Point-by-Point Reply: CELL-D-25-05586

We would like to thank our reviewers for their thoughtful and constructive comments! We appreciate their precious time and efforts. In our view, the paper benefited greatly from considering all the recommendations. Below, we provide point-by-point responses to all queries. All changes in the revised manuscript are indicated with purple text.

Reply formatting guide:

Reviewer remarks: blue box, black text

Author response: black text, Arial

Changes in the manuscript: **purple, Times New Roman**

Reviewer #1:

General remarks

In this work the authors investigated Spanish and English word representations in bilingual speakers using single cell recordings from the hippocampus taken during a variety of listening and speaking tasks. The core finding is that though translated words in English and Spanish evoke different population activity, the relational distances between words tend to be preserved across languages. Words in English and Spanish have distinct codes but largely similar geometries. Overall, I found the results and analysis here compelling. They build nicely off a long line of work in natural language processing and language neuroscience which emphasizes representational distances as a neural underpinning for semantic similarity. The results here bolster this view by showing roughly the same semantic geometry holds in a translated set of words with the same meaning, though the overall map may be rotated across languages. I have a few suggestions that I believe would improve the manuscript.

We appreciate Reviewer 1's kind words; below, we provide a point-by-point response to concerns raised.

1. Translation Neurons

Overall, I'm fairly unclear about what the analysis of the "translation neurons" is adding to the findings. The authors name this group of neurons and then perform a set of analyses which show they have no impact on the basic representational scheme purported to support translation. This is more or less stated outright in the discussion "These neurons may play an important role in translation... However, even when we remove these neurons from our analysis, the semantic geometry is still robustly maintained, suggesting that these neurons do not play a necessary role in translation." Given the evidence presented it's hard to see how they would be sufficient for translation either. Moreover, it's a bit strange to do a single neuron analysis on words paired by their meaning, and then go on to posit that semantic maps which encode meaning exist in the structure of population activity. At the very least I would change the name here, and maybe front load the fact that though this phenomenon exists in single neurons it doesn't play a significant role in the semantic maps that serve as the main mechanism of translation.

Both reviewers raise similar concerns about our claims regarding translation neurons. This shared concern reflects our failure to make our scientific position about these neurons clear in the writing of the original manuscript.

What our position is: We agree with Reviewer 1 that the translation neurons have no measurable impact on the population-level geometry that we believe are the foundation of translation.

The reason we led off with the translation neuron finding is that we reasoned that many readers would not be familiar with the geometric approach to analysis of physiological data. In the narrative flow of our paper, the translation neurons are, in essence, the stepping stone to the core message, which is the geometric one.

It is clear, though, that in the mind of the reader (including both reviewers) this narrative organization results in seemingly contradictory messages.

Our proposed solution is what Reviewer 1 suggests: to "front load the fact that though this phenomenon exists in single neurons it doesn't play a significant role in the semantic maps that serve as the main mechanism of translation".

Both reviewers also suggest we use a different term instead of "translation neurons." We have renamed these neurons "cross-language neurons".

Now we discuss this more straightforwardly in the revised Abstract, Results, and Discussion. (And we have removed mention of these neurons from the Introduction). For example, the revised Abstract now says "However, neurons' semantic tunings differed substantially by language, suggesting language-specific neural implementations. Instead, the crucial driver of translation was a preserved geometric organization of neural responses between the two languages, one that did not depend on neuron level functional overlap."

Moreover, we now cover these neurons more briefly in the revised Results and move on much more quickly to the geometry findings.

And, from the revised Discussion:

Cross-language neurons. We identify a new class of what we call "cross-language neurons," whose responses to co-translated words are correlated. Such neurons would have similar responses to, for example, "*maybe*" and "*quizás.*" These neurons may contribute to translation; however, they do not play an essential, or even a major role: even when we exclude them from our analysis, cross-language semantic geometry is robustly maintained (Table S3). Moreover, while their tuning functions are more aligned than other neurons, they are not perfectly aligned, indicating that they are not simply amodal conceptual representations. Consistent with this, their response fields span multiple semantic categories (e.g., Figure 2C), which differentiates them from the canonical amodal concept-related neurons, concept cells (Quian Quiroga et al., 2005; Quian Quiroga, 2012). Word-level decomposition confirms that their cross-language correlations are driven by a relatively sparse subset of words. These words are constrained along fewer latent semantic axes in mBERT space than other neurons which did not show cross-language correlation (Table S1-S2). We speculate that they may be the extreme end of a larger distribution of neurons whose shared population-level geometry contributes to translation.

I also wonder if a more fine grained analysis might clarify things. For example, in purported translation neurons, is a particular subset of words primarily contributing to the positive correlation effect, and do all of these words have a discernible semantic class? If so, this would presumably imply that the neuron in question contributes in an outsized way to a given semantic axis, which meshes more coherently with subsequent arguments.

This is a very insightful suggestion. We conducted the following analyses and listed the results below. In summary, the cross-language neurons are driven by a relatively sparse subset of words. These driver words did not cluster into discernible semantic categories. These words are constrained along fewer latent semantic axes in mBERT space.

Because cross-language neurons are not the key findings supporting across-language translation (preserved geometry is), we only briefly referenced the Table S1 and Table S2 in the main text:

"Word-level decomposition confirms that their cross-language correlations are driven by a relatively sparse subset of words. These words are constrained along fewer latent semantic axes in mBERT space than other neurons which did not show cross-language correlation (Table S1-S2). We speculate that they may be the extreme end of a larger distribution of neurons whose shared population-level geometry contributes to translation."

See **Table S1** and **Table S2** below.

Table S1. Driver word properties of cross-language versus non-cross-language neurons (Study1).

Cross-language neurons required significantly fewer words to explain their cross-language correlations, confirming that a particular subset of words disproportionately contributes. However, these driver words did not cluster into discernible semantic categories. Indeed, the mean semantic cluster index was identical between groups. That is, driver words of cross-language neurons were no more semantically similar to one another than those of non-cross-language neurons, which argues against the possibility that these neurons simply respond to a single named category (e.g., "animals" or "emotions"). Critically, despite equivalent pairwise semantic similarity, the driver words of cross-language neurons spanned fewer independent semantic dimensions. Because pairwise cosine similarity is a scalar average that collapses all directional information, it is insensitive to the anisotropy of the distribution; two sets of words can have identical average pairwise distances yet differ in how their variance is distributed across principal axes. Therefore, we calculated the participation ratio (PR), which provides a continuous measure of how many independent semantic dimensions the driver words effectively span. A lower PR indicates that the driver words cluster along fewer embedding dimensions (i.e., occupy a more semantically constrained subspace), whereas a higher PR indicates that they spread across a broader range of semantic directions. We found the driver words have lower participation ratio than non-cross-language neurons.

Commented [1]: can we just state clearly what the answer to the question is? the reviewer doesnt want to have to read all this - they just want a simple yes or no

Commented [2]: Done

Stimulus	*n* (cross-language)	Selectivity (K80%)			Cluster Index *(similarity)*			Participation Ratio *(dimensionality)*		
		Cross-Language neurons	Non-cross-language neurons	*d*	Cross-Language neurons	Non-cross-language neurons	*d*	Cross-Language neurons	Non-cross-language neurons	*d*
K	6	21.6 ± 6.4	18.3 ± 5.9	0.56	0.663 ± 0.002	0.664 ± 0.003	-0.092	66.5 ± 3.4	65.6 ± 3.4	0.27
A1	20	16.3 ± 5.8	25.6 ± 14.1	-0.66**	0.666 ± 0.011	0.667 ± 0.004	-0.295	55.3 ± 12.6	62.7 ± 6.5	-1.13***
A2	11	19.2 ± 9.1	24.3 ± 14.6	-0.35	0.696 ± 0.004	0.691 ± 0.005	0.927**	47.4 ± 4.4	47.9 ± 5.3	-0.10
A3	8	17.2 ± 11.1	30.0 ± 17.1	-0.75*	0.693 ± 0.019	0.690 ± 0.006	0.536	41.0 ± 17.4	49.7 ± 6.7	-1.28***
KN-neuropixel	9	26.6 ± 3.7	28.4 ± 5.2	-0.35	0.651 ± 0.004	0.652 ± 0.004	-0.294	56.9 ± 2.6	57.6 ± 3.4	-0.19
Pooled	**54**	**19.3 ± 8.0**	**25.7 ± 13.4**	**-0.48*** **	**0.673 ± 0.019**	**0.673 ± 0.015**	**0.017**	**53.4 ± 12.4**	**57.2 ± 8.7**	**-0.44****

Table Note. Values in each cell are mean ± SD.
d = Cohen's d (negative indicates cross-language neurons < non-cross-language neurons).
Selectivity (K80%): percentage of words needed to explain 80% of positive cross-language correlation mass.
Cluster Index: mean pairwise cosine similarity among driver words in mBERT embedding space.
Participation Ratio: effective dimensionality of driver word embeddings.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table S2. Driver word properties of cross-language versus non-cross-language neurons (read aloud task, Study 2, word level).

In the read aloud task, we performed the same driver word analysis as in Study 1 (listening). Because most results were reported at the word level, we did not extend this analysis to the phrase level. Of 73 valid neurons, 61 were classified as cross-language neurons (83.6%), leaving only 12 non-cross-language neurons.

Because cross-language neurons vastly outnumbered non-cross-language neurons (61 vs. 12), the bootstrap procedure was inverted: in each of 100 iterations, 12 cross-language neurons were randomly subsampled (without replacement) and compared to the fixed set of 12 non-cross-language neurons. Although the cluster index differed significantly between groups, the absolute difference was negligible (Δ = 0.010), suggesting that the semantic similarity structure of driver words was virtually identical across neuron types; the statistical significance likely reflects the large effective sample size in the bootstrap rather than a meaningful biological difference.

Stimulus	*n* (cross-language)	Selectivity (K80%)	Cluster Index *(similarity)*	Participation Ratio *(dimensionality)*

		Cross-Language neurons	Non-cross-language neurons	*d*	Cross-Language neurons	Non-cross-language neurons	*d*	Cross-Language neurons	Non-cross-language neurons	*d*
Word-speech	61	17.4 ± 3.0	33.8 ± 16.4	-2.35***	0.807 ± 0.006	0.797 ± 0.004	1.722***	28.4 ± 4.0	34.7 ± 5.6	-1.47***

Table Note. Values in each cell are mean ± SD.
d = Cohen's d (negative indicates cross-language neurons < non-cross-language neurons).
Selectivity (K80%): percentage of words needed to explain 80% of positive cross-language correlation mass.
Cluster Index: mean pairwise cosine similarity among driver words in mBERT embedding space.
Participation Ratio: effective dimensionality of driver word embeddings.
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Revised Method (see below)

Word-level contribution analysis for cross-language neurons (Table S1 and Table S2).

For each cross-language neuron in Study 1, across all stimulus sets, we decomposed the cross-language Pearson correlation into per-word contributions and defined "driver words" as the minimal set accounting for 80% of the positive contribution mass. We compared across-language neurons to size-matched bootstrap samples of non-across-language neurons (100 iterations per group) on three properties of these driver words: (1) selectivity (the percentage of the driver words), (2) semantic clustering (mean pairwise cosine similarity of driver word mBERT embeddings, tested against permutation null, permutation=1000, we named it as cluster index), and (3) effective dimensionality, the participation ratio (PR) of the driver words' embedding matrix. For each neuron, we extracted the mBERT embeddings (Layer 11) of its driver words, performed PCA via singular value decomposition, and computed PR = $(\Sigma\lambda\_i)^2 / \Sigma\lambda\_i^2$, where λ_i are the eigenvalues. PR ranges from 1, when all variance is concentrated along a single principal axis, to the number of non-zero eigenvalues, when variance is uniformly distributed across all axes. It thus provides a continuous measure of how many independent semantic dimensions the driver words effectively span. A lower PR indicates that the driver words cluster along fewer embedding dimensions (i.e., occupy a more semantically constrained subspace), whereas a higher PR indicates that they spread across a broader range of semantic directions. To ensure a fair comparison, PR values of cross-language neurons were compared against size-matched bootstrap samples of non-cross-language neurons (100 iterations per group), so that any difference in participation ratio reflects genuine differences in the semantic geometry of driver words rather than differences in sample size.

In the read aloud task, we performed the same driver word analysis as in Study 1 (listening). Because the majority of results were reported at the word level, we did not extend this analysis to the phrase level. Because cross-language neurons vastly outnumbered non-cross-language neurons (61 vs. 12), the bootstrap procedure was inverted: in each of 100 iterations, 12 cross-language neurons were randomly subsampled (without replacement) and compared to the fixed set of 12 non-cross-language neurons.

Summary of revisions related to comment 1:

(1) We renamed the “translation neurons” to “cross-language neurons”

(2) We revised Abstract, Results, and Discussion to de-emphasize these neurons, and removed mention of these neurons from the Introduction. In general, we now cover these neurons more briefly in the revised Results and move on much more quickly to the geometry findings.

(3) We conducted the additional analyses for driver words and listed the results in Table S1 and Table S2. We added the analysis procedure to the Methods.

(4) We added one Discussion paragraph addressing the cross-language neurons

2. LLM Analysis and Implications

I think the LLM analyses need a bit more scope and I don't think the evidence is wholly aligned with all of the authors interpretations. The crux is that "both the brain and mBERT achieve shared representational geometries (same computational goal) through similar algorithms (language-specific semantic readout axes)" seems only half correct. While I think the identification of the algorithmic basis of semantics is well founded, the paper overall is about translation of semantics. The neural data suggests this occurs because the geometry of population activity across languages is preserved despite the fact that word-by-word representations are quite different. The analyses presented do not demonstrate that this is happening in mBERT. As the authors probe deeper into mBERT layers they found that word representations for Spanish and English tended to converge (Fig. 3B.). This is consistent with the hypothesis that mBERT performs translation by mapping different languages to a shared, language agnostic semantic space. In this case translating through shared geometry is trivially true, since pushing translated words towards the same representations leads to absolutely equivalent geometries across language (as opposed to equivalence up to rotation).

The reviewer here is talking about our side-by-side comparison between mBERT and the neural data. We show that both mBERT and hippocampus exhibit shared cross-language geometry at the population level. However, the reviewer points out that our data broadly suggest inconsistent single neuron tuning in the hippocampus, while mBERT shows consistent unit-level tuning. By this logic, we are wrong to imply that the brain recapitulates mBERT, at least at the neuron/unit level.

We fully agree with the reviewer, and, on reflection, we think that this shows what appears to be a disjunction between the solution used by mBERT and the one used by hippocampus. In short, while both systems use a shared geometric solution, they diverge in whether this shared geometry is a direct consequence of the underlying abstract tuning of the units (mBERT) or whether it occurs despite a lack of alignment at the neuron level (hippocampus).

In other words: mBERT solves the problem of translation by building out a bank of translation units (equivalent to what we now call our cross-language neurons), while the brain does it (for the most part) without relying on any specialized cross-language neurons (see above). We now make this clear in the following new text added to the revised Discussion:

We also consider the relationship between mBERT and the brain. At the population level, the representational geometry of mBERT (layers 9-11; **Figure S2A-S2C**) resembles that of the hippocampus. Both represent words as high-dimensional vectors with linearly decodable structure and exhibit cross-language geometric alignment in which translation equivalents occupy similar relative positions (Devlin et al., 2019; Pires et al., 2019; Wu et al., 2022). Importantly, this alignment is layer-selective: brain-mBERT correlations were negative in early layers, shifted positive around layers 4-6, and peaked in layers 9-11 (**Figure S2**), indicating that the hippocampus aligns specifically with higher semantic layers

rather than mBERT's full hierarchy. Both systems achieve shared geometry through language-specific subspaces constructed from overlapping populations (neurons in the brain, embedding dimensions in mBERT; **Figure S4**), though mBERT's cross-language subspace alignment far exceeds that observed in the brain at every layer. At the single-unit level, a fundamental difference emerges: the vast majority of 'units' are cross-language correlated units, 100% units of higher layers of mBERT (>=layer 6) are cross-language correlated units (**Table S5**). Meanwhile, neurons with similar tuning functions ("cross-language neurons") are rare in our dataset (5.2%-19.4%, but 83.6% in speech production only). What might account for this difference? One potential reason is a scale mismatch: mBERT similarities are computed over thousands of dimensions that average across many representational factors, whereas a single neuron may reflect a narrow, idiosyncratic subset of features. A second possibility is that mBERT is partly driven by surface form regularities—shared subword structure, orthography, and distributional cues (for example, "universe" and "universo") whereas neurons in semantic regions may be more strictly organized around conceptual content. Finally, there may be a deeper representational difference: mBERT encodings are relatively linear and factorized, so related words align in clean vector subspaces, whereas neural codes may be more curved, multi-lobed, and context-dependent (Perich et al., 2025).

This is also partly borne out by the analysis in Fig. 5c. where the authors correctly point out across vs. within language semantic subspace alignment is significantly different in both hippocampus and mBERT, but fail to mention that the absolute alignment in mBERT is vastly larger than values observed in the brain.

First we apologize that we did not make this point clearer. We do so now in the revised text.

"Despite the similar semantic geometry, the manifold constructed from semantic axes for English and Spanish were less aligned than the within-language subspaces both for mBERT (K, A1, A2, A3, KN, speech-word, all $p<0.001$) and hippocampus (K, A1, A2, A3, speech-word, all $p<0.001$; KN-Neuropixel, $p>0.05$, **Figure 6C**). But the absolute across-language subspace alignment value in mBERT is substantially larger than values (both within and across-language subspace alignment) observed in the brain (**Figure S4A**). "

Moreover, the reviewer's larger point is true: mBERT and the brain are not identical, especially at the neuron level, and this is worth noting. The paragraph added to the Discussion (see previous response) now makes this point clearly.

Taking the layer with maximum across language similarity seems like a reasonable thing to do for the brain-machine alignment in Fig. 4. and the regression analyses presented in Fig. 3. But again, using this layer as a model for how the hippocampus represents semantics seems doomed from the start precisely because these regression analyses show that across language word representations in the brain are not similar.

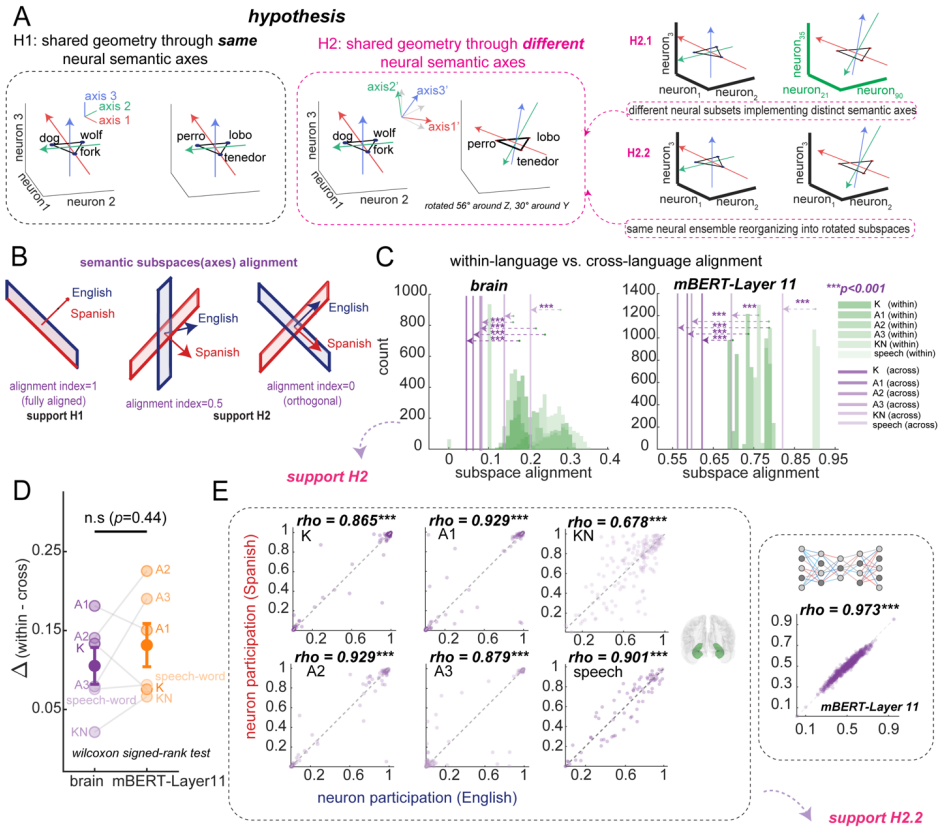
In summary, mBERT is used well as an analysis tool to establish facts about the semantic space, but the authors have yet to demonstrate that it is a good model of translation based on the evidence presented.

The reviewer makes a key distinction - while mBERT is a good tool for investigating the brain, our original manuscript was not careful enough in accounting for differences between mBERT and the brain in how translation is implemented. In particular, as discussed above, there appears to be a qualitative difference between what is going on at the unit level, even if the population / geometric level is similar. We now have revised the manuscript to emphasize this point. In particular, the following paragraph in the Discussion deals directly with this question:

We also consider the relationship between mBERT and the brain. At the population level, the representational geometry of mBERT (layers 9-11; **Figure S2A-S2C**) resembles that of the hippocampus. Both represent words as high-dimensional vectors with linearly decodable structure and exhibit cross-language geometric alignment in which translation equivalents occupy similar relative positions (Devlin et al., 2019; Pires et al., 2019; Wu et al., 2022). Importantly, this alignment is layer-selective: brain-mBERT correlations were negative in early layers, shifted positive around layers 4-6, and peaked in layers 9-11 (**Figure S2**), indicating that the hippocampus aligns specifically with higher semantic layers rather than mBERT's full hierarchy. Both systems achieve shared geometry through language-specific subspaces constructed from overlapping populations (neurons in the brain, embedding dimensions in mBERT; **Figure S4**), though mBERT's cross-language subspace alignment far exceeds that observed in the brain at every layer. At the single-unit level, a fundamental difference emerges: the vast majority of 'units' are cross-language correlated units, 100% units of higher layers of mBERT (>=layer 6) are cross-language correlated units (**Table S5**). Meanwhile, neurons with similar tuning functions ("cross-language neurons") are rare in our dataset (5.2%-19.4%, but 83.6% in speech production only). What might account for this difference? One potential reason is a scale mismatch: mBERT similarities are computed over thousands of dimensions that average across many representational factors, whereas a single neuron may reflect a narrow, idiosyncratic subset of features. A second possibility is that mBERT is partly driven by surface form regularities—shared subword structure, orthography, and distributional cues (for example, "universe" and "universo") whereas neurons in semantic regions may be more strictly organized around conceptual content. Finally, there may be a deeper representational difference: mBERT encodings are relatively linear and factorized, so related words align in clean vector subspaces, whereas neural codes may be more curved, multi-lobed, and context-dependent (Perich et al., 2025).

We revised Figure 6 to remove the claim "brain-machine similarity", as well as the Figure caption.

revised Figure 6:



The clear question based on these observations is whether or not the earlier layers in mBERT, which exhibit lower word representation correlations, also exhibit the same geometries across languages. It would be nice if the authors repeated some of the analyses in Fig. 5 for earlier layers in mBERT to establish this fact either way. To be clear either result would be a valuable contribution. Even if mBERT doesn't exhibit these computational features, it would clearly suggest two valid schemes for translation (across language representational convergence in mBERT, and rotational equivalence in the hippocampus), and future theoretical and complimentary language experiments might adjudicate the advantages of each.

The reviewer helpfully suggests a new analysis that can help us get a handle on these issues: examination of the lower layers of mBERT.

We have now performed these analyses. In summary:

1. At the single-unit level: in mBERT, the vast majority of 'units' are cross-language correlated units. And essentially all units of higher layers of mBERT (>=layer 6) are cross-language correlated units
2. At the population level, neural and semantic geometry (mBERT) aligned well in higher layers 9-11.
3. mBERT's cross-language subspace alignment at every layer far exceeds that observed in the brain

In detail:

We conducted three analyses to better understand the features of lower layers of mBERT.

1. At the single-unit level: in mBERT, the vast majority of 'units' are cross-language correlated units. All units of higher layers of mBERT (>=layer 6) are cross-language correlated units (see **Table S5**).

Table S5

Layer	K	A1	A2	A3	KN	Speech-word
layer0	99.7%	99.5%	97.3%	97.0%	99.2%	83.1%
layer1	99.9%	100%	99.2%	100%	99.9%	90.4%
layer2	100%	100%	100%	100%	100%	94.1%
layer3	100%	100%	100%	100%	100%	97.8%
layer4	100%	100%	100%	100%	100%	99.5%
layer5	100%	100%	100%	100%	100%	99.7%
>=layer6	100%	100%	100%	100%	100%	100%

As described above, this is very different from how the brain works, where such units are rare.

2. The hippocampal populational geometry aligns well with mBERT semantic geometry in layers 9-11.

We repeated the brain-mBERT RSA (Figure 4) across all 13 mBERT layers (layers 0-12) for each stimulus set. Across all five stimulus sets (K, A1, A2, A3, and KN) in Study 1, speech production (word-level; Study 2) as well as the conversation task (Study3), the correlation between neural and semantic RDMs was negative in layers 0-1, progressively shifted toward positive values around layers 4-6, and peaked with statistical significance in layers 9-11. This trajectory mirrors the known progression from subword tokenization to contextualized semantic representations in transformer architectures (**Figure S2**, see below).

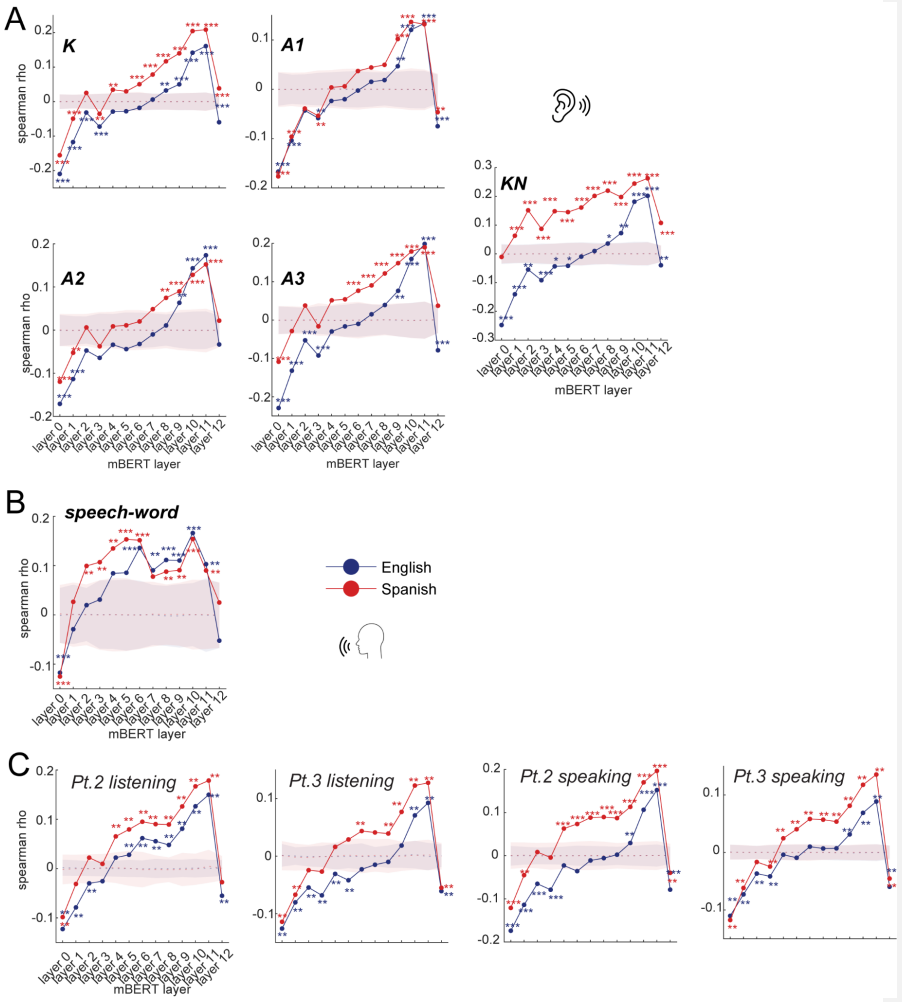


Figure S2. Brain-mBERT alignment at each mBERT layer.

We repeated the brain-mBERT alignment analyses (Figure 4) across all 13 mBERT layers (layers 0-12) for each stimulus set. Across all five stimulus sets (K, A1, A2, A3, and KN) in Study1, speech production (word-level) in Study 2, as well as the conversation task in Study 3, the correlation between neural and semantic RDMs was negative in layers 0-1, progressively shifted toward positive values around layers 4-6, and peaked with statistical significance in layers 9-11.

*$p < .05$, **$p < .01$, ***$p < .001$, permutation test (1000 iterations).

Related to Figure 4.

3. Across all mBERT layers, the mBERT semantic geometry is similar between English and Spanish with rotated subspaces.

We repeated the subspace alignment analyses (Figure 6) across all 13 mBERT layers using stimulus set K (3457 words). Cross-language subspace alignment was significantly lower than within-language alignment at every layer (all $p < 0.001$; **Figure S4A**), confirming that English and Spanish semantic subspaces are consistently rotated relative to each other throughout the model hierarchy. Meanwhile, per-neuron participation correlations between languages were significantly positive across all layers (all $p < 0.001$; **Figure S4B**), with the strongest correlations emerging in higher layers (layers 5-12, Spearman rho > 0.8).

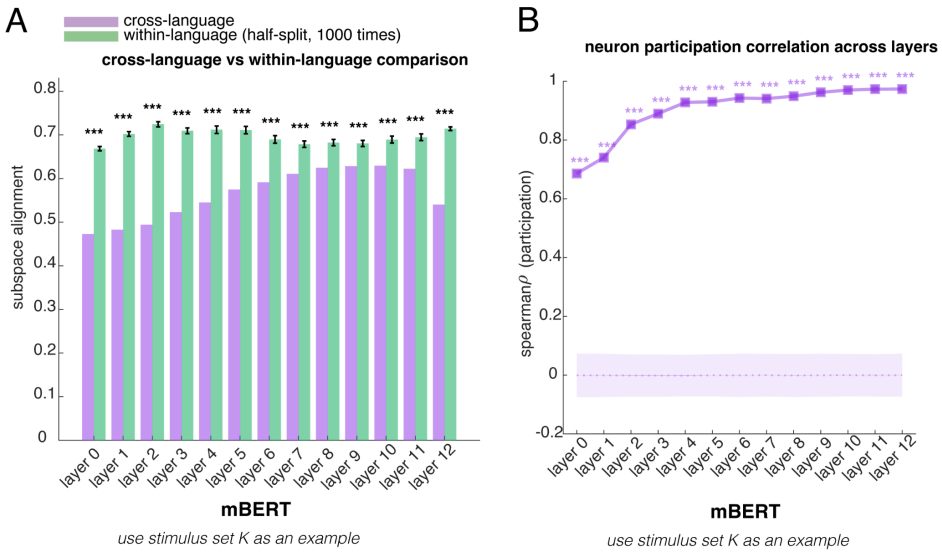


Figure S4. Cross-language subspace alignment and "neuron" participation in subspace construction at every layer of mBERT.
(A) Cross-language subspace alignment (using stimulus set K as an example) is significantly smaller than within-language half-split baselines at every layer of mBERT.
(B) If we treated the embedding dimension of mBERT as a neuron, the per-artificial neuron participation strength in the top subspace (explaining > 95 % variance) was highly correlated between languages (all rho>0.650, all p<0.001) at every layer of mBERT, showing that mBERT utilized the same "hardware" to construct language-specific semantic readout axes to achieve the shared meaning.
*$p < .05$, **$p < .01$, ***$p < .001$, permutation test (1000 iterations)
Related to Figure 6.

Summary of revisions related to comment 2
(1) We revised the Results Section related to Figure 6

(2) We revised Figure 6 to remove the claim “brain-machine similarity”, as well as the Figure caption.
(3) We added Supplemental Figure S2 and Figure S4.
(4) We added one Discussion paragraph to discuss the relationships between mBERT and hippocampus for across-language representation.

3. More Global Measures of Rotation

The ability to predict translated word representations via the Procrustes analysis as presented in Fig. 6 G-H is persuasive, and gives strong validation to the rotated geometry picture presented throughout the paper. However, it left me wanting some more global measures of how cleanly the semantic space as a whole rotates from one language to another.

As a start, it would be nice to see how the prediction accuracy for target word representation changes based on which neighboring words are selected. The authors might plot prediction accuracy as a function of the nearness of the 5 anchor words (accuracy for the 5 nearest as shown, but also for the 2th-6th nearest as anchors, the 3rd-7th, and sliding the window onwards). This would be a quick way to show how local truly rotational translation is.

The reviewer asks a fascinating question about whether geometric rotations are global or local and suggests a method for testing this idea. We agree that this question is important and we have now implemented the suggestion. In short, our new results support the global hypothesis.

In the original analysis, the target word was included in the local neighborhood used to estimate the Procrustes transformation. Our intention was to test whether local neighborhoods are geometrically alignable across languages, using the target as part of that local structure. However, we acknowledge that describing this as “prediction” was misleading, since the target contributed to fitting its own transformation. We apologize for the lack of clarity and have revised the analysis accordingly.

In the revised analysis, we explicitly exclude the target word from the anchor neighborhood, so the rotation is learned entirely from neighboring words and then applied to predict the *held-out* target; and we test predictions for all words in the vocabulary rather than a random subsample.

We first systematically determine the optimal number of anchor words by systematically varying *K* from 3 to 100. For each target word, we identified its k nearest semantic neighbors in English neural space (excluding the target itself), learned a local Procrustes rotation from these anchor pairs, and evaluated prediction accuracy on the held-out target. Then we apply the sliding window to see the prediction accuracy change. While Reviewer 1 did not specifically request these changes, they clearly are in line with the spirit of the reviewer’s request. We then performed the specific analyses suggested by the reviewer.

We revised the Results for Figure 5:

"To test whether preserved semantic structure could support single-word neural prediction, we used local Procrustes alignment (**Figure 5G**, **Methods**). For each target word, we identified its k nearest neighbors in English neural space (excluding the target itself for held-out prediction aim), learned a rotation matrix from these neighbors' cross-language response pairs, and predicted the target's Spanish response. Prediction accuracy was quantified as angular error between predicted and actual vectors. We optimized anchor size (k = 15-19 across datasets; **Figure S3A**) and found median angular errors of 30-61° for listening and word-speech tasks (see example for stimulus set *K* in **Figure 5H,** all results see **Figure S3E**), significantly exceeding null (all $p < 0.001$). Sliding window analysis showed that even distant neighbors (positions 50-100) predicted nearly as well as the nearest neighbors (<1° decay), indicating a near-global rather than strictly local transformation (**Figure S3B & C**)…"

Revised Figure 5G-J:

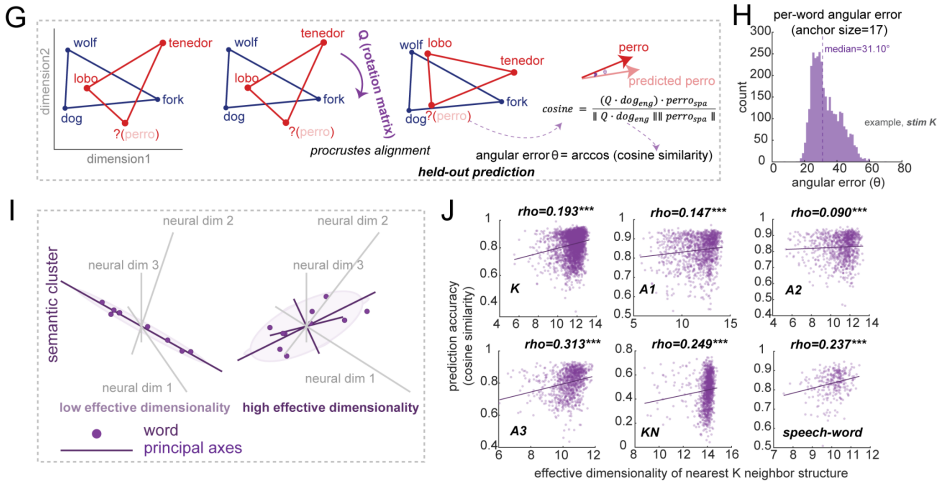


Figure S3-A-B-C:

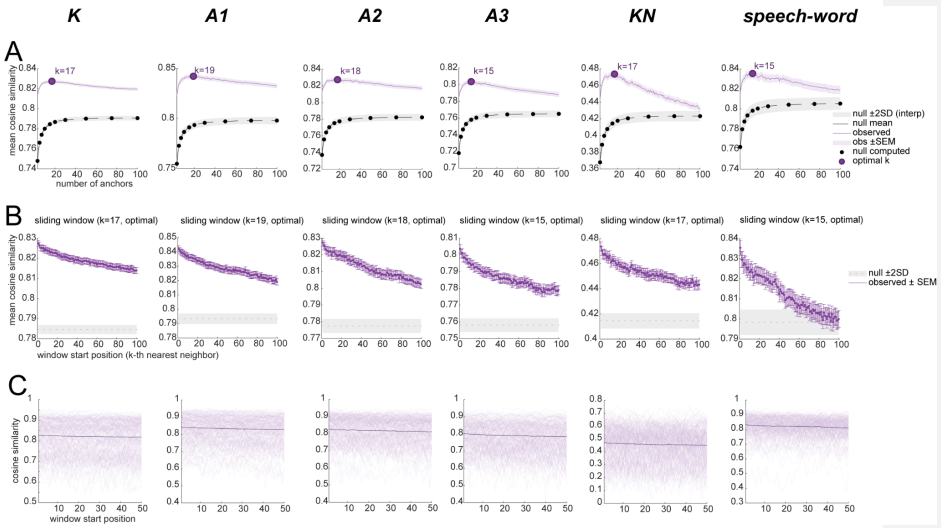


Figure S3. Extended results for local geometric structure enables cross-language prediction
(A) Optimal anchor size. Mean cosine similarity between predicted and actual Spanish neural response vectors as a function of anchor set size (k = 3-100). Purple line shows observed prediction accuracy (±SEM); black dashed line shows null distribution mean with ±2 SD band (gray shading). Black dots indicate anchor sizes where null was explicitly computed (1000 permutations). Large purple dot marks optimal k for each dataset. Prediction accuracy peaks at k = 15-19 depending on dataset and remains stable across anchor sizes. The high null baseline reflects shared global covariance between the two neural spaces; the consistent observed advantage above null reflects the additional precision gained from true semantic correspondence.
(B) Sliding window analysis. Mean cosine similarity as a function of window start position at optimal k. Position 1 uses neighbors ranked 1-k; position n uses neighbors ranked n to n+k-1. Accuracy decays by less than 1° across 100 window positions and remains far above null throughout, indicating the cross-linguistic transformation approximates a near-global rotation rather than a strictly local one.
(C) Per-word sliding window trajectories. Individual word trajectories across window positions. Near-horizontal profiles indicate gradual decay; absence of clustered sharp drops suggests no discrete neighborhood boundaries.

We added the analysis description to Method:

Local geometric cross-language prediction (Figure 5G-J)

To test whether preserved semantic geometry could support single-word neural decoding across languages, we implemented a local Procrustes alignment approach. All neural firing rate vectors were L2-normalized to unit length, which ensures that prediction accuracy reflects directional alignment in neural state space rather than differences in firing rate magnitude.

Anchor size optimization. We first determined the optimal number of anchor words by systematically varying *K* from 3 to 100. For each target word, we identified its k nearest semantic neighbors in English neural space (excluding the target itself), learned a local Procrustes rotation from these anchor pairs, and evaluated prediction accuracy on the held-out target.

Held-out prediction. For each target word with English response vector $\mathbf{e}_{\text{target}}$, we identified its nearest *K* semantic neighbors (*K* was the optimal anchor size from above analysis step) in English neural space using cosine distance (explicitly excluding the target word itself), then extracted these neighbors' responses in both languages to form matrices $\mathbf{X}_E$ (English, $n_{\text{neurons}} \times k$) and $\mathbf{X}_S$ (Spanish, $n_{\text{neurons}} \times k$). After centering both matrices by subtracting column means ($\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_S$), we computed the cross-covariance matrix $\mathbf{M} = (\mathbf{X}_S - \boldsymbol{\mu}_S)(\mathbf{X}_E - \boldsymbol{\mu}_E)^T$ and performed singular value decomposition $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. The optimal orthogonal rotation matrix was obtained as $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$, and the predicted Spanish response was computed as $\hat{\mathbf{y}} = \boldsymbol{\mu}_S + \mathbf{Q}(\mathbf{e}_{\text{target}} - \boldsymbol{\mu}_E)$, then L2-normalized to unit length. Prediction accuracy was quantified as cosine similarity and angular error $\theta = \arccos(\hat{\mathbf{y}} \cdot \mathbf{y}_{\text{true}})$ in degrees (arccos of cosine similarity).

Statistical significance. We tested predictions for all words in each dataset (datasets with matched English-Spanish words, including all stimulus sets in Study 1 and words from short phrases in Study 2). Statistical significance was assessed against 1000 null permutations where we (1) randomly shuffled English-Spanish word pairings using derangement (ensuring no correct translations), (2) randomly selected Spanish prediction targets, and (3) used randomly selected words (excluding the target) rather than true nearest neighbors as anchors for learning rotation matrix Q, thereby destroying both semantic correspondence and neighborhood structure while preserving neural response distributions. P-values were computed as the proportion of null means achieving equal or greater accuracy than the observed mean.

Sliding window analysis. To test whether the cross-linguistic rotation is local or global, we applied a sliding window approach. Instead of using the fixed nearest neighbors (positions 1 to *K*, where *K* represents the optimal anchor size), we shifted the window to use progressively more distant neighbors (for example, if *K*=17, then we have positions 2-18, 3-19, ..., up to positions 84-100). If the rotation were strictly local, prediction accuracy should decay sharply beyond the immediate neighborhood.

I would also suggest using a method which can determine the extent to which decaying accuracy is gradual or sharp (averaging over all word predictions would smooth potentially sharp drops in prediction accuracies for individual words), as sudden decays in accuracy would be indicative of a definitive local semantic neighborhood where rotations are valid.

To address the reviewer's specific concern about whether averaging across words might mask sharp drops for individual words, we examined per-word sliding window trajectories. The evidence uniformly supports gradual rather than sharp decay (see revised **Figure S3C** below).

First, the total per-word change was negligible: for each word, we fit a linear slope to its cosine similarity across the 50 sliding windows. The mean slope was -0.00016 per window, corresponding to a total decline of only Δ = 0.009 cosine similarity (< 1°) across the full range of neighborhood distances. Second, the largest single-step fluctuations for individual words showed no consistent boundary: the positions of maximum drops were broadly scattered across all window positions (mean position = 26.1 out of 49, SD = 14.1), closely matching the theoretical expectation of a uniform distribution (mean = 25.0, SD = 14.1). If definitive local neighborhoods existed, we would expect drops to cluster at a consistent distance; instead, the results suggest stochastic fluctuation rather than systematic boundary crossings.

Figure S3C

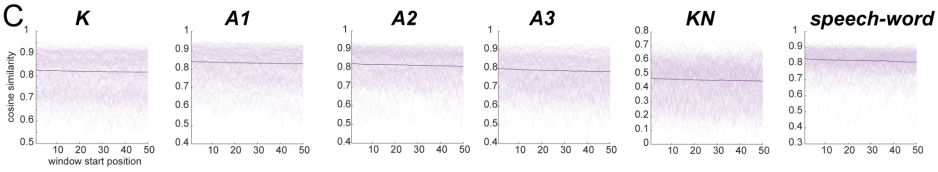


Figure S3

(C) Per-word sliding window trajectories. Individual word trajectories across window positions. Near-horizontal profiles indicate gradual decay; absence of clustered sharp drops suggests no discrete neighborhood boundaries.

We added the analysis to the **Method** accordingly.

Per-word trajectory analysis. To determine whether accuracy decay was gradual or sharp at the individual word level, we computed sliding window trajectories for each word separately. For each word, we calculated: (1) the overall slope of accuracy versus window position, and (2) the maximum single-step drop across adjacent windows. Sharp drops were defined as single-step decreases exceeding 2 standard deviations below the mean derivative across all words and windows.

Even keeping the the 5 closest words as anchors, it would be interesting to see how the absolute closeness of the anchors effects accuracy (i.e. prediction accuracy as a function of the mean proximity of the 5 anchors). It may be that extremely tight clusters over synonyms (e.g. happy, elated, overjoyed…) have less coherence then clusters with conceptual structure (e.g. car, plane, train, bus…).

The reviewer asked a very insightful question. We did this investigation, and the results clearly supported the reviewer's intuition.

First, words embedded in coherent, tight neighborhoods, where neighbors are similar to each other, are predicted most accurately. Words in sparse or irregular regions of the neural space are predicted least accurately (**Figure S3D**).

Figure S3D

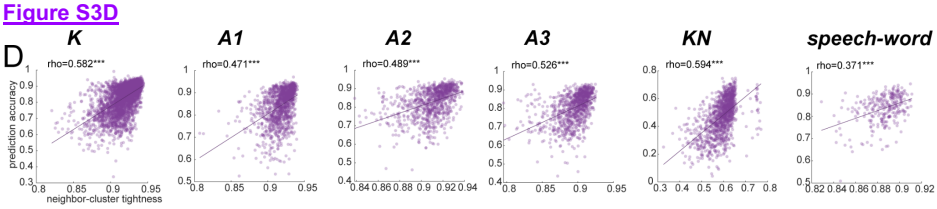
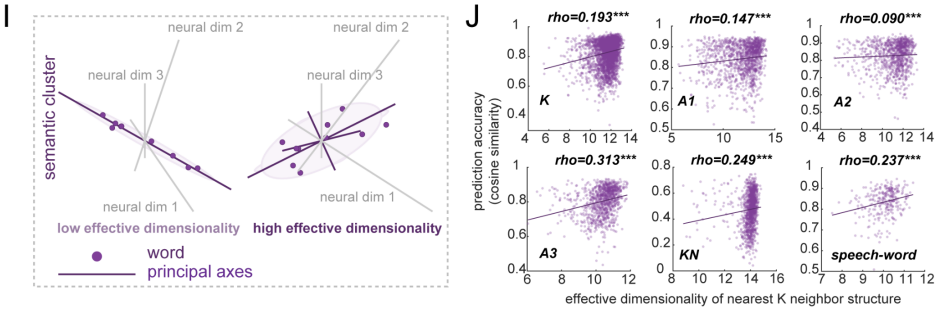


Figure S3.
(D) Neighborhood tightness predicts accuracy. Relationship between neighbor-cluster tightness (mean pairwise neural cosine similarity among k anchors) and prediction accuracy for each word. Strong positive correlations (Spearman rho = 0.37-0.59, all $p < 0.001$) indicate that words in tighter, more coherent neighborhoods are predicted more accurately.

We also did this analysis based on the pairwise semantic cosine similarity (original mBERT embedding space), the conclusions are the same (Spearman rho = 0.20-0.38, all $p < 0.001$).

Second, we found that a more structurally rich organization would perform held-out prediction better. We computed the effective dimensionality (participation ratio from PCA) of each anchor set. The effective dimensionality reflects the geometric structure of the local neighborhood: low values indicate redundant anchors that span a low-dimensional neural subspace, while high values indicate structured anchors that span multiple dimensions. This metric allowed us to test the hypothesis that highly structured neighborhoods might predict better than redundant ones (**Figure 5I-5J**).

Figure 5I-5J



We revised the Results (for Figure 5) as:

"…We further found that the neighbor-cluster tightness strongly predicted accuracy (Spearman rho= 0.37-0.59; **Figure S3D**). Most importantly, higher effective dimensionality of anchor sets, indicating rich structured rather than redundant neighborhoods (**Figure 5I**), was associated with better prediction (Spearman rho = 0.09-0.31, all p<0.001; **Figure 5J**). This aligns with work showing that high-dimensional neural geometry supports cross-condition generalization (Fusi et al., 2016; Bernardi et al., 2020). Structured neighborhoods spanning multiple neural dimensions provide richer geometric constraints for estimating the rotation, enabling better generalization to held-out words than redundant low-dimensional clusters…"

We added the analysis to the Method accordingly.

Local geometry features. To examine what properties of local neighborhoods predict accuracy, we computed two metrics for each target word at optimal *K* (neighbors at positions 1 to k, excluding the target itself): (1) neighbor-cluster tightness, defined as the mean pairwise neural cosine similarity among the *K* neighbors. (2) The effective dimensionality of each anchor set using the participation ratio from PCA on the centered anchor matrix. The effective dimensionality reflects the neural geometric structure of the local neighborhood: low values indicate redundant anchors that span a low-dimensional neural subspace, while high values indicate structured anchors that span multiple dimensions. This metric allowed us to test the hypothesis that highly structured neighborhoods might predict better than redundant ones.

Summary of revisions related to comment 3

(1) We conducted additional analyses including anchor size optimization, sliding window analysis, per-word sliding window trajectories, neighbor-cluster tightness, and anchor set effective dimensionality.
(2) We revised the Results Section related to Figure 5
(3) We revised Figure 5H and added Figure 5I and Figure 5J.
(4) We added Supplemental Figure S3.
(5) We added the analysis procedures to the Methods.

Minor Issues

It's tempting to reach for the cognitive maps analogy in the discussion but I wasn't totally convinced by this. The results aren't really a semantic place code, nor do they represent an invariant semantic basis as grid cells do for space.

Given that this was speculation, and given the reviewer's concerns, we have removed this paragraph from the Discussion.

I found the arrow pointing from Fig 4F. to G. a bit confusing because in F the graphic talks about spearman correlation and then G talks about pairwise distances. It's eventually clear what's happening if one follows the text closely but I feel like the graphic only confuses things as is.

We have revised this figure and its related figure caption for clarity. First, we removed the confusing arrow. And we added the results from speech production (word level, from Study 2: Read Aloud task) based on Reviewer #2's comments, where we used mBERT to extract the embeddings for the words in phrase, therefore the results from all studies can be comparable. Lastly, we explicitly labeled the correlation with *rho* (Spearman correlation, to avoid possible confusion with Pearson correlation r).

The revised **Figure 4** is reproduced below.

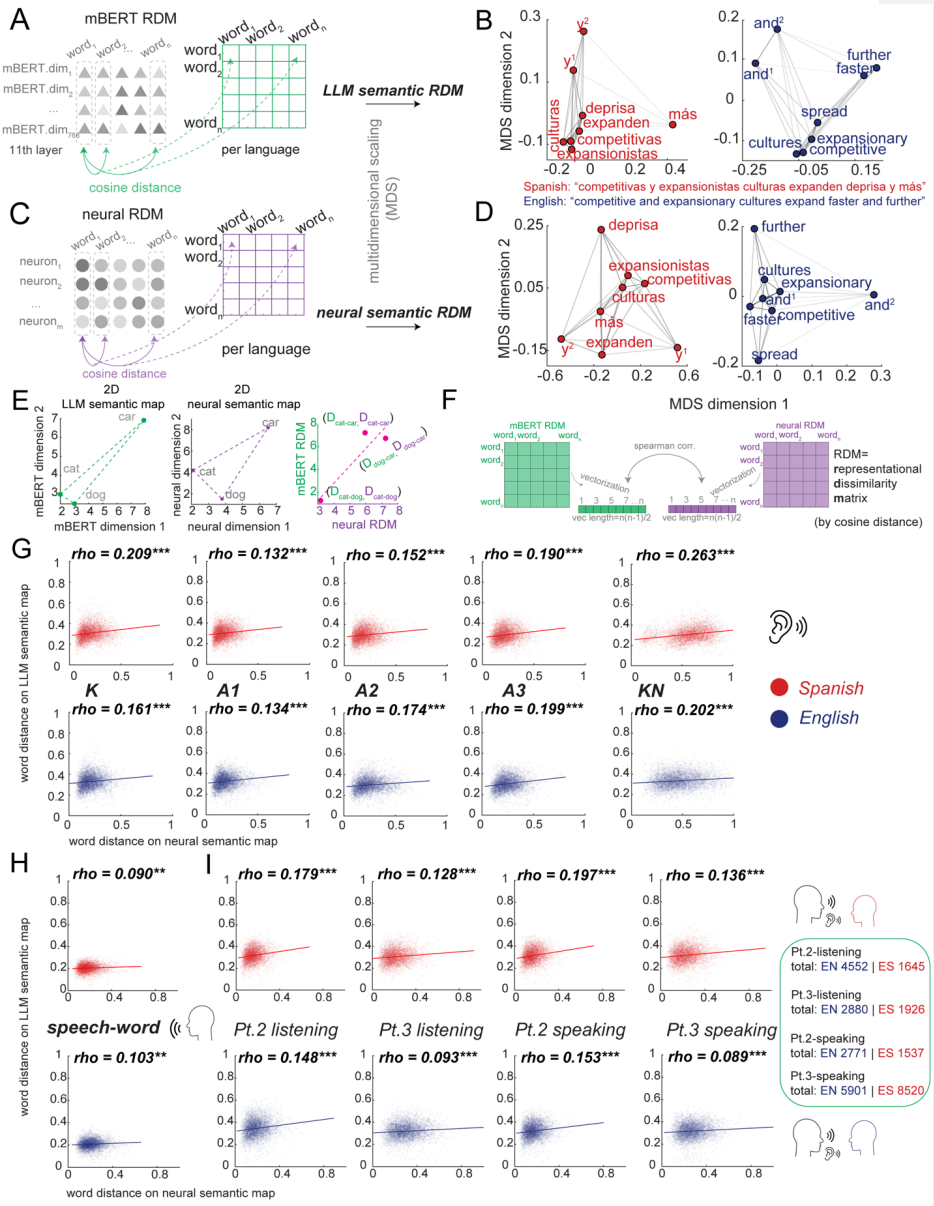

A
mBERT RDM
LLM semantic RDM
multidimensional scaling (MDS)
11th layer
per language
cosine distance
B
Spanish: "competitivas y expansionistas culturas expanden deprisa y más"
English: "competitive and expansionary cultures expand faster and further"
MDS dimension 2
C
neural RDM
neural semantic RDM
D
MDS dimension 1
E
2D LLM semantic map
2D neural semantic map
mBERT RDM
neural RDM
F
spearman corr.
vectorization
RDM= representational dissimilarity matrix
(by cosine distance)
G
rho = 0.209*** rho = 0.132*** rho = 0.152*** rho = 0.190*** rho = 0.263***
K rho = 0.161*** A1 rho = 0.134*** A2 rho = 0.174*** A3 rho = 0.199*** KN rho = 0.202***
word distance on LLM semantic map
word distance on neural semantic map
Spanish
English
H
I
rho = 0.090** rho = 0.179*** rho = 0.128*** rho = 0.197*** rho = 0.136***
speech-word
Pt.2 listening
Pt.3 listening
Pt.2 speaking
Pt.3 speaking
rho = 0.103** rho = 0.148*** rho = 0.093*** rho = 0.153*** rho = 0.089***
Pt.2-listening total: EN 4552 | ES 1645
Pt.3-listening total: EN 2880 | ES 1926
Pt.2-speaking total: EN 2771 | ES 1537
Pt.3-speaking total: EN 5901 | ES 8520

Fig. 5H. should be fleshed out a bit more. It's not clear that the angles shown are an average over many words and we have no idea about the spread of prediction accuracies across words based on this graphic.

We have revised **Figure 5H** for clarity. The revised figure is reproduced here. Also see **Figure S3E** for more details for every dataset with matched English-Spanish words.

Figure 5H:

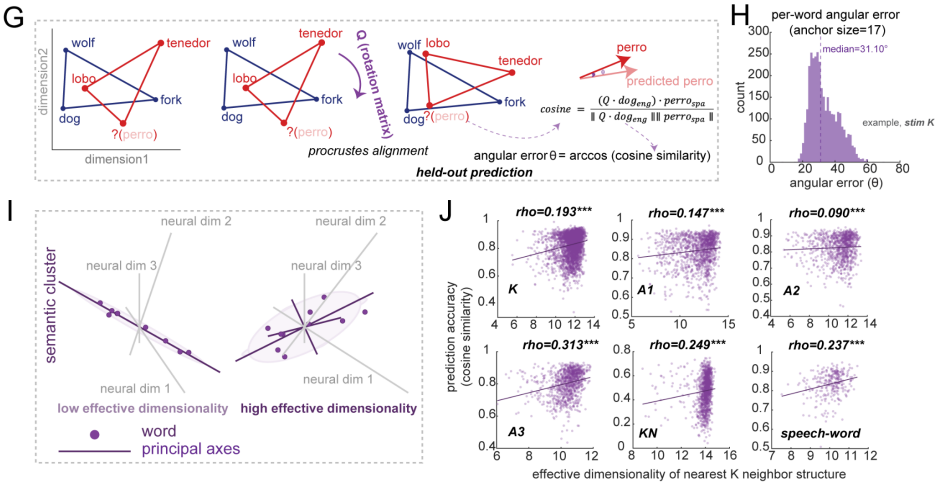
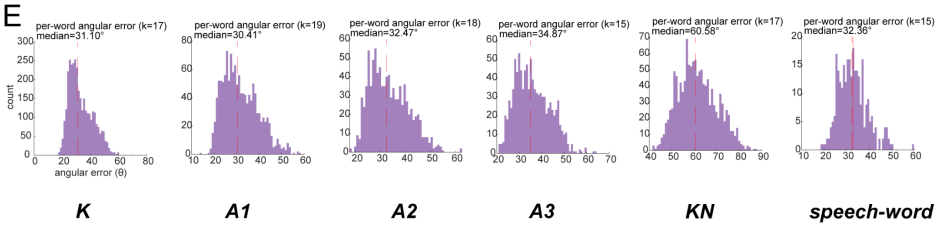


Figure S3E:



In addition to Fig 3F. it would be nice to have a summary plot of the degree of misalignment in beta coefficients across all words.

We thank the reviewer for this suggestion. We now provide the requested summary in Figure S1.

For each neuron, we fit a ridge regression model where the firing rate across all words is predicted by the 100 principal components of semantic embeddings. This yields a single 100-dimensional tuning vector (beta coefficients) per neuron per language, representing that neuron's sensitivity to different semantic features.

To provide a more comprehensive summary, we now include **Figure S1** (see below), which displays the distribution of cross-language tuning similarity (Pearson correlation between English and Spanish tuning vectors) across all neurons for each study.

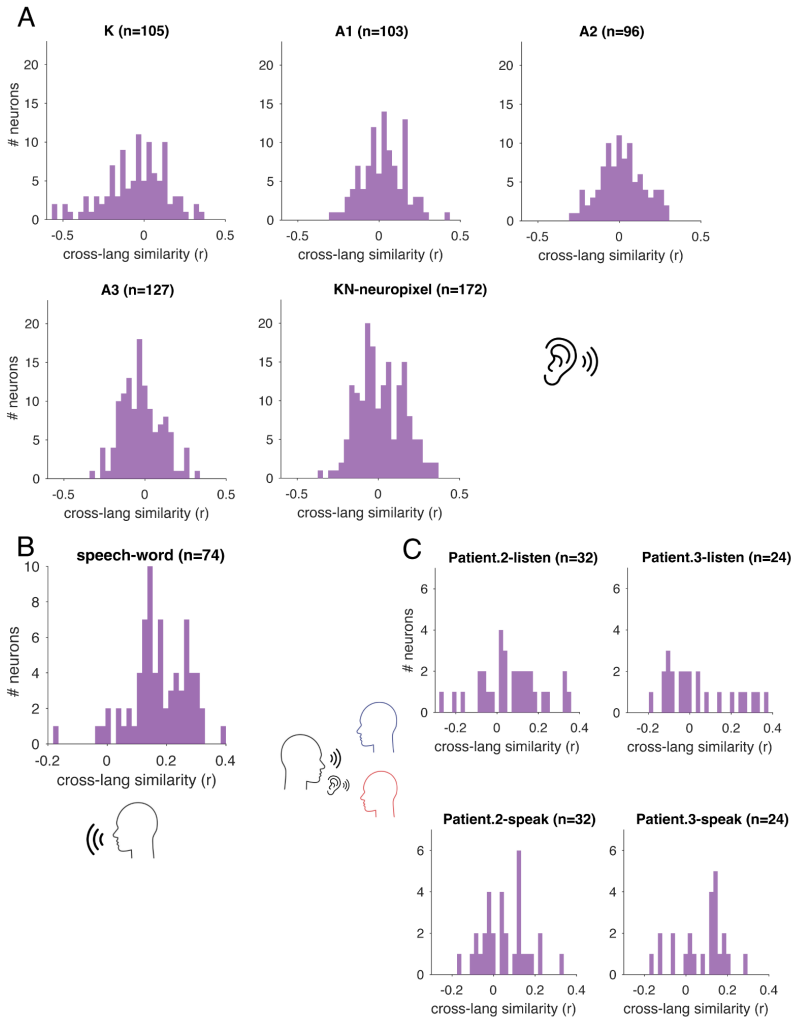


Figure S1. The distribution of cross-language tuning similarity (Pearson correlation between English and Spanish tuning vectors) across all neurons for each study. Related to Figure 3.

Reviewer #2

General remarks

Yan and colleagues use a novel recording technique (neurons in humans) to characterize semantic representation of a novel population (bilingual speakers of English and Spanish). Specifically, they had 4 subjects complete up to three distinct tasks in both English and Spanish: a listening task, a reading aloud task, and a naturalistic conversation task while neurons were recorded from their hippocampus either in the epilepsy monitoring unit or during resective surgery. The authors compare the firing properties of neurons during both languages.
Overall, they find that there is a small amount of overlap between neuronal firing for words/phrases that mean the same thing in both languages, but that the patterns of population activity as differences between word pairs are maintained. Further, the authors show similar results in the activation patterns of an LLM (mBERT), suggesting similarities in how information is represented in both brains and LLMs.
Overall, this is a very interesting study that records from a very novel population. I have some questions/concerns listed below.

We thank the reviewer for these kind and supportive words, below, we provide a point-by-point response to concerns raised.

1. In general, it would be helpful to motivate the behavior a bit more clearly: Why is it important to not have a completely separate or completely overlapping semantic space? What is the behavioral evidence against these positions? How does the neural encoding shown here support this view?

We appreciate the opportunity to add additional clarification on this core conceptual aspect of our paper.

In short:

- The value of having some overlap between semantic spaces is the ability to translate.
- The value of having some difference between semantic spaces is the ability to use a language appropriate to a situation, and to remain stably in that language.

For example, a bilingual speaker needs to conjure up a particular word based on its meaning, but also to do so in the language appropriate for the circumstance. When speaking with Spanish-speaking friends, they would need to conjure the word "*hermana*" and with English-speaking friends they need to conjure the word "*sister*." Thus, while the words may have identical (or nearly identical) meanings, the circumstances where a particular word is

appropriate are different. But if the semantic space is fully identical, with identical readouts, then there is no principled way to differentiate them.

So the encoding problem is that the brain needs BOTH at the same time: (1) translability, (2) context-specificity. Matched geometry with orthogonal readout axes is a straightforward way to solve this problem.

We have now added the following paragraph to the Discussion to clarify this important point:

> *How geometry provides simultaneous shared meaning and language-specific readout.* Bilinguals need to resolve a control problem that monolinguals do not face: they must not only select a word that matches their intended meaning but also select between available equivalent words in different languages. That is, a bilingual speaker who wants to order a glass of water may choose a different equivalent term in El Paso and Ciudad Juárez. That may feel effortless, but it requires a specialized representational scheme that allows for simultaneous, readout-dependent semantic coding, while also allowing for translatability. In other words, bilingual representations are a manifestation of a broader category of generalization/differentiation problems. The need for simultaneous generalization and differentiation is one that is difficult to solve with single neurons but that is readily solved with population codes (Barak et al., 2013; Bernardi et al., 2020; Johnston et al., 2024; Libby & Buschman, 2021; Tang et al., 2020; Tian et al., 2024; Xie et al., 2022). We speculate that language context (provided by surrounding words, interlocutor cues, and even general knowledge) could gate which readout axes' rotation is active (Rigotti et al., 2013).

2. The first sentence of the abstract is grammatically incorrect: "We can understand and express similar concepts in multiple languages. To understand how it does so, …"

Fixed. Now the first sentence of abstract is:

> "The human brain has the remarkable ability to comprehend and express similar concepts in multiple languages."

3. I am not sure that 'translation neurons' is the right term for what you show. I think they are in fact more akin to semantic neurons. While the authors point out that the firing patterns are not entirely identical, it is unclear if this is a fair benchmark. I think that in order to show that these are 'translation' neurons, you would have to add an actual translation task (hear language 1, say language 2), which would be super interesting but also probably beyond the scope of this work.

Reviewers 1 and 2 both raised similar concerns about both our presentation of these neurons and the name we chose for them. In the revised version, we term "*cross-language*

neurons". Following both reviewers' suggestions, we have also put extra work into explaining what these neurons do and do not account for. In short, we now make **it clearer that the bulk of the brain's translation capacity comes from shared geometry, not from special neurons**. We do this through revisions to the Abstract, Introduction (where we now do not discuss them), Results (where we greatly shorten this section, to de-emphasize them), and Discussion, where we now discuss this this way:

Cross-language neurons. We identify a new class of what we call "cross-language neurons," whose responses to co-translated words are correlated. Such neurons would have similar responses to, for example, "*maybe*" and "*quizás.*" These neurons may contribute to translation; however, they do not play an essential, or even a major role: even when we exclude them from our analysis, cross-language semantic geometry is robustly maintained (Table S3). Moreover, while their tuning functions are more aligned than other neurons, they are not perfectly aligned, indicating that they are not simply amodal conceptual representations. Consistent with this, their response fields span multiple semantic categories (e.g., Figure 2C), which differentiates them from the canonical amodal concept-related neurons, concept cells (Quian Quiroga et al., 2005; Quian Quiroga, 2012). Word-level decomposition confirms that their cross-language correlations are driven by a relatively sparse subset of words. These words are constrained along fewer latent semantic axes in mBERT space than other neurons which did not show cross-language correlation (Table S1-S2). We speculate that they may be the extreme end of a larger distribution of neurons whose shared population-level geometry contributes to translation.

> 4. Lines 64-65: "However, these findings do not tell us how the brain matches corresponding concepts between languages while simultaneously maintaining a functional separation." I would like to know more about this functional separation? What is the purpose here? What would happen if the semantic maps for each language were the same? Or perhaps, how are they not the same behaviorally?

This comment raises a similar point to Comment 1. In short, the functional separation is derived from the rotation in the embedding geometry. Our data do not directly tell us why the embedding geometry is rotated, but the literature on coding geometry provides a strong suggestion: geometric rotation often serves a gating function. In this case, speakers need to select not just a word that has a meaning, but a word that is appropriate to the current context. If a speaker is talking to their American friends or their Mexican friends, they may conjure different words with the same meaning. And selecting the right word requires linguistic gating. We have added the following paragraph to the Discussion to clarify this point:

How geometry provides simultaneous shared meaning and language-specific readout. Bilinguals need to resolve a control problem that monolinguals do not face: they must not only select a word that matches their intended meaning but also select between available equivalent words in different languages. That is, a bilingual speaker who wants to order a glass of water may choose a different equivalent term in El Paso and Ciudad Juárez. That may feel effortless, but it requires a specialized representational scheme that allows for simultaneous, readout-dependent semantic coding, while also allowing for translatability. In other words, bilingual representations are a

manifestation of a broader category of generalization/differentiation problems. The need for simultaneous generalization and differentiation is one that is difficult to solve with single neurons but that is readily solved with population codes (Barak et al., 2013; Bernardi et al., 2020; Johnston et al., 2024; Libby & Buschman, 2021; Tang et al., 2020; Tian et al., 2024; Xie et al., 2022). We speculate that language context (provided by surrounding words, interlocutor cues, and even general knowledge) could gate which readout axes' rotation is active (Rigotti et al., 2013).

5. Lines 75-88: What brain regions were used in the cited works? This would be important to situate and motivate the current study.

These studies focus on classic language regions on the lateral surface of the cerebral cortex. In other words, not the hippocampus.

There are two reasons why the hippocampus has generally been ignored: (1) these lateral areas are well-established regions for language, so there is a bias towards studying them, and (2) the hippocampus is tricky to image due to well-known susceptibility artifacts and its structural anatomy, so it is often bypassed. Nonetheless, the close association of the hippocampus with conceptual representation provides a strong motivation for studying it in this translation study, which is focused on cross-modal conceptual representation.

In any case, we have added the following sentence to the revised Introduction:

Regions that elicit overlapping responses include classic language regions, such as the inferior frontal gyrus (IFG) and posterior temporal cortex, as well as several others not as closely associated with language.

And this text into the revised Introduction:

The hippocampus has often been ignored in language research, in part due to difficulty imaging it; however, our laboratory has demonstrated that single neurons in the hippocampus track meanings of words during both listening and monolingual speech production

6. The authors use the words language and meaning/semantics interchangeably. I presume they are focused on the latter (i.e. if you presented pictures the space would be the same)? If so, this should be clarified and the usage of the term language should be reserved for the appropriate context in the manuscript.

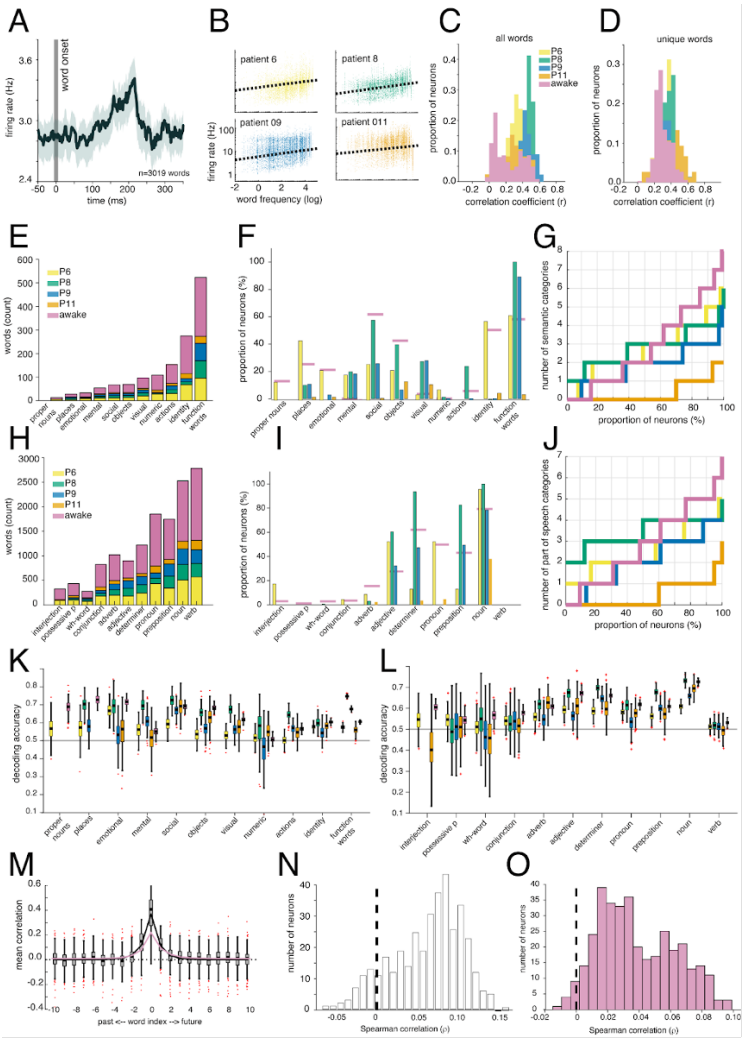
We apologize for this. We have now revised the manuscript to ensure consistent correct usage of these terms.

7. Lines 141-142: How would the removal of the anterior temporal lobe affect semantic processing in the hippocampus (for the neuropixel recordings)? This is especially important given the role for semantics in the ventral temporal stream.

The reviewer raises a fascinating question. As the reviewer correctly notes, our Neuropixel recordings (n=1 patient) occurred in the hippocampus following the resection of the anterior temporal lobe. That in turn raises questions about whether semantic processing is modified. Perhaps surprisingly, the answer seems likely to be no. We have another manuscript about this, with several more patients, using English speech only (Katlowitz et al., 2025).

https://www.biorxiv.org/content/10.1101/2025.04.09.648012v1.full

In that manuscript, we show strong quantitative overlap between semantic encoding in awake and anesthetized patients. The key figure (Figure 4 in that paper) is reproduced below for convenience. In this figure, the pink data refers to the awake patients while the other colors refer to anesthetized. Overall, these two datasets overlap surprisingly closely.



Moreover, our results collected here show a similar pattern. We find very little difference between the codes for semantics in the awake patients and the resected (and anesthetized) patient. In any case, we have added the following new text to the revised Discussion (in the paragraph on Limitations of the Study):

"Finally, some of our data come from an anesthetized patient whose anterior temporal lobe had been resected prior to recording (but the Spanish data from the same patient have never been published or

analyzed). Both the anesthesia and the resection may have altered patterns of responding (but see Katlowitz et al., 2025), although we find largely similar results between this patient and our other three"

> 8. Line 224: Introduction is misspelled.

We thank the reviewer for pointing this out. Given other revisions, this passage no longer appears in the revised manuscript.

> 9. I think the contrast of interest here is not translation vs. anti-translation, but translation vs. separate firing. Are there neurons that respond to a word in one language but not in the other?

The reviewer raises an interesting question: whether there are language-specific neurons. This is different from a question we did address, which is whether the translation neurons (called "cross-language neurons" in the revision) are real or a statistical artifact, which we did by comparing their preponderance to ostensible anti-translation neurons. While we did not answer the reviewer's question in our original manuscript, we agree with the reviewer that it is a good question to answer.

In short, the answer to the reviewer's question is no. Neurons, in general, have complex, mixed tuning, and respond, on average, with equal intensity for words in both languages. Although a very small number (below chance) of neurons appeared responsive in only one language in some stimulus sets, these counts did not exceed chance expectations (binomial test, all $p > 0.300$). In contrast, cross-language overlap was significantly enriched across the population for all stimulus sets (Wilcoxon signed-rank test, all $p < 0.001$). In other words, it appears that the language separation, like the language linkage, occurs through high-dimensional population geometry rather than through specialized neurons.

In detail:

To address this question, we conducted the following analyses and reported our results below.

From the revised **Methods**:

Whether there are language-specific neurons (Table S4)

To test whether individual hippocampal neurons were driven by the same words across languages or by different words in each language (language-specific or "separate" firing), we performed a word-level overlap analysis. For each neuron, firing rates were z-scored within each language across all words. A neuron was considered responsive to a word if its firing rate exceeded a fixed threshold ($z > 1$, that is, ~top 16% responsive words). This yielded two binary response sets per neuron: English-responsive words and Spanish-responsive words. Responsive words were partitioned into three mutually exclusive categories: words eliciting responses in both languages (shared), words eliciting responses only in English, and words eliciting responses only in Spanish. To assess whether the observed overlap between

English- and Spanish-responsive word sets exceeded chance expectations given each neuron's marginal response rates, we used an exact hypergeometric test (without replacement). For each neuron, the null distribution assumed random overlap between English-responsive and Spanish-responsive word sets, conditioned on the total number of words, the number of English-responsive words, and the number of Spanish-responsive words. We computed the probability of observing an overlap greater than or equal to the empirical overlap under this null. Neurons with $p < 0.05$ were classified as showing overlap enrichment beyond chance. Based on this analysis, neurons were categorized as *pure-shared* (only shared responses), *pure-separate* (only language-specific responses), or *mixed* (both shared and language-specific responses).

From the revised **Results**:

Table S4. No language-specific neurons

In the stimulus set K and KN (Study1), at the single-neuron level, no neuron was selectively active in only one language. Every neuron exhibited a mixture of shared and language-specific word responses. In the stimulus set A1, A2, A3 (Study1) and read aloud task (speech production, word level, Study2), although a small number of neurons appeared responsive in only one language, these counts did not exceed the rate expected by chance (binomial test, all $p > 0.300$). For all stimulus sets, across the population, cross-language overlap was significantly enriched relative to chance (Wilcoxon signed-rank test, all $p < 0.001$; Cohen's $d > 0.30$). This matches our broader conclusion: shared meaning is not implemented by special neurons, but by population geometry.

	stimulus-K	stimulus-A1	stimulus-A2	stimulus-A3	stimulus-KN	Speech-word
mixed	105	101	92	119	172	72
Pure shared	0	0	0	0	0	0
Pure selective	0	2	4	8	0	2
selective-neurons above chance?	p=1.000	p=0.967	p=0.713	p=0.303	p=1.000	p=0.890
cross-language overlap exceeded chance	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001	p<0.001

10. It is interesting that for production, the correlation at the phrase level is so much lower than that at the word level. Why is this? I presume it is because these neurons are tuned at the word level, but it would be useful to explain further. Is there something about the phrases that are correlated that drive similar firing compared to those that are not (but 'should be')?

The reviewer accurately describes a pattern that is observed in our data, but that we did not dwell on in our original manuscript. Indeed, for production, our data indicate that hippocampal neurons are most strongly tuned at the word level (also see our response to comments #13-#14 below). We have now done some additional analyses that confirm and modestly extend the original observation. The reviewer's surprise is therefore well justified, and we agree it is worth noting this point more clearly. We do not know why but we are inclined to

agree with the reviewer's intuition that these neurons are simply more selectively tuned for meaning at the word level than the phrase level.

One reason this is notable is that, in the listening data (reception), the word and phrase level are very similar in terms of both encoding model performance and in terms of decoding model. That vast majority of data, of course, comes from the listening condition. But it is often assumed that production will mirror it. However here we report a clear dissociation in the production level (when words are produced in isolation). We note that the neural embeddings for production are not nearly as well understood as for reception, so it is perhaps not as surprising that things are different between the two modalities.

Our extended analyses:

Phrase-level cross-language neurons are a subset of word-level cross-language neurons. All 4 phrase-level cross-language neurons were also identified at the word level, with zero phrase-only cross-language neurons, indicating that phrase-level detection captures the strongest end of a continuous distribution rather than a distinct population.

Word-level cross-language neurons show above-chance phrase-level correlations. The 58 word-level-only cross-language neurons do show cross-language similarity at the phrase level, but at sub-threshold effect sizes. Using a permutation baseline (1000 shuffles of Spanish phrase assignments), 67.2% (39/58) of these neurons showed phrase-level correlations above their individual permutation null means, significantly exceeding the 50% expected by chance (binomial test, $p = 0.006$). However, only 3/58 (5.2%) exceeded their individual permutation null at $p < 0.05$, confirming that the per-neuron effects are real but too small to detect individually with 99 phrases.

The dominant factor is effect size reduction from phrase-level averaging. Observed effect sizes were dramatically smaller at the phrase level (mean $r = 0.03$) compared to the word level (mean $r = 0.29$; paired sign-rank $p<0.001$). We believe this may occur because these neurons are tuned at the word level, and averaging firing rates across 3-4 words within each phrase compresses word-specific semantic signals. In particular, English and Spanish translations of the same phrase often differ in word order and in the number of content versus function words (e.g., "let's have fun" / "vamos a divertirnos"), so the semantic contribution of individual words does not align temporally within the phrase-level averaging window. Combined with the smaller sample size (99 phrases vs. 287 matched words), this effect size reduction could account for the large difference in detection rates between phrase-level and word-level analyses.

Therefore, we added the following text into the Results:

"A subset of neurons met the cross-language neuron criterion for both listening (**Figure 2E**) and speaking (**Figure 2F**; also see **Table S1-Table S2**). Note that sessions of the listening task (Study1) and speech production (Study2) were collected on different days, so we cannot tell if the same cross-language neurons contribute to speaking and listening. The difference between phrase level and word level likely reflects signal attenuation from averaging across words within each phrase, as phrase-level effect sizes were substantially smaller than word-level effect sizes (mean r

= 0.03 vs. 0.29; Wilcoxon signed-rank test, $p<0.001$). All phrase-level cross-language neurons were also identified at the word level, and 67.2% of word-level-only cross-language neurons showed phrase-level correlations above their permutation null (binomial test, $p = 0.006$), confirming that the cross-language signal exists at the phrase level but maybe too weak for per-neuron detection..."

> 11. How was the initial number of neurons established? Are these just isolated neurons, or are they neurons that fire to any stimuli? Please clarify.
> 12. Relatedly, are these single isolated neurons or multi-units? This should be clarified.

We apologize for neglecting these methodological points.

We used all neurons, not just responsive neurons. (This is a consistent lab policy, and we believe it is important because preselecting responsive neurons can introduce a variety of statistical problems, many of them subtle).

For this study, we focused on isolated neurons. While using multi-unit activity has a great deal of utility in many cases, here we adopted the more conservative approach of limiting our analyses to isolated units.

The following text now appears in the revised Methods:

"The initial number of neurons was simply the total number of neurons that we collected. All preprocessing was done by a lab member blind to the goals of the study. In preprocessing (spike sorting), we only kept the isolated neurons. We analyzed all collected neurons, and did not restrict analyses to task-responsive neurons."

> 13. It would be helpful to use word2vec to analyze the data from the listening experiments to better compare to the production experiments. I understand the motivation, but now there is a confound of model between the two experiments. Using word2vec would also allow for the analysis of all datasets for your analyses in each use case.
> 14. Lines 322-323: Did the variance explained by 100 components differ substantially between the two different models?

Since these two comments are related, we combined them and addressed them together. Though our central aim is not to compare or contrast results between listening and production experiments, we agree that it would be informative to use the same LLM model in the production study as the other experiments did.

We unified analyses with mBERT embeddings

After considering the reviewer's suggestion for consistency, we now use mBERT to study the read-aloud task. Since most of the experiments were done in contextual situations, to

better capture the across-language understanding, we re-did the analysis by using the mBERT for production (read aloud task). For the following reasons:

(1) mBERT is trained on 104 languages simultaneously with a shared vocabulary and shared transformer parameters. This results in semantically equivalent words across languages (e.g., “dog” and “perro”) being naturally embedded in the same representational space without requiring any post-hoc alignment procedures. In contrast, word2vec models are trained separately for each language, producing embedding spaces that are not inherently aligned. Aligning word2vec spaces across languages requires additional steps, such as learning projection matrices from bilingual dictionaries or performing joint PCA (like our previous version), which introduces methodological assumptions.

(2) The speech production experiment still required some context because of the short phases we chose. e.g., “your sister is tall” and “now I am hungry” (see Supplementary Table 3 in Silva et al., 2024). Since the task presented the English and Spanish in interleaved order, and fortunately, mBERT can still capture within-phrase contextual information through its transformer architecture. Therefore, we can extract embeddings for isolated phrases (without cross-phrase context). This provides richer semantic representations compared to the static word embeddings from Word2Vec.

(3) Within the mBERT (original dimensions=768), the first 100 PCs explain 69.71% of total variance in K, 72.99% in KN, 70.65% in A1, 74.49% in A2, 74.49% in A3 and 82.94% in speech-word level (read aloud task). Now the 100 PCs for mBERT embeddings for the read aloud task (speech production) explained 82.94% of total variance. The higher variance explained in the speech production task reflects the smaller vocabulary size (288 vs. 3457 matched words in K, for example), which constrains the effective dimensionality of the embedding matrix.

Related revisions in METHODS

mBERT embeddings extraction for short phrase in read aloud task:

We extracted mBERT embeddings using the pretrained *bert-base-multilingual-cased* model from HuggingFace Transformers. Each phrase was tokenized and passed through the model independently (i.e., without cross-phrase context), with input formatted as [CLS] phrase [SEP]. We extracted hidden states from layer 11 (of 13 transformer layers), as middle-to-late layers have been shown to capture more semantic rather than syntactic information. For phrases containing multiple words, we obtained word-level embeddings by averaging the hidden states of all subword tokens belonging to each word. The resulting embeddings were 768-dimensional vectors for each word.

Updated results for the read-aloud task (Figure 3B and Figure 3H)

Accordingly, we updated the text in Results as:

“For Study 2 (read aloud task), words were produced in short phrases without narrative context. We therefore extracted mBERT embeddings for each phrase independently (Methods). Both Layer 10 and Layer 11 showed the highest cross-language similarity for matched translation pairs (Figure 3B, blue line); we used Layer 11 for consistency with Studies 1 and 3.”

We have also updated the manuscript to report validation statistics (for joint PCA) aggregated across all analyses. Because all analyses used unified mBERT embeddings, we updated the following text as:

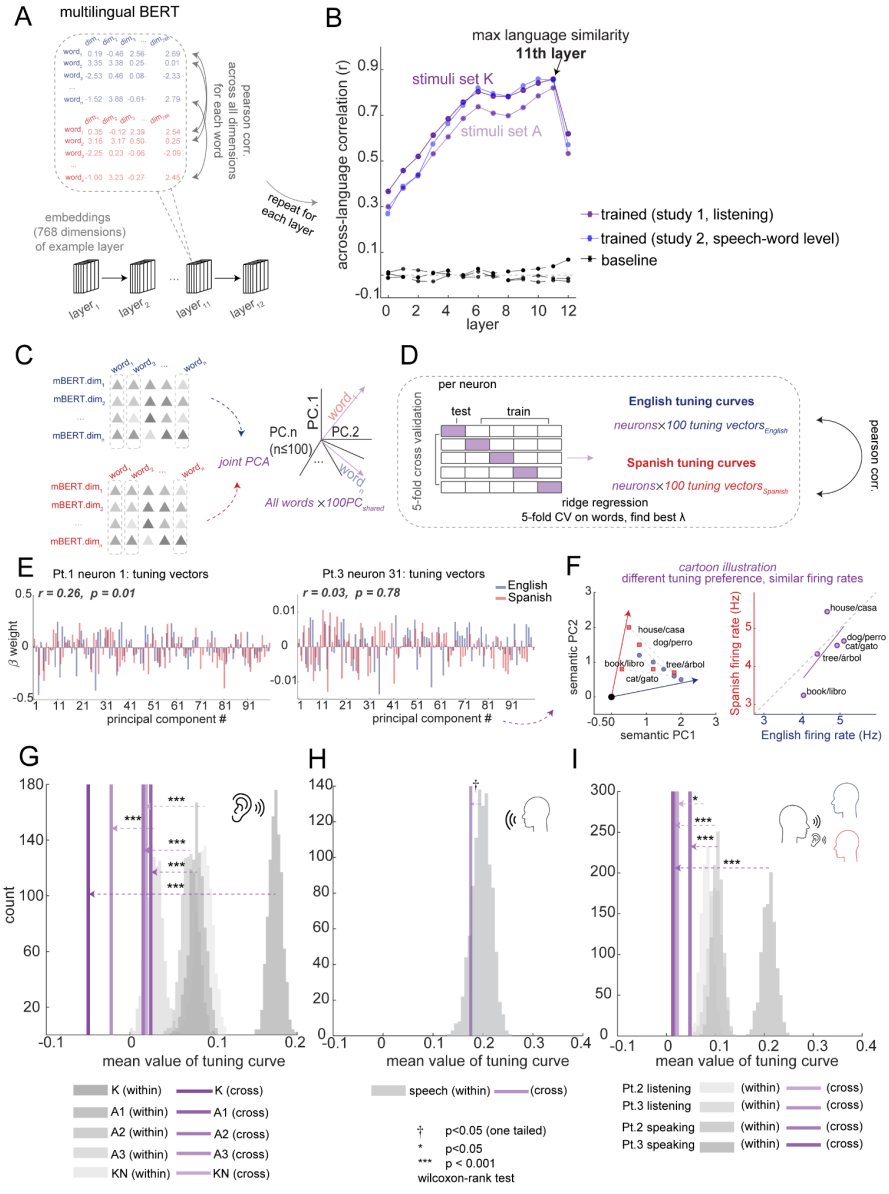
> "..To validate this shared PCA space, we conducted two control analyses. First, for each shared component, we computed the fractional contribution of each language to the total variance explained by that component, which yielded nearly balanced contributions (English = 0.509; Spanish = 0.491). Second, we confirmed that each principal component derived from the joint PCA showed a strong correlation between English and Spanish scores (average correlation, $r = 0.501$, all PCs, $p < 0.001$), indicating that both languages aligned along the same semantic axes. These results confirm that the shared PCA captured common axes for both languages."

The across-language tuning similarity is marginally significantly lower than the within-language tuning similarity (paired Wilcoxon (within > cross): $p = 0.06$). We updated **Figure 3H**. We explained the results as consistent with the high proportion of cross-language neurons at the word level in this condition (83.6%), indicating that individual neurons' semantic tuning is largely shared across languages when processing words during speech production (isolated from the broader semantic contexts).

Accordingly, we updated the text in Results as:

> "For Study 2 (speech production) at the word level, however, the cross-language tuning correlation did not significantly differ from within-language tuning correlations (Wilcoxon signed-rank test, $p = 0.06$, Figure 3H). This result is consistent with the high proportion of cross-language neurons observed at the word level in Study 2 (83.6%, Figure 2F). Cross-language neurons ($n = 115$) showed significantly higher cross-language tuning similarity (mean = 0.14, SD = 0.11) than other neurons ($n = 560$, mean = -0.01, SD = 0.15; Welch's t-test: $t(204) = 11.83$, $p < 0.001$), suggesting that when words are produced in isolation, hippocampal neurons may exhibit largely language-invariant semantic tuning. Taken together, semantic tuning profiles generally differ across languages, with the notable exception of isolated word production, where cross-language neurons maintain similar tuning across English and Spanish."

Revised Figure 3 (see below):


A
multilingual BERT
embeddings
(768 dimensions)
of example layer
pearson corr.
across all dimensions
for each word
repeat for
each layer
B
max language similarity
11th layer
stimuli set K
stimuli set A
across-language correlation (r)
layer
trained (study 1, listening)
trained (study 2, speech-word level)
baseline
C
joint PCA
PC.1
PC.2
PC.n
(n≤100)
All words ×100PC
D
per neuron
test
train
5-fold cross validation
English tuning curves
neurons×100 tuning vectors
Spanish tuning curves
neurons×100 tuning vectors
ridge regression
5-fold CV on words, find best λ
pearson corr.
E
Pt.1 neuron 1: tuning vectors
r = 0.26, p = 0.01
Pt.3 neuron 31: tuning vectors
r = 0.03, p = 0.78
English
Spanish
β weight
principal component #
F
cartoon illustration
different tuning preference, similar firing rates
semantic PC2
semantic PC1
house/casa
dog/perro
book/libro
tree/árbol
cat/gato
Spanish firing rate (Hz)
English firing rate (Hz)
G
count
mean value of tuning curve
K (within)
K (cross)
A1 (within)
A1 (cross)
A2 (within)
A2 (cross)
A3 (within)
A3 (cross)
KN (within)
KN (cross)
H
mean value of tuning curve
speech (within)
(cross)
† p<0.05 (one tailed)
* p<0.05
*** p < 0.001
wilcoxon-rank test
I
mean value of tuning curve
Pt.2 listening
Pt.3 listening
Pt.2 speaking
Pt.3 speaking
(within)
(cross)

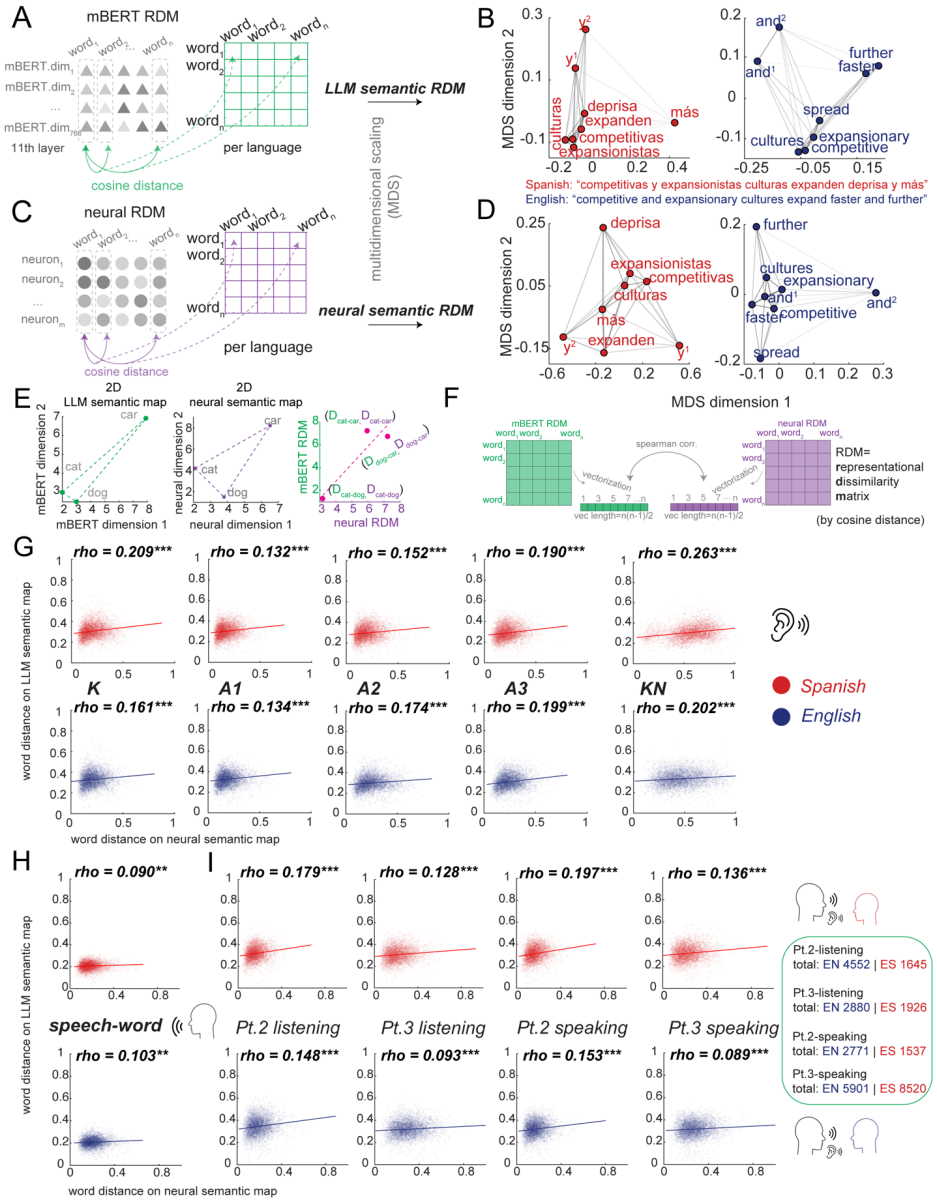
Added brain-LLM alignment analysis for Study 2 (Figure 4H)

We added a new panel as **Figure 4H** to show the brain-LLM alignment for the read aloud task, and updated the Results accordingly (see below). Therefore, all studies (study1, study2, study3) have the results (we used Layer 11, because based on the mBERT embeddings for the read aloud task, both Layer 10 and Layer 11 show the highest cross-language similarity across the matched words)

Related revisions in RESULTS for Figure 4

"... This alignment demonstrates that word pairs that are distant in the *LLM semantic map* (e.g., "*human*" and "*galaxy*") are also distant in the *neural semantic map*, while close word pairs (e.g., "*cat*" and "*dog*") in the *LLM semantic map* would be close in the *neural semantic map*. These alignments are robust in both English and Spanish. Study 2 (read aloud task) also showed similar brain-model alignment (**Figure 4H**; Spearman's rho = 0.090-0.103, all $p < 0.01$, permutation test). Study 3 (conversation) showed … "

Revised Figure 4 (see below):


A
mBERT RDM
11th layer
cosine distance
per language
LLM semantic RDM
multidimensional scaling (MDS)
B
MDS dimension 2
culturas
deprisa
expanden
competitivas
expansionistas
más
spread
further
faster
expansionary
competitive
cultures
Spanish: "competitivas y expansionistas culturas expanden deprisa y más"
English: "competitive and expansionary cultures expand faster and further"
C
neural RDM
neural semantic RDM
D
MDS dimension 2
MDS dimension 1
E
2D
LLM semantic map
neural semantic map
mBERT dimension 1
mBERT dimension 2
neural dimension 1
neural dimension 2
mBERT RDM
neural RDM
car
cat
dog
F
mBERT RDM
neural RDM
spearman corr.
vectorization
vec length=n(n-1)/2
RDM= representational dissimilarity matrix
(by cosine distance)
G
word distance on LLM semantic map
word distance on neural semantic map
rho = 0.209***
rho = 0.132***
rho = 0.152***
rho = 0.190***
rho = 0.263***
K
A1
A2
A3
KN
rho = 0.161***
rho = 0.134***
rho = 0.174***
rho = 0.199***
rho = 0.202***
Spanish
English
H
I
rho = 0.090**
rho = 0.179***
rho = 0.128***
rho = 0.197***
rho = 0.136***
speech-word
Pt.2 listening
Pt.3 listening
Pt.2 speaking
Pt.3 speaking
rho = 0.103**
rho = 0.148***
rho = 0.093***
rho = 0.153***
rho = 0.089***
Pt.2-listening total: EN 4552 | ES 1645
Pt.3-listening total: EN 2880 | ES 1926
Pt.2-speaking total: EN 2771 | ES 1537
Pt.3-speaking total: EN 5901 | ES 8520

Additional subspace and neuron participation analyses for Figure 6C-6D-6E

We added the speech production results (word level) into Figure 6 for the subspace alignment and neuron participation ratio analyses.

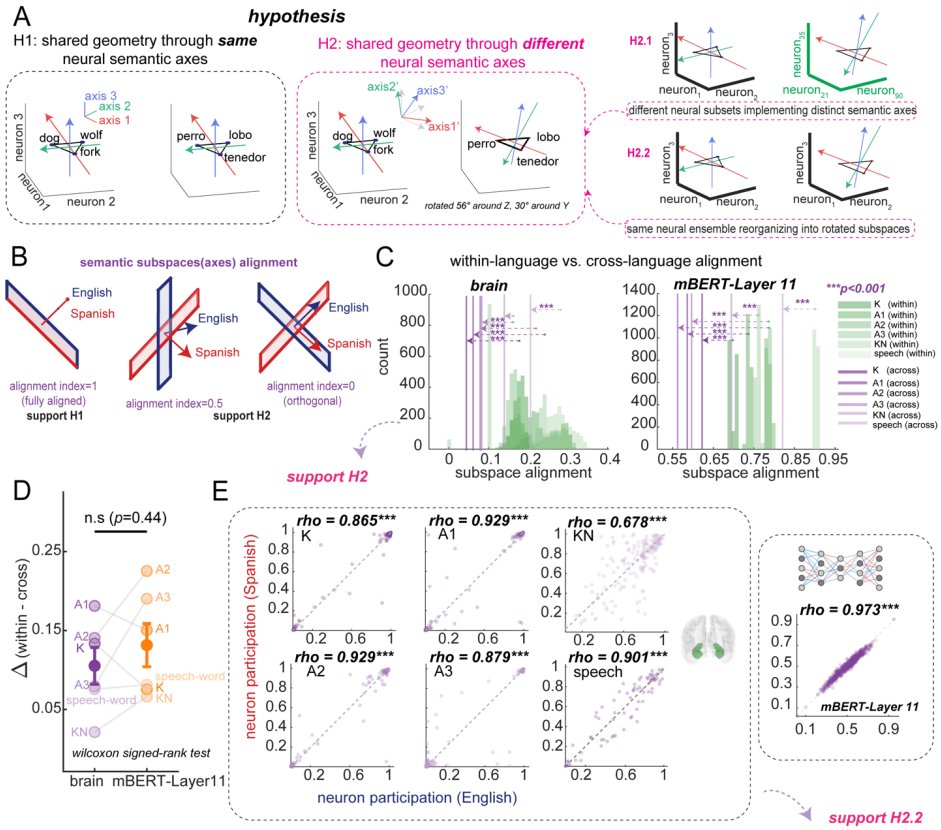
Of note, we reported Spearman correlation value in Figure 6E in the revised manuscript, which is robust to outlier-driven effects (also see our responses for *comment#20*)

Related revisions in RESULTS for Figure 6

"...Despite the similar semantic geometry, the manifold constructed from semantic axes for English and Spanish were less aligned than the within-language subspaces both for mBERT (K, A1, A2, A3, KN, speech-word, all $p<0.001$) and hippocampus (K, A1, A2, A3, speech-word, all $p<0.001$; KN-Neuropixel, $p>0.05$, **Figure 6C**). But the absolute across-language subspace alignment value in mBERT is substantially larger than values (both within and across-language subspace alignment) observed in the brain (**Figure S4A**). The reduction in alignment from within-language to cross-language conditions was comparable between the hippocampus and mBERT (hippocampus=0.11±0.06 vs. mBERT-Layer11=0.13±0.07, $p = 0.44$, Wilcoxon signed-rank test; **Figure 6D**) ..."

"… Each neuron's participation strength (the sum of squared loadings across the top eigenvectors, reflecting how heavily it contributes to these coordination modes), was highly correlated between languages (**Figure 6E**; K: rho=0.87; A1: rho=0.93; A2: rho=0.93; A3: rho=0.88; KN-Neuropixel: rho=0.68; speech-word, rho=0.90, all $p<0.001$; mBERT all layer results, see **Figure S4B**) ..."

Revised Figure 6 (see below):



15. Line 333: I think there is a word missing ('to' after regression?)

We thank the Reviewer for pointing this out. Now we have fixed it.

16. Lines 563-570: did you statistically compare the embedding match with and without the translation neurons (as opposed to simply showing that the effect is still present without the translation neurons)? This would be useful to establish the relative importance of these neurons in this alignment.

We agree with the reviewer that adding this information would be helpful. We now do so in the revised manuscript, summarized in the following table.

Specifically, we chose the Steiger's Z-test for dependent correlation difference significance test. To test whether cross-language neurons (previously called "translation neurons") disproportionately contribute to cross-language representational alignment, we compared the observed drop in English-Spanish RDM (as well as in neural-LLM correlation) upon removing cross-language neurons against a null distribution generated by randomly excluding the same number of neurons 1000 times. The results show no differences (all p>0.400), indicating that cross-language neurons do not contribute more to cross-language geometric similarity than a random subset of equal size. We repeat such analysis for every stimulus set.

Revised in paper:

(1) We added **Table S3**.

Table S3. Cross-language neurons did not affect shared representation geometry.
We chose the Steiger's Z-test for dependent correlation difference significance test. To test whether cross-language neurons disproportionately contribute to cross-language representational alignment, we compared the observed drop in English-Spanish RDM (as well as in neural-LLM correlation) upon removing cross-language neurons against a null distribution generated by randomly excluding the same number of neurons 1000 times. The results show no differences (all p>0.400), indicating that cross-language neurons do not contribute more to cross-language geometric similarity than a random subset of equal size. We repeated such analysis for every stimulus set.

neural-LLM		K	A1	A2	A3	KN
r_with	spanish	r=0.209***	r=0.132***	r=0.152***	r=0.190***	r=0.263***
	english	r=0.161***	r=0.134***	r=0.173***	r=0.199***	r=0.202***
r_without	spanish	r=0.208***	r=0.164***	r=0.153***	r=0.188***	r=0.259***
	english	r=0.160***	r=0.174***	r=0.171***	r=0.199***	r=0.199***
Steiger's Z (r_without vs. r_with)	spanish	z = 0.329, p = 0.742	z=-21.751, p<0.001	z=-0.563, p=0.573	z = 0.846, p = 0.397	z=2.105, p=0.035
	english	z = 0.675, p = 0.500	z=-28.278, p<0.001	z=1.029, p=0.303	z = 0.723, p = 0.469	z=2.136, p=0.033
Permutation p-value for (Δr vs. Δr_random)	spanish	p=0.751	p=1.00	p=0.764	p=0.511	p=0.206
	english	p=0.747	p=1.00	p=0.636	p=0.506	p=0.041

English-Spanish neural geometry similarities	K	A1	A2	A3	KN
r_with	r=0.477***	r=0.255***	r=0.281***	r=0.358***	r=0.385***
r_without	r=0.474***	r=0.341***	r=0.223***	r=0.316***	r=0.379***
Steiger's Z (r_without vs. r_with)	Z = 3.758, p = 0.0001	Z = -62.502, p<0.0001	Z = 2.447, p = 0.01	Z = 3.963, p < 0.0001	Z = 3.656, p = 0.0002
Permutation p-value for (Δr vs. Δr_random)	p=0.773	p =1.000	p=0.736	p=0.486	p=0.606

Table Note. r_with = Cross-language RDM similarity (with cross-language neurons);
r_without = Cross-language RDM similarity (without cross-language neurons);
Δr = r_with − r_without;
Δr_random = mean Δr from random neuron exclusion (of equal size) (mean ± SD)
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

(2) We added the details into the related Results as below

Cross-language neurons are not essential for cross-language neural semantic geometry

We next confirmed that the previously identified cross-language neurons are not responsible for the cross-language geometry match. To test this, we repeated the analyses above excluding the cross-language neurons (that is, the ones that showed similar English-Spanish firing profiles, identified in **Figure 2**, n=115). We found that even after excluding these neurons, we still observed brain-model alignment within languages (all p <0.001) and cross-language neural semantic map correlations (all p <0.001). Critically, the drop in correlation upon removing cross-language neurons did not exceed what would be expected from removing a random subset of equal size (**Table S3**).

17. I'm a bit confused about the relationship between the translation neurons and the overall semantic tuning. On the one hand, the first part of the paper seems to highlight the existence of these neurons, but then in the rest of the paper, there is the view that they place a small or even inconsequential role in the overlapping semantic space is eschewed. It would be helpful to clarify their role (although granted, this is talked a bit about in the discussion).

We apologize for this confusion, which is our fault. We note that Reviewer 1 raised a similar point. In brief, we believe that the translation neurons (which we now call "cross-language neurons") have a very small, if any, role. Why then did we begin our story by focusing on them? We did so because we thought it would be a kind of narrative stepping stone to the core analyses of the paper, the geometric analyses. In retrospect, we can see that our approach was confusing. We have now opted to keep these neurons in the paper, but to de-emphasize them, and to discuss their limitations more clearly, including in the Abstract, Results (we shortened this section considerably), and Discussion, and by not mentioning them in the Introduction.

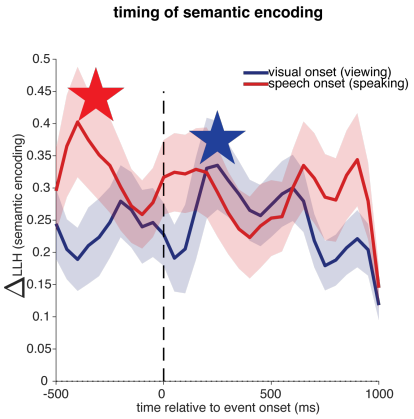
From the **revised Discussion**:

Cross-language neurons. We identify a new class of what we call "cross-language neurons," whose responses to co-translated words are correlated. Such neurons would have similar responses to, for example, "*maybe*" and "*quizás.*" These neurons may contribute to translation; however, they do not play an essential, or even a major role: even when we exclude them from our analysis, cross-language semantic geometry is robustly maintained (Table S3). Moreover, while their tuning functions are more aligned than other neurons, they are not perfectly aligned, indicating that they are not simply amodal conceptual representations. Consistent with this, their response fields span multiple semantic categories (e.g., Figure 2C), which differentiates them from the canonical amodal concept-related neurons, concept cells (Quian Quiroga et al., 2005; Quian Quiroga, 2012). Word-level decomposition confirms that their cross-language correlations are driven by a relatively sparse subset of words. These words are constrained along fewer latent semantic axes in mBERT space than other neurons which did not show cross-language correlation (Table S1-S2). We speculate that they may be the extreme end of a larger distribution of neurons whose shared population-level geometry contributes to translation.

18. There is no information presented about the timing of the tuning of these neurons for perception or production. This is important to better understand the role of these neurons in semantic processing for language. Similarly, it would be useful for the read aloud task to compare the alignment of tuning to the visual presentation of the word versus the utterance.

We followed the reviewer's suggestion and added the following analysis. The results showed the semantic encoding peaked ~200-250 ms after visual onset and ~300-400 ms before speech onset (see the figure below), indicating that hippocampal neurons encode semantic content during both retrieval and pre-articulatory preparation. These numbers, which we now

report, are not particularly surprising, and are broadly in line with numbers from other studies (including our own, Chavez et al., 2025).



We added the details and results into the revised text in the Methods

Temporal dynamics of semantic encoding (Study 2)

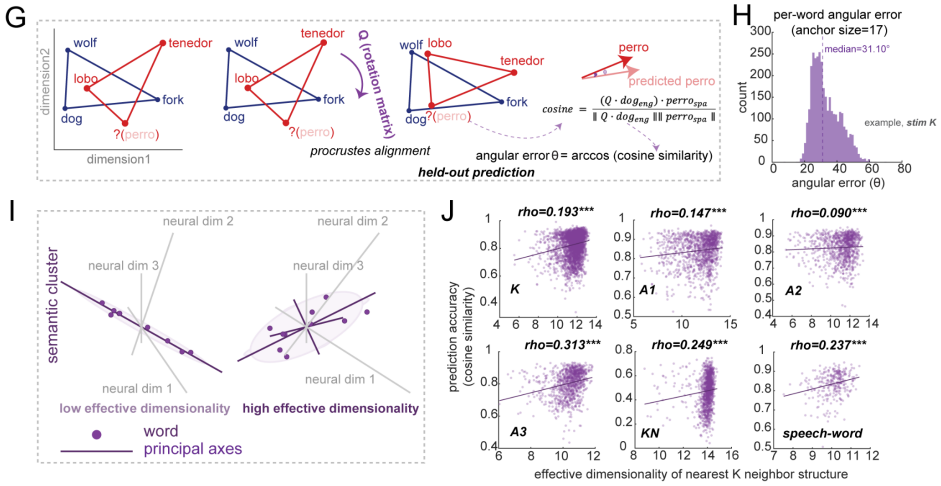
To characterize the timing of semantic encoding during reading aloud, we performed a sliding window analysis aligned to two events: visual onset (when the phrase appeared on screen) and speech onset (when the participant began speaking). We used 50-ms non-overlapping windows spanning -500 to 1000 ms relative to each phrase and each patient. Following previous research (Chavez et al., 2025), for each window, we extracted spike counts and fit Poisson generalized linear models (GLM) with ridge regularization to predict neural responses from semantic features. Predictors included the first 100 principal components of mBERT embeddings (z-scored), phrase duration (z-scored), and their interactions, yielding 201 features total. Model performance was quantified as ΔLLH, the difference between the actual model log-likelihood and the median log-likelihood from 100 null models in which phrase-neural correspondence was shuffled. The analysis revealed that semantic encoding peaked ~200-250 ms after visual onset and ~300-400 ms before speech onset, indicating that hippocampal neurons encode semantic content during both retrieval and pre-articulatory preparation. Although our timing analysis (based on Patient 2 and Patient 3) suggested peaks at 300-400 ms before speech onset, we adopted the more conservative 200 ms pre-shift established by Chavez et al. (2025) using a larger cohort (n = 10 patients) to ensure robustness across patients.

19. For the analysis in Figure 5, why was the closest 5 words chosen? What are the parametric properties of using k words instead of 5?

The reviewer raises a fascinating question (the similar question as Reviewer#1 did).

In short, the original choice of k=5 was arbitrary, and we have now systematically optimized this parameter.

We varied anchor size from k=3 to k=100 and found that prediction accuracy peaks at k=15-19 depending on the dataset, albeit with minimal variation across the full range (**Figure 5G-J**; see below). This parameter insensitivity supports the hypothesis that the cross-linguistic transformation approximates a near-global rotation rather than depending critically on a specific neighborhood size.



We revised the Results for Figure 5:

To test whether preserved semantic structure could support single-word neural prediction, we used local Procrustes alignment (**Figure 5G**, **Methods**). For each target word, we identified its k nearest neighbors in English neural space (excluding the target itself for held-out prediction aim), learned a rotation matrix from these neighbors' cross-language response pairs, and predicted the target's Spanish response. Prediction accuracy was quantified as angular error between predicted and actual vectors. We optimized anchor size (k = 15-19 across datasets; **Figure S3A**) and found median angular errors of 30-61° for listening and word-speech tasks (see example for stimulus set *K* in **Figure 5H,** all results see **Figure S3E**), significantly exceeding null (all $p < 0.001$).

We added the analysis description to Method:

Local geometric cross-language prediction (Figure 5G-J)

To test whether preserved semantic geometry could support single-word neural decoding across languages, we implemented a local Procrustes alignment approach. All neural firing rate vectors were L2-normalized to unit length, which ensures that prediction accuracy reflects directional alignment in neural state space rather than differences in firing rate magnitude.

Anchor size optimization. We first determined the optimal number of anchor words by systematically varying *K* from 3 to 100. For each target word, we identified its k nearest semantic neighbors in English neural space (excluding the target itself), learned a local Procrustes rotation from these anchor pairs, and evaluated prediction accuracy on the held-out target.

Held-out prediction. For each target word with English response vector $\mathbf{e}_{\text{target}}$, we identified its nearest *K* semantic neighbors (*K* was the optimal anchor size from above analysis step) in English neural space using cosine distance (explicitly excluding the target word itself), then extracted these neighbors' responses in both languages to form matrices $\mathbf{X}_E$ (English, $n_{\text{neurons}} \times k$) and $\mathbf{X}_S$ (Spanish, $n_{\text{neurons}} \times k$). After centering both matrices by subtracting column means ($\boldsymbol{\mu}_E$ and $\boldsymbol{\mu}_S$), we computed the cross-covariance matrix $\mathbf{M} = (\mathbf{X}_S - \boldsymbol{\mu}_S)(\mathbf{X}_E - \boldsymbol{\mu}_E)^T$ and performed singular value decomposition $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. The optimal orthogonal rotation matrix was obtained as $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$, and the predicted Spanish response was computed as $\hat{\mathbf{y}} = \boldsymbol{\mu}_S + \mathbf{Q}(\mathbf{e}_{\text{target}} - \boldsymbol{\mu}_E)$, then L2-normalized to unit length. Prediction accuracy was quantified as cosine similarity and angular error $\theta = \arccos(\hat{\mathbf{y}} \cdot \mathbf{y}_{\text{true}})$ in degrees (arccos of cosine similarity).

Statistical significance. We tested predictions for all words in each dataset (datasets with matched English-Spanish words, including all stimulus sets in Study 1 and words from short phrases in Study 2). Statistical significance was assessed against 1000 null permutations where we (1) randomly shuffled English-Spanish word pairings using derangement (ensuring no correct translations), (2) randomly selected Spanish prediction targets, and (3) used randomly selected words (excluding the target) rather than true nearest neighbors as anchors for learning rotation matrix Q, thereby destroying both semantic correspondence and neighborhood structure while preserving neural response distributions. P-values were computed as the proportion of null means achieving equal or greater accuracy than the observed mean.

20. For figure 6E, it appears that the effect is being largely driven by a small population of neurons (at least for K and A1). It is hard to see because of the size of the figure, but is this in fact the case? If so, it would be useful to characterize the properties of these neurons that enable them to facilitate this effect.

The reviewer raised a concern about whether the results were driven by extreme values, as the figures in our previous version were indeed small (we have now adjusted the figure size and did control analyses, see below).

To test whether the participation correlation was driven by a small subset of high-participation neurons, we performed two additional analyses. First, we replaced Pearson correlation with Spearman rank correlation, which is robust to outlier-driven effects. Spearman correlations remained high across all stimulus sets (all Spearman rho > 0.677, $p < 0.001$), confirming that the effect is not driven by a few extreme values.

Second, we progressively removed the top 5-30% highest-participation neurons and recomputed the correlation at each step. The Spearman rank correlation remained very significant even after removing the top 30% of neurons (all stimulus sets, K, A1, A2, A3, KN-

neuropixel and word speech, p<0.001; **Figure S5**, purple line). We then progressively removed the bottom 5-30% lowest-participation neurons and recomputed the correlation at each step. The Spearman rank correlation still remained very significant (all stimulus sets, K, A1, A2, A3, KN-neuropixel and word speech, p<0.001; **Figure S5**, green line). Together, these results demonstrate that the effect is not driven by a small subset of extreme neurons.

Revised in the paper:

(1) We have added these results to the **Figure S5** and mentioned it in the main text.

> "Each neuron's participation strength (the sum of squared loadings across the top eigenvectors, reflecting how heavily it contributes to these coordination modes), was highly correlated between languages (**Figure 6E**; K: rho=0.87; A1: rho=0.93; A2: rho=0.93; A3: rho=0.88; KN-Neuropixel: rho=0.68; speech-word, rho=0.90, all p<0.001; mBERT results, see **Figure S4B**). These effects are not driven by a small subset of extreme neurons (see control analysis in **Figure S5**)."

Figure S5

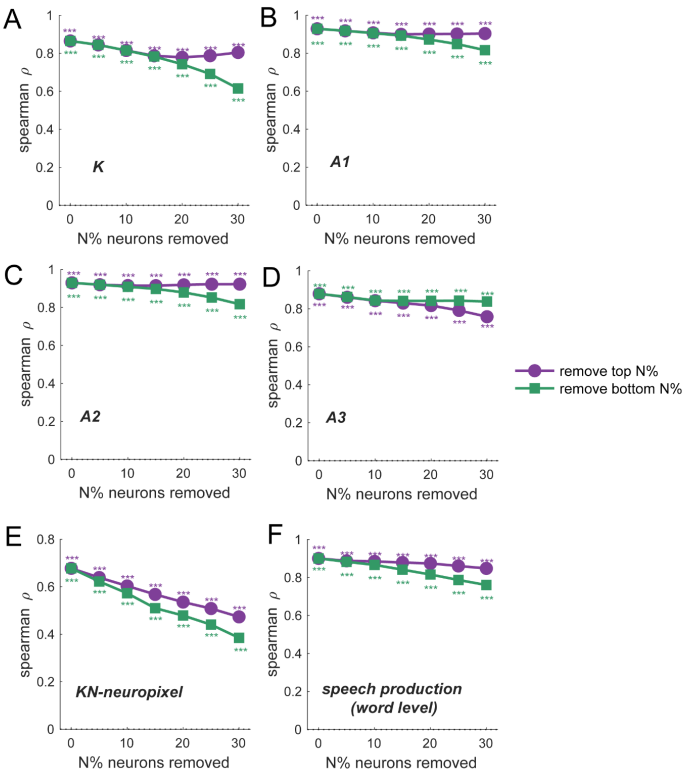


Figure S5. The robustness of per-neuron participation strength.

(A-E) We progressively removed the top 5-30% highest-participation neurons and recomputed the Spearman correlation at each step. The Spearman rank correlation remained very significant even after removing the top 30% of neurons (all stimulus sets, K, A1, A2, A3, KN-neuropixel in Study 1 and word speech in Study 2, $p<0.001$; purple line). We then progressively removed the bottom 5-30% lowest-participation neurons and recomputed the correlation at each step. The Spearman rank correlation remained very significant (all stimulus sets, K, A1, A2, A3, KN-neuropixel in Study 1, and word speech in Study 2, $p<0.001$; green line). Together, these results demonstrate that the effect is not driven by a small subset of extreme neurons.

$*p < .05$, $**p < .01$, $***p < .001$, permutation (1000 iterations)

Related to Figure 6.

(2) And we reported the Spearman rho correlation in **Figure 6E** (see below) in the revised manuscript.

E