is especially confusing to use the term "trained" when learning is implicit and they are doing an unrelated cover task.

Thank you for the suggestion, we agree that the term "trained" is misleading. We now clarified this point as follows:

"...participants were exposed to object sequences in an implicit learning task. The object transitions followed a pseudo-random walk along a graph (Figure 1A) which was unknown to the participants. This means that each object could only be followed by an immediate neighbor in the graph structure." (Methods, page 4)

-The cover task itself was not well explained. It was written that they must "associate a random stimulus orientation with a button press" - this is very unclear. Which orientations of which stimuli were associated with what buttons? It will help to walk the reader through concretely was the task was.

Together with our response to the reviewer’s comment 2.1, we have now added a more detailed description of the cover task to the manuscript. Specifically regarding the question of stimulus orientation: Participants learned to associate a random stimulus orientation with a specific button press. For example, the left-facing motorcycle was linked to button F, while the right-facing motorcycle corresponded to button J. We now say:

“On the day of training (day 1), participants were exposed to object sequences in an implicit learning task. The object transitions followed a pseudo-random walk along a graph (Figure 1A) which was unknown to the participants. This means that each object could only be followed by an immediate neighbor in the graph structure. Participants performed a behavioral cover task, in which they learned to associate a random stimulus orientation with a specific button press. For example, the left-facing motorcycle was linked to button F, while the right-facing motorcycle corresponded to button J.” (Methods, page 4)

-"the minimum number of links between the pairs of objects" -- were objects pre arranged in specific pairs? Or perhaps "any pair of objects" in the set?

Sorry for the misunderstanding. After randomly assigning objects to each participant, we computed the link distance for every two objects on the given graph. Indeed, we were referring to “any pair of objects” in the set. We have clarified it in the manuscript.

“The graph structure was the same for all participants. The link distance between any pair of objects in the graph is defined as the minimum number of links between this pair of objects (e.g., in the example displayed in Figure 1A, the link distance between the rabbit and the leaf is two), which ranges from one to four.” (Methods, page 4)

- "objects within the same semantic categories were not assigned to the same participant". It will help to first introduce what are the semantic categories, how many exemplars of each in the set, and then how they were assigned. Was it that all objects for a given subject came from different semantic categories?

The 31 items used in the original study (Garvert et al., 2017) cover a wide range of semantic categories (e.g., furniture, plants, body parts, animals; see Figure 2B for the full set of items). The assignment of items to participants (a set of 12 items per participant) is pseudo-randomized: objects within the same semantic categories were not assigned to the same participant and each participant received a unique set of objects. Below we provide two example sets of objects together with their semantic categories:

Objects - Semantic Categories (participant 1):	**Objects - Semantic Categories (participant 2):**
Chair - Furniture	Table - Furniture
Umbrella - Accessory	Bag - Accessory
Basket - Container	Bin - Container
Lightbulb - Electrical Device	Lightbulb - Electrical Device
Key - Tool	Broom - Tool
Book - Reading Material	Book - Reading Material
Bus - Vehicle	Truck - Vehicle
Flower - Plant	Leaf - Plant
Dog - Animal	Rabbit - Animal
Carrot - Vegetable	Strawberry - Fruit
Ear - Body Part	Ring - Jewelry
Blouse - Clothing	Shoe - Footwear

We have clarified this part of the experiment design in the manuscript. We now say:

"The objects covered a wide range of semantic categories (e.g., furniture, plants, body parts, animals; see figure 2B, top rows for the full set of objects used). Only one object within a semantic category was assigned to a participant (e.g. either banana or strawberry, but not both) and each participant was assigned a unique set of objects.” (Methods, page 5)

> 2.3 Distance measures for relational links vs transition probabilities.
>
> A major question I had was regarding how link distance relates to transition probability or even future discount state occupancy (as per SR). In Figure 1, it would help if the graph showed the transition probabilities between items. But furthermore, it would be ideal to have a probability-weighted distance matrix to understand how it compares with the link distance matrix. For example, if a 2-link path had probabilities .33 and .16, that is different than .16 and .25. It would be very important to understand why link distance is the right measure and how it compares to others. Moreover, these are quite different than values in an SR representation, which is also not reported.
>
> On the other hand, it was noted that "Indeed, the distance metric that best explained BOLD responses in Garvert et al. (2017) was “communicability”, a graph-theoretic measure capturing the distribution of future states in a graph that is closely related to the successor representation " -- could this metric be reported here? was it used? why was this notion only raised in the discussion, and is it the same or different as link distance? In short it would help to have a lot more quantified measures of temporal distance reported in the manuscript.

We would like to thank the reviewer for this suggestion. We believe that there are two separate points to address here.

1) **Transition probabilities between items.** It is indeed the case that transition probabilities between pairs of nodes differ even for nodes that are direct neighbors on the graph, a fact that is not well captured by the graph structure and the corresponding adjacency matrix, or the link distance measure we use here. However, we think it is a challenge to reflect this in the depiction of the graph itself since transition probabilities are directional. For example, the probability of transitioning from the ear to the motorcycle is not the same as the probability of transitioning from the motorcycle to the ear, because in the latter case there are more objects to transition to (Figure 1 A).

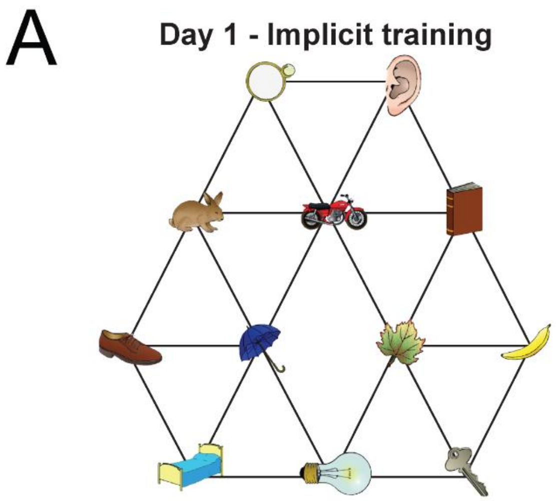
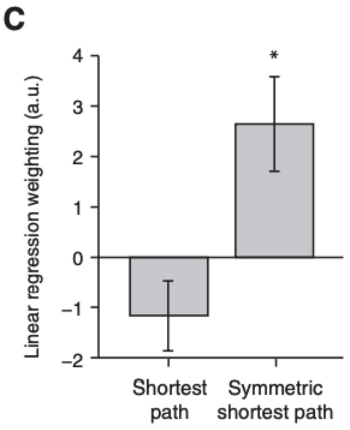


Figure 1A. An example graph structure used to generate stimulus sequences on day 1.

We would also like to refer the reviewer to an analysis that was reported in the original paper, where we explored the nature of the representation of the newly learnt graph structure in depth (Garvert et al. 2017). For example, we explicitly tested whether the symmetric link-distance measure or a non-symmetric shortest path measure based on actually experienced transitions explained our data better. We found clear evidence for a symmetric representation of the map:

Analysis reported in Garvert et al. 2017:



> “Furthermore, relationships between items arranged in a map-like structure are non-directional. Our subjects were not constrained to experience each pair of transitions an equal number of times (Figure 3B). Based upon this, we could test whether the fMRI signal was better predicted by the true or symmetrised distance between any two objects. We constructed a measure of the shortest path between each pair of objects according to the actual number of times each transition was experienced by a subject during training (see ’Materials and methods’ section). When allowing this measure to compete with its symmetrised, and thereby non-directional, self in a linear model, it was the symmetrised version alone that predicted the fMRI suppression effect (Figure 3C, t(22) = 2.78, p=0.01 and t(22) = 0.11).

We would therefore maintain our conclusion that the representations we find here are map-like.

2) **Probability-weighted distance matrix.** The differences in transition probabilities between pairs of nodes of course propagate especially for longer sequences along the graph structure. This can be captured by graph-theoretic measures we introduced in our previous paper such as communicability and the successor representation. Simply put, the communicability between two nodes in a graph captures the ease with which information can flow between them, taking into account not just the direct connections, but also the indirect connections through all possible paths. This measure considers all paths between nodes, exponentially discounting longer paths.

   Traditionally, communicability is computed from the unweighted adjacency matrix A of an unweighted graph as the (i,j)-th entry of the matrix exponential exp(A). However, it is equally possible to compute the same measure based on a matrix P reflecting transition probabilities, which could provide a more nuanced understanding of communicability in systems where the strength and directionality of connections (as given by transition probabilities) matter.

   The matrix exponential grows with the contributions of paths of increasing lengths in a super-linear manner. If two nodes have high communicability, it implies that there are many paths (both direct and indirect) connecting them. If the communicability is low, it indicates that the nodes are more isolated from each other. This measure can therefore capture the entire structure of a graph and all possible paths **in a manner similar to what the reviewer suggests here**. For example, it can differentiate between two nodes connected by a single long path and two nodes connected by many short paths.

The successor representation, on the other hand, reflects expected future state visitations for any given state given a transition probability matrix. A random-policy SR can be computed from the transition probability matrix using the equation $M=(I-\gamma T)^{-1}$ where M is the successor representation matrix, I is the identity matrix, T is the transition probability matrix and γ is the discount factor (here set to 0.85).

The different measures (shortest path, transition probabilities, communicability and successor representation) capture slightly different aspects of the graph structure, as visualized in Figure R1:

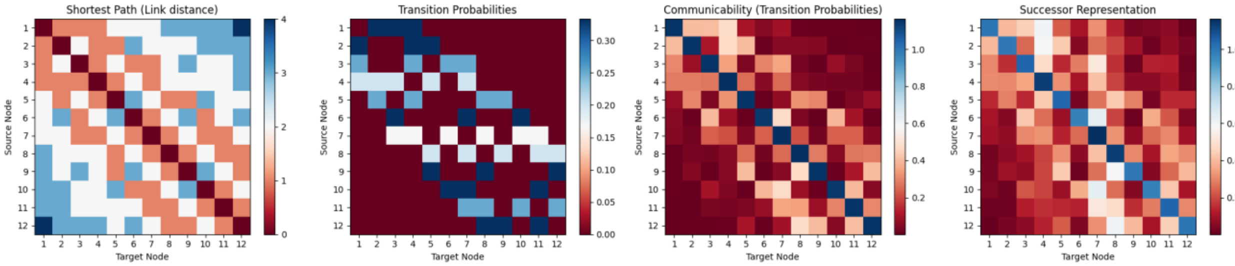


Figure R1. Different distance measures (shortest path, transition probabilities, communicability and successor representation) capture slightly different aspects of the graph structure

However, due to the high correlations among the different distance metrics (Spearman r for correlations between link distance and communicability/SR > -0.9, p < 0.0001, Spearman r for correlation between communicability and SR = 0.97, p < 0.0001), it is impossible to isolate the contribution of each precise measure to the neural representation. Indeed, when we include communicability instead of link distance as a regressor in the same GLM as semantic distance, we find suprathreshold clusters in exactly the same region as when link distance effects are included (Figure R1), but this effect does not survive correction for multiple comparisons.

Since we explored communicability and SR measures in our previous article in depth, we decided not to include those analyses here. While maintaining link distance as the measure of the learnt graphs structure for simplicity, we toned down our statement about the distance metric that “explained our data best”. In addition, we added a paragraph in the discussion pointing out that there are other plausible graph-theoretic measures such as communicability and SR that might be represented neurally, but that are challenging to isolate. We now say:

“It is also worth noting that there are other plausible measures that might better characterize the neural representation of the transition structure, which are discussed comprehensively in Garvert et al. (2017). In the domain of reinforcement learning, the utility of a cognitive map is greatly enhanced when the representation of a state not only embodies the present but is also predictive, encompassing a spectrum of probable future states. This concept is encapsulated in the successor representation (Dayan, 1993; Momennejad et al., 2017; Russek et al., 2017), which is suggested to be encoded by hippocampal place cells (Stachenfeld et al., 2014, 2017). From this perspective, hippocampal place cells are posited to encode not the immediate location of an animal, but a predictive array of forthcoming locations. Such a representation is advantageous for reinforcement learning, as it amalgamates predictive insights of future states with reward information, thereby facilitating the swift computation of potential navigational paths (Baram et al., 2017; Dayan, 1993; Momennejad et al., 2017; Russek et al., 2017). Analogous to the successor representation, graph theory introduces the matrix resolvent as a means to quantify 'communicability' or the closeness between nodes. Similarly, the matrix exponential, another graph theory measure, computes a weighted summation over future states and exhibits versatility across various dimensions and contexts (Estrada & Hatano, 2008). Both the successor representation and these graph-theoretic measures explain the fMRI adaptation effects observed by Garvert et al. (2017). Nonetheless, disentangling their unique neural contributions presents a challenge, primarily due to the high intercorrelations among these distinct distance metrics.” (Discussion, pages 24-25)

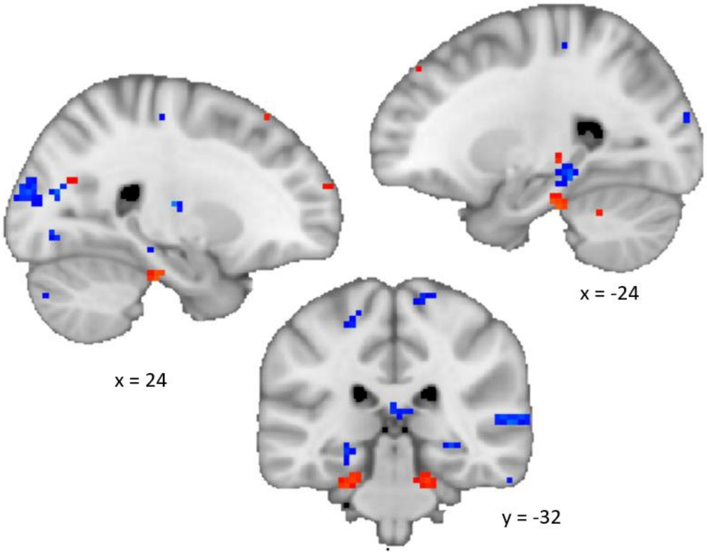


Figure R2. Whole-brain analysis shows a decrease in fMRI adaptation with semantic distance (blue) and communicability (red). These two effects also form an anatomical gradient along the anterior-posterior axis of the hippocampal formation. Communicability - instead of link distance - , semantic distance and residual distance are included in the model. Figure thresholded at p < .01, uncorrected for visualization.

References:

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. Nature human behaviour, 1(9), 680-692. https://doi.org/10.1038/s41562-017-0180-8

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS computational biology, 13(9), e1005768. https://doi.org/10.1371/journal.pcbi.1005768

Baram, A. B., Muller, T. H., Whittington, J. C., & Behrens, T. E. (2018). Intuitive planning: global navigation through cognitive maps based on grid-like codes. BioRxiv, 421461. https://doi.org/10.1101/421461

Estrada, E., & Hatano, N. (2008). Communicability in complex networks. Physical Review E, 77(3), 036111. https://doi.org/10.1103/PhysRevE.77.036111

2.4 The semantic distance matrices varied more between subjects than link distances; it should be tested how much more variance there was in the semantic model across subjects, and within the models within subjects. This affects the detectability of model fits. Furthermore, what was the range of correlation values between semantic disntance matrices and link disatnce matrices across subjects? This kind of information is important in understanding why semantic distance effects were not seen in some expected areas.

Thank you for pointing out the need for further analysis regarding the variance in semantic distance measures across subjects. Indeed, unlike the link distance, the semantic similarity measure depends on the specific set of objects participants received and therefore differs across subjects. We agree that understanding this variance can provide insights into the detectability of model fits.

Inter-subject variance: To address the difference in variance between participants, we now performed a Bartlett variance test on participants' object dissimilarity measures. The results show that the variance across subjects is not significantly different (chi square (22) = 11.99, p = .958). The semantic distances used in our analysis are computed as the **z-scored**, shared variances between the object similarity measures of two datasets. The variance of semantic distance is also not different across participants (chi square(22) = 4.65, p = .1). The figure below shows the object similarities across participants and link distances which are identical for all participants.

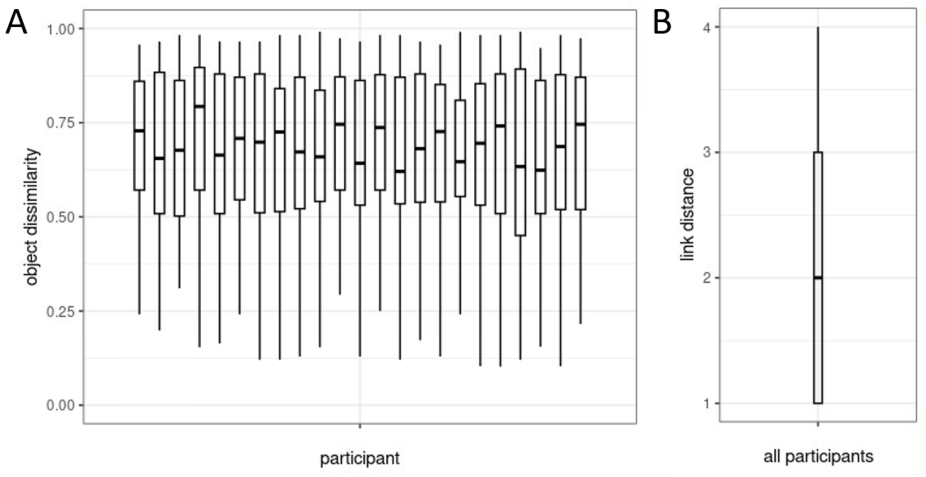


Figure R3. Dissimilarity measures. (A) Object dissimilarities depend on the unique set of items assigned to individual participants, therefore varied slightly across participants. Each box represents one participant. (B) link distances range from 1-4 and are identical for all participants.

Correlation between matrices: To determine the range of correlation values between semantic distance matrices and link distance matrices across participants, we also computed pairwise correlation between the link distance and semantic distance for each participant. The range of correlation coefficients (Spearman's Rho) was found to be between -0.25 and 0.30 (mean = .03, SD = .12, t(22) = 1.04, p = .31).

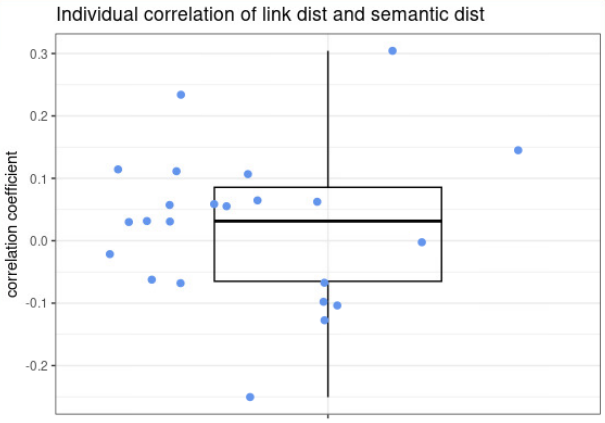


Figure R4. Correlation between semantic distance matrices and link distance matrices across participants. Each dot represents one participant.

Intra-subject variance: When modelling our fMRI data, we z-scored both dissimilarity matrices. Therefore, the standard deviation across all matrix entries within individual subjects is 0 or close to 0.

Relationship between matrix correlation and observed effects: To further explore how the correlation between semantic and link distance matrices might relate to the observed effects, we conducted an analysis comparing the strength of the correlation with the magnitude of effects we identified in our original analyses for each participant. Specifically, we computed two linear regressions with the correlation values (Figure R4) as the predictor and the magnitude of the semantic and link distance effects extracted from the ROIs (Figure 4A) as the dependent variable. Our findings indicate that neither the semantic distance effect (t(22) = 1.53, p = .140) nor the link distance effect (t(22) = 0.59, p = .560) can be explained by the correlation between the two distance matrices for individual participants.

Together, we think that these analyses demonstrate that the larger variance in the semantic matrices does not influence the detectability of effects.

We also reported this in the manuscript:

"Neither the semantic distance nor the residual distance was correlated with link distance (semantic: Spearman's Rho mean = .03, SD = .12, range = -.25 – .30, t_{22} = 1.04, p = .31; residual: Spearman's Rho mean = -.03, SD = .09, range = -.22 – .15, t_{22} = -1.52, p = .14)." (Results, page 13)

3. Results - several clarifying questions.

3.1 Why is an initial thershold applied at the cluster level while the FWE correction is applied to peaks?

We thank the reviewer for pointing this out. Indeed, given that we applied FWE correction on the peak level, the final p-value is independent of the cluster forming threshold. The initial threshold is purely for visualization purposes. We have now corrected this in the manuscript:

"We expected both the semantic information and the transition structure to be mapped in the hippocampal formation. Therefore we focused our analysis on this region ... We consider our results significant if they survived family-wise error (FWE) correction at the peak-level of $p < .05$ within this anatomically defined mask (small volume correction, SVC)... Activations in other brain regions were only considered if they survived whole-brain peak-level FWE correction at $p < .05$. All statistical parametric maps visualized in the manuscript are thresholded at $p < .01$ uncorrected and unmasked for illustration." (Methods, page 10)

3.2 It is notable that prior work has identified semantic representations, specifically about objects, in perirhinal cortex. It would be crucial to cite this work and point out that PRC is very close to the ROI used here. Could the authors comment on how the HC-EC complex was defined, and whether it spanned or excluded PRC? In general, more detail about ROI definition would be helpful.

Thank you for this suggestion, we concur that the perirhinal cortex's significance in semantic representations, especially related to objects, warrants discussion.

The entorhinal-hippocampal mask is created using Freesurfer segmentation in MNI space, combining bilateral hippocampus and bilateral entorhinal cortex. Due to the fine resolution of the entorhinal label in Freesurfer, it was further dilated by one voxel to have non-disrupted substructures. The perirhinal cortex is however not included in this mask, even though it is indeed proximate.

A significant challenge with imaging the perirhinal cortex is its location beneath the hippocampal formation and adjacency to the air-brain tissue interface. This position leads to pronounced fMRI signal drop-out, making it very difficult to obtain reliable data. As a consequence, we could not examine our effects of interest in this ROI and are cautious about making claims regarding semantic representations in the PRC.

In the methods section, we have now clarified our ROI definition both for the hippocampal ROI and the cortical ones:

"We expected both the semantic information and the transition structure to be mapped in the hippocampal formation. Therefore we focused our analysis on this region. The anatomical mask is created using Freesurfer (Fischl, 2012) segmentation in MNI space, combining bilateral hippocampus and bilateral entorhinal cortex (Supplementary Material S2). " (Methods, page 10)

"To explore the cortical semantic representation, we performed additional SVC using two anatomically defined masks: the left anterior temporal lobe and the left angular gyrus, two regions previously reported to be important in semantic processing (Visser et al., 2010; Humphreys et al., 2021). Both masks are defined using the Harvard-Oxford cortical structural atlas with a probabilistic threshold of 30%." (Methods, page 10)

Lastly, we integrated the mention of prior work on object-specific semantic representations in our discussion:

"It is also worth noting that object-specific semantic representations have been identified previously in the perirhinal cortex (Clarke & Tyler, 2014), a cortical region close to the hippocampal formation. However, due to fMRI signal drop-out in this region, we could not examine whether our effects of interest are also represented there." (Discussion, page 22)

4. Discussion: a few limits on interpretation

4.1 Several factors differed between the relation types: recency of learning, kind of relation/distance (semantic/taxonomic vs temporal association), and whether it was explicitly known or implicitly known. For example it is not just that semantic knowledge is older, but it is also not neccesarily based on temporal contingiuty (e.g., taxonomic object categories). Which is mostly likely the driver of the localization differences? It may be worth refering to prior related findings comparing such factors. For example, perceptual feature information and newly learned temporal relations show effects in adjacent parts of lateral temporal areas, when both are explicitly known:

Leshinskaya, A., & Thompson-Schill, S. L. (2020). Transformation of event representations along middle temporal gyrus. Cereb Cortex, 30(5), 3148–3166. https://doi.org/10.1167/19.10.91a

Other work compares taxonomic vs semantic relationships in the brain:

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and Thematic Semantic Systems. Psychological Bulletin, 143(5), 499–520. https://doi.org/10.1037/bul0000092

And I am less familiar on contrasts between implicit vs explicit knowledge but it might help to refer to other principles of HC long-axis organization:

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z. X., Grady, C., Rosenthal, C. R., Winocur, G., Barense, M. D., & Moscovitch, M. (2018). Multiple Scales of Representation along the Hippocampal Anteroposterior Axis in Humans. Current Biology, 28(13), 2129-2135.e6. https://doi.org/10.1016/j.cub.2018.05.016

Perhaps there is a broader principle behind such organization across the brain. It would help if the authors could elaborate on what they think, based on prior literature, is the most likely driver of the spatial segregation they see.

We would like to thank the reviewer for their thoughtful feedback and for providing relevant literature to further contextualize our findings.

We agree with the reviewer that the two types of knowledge structure investigated in the current study differ in many aspects, including the recency of learning, the nature of the relation/distance, and the degree to which knowledge is explicit or implicit. While it is exciting to see two knowledge structures that differ vastly to be encoded using the same cognitive mapping principle, it is also hard to isolate the driving factor behind the anatomical separability of the representations. Therefore, we concur that these factors could have contributed individually or collectively to the observed spatial segregation.

We now discuss these different potential explanations in depth in light of the previous literature to provide a more nuanced and informed discussion. We believe that integrating these prior findings enhance our understanding of the underlying neural mechanisms and the broader organizational principles at play. We now say:

“Several previous investigations should also be mentioned here. Leshinskaya & Thompson-Schill's (2020) suggested that perceptual features, newly acquired associations as well as generalizable relational knowledge manifest in neighboring regions of the lateral temporal areas. However, in contrast to our own observations, the authors did not find any evidence of associative coding in medial temporal lobes or the hippocampus. In addition, Mirman et al. (2017) report a neural dissociation between taxonomic and thematic semantics across a set of studies (e.g., Davey et al., 2016; Kalénine & Buxbaum, 2016; Schwartz et al., 2011). These studies suggest that anterior temporal lobes (ATL) predominantly encode taxonomic semantic knowledge and the temporo-parietal cortex (TPC) encodes thematic semantic processing. However, the literature on this neural observation is by no means conclusive, and many studies, including our own, do not echo this ATL-TPC dissociation.

Our findings suggest a more integrative role for the hippocampus, accommodating various types of relational knowledge, both taxonomic (semantic) and associative/temporal (transition structure)(Peer et al. 2021), underscoring the dynamic and flexible nature of hippocampal codes. This is further supported by the anatomical gradient reminiscent of the gradient observable in the scale of hippocampal spatial codes, where anterior parts of the hippocampus display coarser spatial codes than posterior parts of the hippocampus (Strange et al. 2014, Brunec et al., 2018; Poppenk et al., 2013). This hints at broader organizational principles within the hippocampus.

The anatomical separability we report could also be attributable to the temporal disparity in the acquisition and consolidation of semantic relationships versus newly learned relations. Semantic relationships, built and reinforced over a lifetime, have undergone extensive consolidation processes, perhaps resulting in more stable and distinct neural representations within the hippocampus. In contrast, relationships acquired over a short duration, such as those from a single training session, might still be in the early phases of consolidation (Squire et al., 2015; Walker & Stickgold, 2004). In short, several features differ between the two relational structures in our study, including the recency of learning, the nature of the type of relational knowledge, and the degree to which knowledge is explicit or implicit. The observed spatial segregation in the hippocampus is likely driven by a combination of these features, potentially reflecting the nature of the encoded information." (Discussion, pages 23-24)

References:

Leshinskaya, A., & Thompson-Schill, S. L. (2020). Transformation of event representations along middle temporal gyrus. Cereb Cortex, 30(5), 3148–3166. https://doi.org/10.1167/19.10.91a

Brunec, I. K., Bellana, B., Ozubko, J. D., Man, V., Robin, J., Liu, Z. X., ... & Moscovitch, M. (2018). Multiple scales of representation along the hippocampal anteroposterior axis in humans. Current Biology, 28(13), 2129-2135. https://doi.org/10.1016/j.cub.2018.05.016

Mirman, D., Landrigan, J.-F., & Britt, A. E. (2017). Taxonomic and Thematic Semantic Systems. Psychological Bulletin, 143(5), 499–520. https://doi.org/10.1037/bul0000092

Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., ... & Jefferies, E. (2016). Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. Neuroimage, 137, 165-177. https://doi.org/10.1016/j.neuroimage.2016.05.051

Kalénine, S., & Buxbaum, L. J. (2016). Thematic knowledge, artifact concepts, and the left posterior temporal lobe: Where action and object semantics converge. Cortex, 82, 164-178. https://doi.org/10.1016/j.cortex.2016.06.008

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., ... & Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. Proceedings of the National Academy of Sciences, 108(20), 8520-8524. https://doi.org/10.1073/pnas.1014935108

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. Neuron, 44(1), 121-133. https://doi.org/10.1016/j.neuron.2004.08.031

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. Cold Spring Harbor perspectives in biology, 7(8), a021766. https://doi.org/10.1101/cshperspect.a021766

4.2 p 19 argues that the findings reveal that newly learned relations and prior semantic knowledge are "organized in similar ways". I am not sure if this claim is warranted - what do the authors mean by "organized"? What is shown is that they are represented in nearby areas. I think it is difficult to argue here and below on the basis of a "shared neural system" for these two relation types. By what criteria are two peaks in the "same neural system" and is there justification for this claim?

Thank you for highlighting this aspect. By stating that the two knowledge structures are “organized in similar ways” in “the same neural system”, we mean both relational structures are represented in the hippocampal-entorhinal system using a cognitive mapping framework, strengthening the notion that cognitive mapping is a universal, domain-general organizing principle for relational information (Eihenbaum & Cohen, 2014). We show that not only are both types of relational information represented in this neural system, but they also follow the same geometric distance coding principles. This implies that the hippocampus might use a consistent framework for representing relational knowledge, regardless of the precise nature of the knowledge and its mode or timescale of acquisition.

However, we recognize that while our findings suggest this commonality, they do not conclusively determine that both types of knowledge are organized identically. To clarify this in our manuscript, we've refined our statement to:

“Specifically, we observed that repetition suppression of signals in the hippocampus scales with semantic distance. This representation aligns with the defining features of a cognitive map: Relationships can be quantified in terms of a metric, this metric is symmetric and it adheres to geometric norms (Bellmund et al., 2018; Gärdenfors, 2004; Gärdenfors & Zenker, 2015). Not only are both knowledge structures mapped in the hippocampal-entorhinal system, they also both adhere to geometric coding principles whereby similar states are represented more similarly. This suggests that different types of relational knowledge, regardless of whether that knowledge was gathered over short durations or over a lifetime, might be structured within a similar cognitive mapping framework in the hippocampus.”

(Discussion, page 20)

Response to Reviewer #2

In their research, Xiaochen et al. examine the representation of objects characterized by various relational aspects within the hippocampal-entorhinal system. Through the reanalysis of fMRI data originally presented by Garvert et al., 2017, they identified distinct cognitive maps within the hippocampal formation. Specifically, one map emphasizes transition structures, whereas a more posteriorly situated map reflects semantic relations. This clear dissociation underscores the capability of the hippocampal-entorhinal system to construct diverse cognitive maps.

The authors present an insightful question. Their exploration of hippocampal long-axis differences is timely and should resonate with those in hippocampal research. While I have noted some queries below, I am confident they can all be addressed.

We would like to thank the reviewer for this positive assessment of our manuscript. We have addressed their comments below point by point.

Major Issues:

1. In the study conducted by the authors, the behavioral task included participants ranging in age from 20 to 70 years old. However, the fMRI task was limited to participants aged 18 to 31 years. This discrepancy in age ranges brings up concerns about potential variations in prior semantic knowledge across the age groups. It would be constructive if the authors could address this aspect. Do they consider it a limitation in their study?

We thank the reviewer for bringing up this point. To address this concern, we analyzed our behavioral data by splitting our participants into an “old” group (age range = 32-70, 49.1 ± 10.0 years, N = 104) and a “young” group consistent with the age group of our fMRI participants (age range = 20-31, 26.8 ± 2.6 years, N = 24). We computed the similarity matrix for each group separately, as well as for the entire cohort (i.e., the similarity measure we used in our manuscript). Our results show that the similarity ratings were highly consistent across the young and old groups and with the full sample (r=0.92, $p < 0.001$ for all vs. young and r=0.88, $p < 0.001$ for young vs. old, Figure R5) demonstrating that similarity ratings are closely aligned across the age groups and the age discrepancy did not bias the results.

We opted to use the full sample for our final analyses for a key methodological reason: in the young group, not every stimulus pair was sampled due to a smaller number of participants. By using the full sample, we ensure that all stimulus pairs were adequately represented, thus providing a more reliable and comprehensive measure of semantic similarity.

Overall, we are confident that our behavioral similarity ratings are robust and provide a valid measure of stimulus similarity across age ranges.

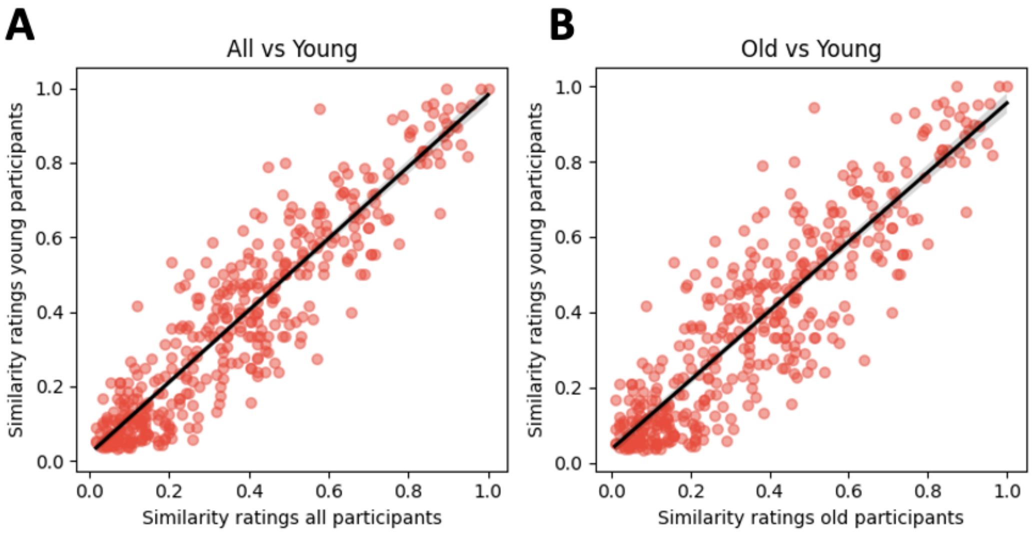


Figure R5. Correlation between similarity ratings computed based on sub-groups. (A) Inter-group similarity consistency: This panel demonstrates the correlation between semantic similarity ratings derived from the entire cohort of participants and those obtained exclusively from the subgroup of young participants (aged 20-31). The high correlation coefficient (r=0.92, p < 0.001) indicates a strong alignment in semantic perceptions across the entire age spectrum, suggesting that the younger participants' semantic judgments are representative of the broader sample. (B) Cross-age comparison of similarity judgments: This panel presents the correlation between the semantic similarity ratings provided by the old (aged 32-70) versus the young participants (aged 20-31). The correlation (r=0.88, p < 0.001) underscores a substantial degree of agreement in semantic assessments between the two age cohorts, despite the inherent variability in experiential background and potential age-related cognitive changes.

2. In the study, on page 9, the authors touch upon the notion of semantic representation at the cortical level. Yet, this concept was not introduced in the introduction. I am left wondering if this analysis was a part of the original plan or if it emerged as a post-hoc investigation. The clarity of the study would be enhanced if the authors integrated details about the semantic representation right from the introduction.

Thank you for raising this point. Our primary objective was to investigate whether the hippocampal formation codes semantic relationship simultaneously with the learned transition structure, in line with the notion of a domain-general relational code in the hippocampal formation (Eichenbaum & Cohen, 2014). Notably, prior research has underscored the representation of semantic relationships at both hippocampal and cortical levels. In light of this, while our main emphasis was on the hippocampal formation, we believed it was complementary to also touch upon known cortical semantic areas.

To provide readers with a holistic understanding from the outset, we have revised our introduction:

"In this situation, besides the newly learned transition structure between objects, participants can be assumed to have explicit knowledge about the semantic relationships between the same objects (e.g., rabbit and dog are both animals). Previous research has provided evidence that semantic relationships are represented in the hippocampus (Pacheco Estefan et al., 2021; Romero, Barense, & Moscovitch, 2019; Solomon et al., 2019) but also across various cortical regions (Bracci et al., 2015; Charest et al., 2014; Clarke & Tyler, 2014; Price et al., 2015; Huth et al., 2016; Frisby et al., 2023)...Here, we ask whether prior semantic knowledge about objects would be simultaneously mapped in the same hippocampal system which also represents knowledge about transition structure." (Introduction, page 3)

3. In the research presented by the authors, tests were conducted on both hemispheres. Notably, with corrections made at .05 for each side of the contrast, there is a potential to double the familywise error rate. It might be advisable for the authors to consider adjusting the significance threshold to p < .025 for each test to address this concern.

We believe that this is a misunderstanding and apologize for not being clear. We have used a single, bilateral entorhinal-hippocampal mask for the SVC. Therefore, no adjustment is needed for the significance threshold.

We have revised our manuscript to avoid any ambiguity. In relation to R1's comment 3.2, we have also added information about the bilateral entorhinal-hippocampal mask we used:

"Therefore, we focused our analysis on this region. The anatomical mask is created using Freesurfer (Fischl, 2012) segmentation in MNI space, combining bilateral hippocampus and bilateral entorhinal cortex (Supplementary Material S2). We consider our results significant if they survived family-wise error (FWE) correction at the peak-level of $p < .05$ within this anatomically defined mask (small volume correction, SVC)." (Methods, page 10)

"The fMRI adaptation analysis showed a cluster bilaterally in the entorhinal cortex (Figure 3A; FWE corrected at peak level, peak $t_{22} = 4.44$, $p = .042$, [-18, -19, -25]). "

(Results, page 15)

"Critically, we also observed a semantic distance effect in the bilateral hippocampus (Figure 3B; peak $t_{22} = 4.69$, $p = .028$, [24, -31, -10])." (Results, page 15)

4. In the paper, the authors characterize some results as being more posterior. Yet, the distinction of the posterior portion of the hippocampus has been established, for instance, by Poppenk et al., 2013. It would be beneficial if the authors could provide clarity on this distinction, especially since a growing body of literature emphasizes this dissociation.

We thank the reviewer for referring us to this work. Indeed, our result of the anterior-posterior gradient (depicted in Figure 4C, also included below) is highly in line with the long-axis hippocampal specialization defined in Poppenk et al. (2013, "we propose that foci at or anterior to $y = -21$ mm in MNI space may be regarded as falling in the aHPC"). Especially in the right hippocampus, we find that this coordinate nicely divides the hippocampus into an anterior region mostly representing the transition structure and a posterior region mostly representing the semantic similarities.

We have added more clarification in the manuscript about this anterior-posterior division:

"These analyses demonstrate that the semantic similarity effect is localized in more posterior regions of the hippocampal formation, whereas the transition structure effect resides in more anterior regions. This difference, found in both hemispheres, suggests the existence of a posterior-anterior gradient along the hippocampal long axis (Poppenk et al., 2013). This effect is particularly pronounced in the right hemisphere where peaks do not overlap." (Results, page 17)

"Notably, we found an anatomical gradient along the anterior-posterior axis of the hippocampus (Poppenk et al., 2013; Strange et al. 2014), with the graph structure represented in more anterior parts of the hippocampal formation and the semantic map in more posterior parts." (Discussion, page 22)

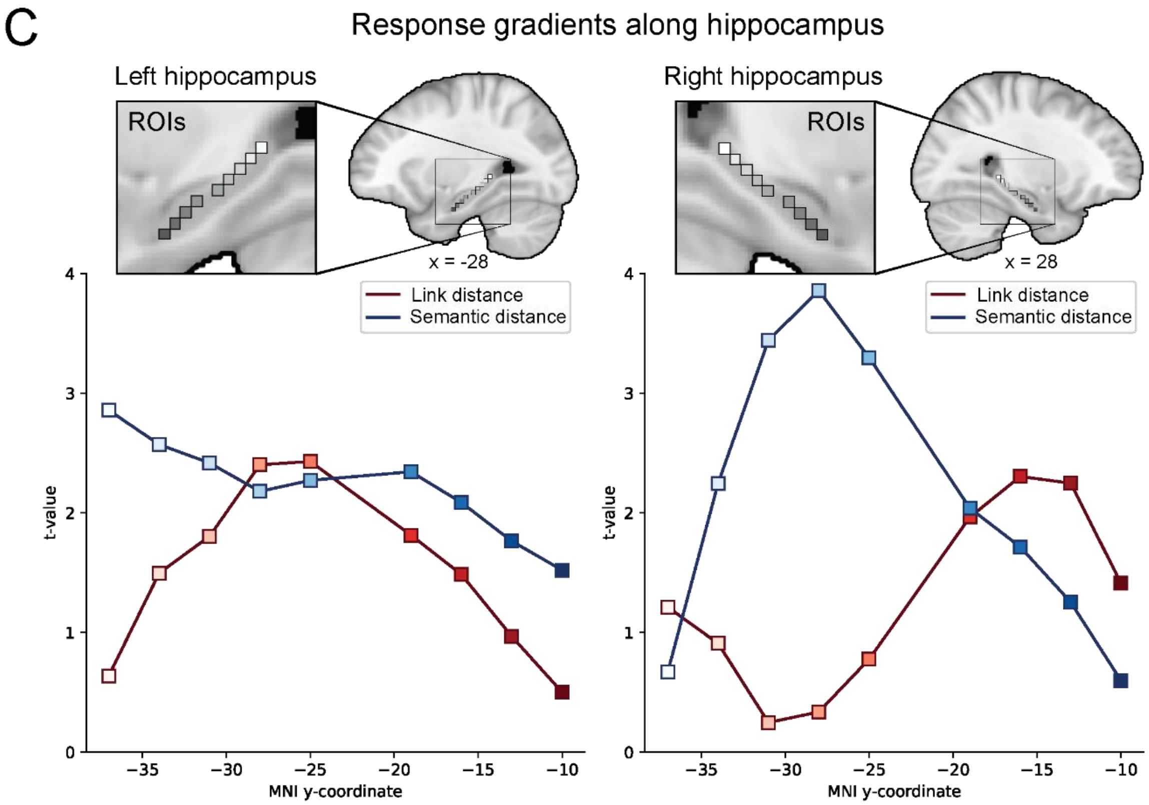


Figure 4C. Visualization of response gradient along the hippocampal long axis. In both the left and the right hippocampus, the semantic distance peaks at more posterior locations compared to the link distance.

Reference:

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. Trends in cognitive sciences, 17(5), 230-240. https://doi.org/10.1016/j.tics.2013.03.005

5. In their findings, the authors describe the results as a 'gradient.' I wonder if it might be more correct to label it as a dissociation of functionality?

We agree with the reviewer that it can be both: a gradient along the hippocampus as we described in Figure 4B and 4C, or alternatively, two functionally dissociated clusters as we tried to depict in Figure 4A. Whether a dissociation or a gradient is a more accurate depiction has important implications for our understanding of hippocampus codes and their role in cognition. However, due to the spatially correlated nature of fMRI data, it is unfortunately difficult to arbitrate between these two possibilities.

We now discuss this possibility explicitly in the Discussion:

"Notably, we found an anatomical gradient along the anterior-posterior axis of the hippocampus (Poppenk et al., 2013; Strange et al. 2014), with the graph structure represented in more anterior parts of the hippocampal formation and the semantic map in more posterior parts. Alternatively, this could also be viewed as two functionally dissociated clusters, with the cluster residing in the entorhinal cortex encoding statistical information about transition structures and the cluster in hippocampus encoding semantic similarities between specific objects. Distinct functional clusters would suggest more specialized processing within the hippocampus, suggesting that different types of knowledge are more rigidly localized, perhaps facilitating categorization of information for more systematic retrieval. A gradient on the other hand suggests a more integrated and potentially overlapping functionality within the hippocampus, perhaps facilitating processing in ambiguous situations and retrieval of information in context-rich situations. Due to spatial correlations inherent to fMRI data, it is not possible to completely disentangle a gradient from two separable clusters. Future studies, potentially employing higher-resolution fMRI or intracranial recordings, can provide more definitive answers.

In either case, the anatomical segregation of the two maps may reflect differences in the nature of the underlying knowledge structures (Peer et al. 2021). The semantic relationships may reflect taxonomic knowledge derived from shared features and properties between objects that participants formed over their lifetimes. The transition structure on the other hand could stem from recent associative learning..." (Discussion, pages 22-23)

Minor Issues:

1. Some participants were assigned to perform 1460 trials, while others only completed 20 trials for the triplet odd-one-out task. It would be informative if the authors could share the statistics, such as the standard deviation (SD) for the number of trials and reaction time (RT). Given this significant disparity in trial counts, I am curious if the authors considered the potential influence of participant motivation on the results, especially in light of possible fatigue from a high number of trials.

We thank the reviewer for this suggestion. Participants engaged in a variable number of trials, ranging from a minimum of 20 to a maximum of 1460. The median number of trials per participant was 50, with a 25th percentile at 20 trials and a 75th percentile at 145 trials. Below we show the number of trials each participant performed in the odd-one-out task (Figure R6A). The distribution is heavily right-skewed, indicating that a large proportion of participants completed a relatively small number of trials, while a small minority completed a large number of trials.

To assess the potential implications of this distribution more systematically, we ran an analysis where we computed the resulting similarity matrix by including only up to the first X trials of each participant. The objective was to discern any potential influences such as wavering participant motivation or increasing fatigue over trials. Our result shows that the reduced matrix converges to really high similarity with the full matrix very quickly (Figure R6B). Therefore, we conclude that our resulting semantic matrix remains rather unaffected by the different number of trials participants performed in the task. In light of these findings, we are confident that our results remain robust despite the inherent disparities in trial counts among participants.

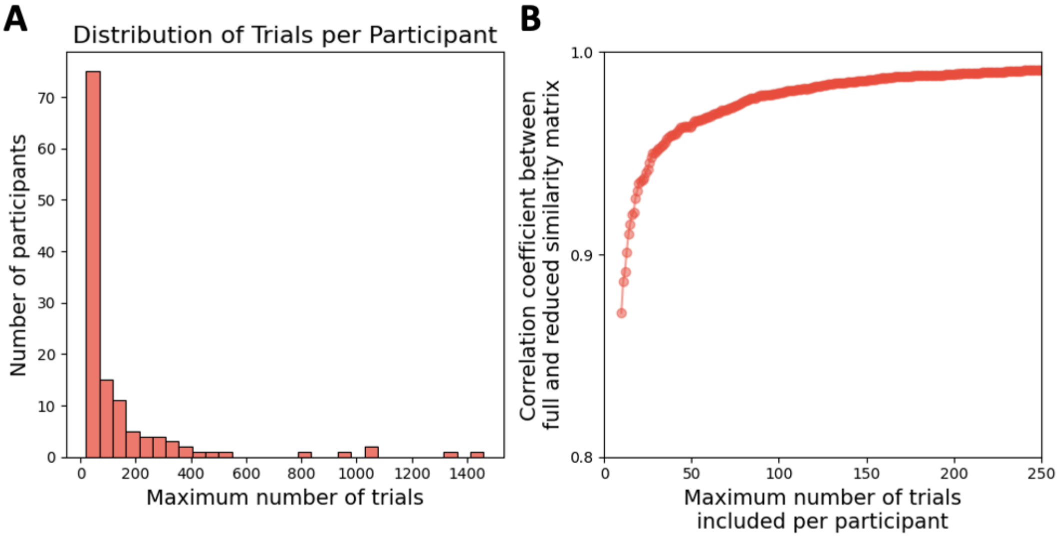


Figure R6. Distribution and impact of trials per participant. (A) Distribution of the maximum number of trials performed by each participant.(B) Correlation analysis showing that when considering a reduced number of trials per participant, the similarity matrix rapidly approaches high congruence with the matrix derived from the full set of trials.

Regarding the influence of fatigue and motivation: We believe that by providing participants with the opportunity to leave the study at any point at their discretion, we actually reduced problems with loss of motivation and fatigue often influencing results in long experiments.

To test engagement and responsiveness during the experiment, we analyzed participants' average response time (RT) over trials. As trials progressed, we observed a decrease in response times that stabilized around trial 400 (Figure R7A, Spearman r = -0.59, p < 0.001). This might either suggest that over time response times decreased, or alternatively that slower participants decided to complete fewer trials. The second hypothesis is confirmed by a negative correlation between average response times and total number of completed trials per participant (Figure R7B, Spearman r = -0.21, p = 0.02).

This suggests that participants who experienced a loss of motivation or fatigue as indicated by long response times terminated the experiment early and only highly motivated participants kept going for many trials.

We tried to exclude trials where response times increase at a certain threshold (Figure R7C). Again, it did not significantly influence similarity ratings even for a relatively short threshold (ie 2 seconds, which excludes a large portion of the data). This confirms our above observation that similarity measures were robust and reliable.

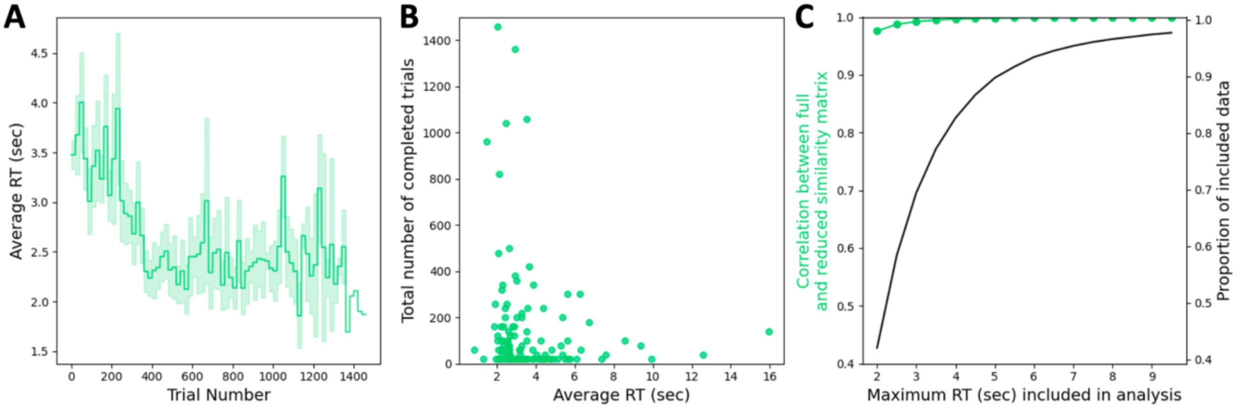


Figure R7. Analysis of response times (RT) in the odd-one-out task. (A) Average RT (in seconds) plotted against trial number, showing variability in participant response times. The shading indicates the standard error of the mean. The data is binned, size of each bin = 20 trials. (B) Scatter plot presenting the

relationship between average RT (in seconds) and the total number of completed trials for individual participants. (C) Illustration of the similarity between full and reduced matrices as a function of the maximum RT (in seconds) included in the analysis (green) as well as proportion of included data relative to the full dataset (black).

We have now added information about the distribution of the number of trials participants completed as well as their response times to the manuscript. More detailed information about these analyses was added to the Supplementary Materials.

“Participants engaged in a variable number of trials, ranging from a minimum of 20 to a maximum of 1460 (median = 50, 25th percentile = 20, 75th percentile = 145), with a median RT of 2221 ms (Supplementary Material S1).” (Methods, page 7)

2. On page 12, there seems to be an issue with Figure 2, as it appears to be cut off.

We thank the reviewer for pointing it out. We have now fixed the cut off. The figure is also included here:

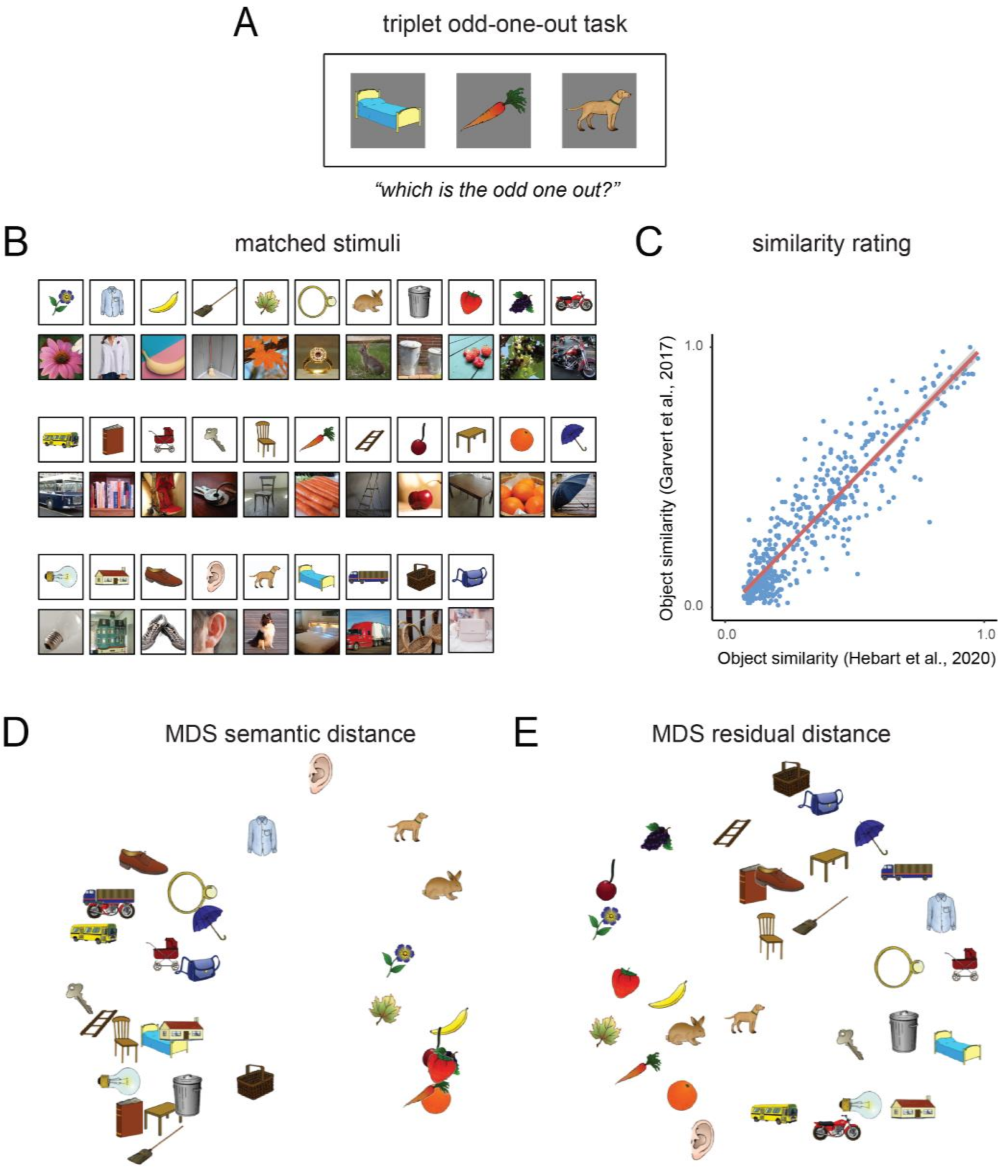


Figure 2. Semantic distance constructed using the triplet odd-one-out task.

> 3. In reviewing the manuscript, I noticed that although both the 2020 and 2022 papers by Whittington et al. are listed in the references, only the 2020 publication is cited within the text. Also, the work by Bedney et al., 2008 seems to be absent from the reference list.

We thank the reviewer for pointing it out. We have now fixed both references.

4. In the provided figure, the authors have opted for an uncorrected threshold for visualization. However, I feel it would be more aligned with best practices to present the actual results, allowing for a transparent depiction of the findings.

We opted for visualization using an uncorrected threshold, because we believe it is important to provide the full picture to the readers. Nevertheless, we agree that aligning the visualization and the statistical outcome is also essential. We now present updated figures (Figure 3 and Figure 4) as full-brain results using an uncorrected threshold, with significant voxels clearly highlighted. We include both figures below.

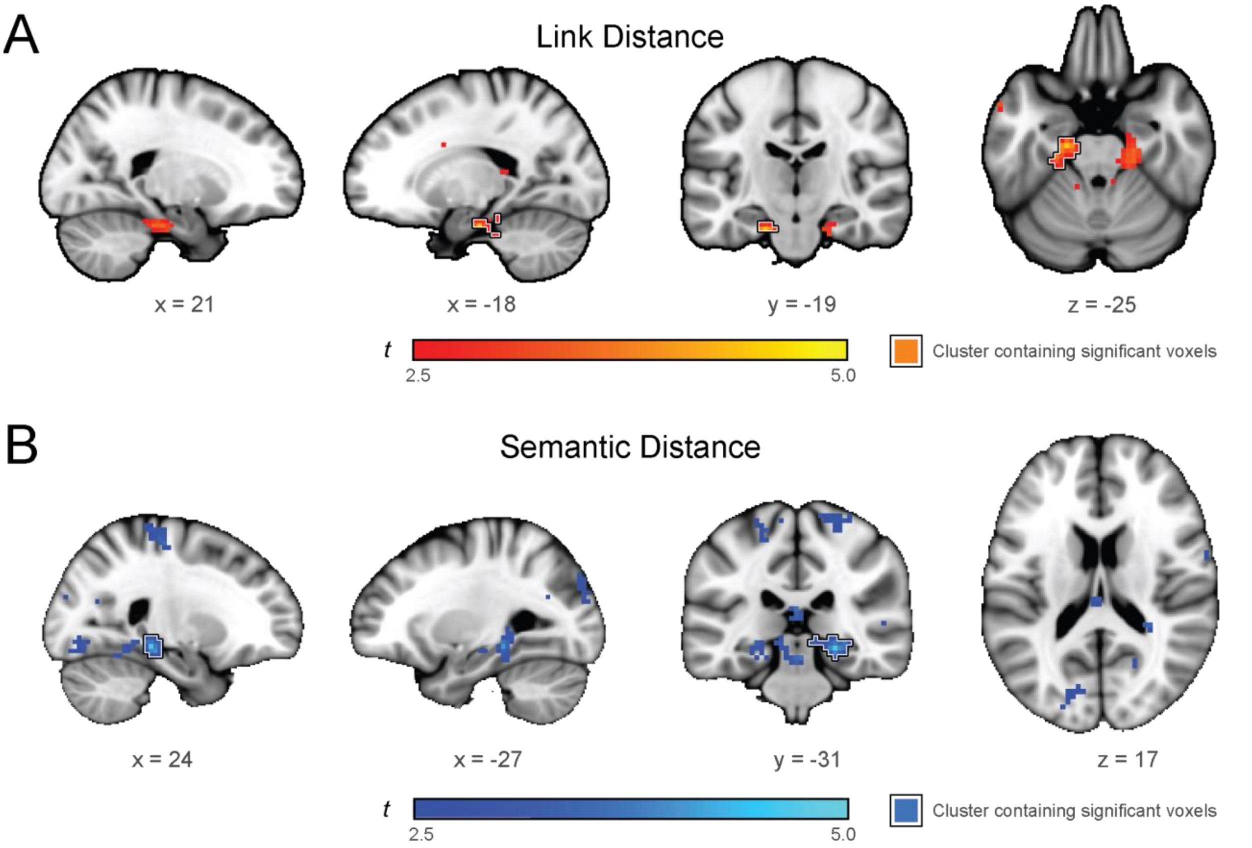


Figure 3. Transition structure and semantic similarities are represented in the hippocampal-entorhinal system. (A) Whole-brain analysis showing a decrease in fMRI adaptation with link distance in the hippocampal formation, when link distance, semantic distance and residual distance are included in the model. (B) Whole-brain analysis showing a decrease in fMRI adaptation with semantic distance in the

hippocampal formation, when link distance, semantic distance and residual distance are included in the model. Both (A) and (B) are thresholded at p < .01, uncorrected for visualization. The clusters containing voxels surviving correction for multiple comparisons (FWE, p < .05) are highlighted in solid black lines.

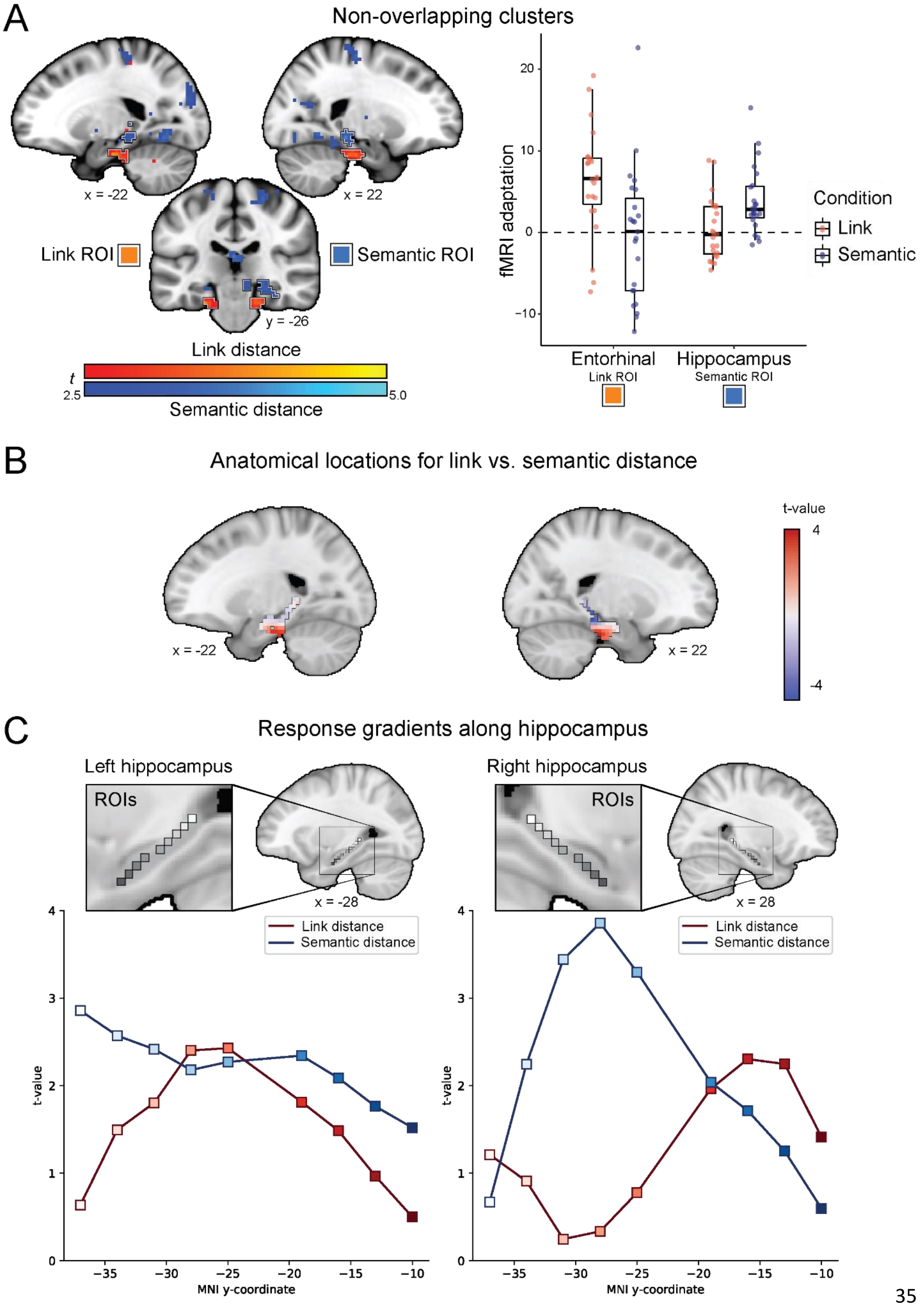

A
Non-overlapping clusters
x = -22
x = 22
y = -26
Link ROI
Semantic ROI
Link distance
t
2.5
5.0
Semantic distance
fMRI adaptation
20
10
0
−10
Entorhinal
Link ROI
Hippocampus
Semantic ROI
Condition
Link
Semantic
B
Anatomical locations for link vs. semantic distance
x = -22
x = 22
t-value
4
-4
C
Response gradients along hippocampus
Left hippocampus
ROIs
x = -28
Right hippocampus
ROIs
x = 28
Link distance
Semantic distance
t-value
0
1
2
3
4
−35
−30
−25
−20
−15
−10
MNI y-coordinate

Figure 4. Anatomical localization of transition structure and semantic similarities. (A) Left: Link distance (red) and semantic distance (blue) are represented in non-overlapping clusters (thresholded at $p < .01$, uncorrected). Two ROIs were defined based on the link distance effect (in red) and the semantic distance effect (in blue, both ROIs highlighted in solid lines) and included voxels exceeding a cluster-defining threshold of $p < .01$, uncorrected). Right: boxplot of the parameter estimates for the link distance and semantic distance effects extracted from these two ROIs. The thick horizontal line inside the box indicates the median, and the bottom and top of the box indicate the first and third quartiles of each condition. Each dot represents one participant. The plot is for visualization only, since the contrast used for defining the ROIs is not independent from the interaction effect of interest here. (B) Anatomical location where the link distance is represented more strongly (red) versus where the semantic distance is represented more strongly (blue). The analysis is restricted to the hippocampal formation (incl. hippocampus and entorhinal cortex). (C) Visualization of response gradient along the hippocampal long axis. In both the left and the right hippocampus, the semantic distance peaks at more posterior locations compared to the link distance.

5. On page 21, there is a mention that the results were found without active attention. It would be essential for the authors to prominently address this point in the methods section for clarity.

We thank the reviewer for their suggestions. Indeed, we argue that “Participants were not even required to pay attention to the objects, as they only had to attend to the presence of a grey patch on the screen” given the cover task. We have added more detailed descriptions of the cover task to the Methods.

“In the scanning session (day 2) …To reduce the motor responses in the scanner, a different behavioral cover task was employed that was orthogonal to the imaging analysis of interest: In 10% of the fMRI trials, participants performed an unrelated cover task, reporting whether a gray patch had been present on the preceding object (Figure 1B). This means that participants were not required to pay active attention to the object identity.” (Methods, page 5)