Mirror manifolds: partially overlapping neural subspaces for speaking and listening

Ana G. Chavez[1], Melissa Franch[1], Elizabeth A. Mickiewicz[1], Will Baltazar[1], James L. Belanger[1], Davin Devara[1], Maya Etta[1], Thomas Hamre[1], Taha Ismail[1], Brad Joiner[1], Yoon Kim[1], Abhinav Kona[1], Kasra Mansourian[1], Amy Nangia[1], Molly Pluenneke[1], Sarah Soubra[1], Thara Venkateswaran[1], Karthik Venkudusamy[1], Assia Chericoni[1] , Katherine E. Kabotyanski[1], Kalman A. Katlowitz[1], Raissa K. Mathura[1], Danika L. Paulo[1], Xinyuan Yan[1], Hanlin Zhu[1], Eleonora Bartoli[1], Nicole R. Provenza[1,2], Andrew Watrous[1], Krešimir Josíc[3], Sameer A. Sheth[1,2], and Benjamin Y. Hayden[1,2]

[1] Neurosurgery, Baylor College of Medicine
[2] Neuroengineering Initiative, Rice University
[3] Departments of Mathematics, Biology and Biochemistry, University of Houston

Correspondence: Benjamin.Hayden@bcm.edu

Funding statement

This research was supported by the McNair Foundation and by NIH R01 MH129439, U01 NS121472, NINDS Research Education Grant Programs for Residents and Fellows in Neurology, Neurosurgery, Neuropathology, and Neuroradiology (UE5), the SNS Allan Friedman RUNN Research Grant, the NLM Training Program in Biomedical Informatics & Data Science for Predoctoral & Postdoctoral Fellows, T15LM007093-33

Competing interests

S.A.S has consulting agreements with Boston Scientific, Zimmer Biomet, Koh Young, and Neuropace. SAS is Co-founder of Motif Neurotech

Acknowledgements

We thank Joshua Adkinson, Justin Fine, Victoria Gates, Suzanne Kemmer, George Kokalas, and Steven Piantadosi for invaluable assistance.

ABSTRACT

Participants in conversations need to associate words with their speakers but also retain those words' general meanings. For example, someone talking about their hand is not referring to the other speaker's hand, but the word "*hand*" still carries speaker-general information (e.g., having five fingers). These two requirements impose a cross-speaker generalization / differentiation dilemma that is not well addressed by existing theories. We hypothesized that the brain resolves the dilemma by use of a vectorial semantic code that blends collinear and orthogonal coding subspaces. To test this hypothesis, we examined semantic encoding in populations of hippocampal single neurons recorded during conversations between epilepsy patients and healthy partners in the epilepsy monitoring unit (EMU). We found clear semantic encoding for both spoken and heard words, with strongest encoding just around the time of utterance for production, and just after it for reception. Crucially, hippocampal neurons' codes for word meaning were poised between orthogonalized and collinearized. Moreover, different semantic categories were orthogonalized to different degrees; body parts and names were most differentiated between speakers; function words and verbs were least differentiated. Finally, the hippocampus used the same coding principle to separate different partners in three-person conversations, with greater orthogonalization between self and other than between two others. Together, these results suggest a new solution to the problem of binding word meanings with speaker identity.

INTRODUCTION

When we converse, we take turns between listening and speaking. That back-and-forth raises an important mirroring problem for neural representations of meaning. When we say or hear a word, we must understand its general meaning, regardless of its speaker; at the same time, we must be able to associate the meaning to the speaker who said it. That in turn influences how we interpret the word. Someone saying "*my sister*" is typically referring to a different person than the conversation partner's sister. However, sisters have some features in common, and speakers need access to general cross-personal representations. Likewise, "*moon*" might mean something different if we are talking to an astronaut or a poet. The need for flexibility raises a generalization / differentiation dilemma for semantic representations. That dilemma is not readily accounted for in standard neural theories of representation, which emphasize the cross-modal aspect of meaning while ignoring the modality-specific aspect. Moreover, the degree of generalization / differentiation may vary by word; a "*leg*" is usually unique to one person; a "*cloud*" is not. The problem of cross-speaker generalization is addressed by mirror neuron theory, which proposes specialized neurons that have identical responses during production and reception of meaning (Arbib, 2005; Corballis, 2010; Rizzolati and Arbib, 1998; Rizzolatti and Craighero, 2004). Aside from its known empirical limitations (Hickok, 2010 and 2014; Heyes and Catmur, 2022), mirror neuron theory resolves only one horn of this generalization / differentiation dilemma.

Recent advances in the topology of multi-neuronal representation offer a potential solution to this dilemma (Barak et al., 2013; Bernardi et al., 2020). Specially constructed neural codes that blend features of collinear codes (which allow for generalization) and orthogonal codes (which allow for differentiation) can have the virtues of both. These semi-orthogonal codes, in other words, resolve the generalization / differentiation dilemma because they allow for simultaneous cross-conditional generalization and differentiation. Populations of neurons with these properties have been found in in complex decision-making tasks (Bernardi et al., 2020), in memory-based decision tasks (Libby and Buschman, 2021; Tian et al., 2024; Tang et al., 2020), and in value-coding regions during choice (Fine et al., 2023; Johnston et al., 2024; Yoo and Hayden, 2020). Here we propose the same principles apply to representations of semantic meaning across speakers during conversations. Because our proposal involves a solution to the mirroring problem using a high dimensional space along which neuronal activity is found (a

manifold, Perich et al., 2025), we call it the *mirror manifold hypothesis*. Critically, while the mirror neuron hypothesis only solves the generalization problem, the mirror manifold hypothesis, which leverages the expressive power of neural populations, can also solve the differentiation problem.

The mirror manifold hypothesis relies on the principle that semantic encoding is vectorized. We have previously shown that, during passive speech listening, hippocampal neurons encode word meanings in distributed patterns of activation, which we have argued is a vectorial coding (Franch et al., 2025; Katlowitz et al., 2025). Indeed, we found that responses of neural populations to semantically similar words are, themselves, similar, suggesting that populations of hippocampal neurons instantiate a semantic embedding space. These patterns are reminiscent of the manner that word meanings are encoded in Large Language Models (LLMs, e.g., Devlin et al., 2019; Mikolov et al., 2013). Previous theoretical work highlights the benefits of vectorial representations for cognitive processes like similarity judgment, analogy, and hierarchical reasoning (Piantadosi et al., 2024). Here we propose another benefit to vectorial representations of meaning, namely, that they endow populations with the ability to solve the generalization / differentiation dilemma imposed by conversational speech.

We were particularly interested in the hippocampus due to its demonstrated role in conceptual representations, as well as generalization and relational reasoning (Bakermans et al, 2025; Bakker et al., 2008; Bernardi et al. 2020; Courellis et al., 2024; Leutgeb et al., 2005). These are complemented by its established role in language, including both semantics and grammar (Binder et al., 2009; Brown-Schmidt et al., 2001; Dijksterhuis et al., 2024; Duff and Brown-Schmidt, 2012; Franch et al., 2025; Katlowitz et al., 2025; Piai et al., 2016; Van de Ven et al., 2020), and its clear role in identity assignment (Danjo et al., 2018; Montagrin et al., 2018; Rey et al., 2020; Tavares et al., 2015). We analyzed speech data collected during unconstrained conversations between patients in the epilepsy monitoring unit (EMU) and members of their family or our research team. We transcribed conversations and computed their semantic embeddings using BERT with speaker tokenization. We aligned each word to firing rates of neurons in the hippocampus and used neural responses to estimate semantic tuning. We found clear encoding of lexical semantics for both production and reception, with strongest encoding around the time of utterance for production, and immediately after it for reception. Critically, tuning functions differed systematically for speaking and listening. Specifically, they reflected

the specific blend of orthogonal and collinear codes that would allow for an efficient solution to the generalization / differentiation problem. Moreover, the degree of orthogonalization varied consistently with semantic category: categories with the greatest generalization included function words and personal pronouns; categories with greatest differentiation included body parts and proper nouns. Finally, the hippocampus used the same coding principle to separate different partners in three-person conversations, with greater rotation between self and other than between two others.

RESULTS

Ten native English-speaking patients undergoing neural monitoring for epilepsy participated in unstructured conversations in the epilepsy monitoring unit (EMU) at Baylor St. Luke's Hospital. All patients had multiple Behnke-Fried electrodes placed in the hippocampus for surgical monitoring (**Figure 1A**). Word counts across all conversations spanned from 720 to 8260 total words (mean: 2042 words). Conversations involved either two (n=3/10) or three-plus (7/10) total participants. Aside from the patient, the other participants were either laboratory members or family members. Conversations were recorded with a directed microphone (Blue Yeti) pointed at the patient and a lavalier microphone (WAYVOSE) on the lapel of the speech partner. All conversation audio streams were recorded directly into the Blackrock recording system, resulting in temporal alignment between speech and neural activity.

To identify speech content, we used a combination of automated transcription and diarization with manual correction (**Figure 1B**). Specifically, we began with speech audio files and passed them to two distinct software systems, a lexical decoder/encoder (such as WhisperX), which identified the words that were spoken and their precise onset and offset times, and, in parallel, a diarizer (Pyannote), which identified who was speaking. However, we found that the quality of transcription and diarization were not accurate enough for neural analysis. Thus, specially trained lab members used Praat, a linguistic annotation software for speech analysis (Boersma and Weenink, 2001) to carefully correct both transcription and diarization outputs. That process took roughly one hour per minute of conversation. Finally, all files were manually re-checked and further fixed by first author AGC.

Conversations in the EMU were typical of those observed in other contexts. For example, conversations consisted of short speech turns consisting of a small number of words (median: 3 words; mean: 9 words for spoken, median: 3 words; mean: 10 words for heard. **Figure 1C and D**). Speakers spoke at a rate of 207-261 words/minute. Word lengths ranged from less than 100 ms to almost a full second (**Figure 1E**). Each participants' speech corpus consisted of an average of 1120 words (**Figure 1F**) and included 325 unique words per patient. Partners' speech turns consisted of an average of 3178 words (**Figure 1F**), including 644 unique words shared across speakers. Across all

ten conversations, 3013 words were used by both the participant and at least one partner; 178 of these were unique. The most common shared words in participant/partners' conversations were "I", "like", "you", "the", and "yeah".

We collected responses of units in hippocampus (HPC, n=442 units, **Figure 1G and H)** from 10 patients (range:19-77 neurons per patient). Neurons showed a clear phasic response to the onset of individual speech turns (two clear examples shown in **Figure 1I**). However, the majority of word-evoked responses had more complex responses. To understand these responses, we took a word-based regression approach (Franch et al., 2025). Specifically, we used a ridge-regularized Poisson regression to predict word-evoked spike trains from static word embeddings (Word2Vec), including both word duration and semantic–duration interaction terms as predictors (**Methods)**. Because our primary interest was in the timing of word-evoked responses, we used a static embedder to avoid contextual bleeding across words and to isolate word-by-word encoding dynamics. Encoding performance was quantified as the difference in log-likelihood between the full model and a control model in which the semantic embedding predictors were shuffled. This analysis revealed distinct temporal dynamics for speaking and listening (**Figure 1J**). Unsurprisingly, responses for spoken words peaked around the time of utterance (red line) while responses for heard word peaked a few hundred ms later (blue line). Guided by these dynamics, we selected a fixed 500 ms window beginning 150 ms before word onset for spoken words and a 500 ms window beginning 200 ms after word onset for heard words. We chose the speaking-related window with the goals of capturing both preparatory and execution-related activity, and because it had a high decoding response. Subsequent analyses confirmed that all results presented are robust to the selection of analysis window. Indeed, other windows tested with a start of +/- 250 ms of the ones we chose produced qualitatively identical results (data not shown).

Once we selected our time windows, we tested the performance on each layer of a contextual embedding model, BERT (Devlin et al., 2019). We chose BERT because it generates context-dependent word embeddings and can incorporate speaker-aware tokens directly into the tokenization process (Gu et al., 2020). We found stronger encoding with higher layers in BERT (**Figure 1K**). Notably, encoding was significantly higher ($p<.001$, **Figure 1L**) for heard than spoken words.

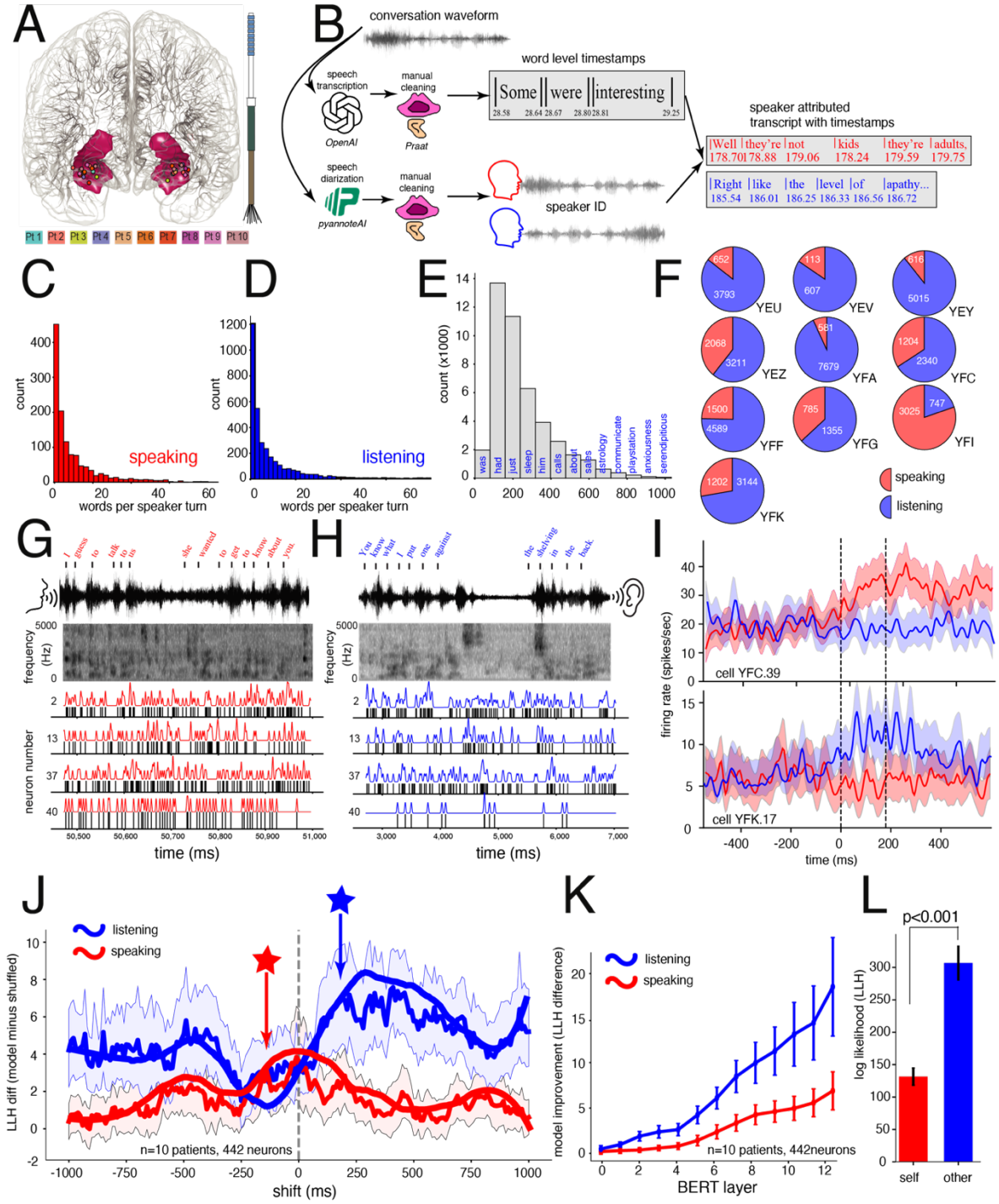

Figure 1. Encoding of semantics of conversational spoken and heard words in the hippocampus. A. Recording locations of hippocampal neurons across ten patients. **B.** Schematic of the speech processing pipeline, from raw audio to millisecond-resolved, speaker-labeled transcripts. In this and future panels, red: speaking; blue: listening. **C.** Conversational statistics: distribution of

words spoken per turn. **D.** Words heard by the patient per turn from other speakers. **E.** Average word lengths across patients. **F.** Proportion of words produced by each speaker. **G.** Example hippocampal neuron responses from patient YFF showing continuous spiking activity aligned to spoken words. **H.** Same but with heard words and corresponding spectrograms. **I.** Peristimulus time histograms (PSTHs) illustrating turn-aligned responses in two neurons from patients YFC (top) and YFK (bottom). **J.** Temporal dynamics of significant difference in log likelihood between the real and shuffled model (**Methods)** across word-aligned shifts: semantic encoding information began to peak ~150 ms before word onset for spoken words (red star) and ~200 ms after word onset for heard words (blue star) according to smoothed data points. **K.** Semantic encoding strength as a function of BERT embedding layer for spoken (red) and heard (blue) words. **L.** Overall encoding strength of the full model for spoken versus heard words.

On the relationship between neural encoding of heard and spoken language

Responses of three example neurons to three example words that were spoken and heard multiple times are shown in **Figures 2A, B, and C**. While responses to these words were different for speaking and listening, they reflect just a subset of the diversity of neural responses observed. The average word-evoked firing rate of all neurons for all patients together was nearly identical ($t=0.268$, $p=0.7885$). Nonetheless, population-level analyses showed that neural activity patterns reliably encoded speaker identity. Both logistic regression and random forest decoders classified speaker identity with an average accuracy of 82% across patients ($p < 0.05$, above the 50% chance level). Importantly, this effect was evident not only in the aggregate but also within individual patients ($p<0.001$ for each of our ten individual patients).

We considered three possible relationships between neuron population responses to the same word for speaking and listening. First, responses to the same word spoken and heard may have no systematic relationship (example neuron, **Figure 2D**). If that pattern were consistent across the population of neurons, then we would expect semantic tuning functions (that is, regression weight vectors) to be uncorrelated. Second, responses could be partially overlapping (example neuron, **Figure 2E**). If that pattern were consistent across the population of neurons, we would expect semantic tuning functions to be partly correlated. Finally, if representations are fully overlapping (cartoon data, **Figure 2F**), the correlation coefficient of tuning functions will be maximal (in the

absence of noise, it would approach 1, although due to noise, it would be lower). Cartoons illustrating the effects of these single neuron solutions on subspace representations are shown in **Figures 2G, H, and I**. These three patterns of encoding correspond to cross-speaker differentiation, differentiation/generalization, and generalization, respectively.

To determine which of these situations our data reflect, we computed the correlation between tuning functions of individual neurons. Here, tuning functions refer to the beta coefficients estimated for each neuron by our ridge regression models fit to spoken versus heard words (**Methods**). Our data are consistent with the intermediate, semi-orthogonalization solution (**Figures 2J and K**). Specifically, the degree of orthogonalization between speaking and listening subspaces (purple violins) is greater than zero in all ten of our patients individually ($p<0.05$ in all cases; one example patient shown in **Figure 2J**), as well as in the group as a whole ($p<0.0001$, **Figure 2K**). The fact that the measured correlation is greater than zero is, itself, notable since in high dimensional spaces two random vectors are overwhelmingly likely to be orthogonal due to cancellation across dimensions. Instead, this finding indicates that semantic representations are at least partially shared between speaking and listening, offering a solution to the generalization problem.

Equally important, we find that the degree of orthogonalization is less than fully collinear. An observed correlation less than 1.0 is mathematically guaranteed, but not scientifically interpretable, because any noise from any source will reduce measured correlations below 1.0. Instead, the critical question is whether the correlation is less than expected by chance given the noise properties of the dataset. We can estimate this using half-split within-condition noise estimates of collinearity. Using these half-split estimates is a powerful control, because it controls against any noise-related process that could cause a spurious positive finding. For example, it may be that fits to the data are not perfect, or that embeddings are suboptimal, or that we chose time windows that are perfectly tailored to our fitting. By comparing real data to half-split within condition data, we control for all of these confounders simultaneously.

We obtained those estimates with both listening (blue violins) and speaking (red violins) data (note that using half-splits reduces our dataset size, thereby reducing

correlations, making our analysis even more conservative). We find, both in the population and consistently across participants, lower half-split correlations with speaking than with listening. This fact is consistent with our result, shown above, that overall fits are better for listening than speaking data (**Figure 1L**). For this reason, we take the most conservative approach and focus on the comparison between speaking and speaking/listening cross-condition. We find, for our example patient as well as in the population, that cross-condition correlations are lower than the half-split speaking-only condition (above the null: $p=0.0002$, below the ceiling $p=0.001$). We find this same effect in 9/10 of our patients (above the null: $p<0.0001$, below the ceiling $p<0.0001$) with the remaining patient showing the same trend (patient YEY, $p=0.1124$). Together, these results indicate that semantic vectors are poised between orthogonality and collinearity, consistent with the mirror manifold (i.e., semi-orthogonalization) hypothesis for semantic representations in the hippocampus.

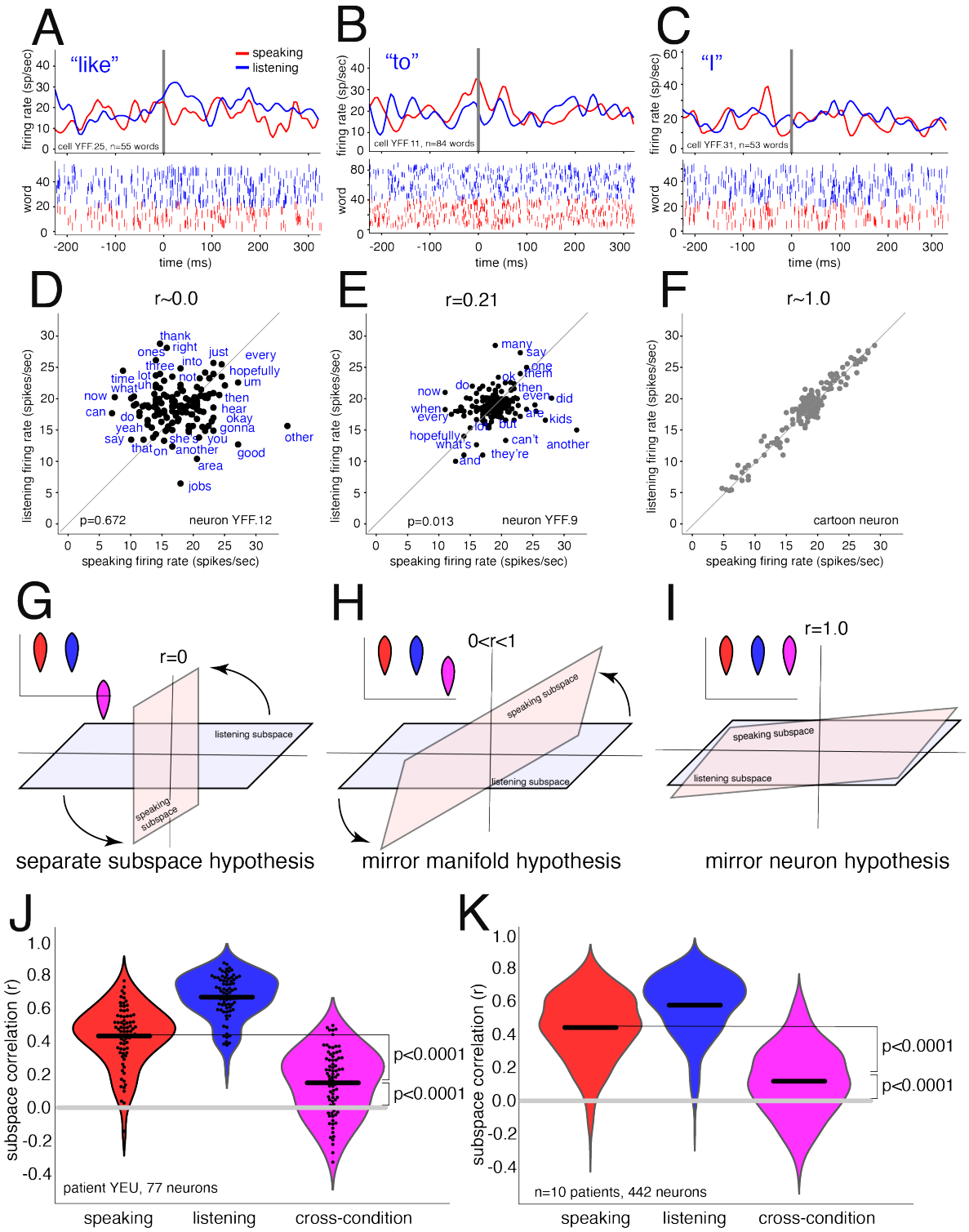

Figure 2. Semi-orthogonality between spoken and heard word representations in the hippocampus. A–C. Example PSTHs and rasters from three hippocampal neurons (patient YFF) during both speaking and listening to the words ***"like"***, ***"to"***, and ***"I"***. These illustrate diverse response patterns across conditions. **D-F** Examples of neuron-

level firing rate correlations between spoken and heard words with conceptual schematics: orthogonal/uncorrelated **D.**, semi-orthogonal/partially correlated **E.**, and collinear/perfectly correlated **F. G-I** Competing hypotheses for how hippocampal populations encode word meaning across speaking and listening: separate subspaces or populations with no overlap **G.**, a mirror manifold with mixed selectivity across neurons **H.**, and mirror neurons that are identical across conditions **I. J.** Example within-patient correlations between semantic encoding weights for spoken and heard words: values are significantly above chance but remain below the within-condition half-split noise ceilings. **K.** Group-level results across ten patients show consistent above-chance but sub-ceiling correlations, supporting a **mirror manifold** mechanism: spoken and heard words are encoded in overlapping but not identical neural subspaces.

Categorical specificity of semantic subspace semi-orthogonalization

If two people are talking about a tree, it is very likely to be the same tree. However, if one person mentions their eyes, they are almost certainly talking about something different from the other's eyes. Different semantic categories, in other words, may have different degrees of cross-speaker generalization and differentiation. We next asked how subspace orthogonalization interacts with semantic categories. We computed categories using a standard approach (Franch et al., 2025; Huth et al., 2016; Jamali et al., 2024; Katlowitz et al., 2025, see **Methods** for details). First, we computed a semantic embedding on all words spoken in all conversations using a non-contextual embedder (Word2Vec, Mikolov et al., 2013). Word embeddings (300-dimensional) were then reduced to two dimensions with UMAP (McInnes et al., 2018), and the resulting space was clustered using HDBSCAN to identify semantic categories in an unsupervised manner. The clustering algorithm yielded 11 semantic categories (**Figure 3A, Methods**). Not surprisingly, function words were the most common (58.6% overall), followed by personal pronouns (13.1% overall). The least used categories were proper nouns, social, and health/medicine (1.3%, 1.6%, and 1.9% of words, **Figure 3B**).

To assess how neural representations of meaning generalize across categories and between speaking and listening, we fit our Poisson ridge regression model (see above and **Methods**) separately within each semantic category, and separately for spoken and heard words. We then measured cosine distance between the resulting tuning vectors. Data from one patient are shown in **Figure 3C**. For this patient, the degree of orthogonality between tuning vectors for self and other varied systematically by semantic category (ANOVA: f-statistic = 3.34; p=0.0002). Function words (example: "*to*") had the most collinear representations between

speaking and listening, followed by action words ("*helping*") and then, surprisingly, personal pronouns ("*mine*"). Conversely, abstract concepts ("*future*"), social words ("*grandparents*"), and name/places ("*California")* were represented least similarly between speaking and listening conditions.

This pattern was common across the population. Indeed, of the 10 patients we tested, all but one showed a significant difference across categories (ANOVA, f-statistic =15.64; $p<0.0001$, **Figure 3D**). Data for the group are shown in **Figure 3E**. Overall, function words were the most similar across participants, followed by personal pronouns and actions. Conversely, body part and clothing terms ("*sweater*") were the most differentiated, followed by objects and names/places.

To confirm these findings using a complementary method, we asked whether this categorical structure could be decoded across conditions (speaking vs listening). Specifically, we trained a multinomial logistic regression classifier to distinguish 11 semantic categories from neural responses to spoken words and tested its performance on responses to heard words. To characterize how similar the underlying representational geometries were, we computed the shared subspace fraction between classifier weight matrices. Across patients, this fraction averaged to 0.46 (0.317-0.652). On this scale, where 0 indicates no overlap and 1 indicates complete overlap, a value of 0.46 reveals that about half of the discriminative structure for semantic categories is shared across conditions. In other words, the geometry is similar enough to support generalization of category membership, but distinct enough to maintain speaker-specific separation.

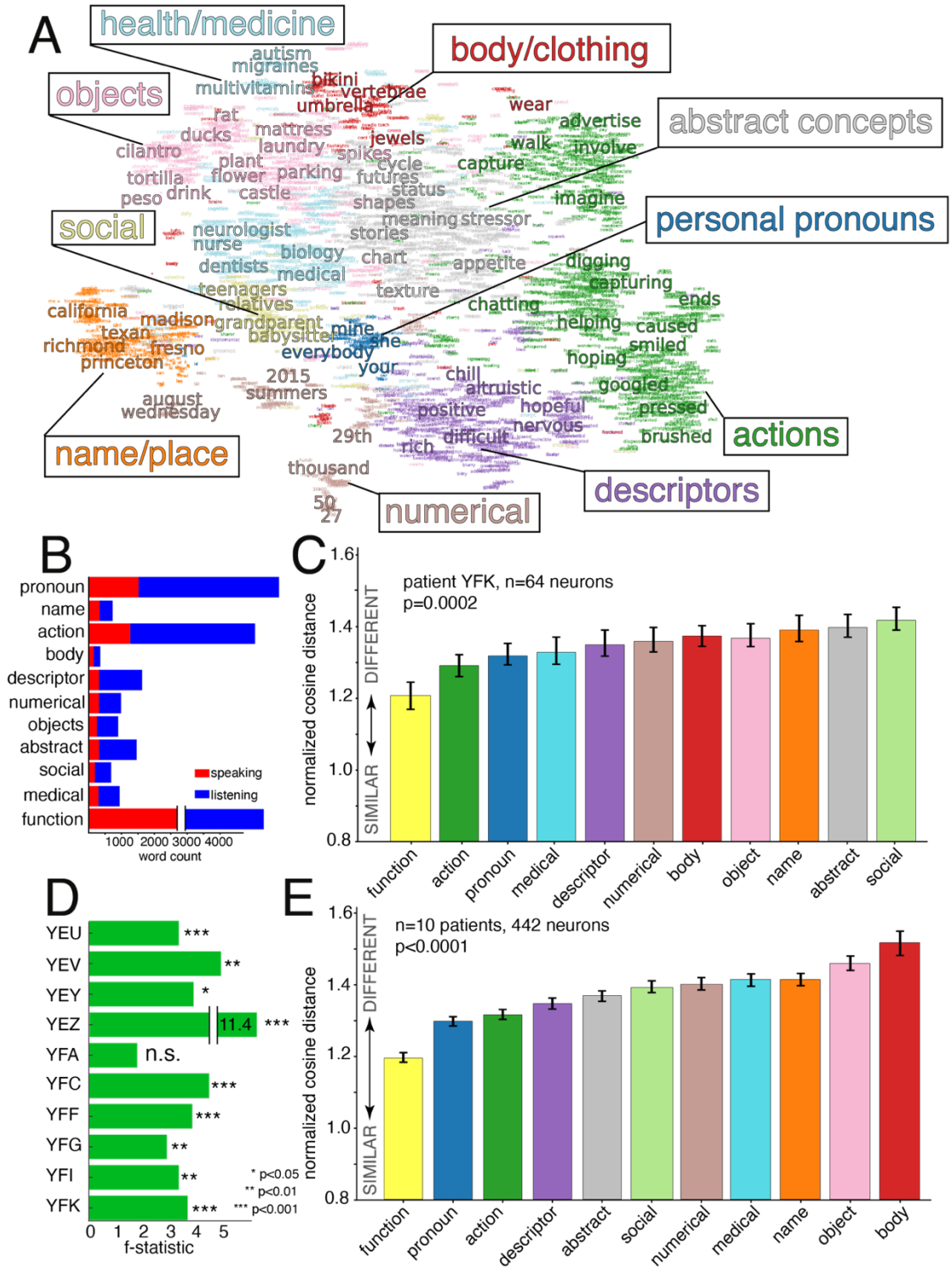

Figure 3. Differential encoding of semantic categories during speaking and listening. A. Clustering map of all words produced across the ten patient conversations, revealing broad semantic categories (e.g., body parts, social relationships, function words). **B.** Distribution of words across categories, separated by condition: words produced by the patient (red) versus words spoken by others and

therefore heard (blue). Function words were the most common in both conditions. **C.** Example from one patient showing cosine distances between encoding weight vectors for each semantic category, separately for speaking and listening, demonstrating systematic differences across categories. **D.** Patient-level significance summaries: one-way ANOVA tests revealed reliable differences between spoken and heard encoding weights across semantic categories. **E.** Group-averaged cosine distances across patients, highlighting that categories differed in their degree of self–other separation: function words (e.g., *"the," "and"*) were encoded most similarly across speaking and listening, whereas self-referential categories such as body-related words showed the greatest divergence.

How the hippocampus differentiates conversation partners

Just as we must distinguish our own words from those of our partners, if we have multiple conversational partners, we must keep them separate from each other. We next asked whether subspace rotation is used for this purpose. In seven of our ten conversations, there were at least three participants. **Figure 4A** shows the number of words spoken by each participant. By convention, we defined the participant as "*speaker 1*" (average of 1120 words), the partner with the most words as "*speaker 2*" (average of 1775 words), and the participant with the next most words as "*speaker 3*" (average of 802 words). We did not analyze data from additional speakers beyond those three; in all cases the number of words was too small to meaningfully analyze.

The key finding is that the brain uses subspace orthogonalization to separate semantics of partners 2 and 3 (**Figure 4B and C**). Specifically, the subspace correlation between semantics for partner 2 and 3 is greater than zero, indicating that there is some overlap in representations crucial for generalization (for example patient YEY, $p=0.0002$; for the population: $p<0.0001$). Moreover, the degree of overlap is less than 1.0, and, more importantly, is less than the amount one would observe with noise (for example patient YEY: $p=0.001$; for the population: $p<0.0001$). As above, we estimated the noise ceiling using a within condition half-split approach (**Methods**).

We next used the same semantic clustering (**Figure 3A**) to assess semantic category-specific orthogonalization across speech partners. As with self-other contrast, we found a consistent pattern. In both our example participant (**Figure 4D**) and in six of seven participants individually, (**Figure 4E**), we found a significant effect of semantic category on subspace angle (ANOVA f-stat: 2.93, $p = 0.0018$). These patterns were also observed when combining across all patients (**Figure 4F,** ANOVA: f-stat: 6.46, $p<0.0001$). As with the self-other distinction (see

above), function words, action words, and pronouns had the most collinearized representations across contexts. In contrast to the self-other distinction, names/titles, objects, and body/clothes were the most distinct for the self-other distinction but were middle positions for the partner 2 vs partner 3 distinction. And, whereas descriptors ("*difficult*"), social ("*grandparents*"), and numerical (*"thousand*") terms were in the middle for self-other, they were maximally differentiated for partner 2 and partner 3.

Together these results indicate that the brain uses subspace rotation as a general principle for separating identity. The presence of three subspaces (for self and the two partners) raises the question of how all three are related. All three subspaces may have some degree of orthogonality relative to each other (**Figure 4G**). More generally, we can imagine four possible configurations of three subspaces (**Figure 4H**). They can all be (i) identical, (ii) two matched and one separated, (iii) all three separated, but organized in a line, or (iv) all three separated, forming the points of a triangle. We find the fourth possible configuration, all three separated and forming the points of a triangle (and therefore, consistent with **Figure 4G**). That is, the separating hyperplane that distinguishes self from other is not unique but generalizes to also separate between different others. Thus, we observe not only reliable separation of self from each other speaker, but also measurable separation among the other speakers themselves. Separation is strongest between self and other speakers, although the degree varies depending on which pair is compared. In contrast, separation between different "other" speakers is present but weaker and demonstrated by the higher correlations (**Figure 4I)**. Specifically, correlations for self/other (1/2 and 1/3) are significantly lower than correlations for other/other (2/3, $p<0.01$, t-test across subjects). This same pattern is observed in 6/7 participants individually ($p<0.05$ in those cases). In other words, the hippocampus uses the same subspace rotation principle to separate different conversation partners, but with less rotation for the two partners than for the self vs. other.

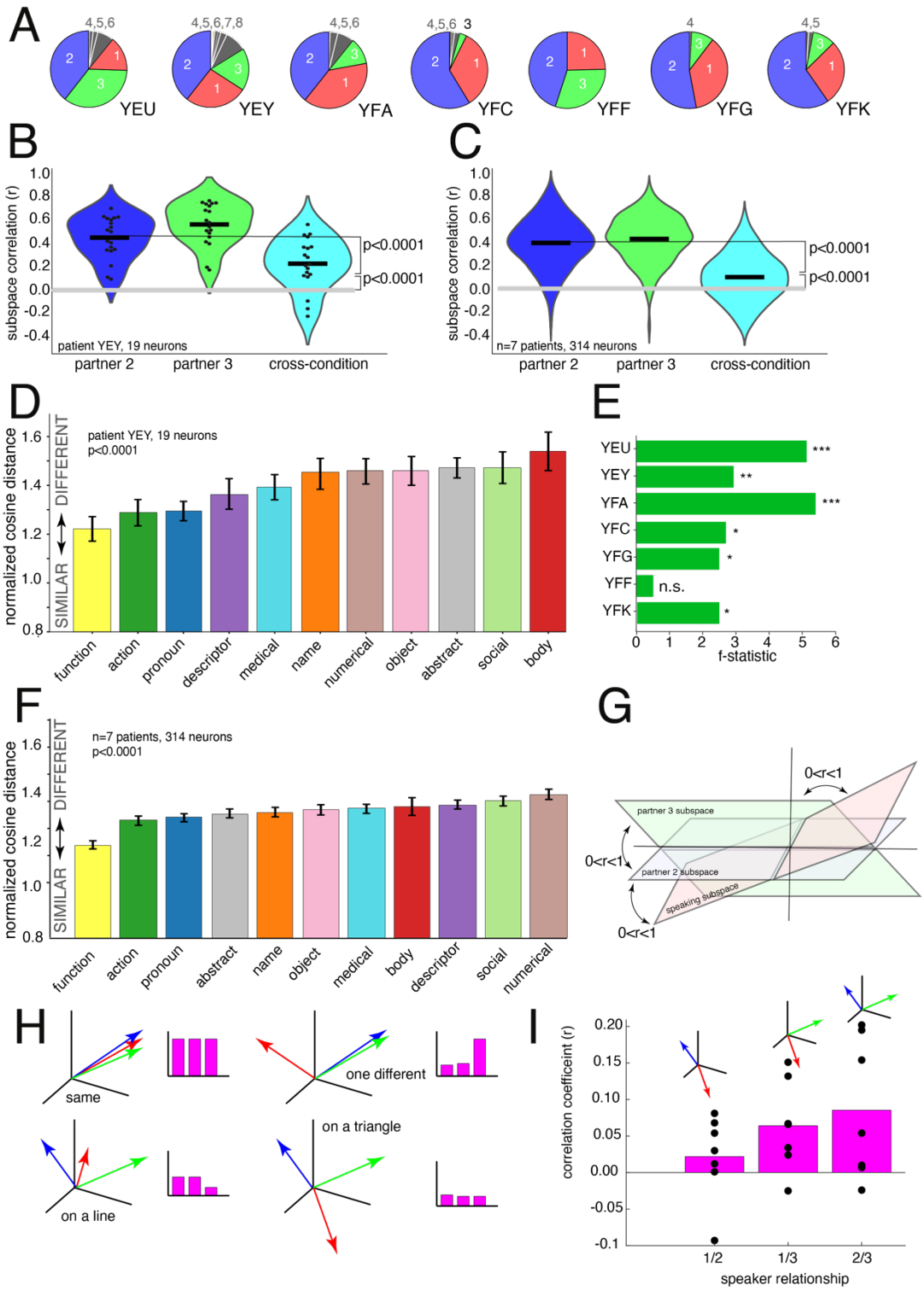

Figure 4. Semantic encoding principles extend beyond self vs. other into differentiation of multiple speakers. A. Pie chart of the number of words spoken by unique speakers in each patient's conversation. Seven patients are highlighted who had

more than one other speaker present. **B.** Example patient correlations between semantic encoding weights for words heard from two different speakers. Correlations were significantly above chance but remained below the within-condition half-split noise ceilings. **C.** Group-level results across these seven patients show that the mirror manifold mechanism generalizes to distinguishing between multiple external speakers. **D.** Example from one patient showing cosine distances between encoding weight vectors for each semantic category when listening to two unique speakers. **E.** Patient-level significance summaries: one-way ANOVA tests revealed reliable differences in encoding weights between words spoken by different external speakers. **F.** Group-averaged cosine distances across patients, highlighting that semantic categories varied in their degree of separation: function words were encoded most similarly across speakers, whereas numerical or time-related words were encoded most differently. **G.** Conceptual schematic extending the mirror manifold hypothesis to multiple speakers. **H.** Competing theoretical models for how semantic encoding might scale across speakers: (i) all speakers share the same manifold, (ii) only self versus other are distinct, (iii) external speakers are differentiated but anchored to a common manifold, or (iv) each speaker is encoded on a unique manifold. **I.** Correlation data from all seven patients support the final model: hippocampal neurons encode meaning in a speaker-dependent manner, such that semantic representations are yoked to speaker identity, and each speaker occupies a distinct manifold.

DISCUSSION

Conversation imposes a difficult generalization/differentiation dilemma for any neural representation of meaning. If one speaker mentions "*my bicycle*", the other speaker should know it's not their own bicycle but can nonetheless leverage their general knowledge of bicycles to make inferences about its features. Our data support the hypothesis that this dilemma can be resolved via rotation of neural population coding subspaces for semantics. Subspace rotation has been proposed as a solution to other generalization/differentiation problems (Bernardi et al., 2020; Chericoni et al., 2025; Elsayed et al., 2016; Johnston et al., 2024; Kaufman et al., 2014; Libby and Buschman, 2021; Tang et al., 2020; Tian et al., 2024; Xie et al., 2022;). Thus, here we propose that it extends to vectorial representations of word meanings in the hippocampus. We use the term "*mirror manifold*" to describe our hypothesis. Like mirror neurons, our proposed mirror manifold can solve the cross-personal generalization problem, but, unlike mirror neurons, can also simultaneously solve the differentiation problem.

Our results may be surprising as we find clear semantic encoding localized to the hippocampus, which is not a classical language structure (Fedorenko et al., 2024). However, our results are aligned with a good deal of research highlighting the role of the hippocampus in representing word meanings (Binder et al., 2009; Brown-Schmidt et al., 2021; Duff and Brown-Schmidt 2012; Franch et al., 2025; Piai et al., 2016; van de Ven et al., 2020) and grammatical processing (Dijksterhuis et al., 2024). Complementary studies of lesions to the hippocampus show clear semantic deficits (Manns et al., 2003; Moscovitch et al., 2005; Keane et al., 2019; Duff et al., 2020). This possibility is consistent with the robust evidence that the hippocampus plays an important role in memory-related processes, which may be foundational for semantics. Moreover, the hippocampus is associated with simultaneous generalization and differentiation (Bernardi et al., 2020), as well as both pattern completion and separation processes, which are conceptually related (Leutgeb et al., 2005; Bakker et al., 2008). Finally, there is evidence that supports a potential role for the hippocampus in social cognition, including identity assignment (Danjo et al., 2018; Tavares et al., 2015; Montagrin et al., 2018; Rey et al., 2020).

Taken together, these properties suggest that the hippocampus is well positioned to address a central question in language research: how production and comprehension, long treated as distinct processes, may nonetheless be linked. Early models drew a sharp boundary between speaking and listening, mapping production to Broca's area and comprehension to Wernicke's

area. More recent functional–anatomical accounts move beyond these two regions as sole language centers but still emphasize that production and comprehension are carried out in partially separable hubs or dual processing streams (e.g., Hickok & Poeppel, 2007; Ben-Shalom & Poeppel, 2008; Castellucci et al., 2022). While such dichotomous frameworks offer a solution to the problem of differentiation, they leave unresolved the complementary need for cross-speaker generalization. Complementary perspectives highlight overlapping neural mechanisms, suggesting that speaking and listening may draw on shared representational systems (Pickering & Garrod, 2013; Tremblay & Dick, 2016; Silbert et al., 2014). Our results support the idea that production and comprehension rely in part on overlapping representations.

Our study complements other recent studies on the neuroscience of conversational language. Zada et al. (2024) showed that contextual embeddings from LLMs approximate a context-sensitive representational space that both speakers map onto during face-to-face conversation. Encoding models revealed that these embeddings predict neural activity before word onset in speakers and after word onset in listeners, creating a temporal flow that mirrors conversational dynamics. Goldstein et al. (2025) extended this approach in a larger dataset of spontaneous conversations using a unified acoustic-to-speech-to-language model (Whisper), showing that the same cortical hierarchy supports both production and comprehension with similar temporal dynamics. In contrast, Cai et al. (2025) used embedding-based encoding models of intracranial LFPs and found that natural conversation recruits a distributed set of channels across cortical and limbic regions. A greater proportion of temporal cortex channels were significantly explained during comprehension, more precentral channels during production, and only ~18% of channels were jointly predictive across both conditions. Finally, using a delayed sentence repetition task, Xu et al., (2024) find evidence for *bridge* and *mirror* neurons in the precentral gyrus. Bridge neurons exhibited sustained firing rate between the same speech elements during listening and speaking. Their mirror neurons implement mirroring functions for specific speech sounds, although they do not participate in speaker differentiation. Our results confirm the idea of cross-speaker mirroring and extend them by (1) showing that cross-speaker mirroring can be identified in natural, free-flowing speech; (2) mirroring is not limited to phonemics, but also works for semantics, and (3) semantic mirroring, at least, can be found outside the classic language network. And, of course, (4) we find simultaneous mirroring and differentiation in the same populations of neurons.

Our subspace alignment proposal relies on the principle that semantic encoding in the hippocampus is vectorial. We and others have proposed that meanings of words in naturalistic contexts are distributed across a large number of neurons and expressed in patterns of neural activity (Franch et al., 2025; Jamali et al., 2024; Katlowitz et al., 2025). Our results here suggest that this vectorial representation includes factors that are outside the domain of semantics; here, the identity of the speaker. In this sense, it resembles semantic representations in BERT, which can incorporate identity with special identity tokens (Gu et al., 2020). In short, rather than separate channels for identity and meaning, the same neurons are used, and identity and meaning are multiplexed in those neurons (Rigotti et al., 2013). This view is in line with modern neuroscience theories, which de-emphasize the expression of function in specialized categories of individual neurons, in favor of distributed, population coding methods (Rigotti et al., 2013; Saxena and Cunningham, 2019; Churchland et al., 2012; Ebitz and Hayden, 2021).

The semantic clustering we observe indicates that subspace rotation is not a one-size-fits-all process (**Figure 3**). While our results do not provide insight into why certain categories have more or less alignment, we can speculate. For example, function words (*the*, for, *and*) and pronouns are the most consistent across both self-other (**Figure 3E**) and partner2-partner3 (**Figure 4F**). These words have both very general meanings that do not vary much by person and, moreover, tend to have very weak semantics on their own (Katlowitz et al., 2025). Verbs are the next most consistent (*helping*, *brushed*, *smiled*). Conversely, body parts (*face*, *eyes*, *arm*) tend to be highly specific to individuals; the speakers' hand is almost certain to be different - and have different conversational relevance - than the partner's hand. That may also explain why objects (*drink*, *laundry*, *mattress*) are highly differentiated - speakers may tend to discuss objects relative to themselves.

More speculatively, our results suggest a circuit-level approach to questions of mind-reading and mental state inference (Mukamel et al., 2010; Koster-Hale and Saxe, 2013; Jamali et al., 2021). As with cross-speaker language, these processes require simultaneously recognizing the similarities between self and other while maintaining their separation (Uddin, 2011; Jamali et al., 2021; Heatherton et al., 2016). For example, empathy requires acknowledging another person's experience while also recognizing that it is not one's own (Decety and Jackson, 2004; Singer et al., 2004; Lamm et al., 2007). This raises what has been termed a superposition problem (Garagnani, 2024). Our findings suggest a complementary solution: representing

person-specific variables in partially overlapping but distinct subspaces. In this framework, neural populations can project information into a shared representational space to capture commonalities, while also aligning along separating axes that preserve individual distinctions. That we observe such a mechanism in the context of language conversations suggests it may provide a foundation for broader cognitive functions, including emotion, theory of mind, or decision making in social contexts.

METHODS

Human intracranial neurophysiology

All experiments were performed in the Epilepsy Monitoring Unit (EMU) at Baylor St. Luke's Hospital using standardized approaches (Xiao et al., 2024). Experimental data were recorded from ten adult patients (seven male, three female, age ranges: 22-54) undergoing intracranial monitoring for epilepsy. The hippocampus was not a seizure focus area of any patients included in the study. Single neuron data were recorded from stereotactic (sEEG) probes, specifically Ad-Tech Medical probes in a Behnke-Fried configuration. Each patient had an average of three probes terminating in the left and right hippocampus. Electrode locations were verified by co-registered pre-operative MRI and post-operative CT scans. Each probe includes eight microwires, each with eight contacts, specifically designed for recording single-neuron activity. Single neuron data were recorded using a 512-channel *Blackrock Microsystems Neuroport* system sampled at 30 kHz. To identify single neuron action potentials, the raw traces were spike sorted using the *Wave_clus* sorting algorithm (Chaure et al., 2018) and then manually evaluated and modified to improve sorting. Noise was removed and each signal was classified as multi or single unit using several criteria: consistent spike waveforms, waveform shape (slope, amplitude, trough-to-peak), and exponentially decaying ISI histogram with refractory period (1 ms) violations below 5%. The analyses here used all single and multiunit activity.

Electrode visualization

Electrodes were localized using the software pipeline intracranial Electrode Visualization (iELVis; Groppe et al., 2017) and plotted across patients on an average brain using Reproducible Analysis & Visualization of iEEG (RAVE; Magnotti et al., 2020). For each patient, DICOM images of the preoperative T1 anatomical MRI and the postoperative Stealth CT scans were acquired and converted to NIfTI format (Li et al., 2016). The CT was aligned to MRI space using FSL (Jenkinson et al., 2002; Jenkinson & Smith, 2001). The resulting coregistered CT was loaded into BioImage Suite (version 3.5β1; Joshi et al., 2011) and the electrode contacts were manually localized. Electrodes coordinates were converted to patient native space using iELVis MATLAB functions (Yang et al., 2012) and plotted on the Freesurfer (version 7.4.1) reconstructed brain surface (Dale et al., 1999). Microelectrode coordinates are taken from the first (deepest) macro contact on the Ad-Tech Behnke Fried depth electrodes. RAVE (Magnotti et al., 2020) was used to transform each patient's brain and electrode coordinates into MNI152 average space. The average coordinates were plotted together on a glass brain with the hippocampus segmentation and colored by patient.

Conversations

Patients with epilepsy and healthy language function engaged in unprompted conversations with researchers, family members, friends, and/or members of the clinical staff. Conversation lengths ranged from 14.7-70.1 minutes, with an average length of 20.2 minutes. Topics of conversation varied across patients. Audio was primarily recorded using two room microphones that were directed toward the location of the patient (Blue Yeti) or personal lapel microphones (WAYVOSE). The audio was synchronized to the neural recording system via analog input going directly into the Blackrock Neural Signal Processor at 30 kHz.

Audio Transcription

After experiments, the audio.wav file was automatically transcribed using WhisperX, a state-of-the-art and commercially available AI model trained to transcribe speech (Bain et al., 2023). Speaker diarization was achieved using an open-source toolkit written in Python named Pyannote (Bredin et al., 2021). The transcribed words, corresponding timestamps, and speaker labels output from these models was converted to a TextGrid and then loaded into Praat, a software for speech analysis. The original .wav file was also loaded into Praat. Trained lab members manually inspected the spectrograms, timestamps, and speaker labels, correcting each word to ensure that onset and offset times were precise and assigned to the correct speaker (this process took about 1 hour per minute of speech). The TextGrid output of corrected words and timestamps from Praat was converted to a .xls and loaded in Python for further analysis.

Contextual and non-contextual embeddings

Contextual embeddings: To obtain contextual embeddings for each word, we used the *bert-base-cased* model (BERT; Devlin et al., 2019) accessed via the Hugging Face Transformers library (Wolf et al., 2020). This model consists of one embedding layer and 12 transformer blocks, each producing a 768-dimensional vector for every token. The conversational transcripts were vectorized into a single sequence of words in their natural temporal order and each word was prepended with a speaker-specific token, allowing the model to differentiate between speakers. This modification was inspired by prior speaker-aware adaptations of BERT for dialogue modeling (Gu et al., 2021). To handle BERT's 512-token input limit, we used a sliding window of 512 tokens with a stride of 256, and averaged embeddings for words that fell in overlapping window segments. This model also allows for extraction of embeddings from each layer (0-12). Embeddings from each of these layers were used as predictors for neural activity (**Figure 1K**), however all other results are from output or final layer.

Non-contextual (static) embeddings: We also extracted static word embeddings using the pre-trained fastText Word2Vec model in MATLAB, specifically the wiki-news-300d-1M vectors (Bojanowski et al., 2017; Grave et al., 2018). This model provides fixed 300-dimensional vectors for over one million words and phrases, trained on Wikipedia and news text to capture broad semantic relationships. Unlike contextual embeddings, which vary depending on sentence position or surrounding words, these static embeddings assign the same representation to each word regardless of context. For our dataset, we mapped every word in the transcripts to its corresponding fastText vector. Words without valid embeddings (e.g., uncommon surnames or rare tokens not present in the model's vocabulary) were excluded from further analyses.

Semantic clustering

To identify natural semantic categories in our word data (**Figure 3A–E**), we clustered all unique words into groups based on distributional similarity in embedding space. We first extracted 300-dimensional static embeddings for each unique word across all conversations using the pre-trained fastText wiki-news-300d-1M model. To reduce noise and focus on semantically informative tokens, we excluded function words, except personal pronouns and their contractions, using the stopword list from the Natural Language Toolkit (NLTK; Bird et al., 2009), and treated them as a separate cluster. Dimensionality reduction was then performed with Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) using the parameters *n_neighbors = 15*, *min_dist = 0.2*, *n_components = 10*, and *metric = cosine*. UMAP produces a lower-dimensional representation in which words with similar meanings are located

nearby. We next applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al., 2013; McInnes & Healy, 2017) to these reduced embeddings. HDBSCAN requires two key parameters: *minimum cluster size* (the smallest grouping to be considered a valid cluster) and *minimum samples* (the number of points required to define a dense region), which together determine the stability and granularity of the resulting clusters. In our analysis, we set *minimum cluster size = 40* and *minimum samples = 8*. Words assigned to the noise cluster were reassigned to the nearest cluster by majority vote of their neighbors. For visualization, we projected the raw embeddings into two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten & Hinton, 2008) and colored words by their assigned cluster IDs.

Turn-aligned PSTHs and rasters

For each neuron, we extracted spike trains aligned to conversational turn onsets defined as the moment when either the patient (self) began speaking or when any other participant in the room initiated speech (listening for self). Each turn onset was treated as time zero, and spikes were segmented into fixed analysis windows spanning a defined pre- and post-onset interval. Within each window, spikes were binned at a fixed bin width (20 ms). Spike counts in each bin were divided by the bin duration (in seconds) to yield firing rates in Hz and averaged across trials to obtain the mean peri-event time course of firing rate (PSTH). Variability across trials was quantified as the standard error of the mean (SEM). For raster visualization, individual spikes from each trial were plotted as vertical ticks at their millisecond timestamps relative to turn onset, forming rasters at the baseline of the axis. To highlight temporal structure, the binary spike trains were also smoothed with a Gaussian kernel ($\sigma \approx 10$ ms) and scaled to approximate firing rate. These smoothed traces were plotted directly above the raster ticks, providing a continuous PSTH trace superimposed on the underlying spike times.

Regressing spike counts on word embeddings

Word embeddings, such as those derived from Word2Vec or similar models, are typically high-dimensional and can exhibit substantial collinearity among dimensions - that is, many of the embedding features are correlated with one another. We first used Principal Component Analysis (PCA, scikitlearn) on the full word embeddings to obtain uncorrelated features that still capture the dominant structure in the embedding space with reduced dimensionality. For each word, we used the first 100 principal components (PC) from Word2Vec or BERT embeddings because they explained at least 60% of the variance of the original embedding vectors. We modeled the spike count responses of individual neurons using a Poisson Generalized Linear Model (GLM) with a log-link function and ridge (L2) regularization (we confirmed that spike counts for each word follow a Poisson distribution). The model aimed to predict the number of spikes a neuron fired in response to each word (summed spikes across word duration), using the top 100 principal components (PCs) of the word embedding vector, the duration of the word, and the interaction between each PC and the word's duration as predictors (z-scored prior to model fitting). This resulted in a 201-dimensional feature vector for each word: 100 PC values, 1 duration value, and 100 PC×duration interaction terms.

The Poisson GLM assumes that the spike count y_i for word i is drawn from a Poisson distribution with mean λ_i where:

$$y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + x_i^T \beta$$

Where $\lambda_{i,n}$,the expected spike count for word (*i)* and neuron n (n), is modeled by the linear regression coefficients (β), one for each of the 201 predictors, 100 PCs from embeddings, word duration, and duration PC interactions, (x_i^T), plus the y-intercept (β_0). For each neuron, we used a 5-iteration nested training loop, where in each iteration, the data was split into training and held-out test sets (80/20 split). Within the training set, 5-fold cross-validation was used to select the optimal regularization strength (alpha) from a log-spaced range based on optimizing the cross-validated log-likelihood. The selected model was then evaluated on the held-out test set using several performance metrics, including log-likelihood, and adjusted R^2. To determine whether the observed performance exceeded chance, we fit matched null models in which embeddings were randomly shuffled across words 100 times, computing LLH from every shuffled model. To determine if the actual model showed significant improvement from the shuffled model, we performed a permutation test comparing the actual model LLH to the shuffled model LLH distribution. We then took the difference between the actual and shuffled model (median) LLH (LLH diff; ΔLLH). A neuron's response was successfully predicted by the model if its ΔLLH was both positive and significant and if McFadden's pseudo-R^2 value was positive ($P < 0.05$). Shuffled model R^2 values were significantly less than those from actual models ($p < 0.05$, Wilcoxon signed-rank test). A predictor was considered significant if its corresponding coefficient had a p-value less than 0.05 in the fitted Poisson model. For population-level encoding analyses, however, we included all neurons and all model coefficients, irrespective of their individual significance.

Hyperparameter Selection:

To choose the regularization strength α, we used a two-stage grid search:

1. **Coarse Search** over: $\alpha \in \{10^{-5}, 10^{-4}, \ldots 10^{3}\}$
2. **Fine Search** around the best coarse value, spanning two orders of magnitude.

For each value of α, we performed 5-fold cross-validation. For each fold, the model was trained on the other four folds and evaluated on the held-out fold by computing the Poisson log-likelihood:

Window selection for speech comprehension and production

To identify the analysis windows for subsequent encoding models, we systematically quantified model performance across a range of candidate window starts and lengths. Candidate windows were defined from –1 s to +1 s relative to word onset, sampled in 10-ms increments, with window lengths ranging from 80 ms to 500 ms. For each [window start × window length] combination, and separately for spoken and heard words, we fit Poisson regression models using Word2Vec embeddings as predictors. Performance was evaluated using ΔLLH, defined as the log-likelihood difference between the full model and a semantic-shuffled model in which only the semantic embedding PCs were permuted (with duration interactions recomputed). Model performance increased monotonically with longer window lengths, with the strongest improvements at 500 ms. Based on this, and to capture preparatory activity prior to speech and integrative activity during comprehension, we selected window starts that were shifted back from the global maximum. For speaking, the window began 150 ms before word onset; for listening, the window began 200 ms after word onset. For visualization, ΔLLH values were smoothed across window positions using a Savitzky–Golay filter (Savitzky & Golay, 1964). For each neuron and word, firing rate was computed as the total spike count within the selected window, divided by the window duration, and expressed in spikes/s. Unless otherwise noted, all analyses

reported in the manuscript used these fixed windows (**Figures 1I and 2A–C** show PSTHs computed using identical windows for speaking and listening).

Reliability and cross-condition comparisons of semantic encoding

To evaluate how neural populations encode semantic information (using BERT embeddings) across speaking and listening, we compared the structure of regression weights obtained under each condition. We fit separate Poisson ridge regression models for spoken and heard words (see above), yielding condition-specific vectors of beta coefficients that define each neuron's semantic tuning curve. However, to dismiss possible explanations of these results coming from an imbalance in the number of words within each condition we made sure to downsample the larger class if the imbalance exceeded 1.5 times the smaller class as implemented in the imbalanced-learn Python package (Lemaitre et al., 2017). These results did not vary significantly from refitting the same models on perfectly balanced data. In our framework, the regression weights learned by the model can be interpreted as the tuning curves. When considered jointly across neurons, these tuning vectors span a neural subspace that captures the population-level encoding of semantics under a given condition (speaking or listening). The alignment of these two subspaces was quantified as the correlation between their coefficient vectors (*r*_cross) (Johnston et al., 2024). To assess whether these correlations exceeded chance levels, we constructed a null distribution by permuting the predictor variables: one condition was modeled on the true data, while the other was modeled on data with semantic predictors randomly shuffled. This procedure was repeated 500 times to estimate the distribution of correlations expected under the null hypothesis of no shared structure. As an upper bound, or noise ceiling, we quantified within-condition reliability by randomly splitting the data in half and re-fitting the model separately to each split. Correlations between the resulting beta vectors provided half-split reliability estimates for spoken (*r*_self) and heard (*r*_other) words. Together, these three metrics allowed us to determine whether cross-condition similarity reflected true shared semantic encoding rather than noise, and to interpret it relative to the maximum reliability attainable in each condition.

Semantic cosine distance

We quantified the separation between semantic encoding during speaking and listening using the cosine distance between regression weight vectors. For two beta coefficient vectors w_1 and w_2 cosine distance is defined as:

$$\text{Cosine distance } (w_1, w_2) = 1 - \frac{w_1 * w_2}{\|w_1\| \|w_2\|}$$

To compute category-specific distances, the regression model described above was fit separately for each semantic category under speaking and listening conditions. This procedure was repeated across all categories, with class counts balanced up to an imbalance ratio of 1.5 to avoid excessive trial loss. To establish a noise ceiling for these distances, we computed half-split reliability estimates within the most populated category (function words). Data from this category were randomly split in half, models were refit, and cosine distances were computed between the two halves. This was repeated with bootstrapping, and all observed category distances were normalized by the mean half-split distance for function words. Finally, differences in normalized cosine distances across semantic categories were evaluated using a one-way ANOVA.

Decoding Analyses

Decoding Speaker Identity

To assess whether hippocampal activity carried sufficient information to distinguish speaker identity, we trained classifiers on neural firing rate data. Two decoding approaches were used: multinomial logistic regression with an elastic net penalty and random forest classifiers. Hyperparameters for both models were optimized using randomized search with cross-validation. Input features were word-by-neuron firing rate matrices constructed from the speaking and listening windows defined above. To probe the robustness of the population code, we also evaluated decoding performance while subsampling random fractions of the available neurons within each patient. Classification tasks included self versus other (patient's own speech vs. all other speakers' speech) and pairwise other–other comparisons (distinguishing between two external speakers). To mitigate class imbalance, we down sampled the majority class so that each class had an equal number of words. Decoding performance was evaluated using accuracy. To confirm that results reflected above chance decodability, we additionally computed null distributions via permutation tests in which class labels were randomly shuffled. Across patients, logistic regression and random forest yielded comparable accuracy, differing by less than ~2% within individuals.

Subspace Correlation of Semantic Decoder Weights

To assess how spoken and heard semantics relate at the population level, we trained a linear decoder to classify words into ten semantic categories (plus function words). We used a multinomial logistic regression model, with hyperparameters tuned by grid search, applied to firing rate matrices (words × neurons) constructed separately for speaking and listening conditions. The target label in each case was the semantic category of the word. To compare the representational geometry between conditions, we quantified the similarity of the semantic subspaces defined by the decoder weights (Johnston et al., 2024). For each condition, the regression weight matrix W was normalized on a per-neuron basis so that each neuron's weight vector had unit length:

$$\hat{\mathrm{w}}_i = \frac{w_i}{\|w_i\|_2}, i = 1, \dots, N$$

where w_i is the row vector of weights for neuron i. These normalized weight vectors were then averaged across neurons to yield a population-level summary vector:

$$\overline{w} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathrm{w}}_i$$

Finally, the subspace correlation between speaking and listening was computed as the cosine similarity between the two summary vectors:

$$\rho = \frac{\overline{w}_{self} * \overline{w}_{other}}{\|\overline{w}_{self}\|_2 \|\overline{w}_{other}\|_2}$$

This measure captures the degree of alignment between semantic subspaces: values near 1 indicate strongly overlapping subspaces, values near 0 indicate orthogonality, and negative values reflect opposing alignment. Across the ten patients, subspace correlations ranged between 0 and 1, with a mean of ~0.46, consistent with a semi-orthogonal relationship between spoken and heard semantic subspaces.

References

Aliko, S., Franch, M., Kewenig, V., Wang, B., Cooper, G., Glotfelty, A., ... & Skipper, J. I. (2023). The entire brain, more or less, is at work:'language regions' are artefacts of averaging. bioRxiv, 2023-09.

Arbib, M. A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and brain sciences*, *28*(2), 105-124.

Bakermans, J. J., Warren, J., Whittington, J. C., & Behrens, T. E. (2025). Constructing future behavior in the hippocampal formation through composition and replay. Nature neuroscience, 1-12.

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. (2008). Pattern separation in the human hippocampal CA3 and dentate gyrus. *science*, *319*(5870), 1640-1642.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Barak, O., Rigotti, M., & Fusi, S. (2013). The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. Journal of Neuroscience, 33(9), 3844-3856.

Ben Shalom, D., & Poeppel, D. (2008). Functional anatomic models of language: assembling the pieces. The Neuroscientist, 14(1), 119-127.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. Cell, 183(4), 954-967.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral cortex, 19(12), 2767-2796.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Boersma, P., & Weenink, D. (2001). PRAAT (Version 4.0)[computer software]. Amsterdam: University of Amsterdam.

Bredin, H., & Laurent, A. (2021). End-to-end speaker segmentation for overlap-awareresegmentation. arXiv preprint arXiv:2104.04045.

Brown-Schmidt, S., Cho, S. J., Nozari, N., Klooster, N., & Duff, M. (2021). The limited role of hippocampal declarative memory in transient semantic activation during online language processing. Neuropsychologia, 152, 107730.

Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Soper, D. J., Xia, T., Wang, A. F., ... & Cash, S. S. (2025). Natural language processing models reveal neural dynamics of human conversation. Nature Communications, 16(1), 3376.

Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 160-172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Castellucci, G. A., Kovach, C. K., Howard III, M. A., Greenlee, J. D., & Long, M. A. (2022). A speech planning network for interactive language use. Nature, 602(7895), 117-122.

Chaure, F. J., Rey, H. G., & Quian Quiroga, R. (2018). A novel and fully automatic spike-sortingimplementation with variable number of features. Journal of neurophysiology, 120(4), 1859-1871.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*.

Chericoni, A., Fine, J. M., Chavez, A. G., Franch, M. C., Mickiewicz, E. A., Mathura, R. K.,...& Hayden, B. Y. (2025). Independent Continuous Tracking of Multiple Agents in the Human Hippocampus. bioRxiv, 2025-03.

Corballis, M. C. (2010). Mirror neurons and the evolution of language. Brain and language, 112(1), 25-35.

Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., ... & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. Nature, 632(8026), 841-849.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179-194.

Danjo, T., Toyoizumi, T., & Fujisawa, S. (2018). Spatial representations of self and other in the hippocampus. Science, 359(6372), 213-218.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, *3*(2), 71-100.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Dijksterhuis, D. E., Self, M. W., Possel, J. K., Peters, J. C., van Straaten, E. C. W., Idema, S.,… & Roelfsema, P. R. (2024). Pronouns reactivate conceptual representations in human hippocampal neurons. Science, 385(6716), 1478-1484.

Duff, M. C., & Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. *Frontiers in human neuroscience*, *6*, 69.

Duff, M. C., Covington, N. V., Hilverman, C., & Cohen, N. J. (2020). Semantic memory and the hippocampus: Revisiting, reaffirming, and extending the reach of their critical relationship. *Frontiers in human neuroscience*, *13*, 471.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. Neuron, 109(19), 3055-3068.

Ebitz, R. B., Sleezer, B. J., Jedema, H. P., Bradberry, C. W., & Hayden, B. Y. (2019). Tonic exploration governs both flexibility and lapses. PLoS computational biology, 15(11), e1007475.

Eisenreich, B. R., Hayden, B. Y., & Zimmermann, J. (2019). Macaques are risk-averse in a freely moving foraging task. Scientific reports, 9(1), 15091.

Elsayed, G. F., Lara, A. H., Kaufman, M. T., Churchland, M. M., & Cunningham, J. P. (2016). Reorganization between preparatory and movement population responses in motor cortex. Nature Communications, 7(1), 13239.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. Nature Reviews Neuroscience, 25(5), 289-312.

Fine, J. M., Maisson, D. J. N., Yoo, S. B. M., Cash-Padgett, T. V., Wang, M. Z., Zimmermann, J., & Hayden, B. Y. (2023). Abstract value encoding in neural populations but not single neurons. *Journal of Neuroscience*, *43*(25), 4650-4663.

Fine, J. M., Chericoni, A., Delgado, G., Franch, M. C., Mickiewicz, E. A., Chavez, A. G., ... & Hayden, B. Y. (2025). Complementary roles for hippocampus and anterior cingulate in composing continuous choice. bioRxiv, 2025-03.

Franch, M., Mickiewicz, E. A., Belanger, J. L., Chericoni, A., Chavez, A. G., Katlowitz, K. A., Sheth, S. A. & Hayden, B. Y. (2025). A vectorial code for semantics in human hippocampus. bioRxiv, 2025-02.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, *37*, 66-74.

Garagnani, M. (2024). On the ability of standard and brain-constrained deep neural networks to support cognitive superposition: a position paper. Cognitive Neurodynamics, 18(6), 3383-3400.

Goldstein, A., Wang, H., Niekerken, L., Schain, M., Zada, Z., Aubrey, B., ... & Hasson, U. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nature human behaviour, 1-15.

Groppe, D. M., Bickel, S., Dykstra, A. R., Wang, X., Mégevand, P., Mercier, M. R., ... & Honey,C. J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. Journal of neuroscience methods, 281, 40-48.

Gu, J. C., Li, T., Liu, Q., Ling, Z. H., Su, Z., Wei, S., & Zhu, X. (2020, October). Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2041-2044).

Hayden, B. Y. (2019). Why has evolution not selected for perfect self-control?. Philosophical Transactions of the Royal Society B, 374(1766), 20180139.

Heyes, C., & Catmur, C. (2022). What happened to mirror neurons?. Perspectives on Psychological Science, 17(1), 153-168.

Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. Language and cognitive processes, 25(6), 749-776.

Hickok, G. (2014). *The myth of mirror neurons: The real neuroscience of communication and cognition*. WW Norton & Company.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature reviews neuroscience, 8(5), 393-402.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453-458.

Jamali, M., Grannan, B., Cai, J., Khanna, A. R., Muñoz, W., Caprara, I., ... & Williams, Z. M. (2024). Semantic encoding during language comprehension at single-cell resolution. Nature, 631(8021), 610-616.

Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. Nature, 591(7851), 610-614.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage, 17(2), 825-841.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. Medical image analysis, 5(2), 143-156.

Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B., & Hayden, B. Y. (2024). Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. Nature Neuroscience, 27(11), 2218-2230.

Joshi, A., Scheinost, D., Okuda, H., Belhachemi, D., Murphy, I., Staib, L. H., & Papademetris,X. (2011). Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics*, *9*(1), 69-84.

Katlowitz, K. A., Belanger, J. L., Ismail, T., Chavez, A. G., Chericoni, A., Franch, M., ... & Sheth, S. A., Hayden, B. Y. (2025). Attention is all you need (in the brain): semantic contextualization in human hippocampus. *bioRxiv*, 2025-06.

Katlowitz, K. A., Shah, S., Franch, M. C., Adkinson, J., Belanger, J. L., Mathura, R. K., ... & Sheth, S. A. (2025). Learning and language in the unconscious human hippocampus. bioRxiv, 2025-04.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., & Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. Nature Neuroscience, 17(3), 440-448.

Keane, M. M., Bousquet, K., Wank, A., & Verfaellie, M. (2020). Relational processing in the semantic domain is impaired in medial temporal lobe amnesia. J Neuropsychol 14 (3): 416–430.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. Neuron, 79(5), 836-848.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. Journal of cognitive neuroscience, 19(1), 42-58.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox totackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, *18*(17), 1-5.

Leutgeb, S., Leutgeb, J. K., Barnes, C. A., Moser, E. I., McNaughton, B. L., & Moser, M. B. (2005). Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science*, *309*(5734), 619-623.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of neuroscience methods, 264, 47-56.

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. Nature neuroscience, 24(5), 715-726.

Magnotti, J. F., Wang, Z., & Beauchamp, M. S. (2020). RAVE: Comprehensive open-sourcesoftware for reproducible analysis and visualization of intracranial EEG data. NeuroImage, 223, 117341.

Manns, J. R., Hopkins, R. O., & Squire, L. R. (2003). Semantic memory and the human hippocampus. *Neuron*, *38*(1), 127-133.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), 205.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Montagrin, A., Saiote, C., & Schiller, D. (2018). The social hippocampus. Hippocampus, 28(9), 672-679

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., ... & Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. Journal of anatomy, 207(1), 35-66.

Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. Current biology, 20(8), 750-756.

Myers, J., Xiao, J., Mathura, R. K., Shofty, B., Gates, V., Adkinson, J., ... & Sheth, S. A. (2025). Intracranial directed connectivity links subregions of the prefrontal cortex to major depression. Nature communications, 16(1), 6309.

Piai, V., Anderson, K. L., Lin, J. J., Dewar, C., Parvizi, J., Dronkers, N. F., & Knight, R. T.(2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing. Proceedings of the National Academy of Sciences, 113(40), 11366-11371.

Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*(9), 844-856.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and brain sciences, 36(4), 329-347.

Pearson, J. M., Hayden, B. Y., & Platt, M. L. (2010). Explicit information reduces discounting behavior in monkeys. Frontiers in psychology, 1, 237.

Perich, M. G., Narain, D., & Gallego, J. A. (2025). A neural manifold view of the brain. *Nature Neuroscience*, 1-16.

Provenza, N. R., Reddy, S., Allam, A. K., Rajesh, S. V., Diab, N., Reyes, G., ... & Sheth, S. A. (2024). Disruption of neural periodicity predicts clinical response after deep brain stimulation for obsessive-compulsive disorder. Nature Medicine, 30(10), 3004-3014.

Provenza, N., Reddy, S., Rajesh, S., Allam, A., Diab, N., Reyes, G., ... & Sheth, S. A. (2024). Neuromodulatory Disruption of Pathological Circadian Neural Rhythms Predicts Clinical Response in OCD. Biological Psychiatry, 95(10), S23.

Provenza, N. R., Rajesh, S. V., Reyes, G., Katlowitz, K. A., Pugalenthi, L. S., Bechtold, R. A., ... & Sheth, S. A. (2025). High beta power in the ventrolateral prefrontal cortex indexes human approach behavior: a case study. Journal of Neuroscience, 45(37).

Rey, H. G., Gori, B., Chaure, F. J., Collavini, S., Blenkmann, A. O., Seoane, P., ... & Quiroga, R. Q. (2020). Single neuron coding of identity in the human hippocampal formation. *Current Biology*, *30*(6), 1152-1159.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in neurosciences*, *21*(5), 188-194.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. Annu. Rev. Neurosci., 27(1), 169-192.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry, 36(8), 1627-1639.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proceedings of the National Academy of Sciences, 111(43), E4687-E4696.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science, 303(5661), 1157-1162.

Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., & Yen, S. C. (2020). Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. Elife, 9, e58154.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. Neuron, 87(1), 231-243.

Tian, Z., Chen, J., Zhang, C., Min, B., Xu, B., & Wang, L. (2024). Mental programming of spatial sequences in working memory in the macaque frontal cortex. Science, 385(6716), eadp6091.

Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classicn model of language neurobiology. Brain and language, 162, 60-71.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

van der Maaten, L., & Hinton, G. (2008). User's Guide for t-SNE Software. *Structure*.

van de Ven, V., Waldorp, L., & Christoffels, I. (2020). Hippocampus plays a role in speech feedback processing. NeuroImage, 223, 117319.

Wang, M. Z., & Hayden, B. Y. (2021). Latent learning, cognitive maps, and curiosity. Current Opinion in Behavioral Sciences, 38, 1-7.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Xiao, J., Adkinson, J. A., Allawala, A. B., Banks, G., Bartoli, E., Fan, X., ... & Hayden, B. Y. (2024). Insula uses overlapping codes for emotion in self and others. *bioRxiv*

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X. J., ... & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. Science, 375(6581), 632-639.

Xu, D., Chung, J. E., Silva, A. B., Metzger, S. L., Greicius, Q. R., Zhang, Y., ... & Chang, E. F. (2024). Human precentral gyrus neurons link speech sequences from listening to speaking. bioRxiv, 2024-12.

Yacoub, E., Grier, M. D., Auerbach, E. J., Lagore, R. L., Harel, N., Adriany, G., ... & Zimmermann, J. (2020). Ultra-high field (10.5 T) resting state fMRI in the macaque. Neuroimage, 223, 117349.

Yang, A. I., Wang, X., Doyle, W. K., Halgren, E., Carlson, C., Belcher, T. L., ... & Thesen, T.(2012). Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage*, *63*(1), 157-165.

Yoo, S. B. M., & Hayden, B. Y. (2020). The transition from evaluation to selection involves neural subspace reorganization in core reward regions. Neuron, 105(4), 712-724.

Zada, Z., Goldstein, A., Michelmann, S., Simony, E., Price, A., Hasenfratz, L., ... & Hasson, U. (2024). A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, *112*(18), 3211-3222.