Representational similarity learning reveals a graded multi-dimensional semantic space in the human anterior temporal cortex

Christopher R. Cox[1※*], Timothy T. Rogers[2※*], Akihiro Shimotake[3*], Takayuki Kikuchi[4], Takeharu Kunieda[4,5], Susumu Miyamoto[4], Ryosuke Takahashi[3], Riki Matsumoto[7*], Akio Ikeda[6†], Matthew A. Lambon Ralph[8†*]

※ Co-first author; † Co-senior author; * Corresponding author

Affiliations:

1. Dept. of Psychology, Louisiana State University, Audubon Hall, Baton Rouge, LA 70803
2. Dept. of Psychology, University of Wisconsin, Brogden Hall, Madison, WI 53706
3. Dept. of Neurology. Kyoto University Graduate School of Medicine, Kawaharacho, Shogoin, Sakyo-ku, Kyoto, 606-8507, Japan
4. Dept. of Neurosurgery, Kyoto University Graduate School of Medicine, Kyoto, Japan
5. Dept. of Neurosurgery, Ehime University Graduate School of Medicine, Shizukawa Toon city, Ehime, 791-0295, Japan
6. Dept. of Epilepsy, Movement Disorders and Physiology, Kyoto University Graduate School of Medicine, Kyoto, Japan
7. Div. of Neurology, Kobe University Graduate School of Medicine, Kusunoki-cho, Chuo-ku, Kobe, 650-0017, Japan
8. MRC Cognition and Brain Sciences Unit, 15 Chaucer Rd., Cambridge, UK, CB2-7EF

Corresponding Authors:

- Christopher R. Cox (chriscox@lsu.edu) 0000-0001-8287-3865
- Timothy T. Rogers (ttrogers@wisc.edu) 0000-0001-6304-755X
- Akihiro Shimotake (smtk@kuhp.kyoto-u.ac.jp) 0000-0003-3777-685X
- Riki Matsumoto (matsumot@med.kobe-u.ac.jp) 0000-0003-3985-9210
- Matt Lambon Ralph (Matt.Lambon-Ralph@mrc-cbu.cam.ac.uk) 0000-0001-5907-2488

Open access: For the purpose of open access, the UKRI-funded author MALR has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Keywords: Semantic memory, ECoG, temporal lobe, decoding, representational similarity analysis, RSA, network RSA

Abstract

Neuro-cognitive models of semantic memory have proposed that the ventral anterior temporal lobes (vATLs) encode a graded, distributed, and multidimensional semantic space—yet neuroimaging studies seeking brain regions that encode semantic structure rarely identify these areas. In computer simulations we show that this discrepancy may arise from limitations of a commonly-used multivariate analysis, and introduce a new approach, *representational similarity learning* (RSL), that resolves these. We then use RSL to decode semantic structure from ECoG data collected from the vATL cortical surface while participants named line drawings of common items. The results reveal a graded, multi-dimensional semantic space encoded in neural activity across the vATL, which evolves over time and simultaneously expresses both broad and finer-grained semantic structure amongst animate and inanimate concepts. The work thus resolves the apparent discrepancy within the semantic cognition literature and suggests a new approach to discovering representational structure in neural data more generally.

Conflict of interest

The Department of Epilepsy, Movement Disorders, and Physiology, Kyoto University Graduate School of Medicine conducts Industry-Academia Collaboration Courses, supported by Eisai Co., Ltd, Nihon Kohden Corporation, Otsuka Pharmaceutical Co., and UCB Japan Co., Ltd.

Introduction

If you encounter a wolf when walking home through a dark wood, your mind readily accomplishes some remarkable feats: it anticipates the animal's likely behavior, perhaps slinking closer toward you; it assigns the thing a name, which you might shout to alert others ("wolf!"); and it directs you to change your own plans, maybe running back down the path. These feats arise from the human ability to discern conceptual structure—to realize that the wolf, despite its resemblance to friendly dogs in town, is nevertheless a quite different sort of animal.

This ability to represent and exploit conceptual structure is central to human semantic cognition. Such structure is *graded* in that similarities vary along a continuum: wolves are highly similar to coyotes, partially similar to elk, and quite distinct from birch trees or canoes. It is also *multidimensional* in that concepts vary along myriad independent components: wolves and dogs are similar in their shapes, parts, movements, and phylogeny, but different in their behaviors, habitats, and diets. To capture these properties of knowledge, computational approaches to semantics often represent concepts with *vector spaces*: the meaning of a word or image is expressed as a point in an *n*-dimensional space (or equivalently as an *n*-dimensional vector) such that the proximity between points expresses the similarity in meaning between the denoted concepts. Cognitive science and machine-learning offer many techniques for estimating semantic vector spaces from natural language [1], feature norms [2], or similarity-judgments [3], and these methods have provided a critical empirical foundation for studying human conceptual knowledge.

A central question for cognitive neuroscience concerns how semantic vector spaces are encoded in neural activity, but progress has been hampered by contradictory findings across different literatures. Convergent evidence from neuropsychology, clinical neurophysiology,

transcranial magnetic stimulation, and computational modeling suggests that the anterior temporal lobes (ATLs) encode graded and multidimensional conceptual similarity structure [4-9], but a direct empirical test of this hypothesis—multivariate decoding of functional neuroimaging data—has yielded decidedly mixed results. A few studies have found ATL areas that encode conceptual structure in neural activity [10-14], but others have observed such structure only in posterior temporal areas [15-22], or within a left-lateralized peri-sylvian network [12,23-25], or even across the entire cortex *except* the ATL [26].

This paper considers the possibility that the gap between imaging and other literatures arises because the most common analytic technique employed in these studies—representational similarity analysis (RSA)—is not actually well-suited to finding <u>*graded*</u> <u>*multidimensional*</u> structure. This may seem surprising, since RSA was developed specifically as a tool for finding cognitive similarity structure in imaging data [27]—yet the reliance of the approach on unsupervised correlation places non-trivial limits on the kinds of signal it can detect. Study 1 reports simulations assessing how these limitations will influence discovery of neuro-semantic representations, and introduces a new approach to neural decoding, *representational similarity learning* RSL; [28], designed to address them. Study 2 considers how RSL can be applied to decoding of very large neural datasets by exploiting patterns of *structured sparsity* present in neural signals. We show how such patterns can inform model fitting in RSL, then use the approach to decode graded multidimensional semantic structure from electro-corticography (ECoG) recorded from the surface of human vATL. The results resolve the contradiction in the literature and suggest a new path for multivariate neural decoding more generally.

Results

Study 1: Simulation experiments.

RSA aims to find sets of neural features—voxels, electrodes, or other measurements of neuro-physiological activity—whose responses to various cognitive events (e.g., the perception of a stimulus) jointly encode an independently-measured target structure. Typically, the target structure is a *representational similarity matrix* (RSM)[1] in which the rows and columns correspond to the different stimuli in an experiment and the entries indicate the cognitive/representational similarities between stimuli. For semantic representation, entries in the target RSM indicate similarity of meaning among pairs of concepts as estimated from behavioral or corpus data. To determine if a set of neural features (e.g., voxels, electrodes, EEG sources, etc.) encodes the target similarities, the experimenter estimates the pattern of neural activity evoked by each stimulus and computes the pairwise similarities between these to create a *neural similarity matrix* (NSM). Correlations between the RSM and NSM are computed separately for each participant, and brain regions where these are reliably greater than zero are interpreted as encoding the target structure.

This approach faces three challenges when used to find *graded*, *multidimensional* structure in neural codes. First, reliable correlations with a continuous-valued target RSM can arise even if the neural response is discrete or categorical (e.g., equally active for items in one semantic category and equally inactive for a different category). Thus, a positive result does not indicate that the neural response encodes *graded* structure even if such structure is present in the target RSM. Second, when computing the correlation, all neural features are equally weighted

[1] Many studies employ the term *representational dissimilarity matrix* or RDM instead. Since the approach is called "representational similarity analysis," and because similarity/dissimilarity are inversely related, we prefer *representational similarity matrix* or RSM.

and no parameters are fit, so the approach may yield a null result even if the target structure *is* encoded by a weighted subset of the neural features under consideration. Thus, a negative result does not indicate that the neural features carry no target information. Third, correlation is inherently one-dimensional, so in itself cannot provide evidence of multidimensional structure in the neural response. Simulation 1 assessed how these characteristics of RSA will influence discovery of neuro-semantic representations by applying the approach to simulated neural data encoding different aspects of a target similarity matrix.

Simulation 1: Limitations of RSA. The target RSM was computed from semantic feature norms for 100 items, half animate and half inanimate, collected in prior work [29]. The feature vectors for each item were aggregated in a matrix with rows indicating words and columns indicating features and entries indicating whether the referent of the word possess that feature. Matrix columns were mean-centered and the RSM was computed as the cosine similarity for all pairs of row vectors.

To understand the latent structure of the RSM, we used singular value decomposition (SVD) to extract three components accounting for 90% of the variance in the full matrix. **Figure 1** shows the coordinates for all 100 items on each latent dimension, color coded by semantic category (see SM). The decomposition reveals graded semantic structure along each dimension. The first separates animate from inanimate items but also individuates subcategories in each domain. The second strongly differentiates subcategories of animals but also weakly distinguishes inanimate subcategories. The third differentiates the inanimate items, though the resulting spread is less “clumpy” than within animate items as commonly found with such data [5].

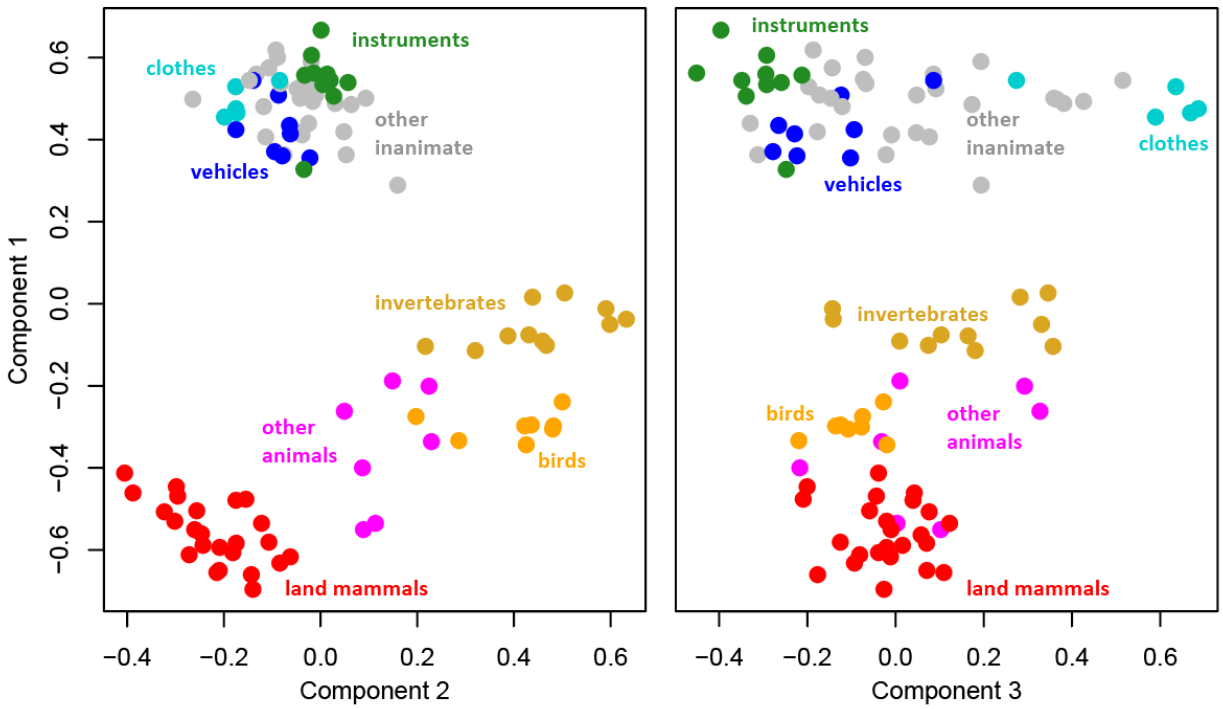

Figure 1. Coordinates of the 100 stimuli in the three-dimensional singular-value-decomposition of the of the semantic similarity matrix. Colors show category membership within each domain. Component 1 separates animate from inanimate items, but also separates animate sub-categories. Component 2 largely separates animate sub-categories, while component 3 largely separates inanimate sub-categories.

We investigated what results RSA would yield when applied to a neural signal known to encode some aspects of this underlying semantic structure. Specifically, we created five simulated neural datasets, each capturing a different aspect of semantic structure in the target RSM. For each dataset we simulated the responses of 24 hypothesized neural features (e.g., the BOLD response over 24 voxels in an ROI, or the mean evoked response at each of 24 electrodes) to the 100 stimuli. Half of the simulated neural features encoded true semantic information about the target items while half adopted random values sampled from a uniform distribution in the same range. The five datasets varied in which aspects of the target matrix they encoded, as illustrated in the schematic plots in **Figure 2A.** In the *binary condition*, all signal-carrying neural features adopted one state (-1 or 1) for animate items and the opposite state for inanimate items. Thus, the code was both one-dimensional and discrete. In three *one-D conditions*, all signal-

carrying neural features only encoded a single latent component of the target RSM—either the first, second, or third singular vector as shown in **Figure 1**. Thus, each encoded *graded* similarity structure but only *one* dimension of variation. In the *full structure condition,* the signal-carrying units jointly encoded all three latent components of variation in the target RMS—thus the code was both graded and multidimensional.

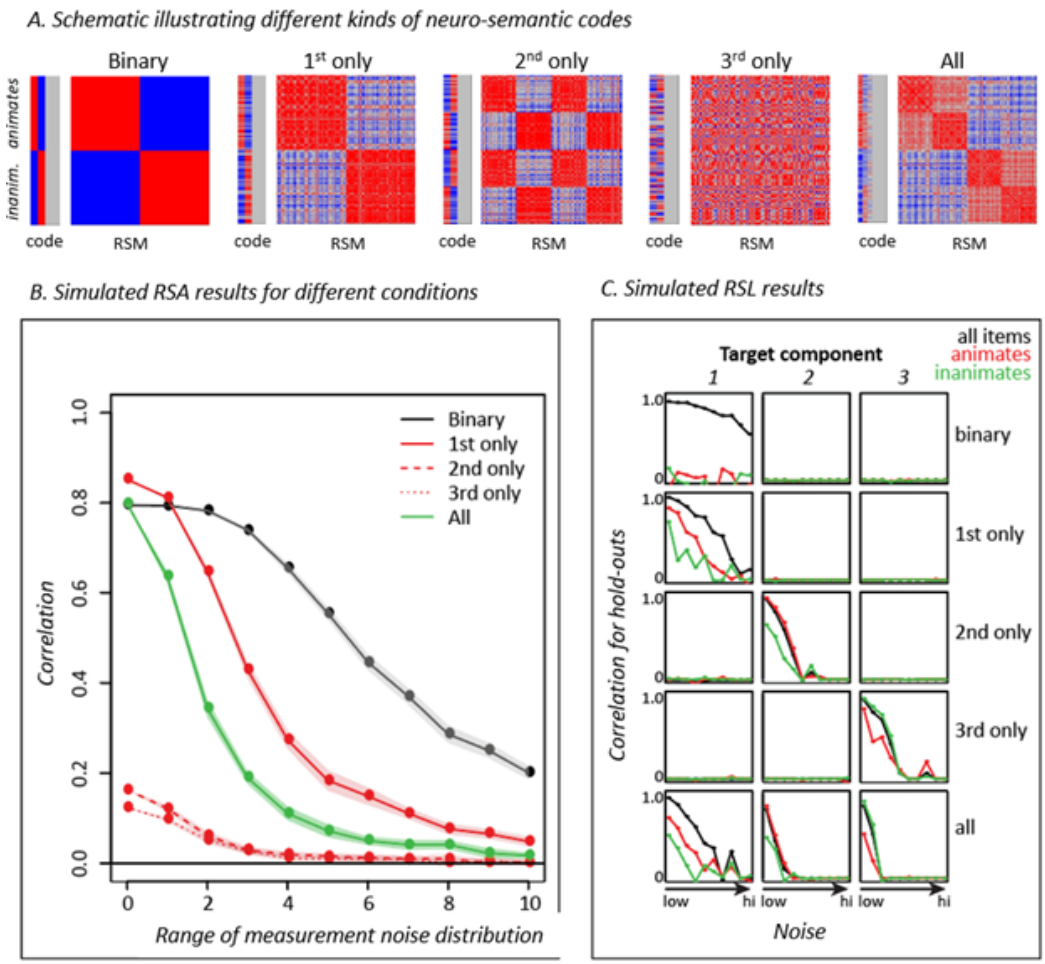

Figure 2. Simulation results. A. Each plot illustrates different ways that a neural response might encode some aspects of a target RSM, schematized in the rightmost plot ("All"). Vertical bars show the hypothesized responses of neural features to various animate and inanimate stimuli (red more active; blue less active; gray random) while the squares show the neural similarity matrices that would then result (red high similarity, blue low). All conditions encode some components of the full RSM and so should be detected by a multivariate method seeking semantic structure. B. Curves showing the expected results of standard RSA when used to decode real semantic structure from the Dilkina norms under different hypotheses about the neural code and increasing amounts of measurement noise. Dots show the mean correlation while ribbons show the 95% confidence interval. The approach can yield robust results when the neural signal is discrete and/or one-dimensional, even under substantial noise, and can yield weak or null results when the neural signal faithfully encodes weaker components of the RSM, even under low-noise conditions. C. Results of RSL applied to the same simulated data, showing the correlation between true and predicted coordinates for held-out items on each latent component of the target matrix (columns), for the five different simulation conditions (rows), and computed across all items

(black lines), animate items (red), or inanimate items (green). RSL shows a positive result in each case, identifies which components of the matrix are present in the neural code, and reveals reliable within-domain decoding only when the neural code expresses continuous similarity structure.

To examine the results RSA would yield for each kind of neural signal, we distorted the simulated neural responses to each stimulus with measurement noise sampled from a uniform distribution centered on 0, then computed the correlation between the simulated NSM and target RSM. **Figure 2B** shows the results across increasing measurement noise and averaged over 20 runs of the simulation. The binary code showed robust correlations across the full noise-spectrum, illustrating that RSA can yield a positive result even when the neural code is neither continuous nor multidimensional. Amongst the one-D codes, RSA showed a strong positive result if the neural response encoded the first singular vector of the target matrix, but much weaker results that decayed to zero with increasing noise when the neural response encoded other components. Thus, RSA can yield null results even when the neural features under consideration reliably encode target information orthogonal to the primary component of the RSM. A strong correlation was observed when the neural features encoded all three components of the target matrix and there was little noise in the signal, but this declined steeply with noise. In high noise conditions, RSA yielded a more reliable result when the simulated neural response encoded a binary domain distinction or just the first component of variation. For these scenarios, all signal-carrying features encode the same information, so noise “cancels out” across features and the signal remains robust. When the same number of features encode three orthogonal components, fewer resources are dedicated to each, and corruption from noise more seriously degrades the signal.

The simulation shows that, when decoding semantic similarity structure of the kind captured by feature norms, RSA can yield a positive result when the neural response is discrete and/or one-dimensional, and a null result when the neural response encodes underlying

dimensions of the target matrix beyond the strongest. Thus RSA, as typically deployed, does not reliably indicate whether a set of neural signals encode the graded, multidimensional semantic structure existing in a target RSM.

Simulation 2: Decoding with RSL. These issues led us to consider an alternative approach designed explicitly to decode graded, multidimensional similarity structure, which we call *representational similarity learning* RSL; [28]. Rather than using pairwise similarities in the RSM as the target values for decoding, RSL first decomposes the matrix into orthogonal latent components, effectively re-representing each stimulus item as a point in a low-dimensional semantic space as shown in **Figure 1**. It then fits *decoding models* to predict the coordinates of each item along each dimension from their evoked neural responses.

Simulation 2 assessed whether this approach can reveal the graded and/or multidimensional structure obscured by RSA. Using the same dataset from Simulation 1, we fit regression models to predict coordinates of each stimulus along each of the three latent semantic dimensions shown in **Figure 1**, then assessed model fit using 10-fold hold-out cross validation. Multidimensional structure in the neural code should be revealed by reliable prediction of coordinates along more the one latent dimension. To assess whether the neural code captures graded similarity, we evaluated models fit separately to animate- or inanimate items only. If the neural signal only categorically differentiates, for instance, animate from inanimate stimuli as observed, for instance, by [30], then predicted and true values should correlate significantly when models are fit to all items, but should not correlate when fit to just the animate or just the inanimate items. Reliable intra- and inter-domain correlations thus indicate that the neural code captures a graded degree of similarity structure.

Figure 3C shows the correlation between true and predicted coordinates for held-out items on each of the three target components (columns), for the five different neural coding scenarios (rows), and computed across all items (black lines), animate items (red), or inanimate items (green). The pattern of results reveals the information encoded in each simulated neural response. With the binary neural signal (top row), predicted and true coordinates correlated strongly for models fit to all items but not separately for animates vs inanimates. With continuous one-D signals, strong correlations were observed for models fit to all items and within each domain separately, but only for the dimension encoded by the neural response. When the neural signal encoded graded multidimensional structure, reliable correlations were observed on all three latent components, for models fit to all items and fit separately to each domain. Thus in simulation, the pattern of results uniquely revealed whether the neural signal encodes graded or discrete structure, and which semantic dimensions it captures.

Study 2: Decoding semantic structure from ECoG using RSL

The simulations show how RSL can test for graded and multidimensional structure where RSA results are ambiguous. Applying the approach to real brain imaging data, however, faces an immediate challenge: such technologies produce many more neural measurements than there are stimuli, so regression models predicting stimulus characteristics from neural data are ill-defined (i.e., there exist infinite solutions that can perfectly predict outcomes even from random input data). To find a unique fit without a-priori feature selection, the regression must be *regularized* to satisfy some additional constraint that will guarantee a single unique solution for model fitting. For instance, the weight optimization might jointly minimize prediction error and the size of the model coefficients, measured as the sum of their absolute values (the *L1-norm*). Since this sum can be minimized by placing zero coefficients on many predictors, this form of

regularization promotes a sparse solution in which only a few predictors have non-zero coefficients in the final model e.g., [31]. Other regularizers (such as the L2 norm or the elastic net) enforce different constraints on model fitting, and thus lead to other solutions when applied to the same dataset [32]. The selection of a regularization function therefore amounts to a prior hypothesis stipulating how the neural signal is likely to be structured [33].

Regularization with structured sparsity. RSL applied to real neural data adopts the *group-ordered-weighted LASSO* or grOWL loss [28], a regularization function designed specifically in prior work to capture three theoretically-motivated assumptions about the structure of neural representations. First, it assumes the signal is *sparse*: of all measurements taken, only a relatively small proportion encode target information of interest, so the regularizer should place zero coefficients on many predictors (as with L1 regularization). Second, it assumes *redundancy* in the true signal: neural populations that encode variation along the same dimension will be correlated in their activity patterns over items. The regularizer should therefore spread similar coefficients across correlated signal-carrying populations (similar to L2 regularization). Third, it assumes that signal-carrying populations are unlikely to be "axis-aligned" with the target semantic space (i.e., where each semantic dimension aligns perfectly and uniquely with individual neural elements), but more likely encode multiple, varying "directions" within the space, such that the neural population carries information about the items' coordinates within the multiple dimensional target space. We will refer to this as the *spanning* assumption, since the selected neural features span the target representational space but are unlikely to only encode distinct, individual dimensions. To enforce this assumption, the regularizer should prefer solutions where coefficients on a given neural feature are either all zero (the feature does not carry any information) or all non-zero (it predicts some variation along all target dimensions).

All three assumptions can be captured in a single loss function, as follows. Rather than fitting a separate regression model for each latent component of the target RSM, we instead predict all latent components (**Figure 3A**) simultaneously by estimating the parameters of a *decoding matrix* β (Figure 3C). Suppose the coordinates of *n* stimuli on *r* latent semantic dimensions are stored in matrix $\boldsymbol{Y}_{n\times r}$ while the responses of *m* neural features to the stimuli are stored in matrix $\boldsymbol{X}_{n\times m}$. RSL models the entries in $\boldsymbol{Y}$ as the matrix product of the neural activations in $\boldsymbol{X}$ and the decoding matrix $\boldsymbol{\beta}_{m\times r}$: $\boldsymbol{Y} = \boldsymbol{X\beta}$. Each row of $\boldsymbol{\beta}$ corresponds to one neural feature (voxel, electrode, etc.), and each column encodes weights for each neural feature when predicting stimulus coordinates on the corresponding dimension of the target matrix $\boldsymbol{Y}$. With this setup, the assumptions about neural signal just listed can be formalized as constraints on the structure of the decoding matrix $\boldsymbol{\beta}$ (**Figure 3D**). To capture the sparsity and spanning assumptions, $\boldsymbol{\beta}$ is constrained to be *row-sparse*: most rows have all zero values (sparsity), and the rest have all non-zero values (spanning). To capture the redundancy assumption, highly correlated and signal-carrying neural features receive similar row-vectors in $\boldsymbol{\beta}$ (**Figure 3D**). A formal description of the decoding model and grOWL regularizer, including a definition of the loss and explanation of how it encourages these properties in the decoding matrix, are provided in the Supplementary Materials. Further analysis and proofs were described by [28]. **Figure 3** shows the full workflow decoding semantic structure from ECoG data.

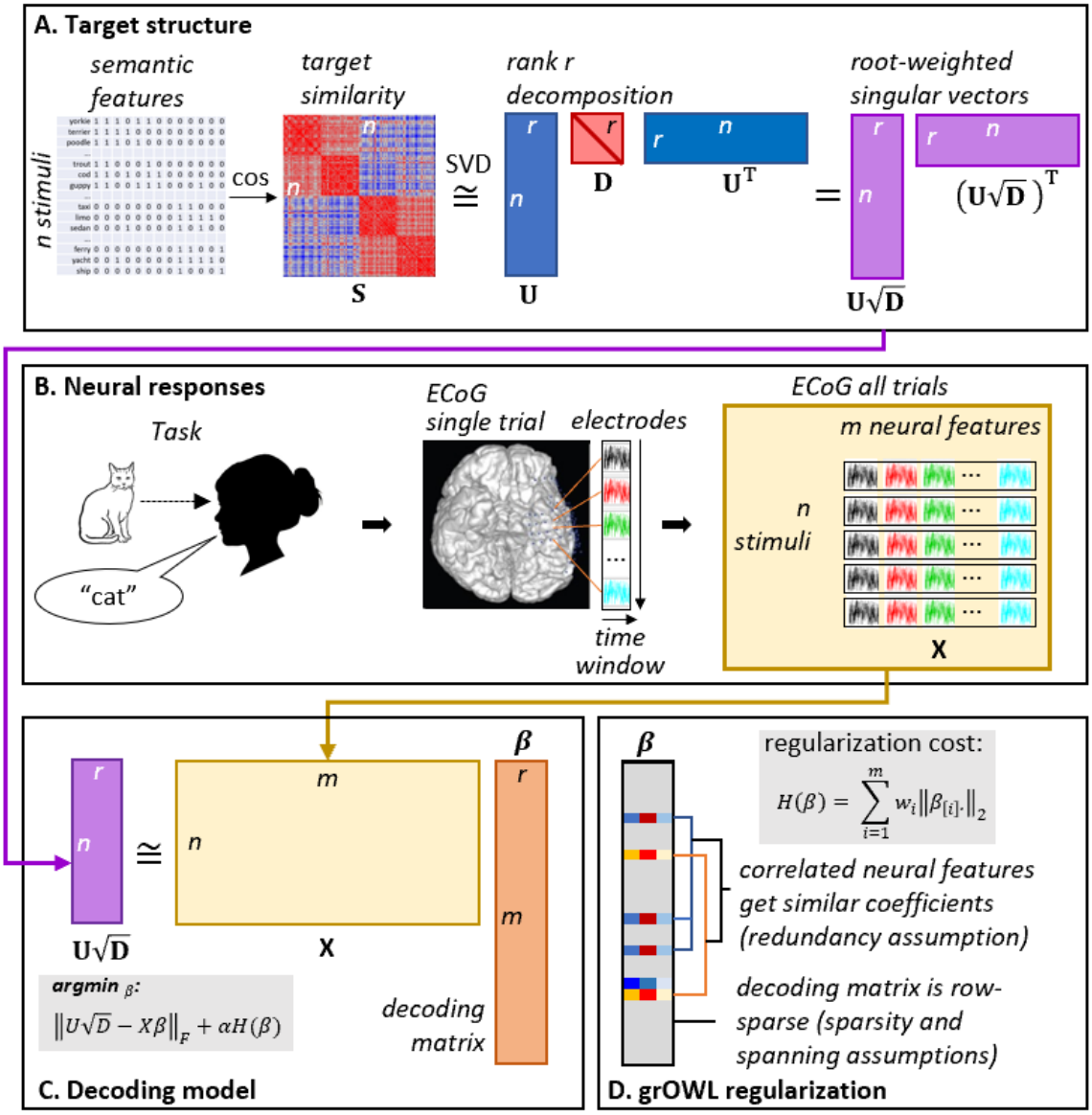

Figure 3. Representational similarity learning. A. To generate the target structure for decoding, semantic feature vectors for each of n items are mean-centered and converted to a cosine similarity matrix $\boldsymbol{S}_{n\times n}$ of approximate rank r. This is decomposed into matrix $\boldsymbol{U}_{n\times r}$ containing r orthogonal singular vectors and diagonal matrix $\boldsymbol{D}_{r\times r}$ containing singular values for each vector. The product $\mathbf{U}\sqrt{\mathbf{D}}$ is an *nxr* matrix whose columns contain *r* root-weighted singular vectors that are the target of the decoding model. Note that the product of this matrix with its transpose provides an estimate of the original matrix $\boldsymbol{S}_{n\times n}$. B. Neural responses in a picture-naming task are recorded for each stimulus as intra-cranial voltage potentials collected at 1000 Hz over electrodes implanted in vATL. Data are concatenated across electrodes to create a single m-dimensional "neural feature vector" for each stimulus, where m is the total number of electrodes times the size of the decoding window in milliseconds. All such vectors are compiled into a neural feature matrix $\boldsymbol{X}_{n\times m}$ which contains the predictors for the decoding model. C. The target structure $\boldsymbol{U}\sqrt{\boldsymbol{D}}$ is then modeled as the product of the neural feature matrix $\boldsymbol{X}$ and a decoding matrix of regression coefficients $\boldsymbol{\beta}$ whose values are chosen via gradient descent to minimize prediction error plus the regularization penalty $H(\boldsymbol{\beta})$ scaled by the weighting parameter α. D. Definition of the group-ordered-weighted LASSO (grOWL) regularization function and illustration of the structure it promotes in the decoding matrix.

With this overview we can consider how RSL can be applied to assess whether neural data encode a multidimensional and graded similarity space. Parameters in $\boldsymbol{\beta}$ are estimated for a batch of training data using regularized regression. The decoding model is then applied to predict the coordinates of held-out items along each dimension in the target representation space. If predicted and true coordinates for held-out items correlate reliably with more than one dimension, the neural signal must encode multidimensional structure. To assess whether the neural code captures graded similarity, we additionally fit and evaluate models separately for animate and inanimate items, as in the simulation. If the neural signal only categorically differentiates animate from inanimate stimuli as suggested by the RSA analysis in [30], then predicted and true values should correlate significantly across all items, but not across just the animate or just the inanimate items. Observation of both inter- and intra-domain correlations thus indicates that the neural code captures a graded degree of similarity structure. Finally, to assess how well the decoder recovers the target semantic similarities, the predicted coordinates for held-out items can be used to estimate semantic distances between stimulus pairs. These estimates can then be correlated with true semantic distances, similar to standard RSA.

ECoG experiment. We applied this approach to an ECoG dataset where prior work using RSA found little evidence for graded multidimensional structure [30]. While the authors did find a reliable correlation between a semantic RSM and the NSM in the anterior ventral temporal cortex, further analysis showed that the NSM correlated equally well with a binary matrix capturing only stimulus animacy (similar to the binary code in Simulation 1) and did not correlate at all with a matrix capturing within-domain semantic structure (similar to the 1D codes in Simulation 1). These results do not seem to be in keeping with the hypothesis that vATL

activity encodes a graded, multi-dimensional semantic space; accordingly, we assessed whether such a space could be recovered from the same data using RSL.

The dataset contains voltages measured from grid electrodes implanted in the surface of left (8) or right (2) vATL in ten patients undergoing preparation for intractable epilepsy (9) or brain tumor (1) while they named line drawings depicting the 100 items described in Study 1. Each patient had between 6 and 32 electrodes (mean of 20) covering vATL. The dataset included, for each stimulus, voltages measured at each electrode over a 1s window from stimulus onset and downsampled to 100 Hz. Thus, for a participant with 20 electrodes, each stimulus was associated with a 2,000-element vector of voltages (20 electrodes x 100 time points). See Methods for further detail about the patients, stimuli and experimental procedure.

All models were fit and evaluated using a nested 10-fold cross-validation procedure: an inner loop performed on 90% of the data was used to select model hyper-parameters; a single model was then fit using the best hyper-parameters and was evaluated on the remaining 10% of data. This procedure was run 10 times for 10 different final hold-out sets. In each fold, the fitted model was used to predict coordinates of the held-out items on the three latent semantic dimensions. The predictions were then aggregated across folds to create a single matrix of predicted coordinates for each item when it was held out from model fitting. Separate models were fit and evaluated independently for each participant, and we assessed whether the mean correlation between predicted and true coordinates was reliably positive across participants. Further model fitting and evaluation details appear in the Methods.

Our analyses closely followed the simulations: correlations between predicted and true coordinates on each dimension were computed for all 100 stimuli, just the 50 animate items, or just the 50 inanimate items. To determine whether the observed correlation was statistically

reliable, we compared it to bootstrapped null distributions generated via permutation testing [34; see Methods], using a familywise error rate of 0.05. Finally, to understand the effects of regularization with structured sparsity (grOWL) vs. more standard techniques, we compared models fit with grOWL regularization to those fit with L1 (LASSO) regularization.

Figure 4(top) shows results from decoding the full 1s time-window. Models fit with grOWL reliably decoded the first and second latent components of the semantic matrix, both when fit to all items and just to animate items. Thus, grOWL reveals graded, multidimensional semantic structure in ECoG signals recorded from ventral ATL. Regularization with the LASSO showed similar decoding of the first latent component but null results for the second and third, suggesting that additional constraints from grOWL aided in the discovery of semantic structure.

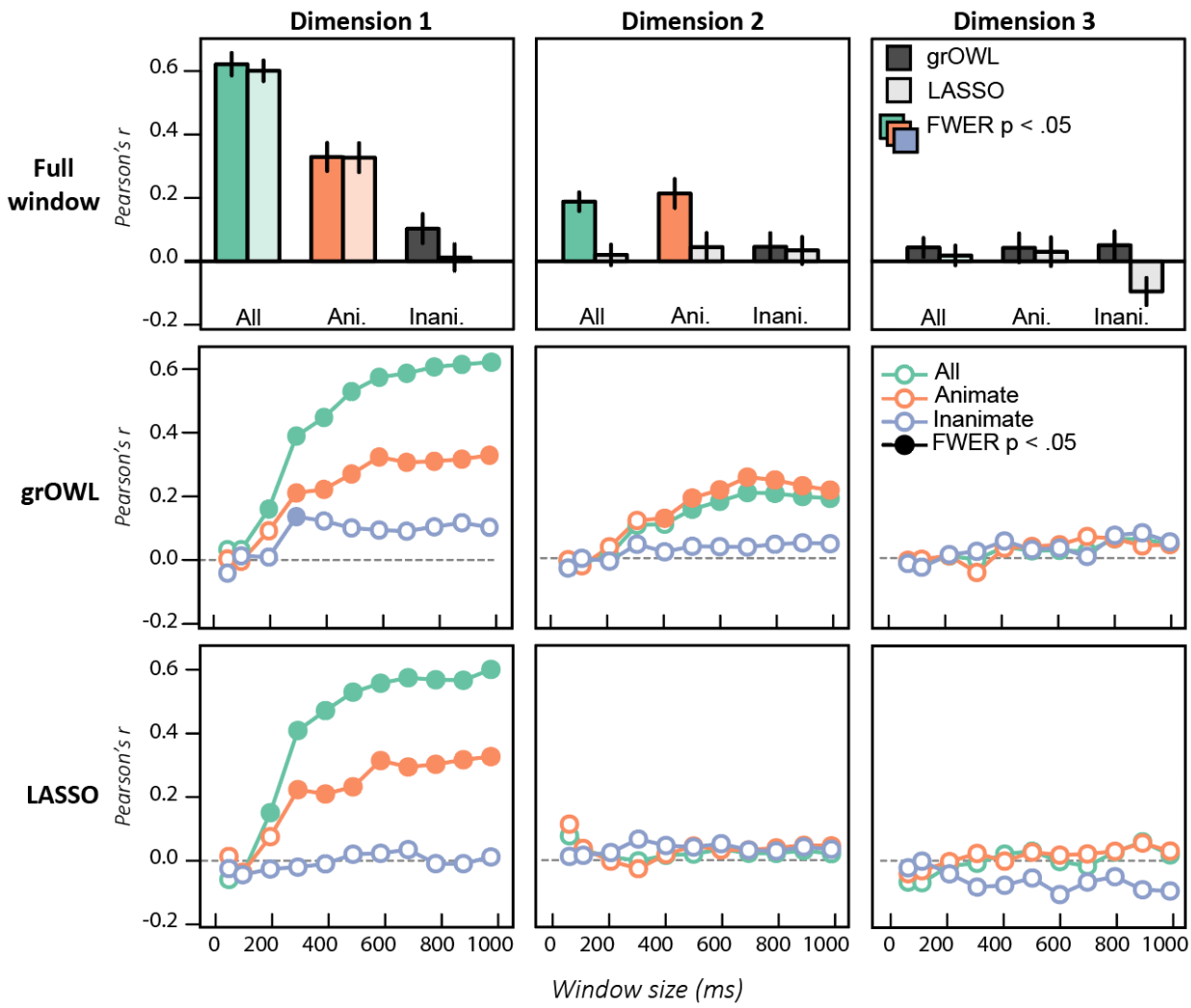

Figure 4. ECoG decoding results. Correlation between true and predicted coordinates along each latent dimension (columns) for models fit to all items (green), animate items only (orange), or inanimate items only (blue), and regularized with grOWL or LASSO. Bars show standard errors. Colored bars / filled circles indicate reliable decoding with Westfall-Young FWER < .05 for all points in a panel. Top row: grOWL and LASSO model performance when trained on the full 1000-ms trial epoch. Middle row: grOWL model performance within opening windows. Bottom row: LASSO model performance within opening windows.

To see how this structure emerges in the ECoG signal over time, we conducted an "opening window" analysis in which the same procedure was applied to an increasingly wide aperture of data, beginning with just the first 50 ms post-stimulus, extending to 100 ms, then growing by 100 ms up to 1000 ms. At each window models were fit and evaluated as just described. The results showed reliable decoding of the first dimension by 200 ms, but only for models fit to all items, suggesting an early differentiation of animate from inanimate items. By

300 ms, models also reliably decoded component-1 variation when fit separately for just animate or just inanimate items. Decoding accuracy along component 1 continually improved with wider windows for models fit with all items or animates alone, but not for models fit only with inanimate items. Decoding along the second component was not observed until 400 ms (for animates only) or 500 ms (for all items). Decoding along the third (weakest) component of the target matrix was not observed at any time point. Models fit with LASSO again revealed reliable decoding of only the first latent dimension, and only for models fit to all data or just animates.

Finally, we considered how well the decoding model recovers true semantic similarities between stimuli by reconstructing the pairwise similarity matrix from the predicted coordinates of held-out items across all three dimensions, then correlating these decoded similarities with the best possible reconstruction (i.e., similarities expressed by the three singular vectors/values of the true RSM; results for the original full-rank matrix are similar and shown in SM). For comparison to the prior RSA analysis of the same dataset [30], models were fit and evaluated on a 100 ms moving window, beginning at 0 ms from stimulus onset and advancing in 50 ms increments (see Methods). At each window, predicted coordinates for all items were generated using the nested cross-validation procedure described earlier and concatenating the predictions for each hold-out set across folds. The predicted coordinates were used to compute predicted similarities between all stimulus pairs. Each row of the resulting matrix contains predicted similarities between an item and all other items. We then computed, for each row, the correlation between the predicted and true similarities, and averaged these correlations across (1) all rows, (2) just the animates or (3) just the inanimates. The full procedure was carried out independently for each participant at each time-window. Correlation coefficients in each condition (all items,

animates, inanimates) were then averaged across participants at each time-window and compared to a bootstrapped permutation-based null distribution using a familywise error-rate of $p < 0.05$.

The results are shown in **Figure 5** (left). Filled circles indicate where predicted/true correlations are reliably non-zero relative to a bootstrapped permutation-based null distribution with a familywise error-rate of $p < 0.05$. Reliable correlations were observed from 100 – 150 ms post-stimulus onward, whether computed across all items, within domain only, or between-domain only. **Figure 5** (left) also shows the mean correlation (across participants) between RSM and NSM at each window reported in Y. Chen et al.'s [30] RSA analyses of the same data. While reliably positive correlations were observed from 200 ms onward, the correlation coefficient was about 1/10th the magnitude observed with RSL.

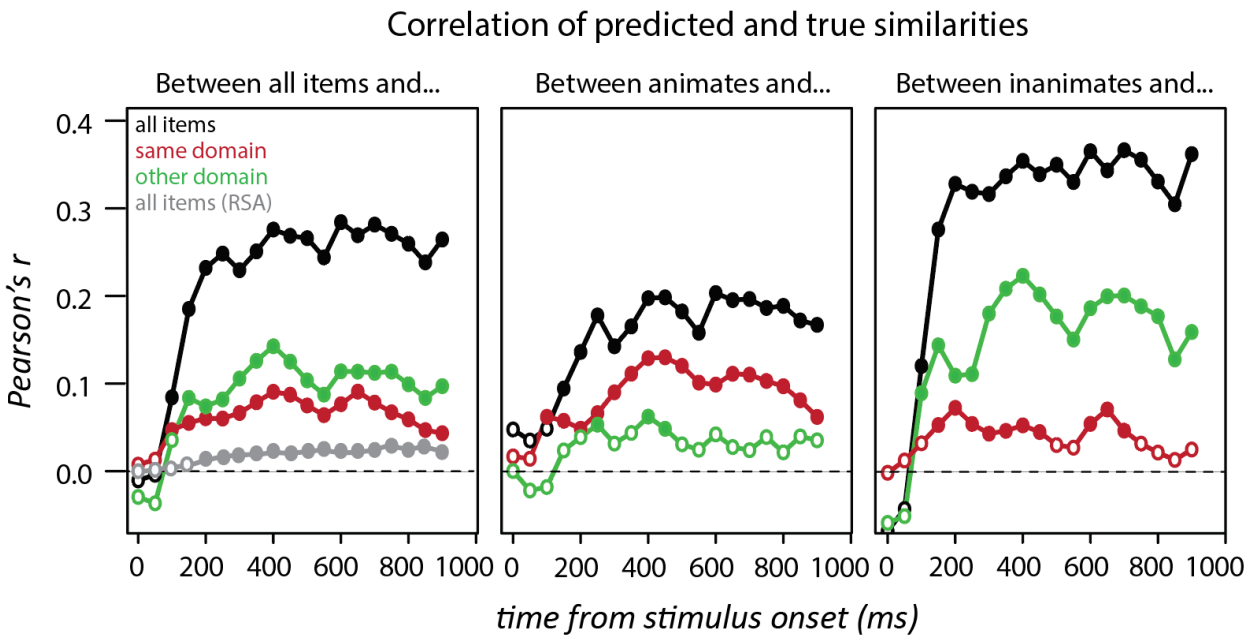

Figure 5 Correlation between predicted semantic similarities and the best-possible similarities reconstructed from the first three singular vectors/values of the true matrix. Considering all pairwise similarities (left panel black line), predictions from RSL show much stronger correlations with the true structure than observed in a previous study using searchlight RSA (left panel gray line; see [30]). Moreover, where the prior study found no evidence for within-domain semantic structure in these data, decoding models fit with RSL reliably predict semantic similarities both within (red lines) and between (green lines) conceptual domains, for both animate (middle panel) and inanimate (right panel) stimuli, from about 100 – 200 ms post-stimulus onward.

To assess whether RSL models capture structure for both animate and inanimate items, we computed the same three metrics separately for each domain. First considering just the animates, we computed the correlation between true and predicted similarities to all other items, other animate items (i.e., same domain), or inanimate items (i.e., other domain). The results in **Figure 5** (middle) show reliable decoding in all three cases, albeit at only a few time-windows for similarities from animates to inanimates. Applying the same analysis to inanimates, we observed reliable decoding of similarities to all items, same-domain items, and other-domain items from about 200 ms post-stimulus. Note that the weakest reliable correlations are still about double the magnitude of those observed in the Y. Chen, et al. [30] study using RSA.

Discussion

We introduced this paper with a puzzle: neuropsychology, clinical neurophysiology, TMS and computational modelling all suggest that the vATLs encode a continuous multidimensional semantic space, but multivariate analysis of brain imaging data often leads to quite different conclusions. In simulation we showed that the puzzle might reflect more general limitations of RSA, which can produce positive results even if the underlying neural code is discrete and one-dimensional, or negative results even if the neural code does capture latent structure in the target matrix beyond the first component. RSL addresses these limitations by using regularized regression to predict coordinates of stimuli along the latent dimensions of the RSM. For application to large neural datasets, we illustrated how hypothesized patterns of structured sparsity in the neural signal can constrain model fit via grOWL regularization and applied this approach to discovery of semantic structure in an ECoG dataset where prior work using RSA found only a binary animacy code. In contrast, RSL uncovered a graded and multidimensional semantic space capturing similarities within and between animate and

inanimate domains—reconciling the apparent discrepancy between neuroimaging and results yielded by other methodologies.

We note that decoding of intra-domain similarities was always more robust for animate compared to inanimate items. Considering each dimension separately, within-inanimate decoding was only reliable for one time window on the first latent dimension. When predicting pairwise distances, correlations with true similarities were always smallest for within-inanimate structure (i.e., the "other domain" correlations in the middle panel of **Figure 5**, and "same domain" correlations in the rightmost panel). We attribute this to differences in the similarity structures of animate and inanimate concepts. Semantic subcategories of inanimate objects are only weakly differentiated, not just in the current dataset, but in norming studies more generally and in other approaches to characterizing semantic structure e.g., [35,36-38]. Indeed, this difference in statistical structure is precisely what leads, under some theories, to apparent category-specific patterns of semantic impairment [5,39] and functional activation [40]. The differences can be seen in **Figure 1**, where animate items fall into somewhat well-differentiated subcategories in dimensions 1 and 2 while inanimates are uniformly distributed with no clear subclusters even along dimension 3. Indeed, pairwise similarities reconstructed from the first three singular vectors of the full semantic matrix are more accurate for animate than inanimate items (see SM3). Thus, the structure of the semantic matrix itself *requires* that decoding be worse within inanimates than within animates, precisely because inanimates are less well-structured. Yet despite this intrinsic disadvantage, within-inanimate similarities were still reliably decoded over many time-windows, illustrating that neural signal in vATL does express some intra-domain semantic structure even for inanimate stimuli. In summary, the results suggest the vATLs encode a multidimensional representation space that captures the conceptual similarities existing

amongst a variety of different concepts including both animate and inanimate items [6,9,41]. Better decoding within animates simply reflects the fact that animate subcategories are better-differentiated, so that intra-domain similarities are better-approximated by a low-rank decomposition.

We further note that multidimensional structure was only discovered with grOWL regularization—that is, when the decoding model was constrained to be sparse overall, with correlated neural features receiving similar weights, and with weights on all dimensions encouraged to be all-zero or all-nonzero. The observation is important for two reasons. First, consistent with other recent work [32], it shows that the form of model regularization matters critically to conclusions about information encoded in the neural signal. Simple regularization with the L1 norm was insufficient to reveal more than one dimension of structure in the neural code. Second, it suggests that different dimensions in the semantic matrix are *not* encoded separately by distinct neural populations at the resolution captured by ECoG. The pressure from grOWL to place either all-zero or all-nonzero coefficients on each neural feature produced reliable decoding of both the first and second latent dimensions of the semantic matrix—thus features that predict variation along the first dimension are also likely to carry information about the second dimension and vice versa. This observation is again consistent with the view that all signal-carrying neural populations jointly contribute to encoding a single semantic space (see also analysis of feature distributions across the cortical surface in SM).

Since we only analyzed electrodes situated in vATL, the current data do not shed light on structure encoded in other parts of cortex. The characteristics of RSA we have identified do, however, carry additional implications for interpretation of the broader literature. A recent review [33] identified 34 studies that have applied RSA to the discovery of semantic

representations in the brain, with positive results observed across multiple cortical regions including posterior temporal cortex [19], angular gyrus [12,42], left perisylvian cortex [25], and prefrontal cortex [22]. Simulation 1 suggests, however, that RSA will yield a positive result for any property correlated with the animate/inanimate distinction, including even discrete binary properties. Many features are confounded with animacy: inanimate items tend to be more familiar, less visually complex, more associated with action plans, less associated with motion, more likely to have lines and corners, less predictable from color or texture, etc. see [43 for a review]. Most RSA papers do not report the magnitude of correlation between RSM and NSM where results are significant, instead focusing on whether the estimated correlation coefficient is reliably non-zero across participants. As shown in **Figure 5**, this can happen even when mean correlations are very small. Moreover, several studies use target RSMs with only a small number of rows/columns—sometimes as few as 5 e.g., [12,19], meaning that correlations are computed across just 10 cells of the matrix (i.e., the lower triangle of a 5x5 RSM). Such small numbers increase the likelihood that a small-but-non-zero correlation is driven by some arbitrary property of the chosen stimuli or categories. Together these observations raise the possibility that the literature contains misleading positive results—brain areas whose responses encode unidimensional characteristics weakly confounded with animacy, rather than multi-dimensional semantic structure. Testing this possibility for various brain areas hypothesized to encode semantic structure will require analyses similar to the ones we have developed here.

Conclusion

In cognitive science, semantic representations are often construed as vector spaces that encoded graded, multidimensional similarity structure amongst the concepts experienced in our verbal and nonverbal world. Through application of a new technique for mapping

494 representational similarity in neural activity, we have shown that neural signals in vATL encode
495 such a space. In so doing, we have identified some limitations of representational similarity
496 analysis, a widespread technique commonly thought to reveal graded and multidimensional
497 representational structure. The work resolves an important discrepancy between behavioral and
498 neuroimaging results in prior work and suggests a new approach to discovering representational
499 structure in neural data more generally.

500 **Methods**

501 **Patients and ECoG Acquisition**

502 Ten patients with intractable partial epilepsy (nine) or brain tumor (one) participated in this
503 study, including two for whom language was not left-lateralized as determined by the WADA
504 language lateralization test. Chen, et al. [30] provide their background clinical information.
505 Platinum subdural electrodes (83 per patient on average; range 56 – 107 electrodes) were
506 implanted in the left (8 patients) or right (2 patients) hemisphere for presurgical evaluation with
507 inter-electrode distance of 1 cm and recording diameter of 2.3 mm (ADTECH, WI). Between six
508 and 32 (mean 20) electrodes covered the vATL in each patient. Data were sampled at 1000 Hz
509 (eight patients) or 2000 Hz (two patients) with a band-pass filter of 0.016 – 300 (eight patients)
510 or 0.016 – 600 Hz (two patients), though our analyses downsampled these to 100 Hz. All patients
511 with epilepsy had seizure onset zones outside the anterior fusiform region, except one patient for
512 whom it was not possible to localize the core seizure onset region. The study was approved by
513 the ethics committee of the Kyoto University Graduate School of Medicine (No. C533).
514 Participants provided written informed consent to participate in the study.

515 **Stimuli and Procedure**

One hundred line drawings (50 animate and 50 inanimate items) were obtained from previous norming studies (Morrison et al., 1997; Snodgrass & Vanderwart, 1980). See Chen, et al. [30] for a complete list. Animate and inanimate stimuli were matched on age of acquisition, visual complexity, familiarity, and word frequency, and had high name agreement. Each stimulus was presented on a computer screen for 5 seconds, one after another with no interstimulus interval, once during each of four sessions (four times in total). Each session proceeded in a different random order. Participants were instructed to name each item as quickly and accurately as possible. Participants were video and audio recorded during the experiment. Video was used to monitor eye fixations and general attention to the task. The mean response latency was 1190 ms.

Semantic similarity structure and dimensionality reduction

Each of the 100 images used in the picture-naming task refer to concepts that have been included in prior semantic feature norming studies (Dilkina & Lambon Ralph, 2013; McRae et al., 2005). From these we constructed binary semantic feature vectors indicating which features were judged to be true of each concept. The feature vectors for each concept were aggregated in a matrix with rows indicating concepts and columns indicating features and entries indicating whether the concept possess that feature. Matrix columns were mean-centered and the similarity between all row vectors was then estimated as the cosine of the angle between them: $\boldsymbol{S_{i,j}} = (x_i^T x_j) / \sqrt{x_i^T x_i \times x_j^T x_j}$. The resulting symmetric cosine similarity matrix is the full-rank semantic similarity structure among concepts. The full-rank structure was then embedded into a 3-dimensional metric space by applying singular value decomposition to the corresponding cosine distance matrix $1 - \boldsymbol{S}$. The low-dimensional embedding, shown in **Figure 1**, accounted for 90% of the variance in the full distance matrix.

ECoG data preprocessing

Preprocessing was performed in MATLAB. Data were downsampled to 100 Hz by averaging measurements within 10 ms boxcars. A prior study applying pattern classification to these data found near identical results when analyzing raw voltages versus voltages referenced to the electrode beneath the galea aponeurotica or to the scalp electrode on the mastoid process contralateral to the side of electrode implantation (Rogers et al., 2021). For this reason, and because the location of the reference electrode varied across patients, we analyzed all voltages without referencing. For each stimulus we retained data for 1000 ms from stimulus onset. While the trial epoch may include the onset of articulation toward the end, the critical results cannot reflect such motor activity since all key phenomena are observed early in the epoch. Baseline correction was not performed. The mean voltage at each electrode for each stimulus was computed across the four repetitions. Voltages from different electrodes were concatenated into a row vector and arranged in a neural data matrix with each row containing the vector of voltages for one stimulus sampled from multiple electrodes over time. We rejected columns and then rows where the marginal mean was and more than five standard deviations from the average marginal mean to censor extreme outliers.

Representational similarity learning (RSL) analysis

RSL decoding models regularized with either grOWL or the L1-norm were fit and analyzed in MATLAB v9.5 (2018b) [44] using the Whole-brain Imaging with Sparse Correlations (WISC) toolbox [45], which implements the approach as a multitask variant of group ordered weighted L1-regularized regression grOWL; [28,46]. Supplementary Materials provide a formal description of the grOWL regularization and its application to neural decoding.

Fitting a decoding model regularized with grOWL requires specification of two hyper-parameters, λ and w, which govern the strength with which the decoding matrix is pressured to be row-sparse (λ) with similar weights across correlated neural features (w). The choice of hyperparameters was determined by ten-fold nested cross-validation for each model fit. The 100 stimuli were randomly divided into ten subsets, each containing five animate and five inanimate items. On each of the ten outer-loop folds, one such set was held out (outer-loop holdout). The remaining nine sets were used to search for good hyperparameter values, using inner-loop cross validation.

On each inner-loop fold, one of the nine remaining sets was held out (inner-loop holdout), and a decoding model was fit to the remaining eight sets using a specified pair of hyper-parameter values. Prediction error (Frobenius norm of true versus predicted coordinates on latent dimensions) for the fitted model was evaluated on the inner-loop holdout set. This procedure was repeated once for each of the nine different inner-loop holdouts, all using the same specified hyperparameter values. We then used mean prediction error across the 9 inner-loop hold outs as the estimated performance of the decoder when fit using the specified hyperparameters. Many different hyperparameter values were evaluated in the inner loop, chosen using the Hyperband procedure [47]. The best-performing values were selected from these and a final inner-loop model was fit using those parameters and all nine data subsets. This final inner-loop model was used to predict coordinates on the latent semantic dimension for the outer-loop holdout set, completing one iteration of the outer-loop cross-validation. The whole procedure was done ten times, using each data subset once as the outer-loop holdout. Across folds, the procedure generated predicted coordinates for all 100 items when those items were completely

583 held out of model fitting. These final predicted, held-out coordinates, computed separately for
584 each time-window in each participant, were the primary data evaluated in the results.

585 **Statistical thresholding with permutation testing.**

586 The correlation metric we computed to assess model accuracy has many useful
587 properties: it standardizes means and variances of the target and predicted vectors to focus on
588 just their covariance. This is especially critical when evaluating animate and inanimate items
589 separately. Because data are mean-centered, however, cross-validated correlation has a negative
590 bias when evaluating held-out items [48], so that a null hypothesis of zero with a t-distributed
591 sampling distribution cannot be assumed. We therefore used permutation testing to generate a
592 group-level null distribution following the prescription of [34] with 10,000 simulated values for
593 each test (i.e., each window, for each subset of items, in all analyses).

594 For every patient, window, and experimental condition, we conducted the full analysis
595 pipeline just described 100 times, randomly permuting the ordering of the rows in the target
596 matrix each time. This yielded 100 correlation values for each patient, representing expected
597 values from our workflow when no reliable relationship exists between neural data and target
598 matrix (due to the permutation procedure). To estimate the group-level null distribution for each
599 window and condition, we randomly sampled one permutation-based correlation value from the
600 100 associated with each patient and averaged these. This was repeated 10,000 times [34]. If m is
601 the number of values in the permutation distribution and b is the number of values in the
602 distribution larger than the true correlation value, then the one-tailed p-value can be computed
603 [49,50] as:

604 $$p = \frac{b+1}{m+1}$$

To control the family-wise error rate, we applied the Westfall-Young procedure [51,52], which involves taking the maximum value over family members for each of permutation and comparing all test statistics in the family to this single distribution of maximal values.

Acknowledgments

Funding: This work was partially supported by MRC Programme grant to MALR and TTR (MR/R023883/1) and an intramural award (MC_UU_00005/18), and by a European Research Council grant GAP: 502670428 - BRAIN2MIND_NEUROCOMP. RM reports grants from MEXT, KAKENHI 22H04777, 22H02945. AS reports grants from MEXT, KAKENHI 22K07537

Author Contributions: Cox conducted all analysis and drafted the manuscript. Rogers and Lambon Ralph contributed to all aspects of the study design, analysis, and manuscript preparation. Shimotake, Kikuchi, Kunieda, Miyamoto, Takahasi, Ikeda and Matsumoto were responsible for ECoG data collection and clinical assessment.

Competing interests: Authors declare no competing interests.

Data and materials availability: All data are available via links in the main text or the Supplementary Information.

Supplementary Information:

S1. Formal description of the RSL approach

S2. Anatomical distribution of decoding model weights across vATL

S3. Reconstructing item similarity from the SVD coordinates

S4. Correlation between predicted and true semantic distances

References

1 Pereira, F., Gershman, S., Ritter, S. & Botvinick, M. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology* **33**, 175-190 (2016). https://doi.org:10.1080/02643294.2016.1176907

2 McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods* **37**, 547-559 (2005). https://doi.org:10.3758/BF03192726

3 Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour* **4**, 1173-1185 (2020). https://doi.org:10.1038/s41562-020-00951-3

4 Rogers, T. T. *et al.* Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation. *Psychological Review* **111**, 205-235 (2004). https://doi.org:10.1037/0033-295X.111.1.205

5 Lambon Ralph, M. A., Lowe, C. & Rogers, T. T. Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain* **130**, 1127-1137 (2007). https://doi.org:10.1093/brain/awm025

6 Patterson, K., Nestor, P. J. & Rogers, T. T. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience* **8**, 976-987 (2007). https://doi.org:10.1038/nrn2277

7 Lambon Ralph, M. A. Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20120392 (2014). https://doi.org:10.1098/rstb.2012.0392

8 Shimotake, A. *et al.* Direct Exploration of the Role of the Ventral Anterior Temporal Lobe in Semantic Memory: Cortical Stimulation and Local Field Potential Evidence From Subdural Grid Electrodes. *Cerebral Cortex* **25**, 3802-3817 (2015). https://doi.org:10.1093/cercor/bhu262

9 Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* **18**, 42 (2017). https://doi.org:10.1038/nrn.2016.150

10 Peelen, M. V. & Caramazza, A. Conceptual object representations in human anterior temporal cortex. *Journal of Neuroscience* **32**, 15728-15736 (2012). https://doi.org:10.1523/JNEUROSCI.1953-12.2012

11 Bruffaerts, R. *et al.* Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *Journal of Neuroscience* **33**, 18597-18607 (2013). https://doi.org:10.1523/JNEUROSCI.1548-13.2013

12 Fairhall, S. L. & Caramazza, A. Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience* **33**, 10552-10558 (2013). https://doi.org:10.1523/JNEUROSCI.0051-13.2013

13 Devereux, B. J., Clarke, A. & Tyler, L. K. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports* **8**, 10636 (2018). https://doi.org:10.1038/s41598-018-28865-1

14 Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L. Y. & Barense, M. D. Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife* **7**, e31873 (2018). https://doi.org:10.7554/eLife.31873

15 Chao, L. L., Haxby, J. V. & Martin, A. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience* **2**, 913-919 (1999). https://doi.org:10.1038/13217

16 Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425-2430 (2001). https://doi.org:10.1126/science.1063736

17 Martin, A. The representation of object concepts in the brain. *Annual Review of Psychology* **58**, 25-45 (2007). https://doi.org:10.1146/annurev.psych.57.102904.190143

18 Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126-1141 (2008). https://doi.org:10.1016/j.neuron.2008.10.043

19 Connolly, A. C. *et al.* The representation of biological classes in the human brain. *Journal of Neuroscience* **32**, 2608-2618 (2012). https://doi.org:10.1523/JNEUROSCI.5547-11.2012

20 Sha, L. *et al.* The animacy continuum in the human ventral vision pathway. *Journal of Cognitive Neuroscience* **27**, 665-678 (2015). https://doi.org:10.1162/jocn_a_00733

21 Martin, A. GRAPES--Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the brain. *Psychonomic Bulletin & Review* **23**, 979 -- 990 (2016). https://doi.org:10.3758/s13423-015-0842-3

22 Carota, F., Kriegeskorte, N., Nili, H. & Pulvermuller, F. Representational Similarity Mapping of Distributional Semantics in Left Inferior Frontal, Middle Temporal, and Motor Cortex. *Cerebral Cortex* **27**, 294-309 (2017). https://doi.org:10.1093/cercor/bhw379

23 Binder, J. R., Desai, R. H., Graves, W. W. & Conant, L. L. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* **19**, 2767-2796 (2009). https://doi.org:10.1093/cercor/bhp055

24 Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends in Cognitive Sciences* **15**, 527-536 (2011). https://doi.org:10.1016/j.tics.2011.10.001

25 Devereux, B. J., Clarke, A., Marouchos, A. & Tyler, L. K. Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *Journal of Neuroscience* **33**, 18906-18916 (2013). https://doi.org:10.1523/jneurosci.3809-13.2013

26 Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453-458 (2016). https://doi.org:10.1038/nature17637

27 Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, 4 (2008). https://doi.org:10.3389/neuro.06.004.2008

28 Oswal, U., Cox, C. R., Lambon Ralph, M. A., Rogers, T. T. & Nowak, R. D. in *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning.* 1041-1049.

29 Dilkina, K. & Lambon Ralph, M. A. Conceptual structure within and between modalities. *Frontiers in Human Neuroscience* **6**, 333 (2013). https://doi.org:10.3389/fnhum.2012.00333

30 Chen, Y. *et al.* The 'when' and 'where' of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. *Cortex* **79**, 1-13 (2016). https://doi.org:10.1016/j.cortex.2016.02.015

31 Rogers, T. T. *et al.* Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. *eLife* **10**, e66276 (2021). https://doi.org:10.7554/eLife.66276

32 Cox, C. R. & Rogers, T. T. Finding Distributed Needles in Neural Haystacks. *Journal of Neuroscience* **41**, 1019-1032 (2021). https://doi.org:10.1523/jneurosci.0904-20.2020

33 Frisby, S., Halai, A. D., Cox, C. R., Lambon Ralph, M. A. & Rogers, T. T. Decoding semantic representations in mind and brain. https://doi.org:10.31234/osf.io/6kjtf

34 Stelzer, J., Chen, Y. & Turner, R. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage* **65**, 69-82 (2013). https://doi.org:10.1016/j.neuroimage.2012.09.063

35 McRae, K., de Sa, V. R. & Seidenberg, M. S. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General* **126**, 99-130 (1997). https://doi.org:10.1037/0096-3445.126.2.99

36 Devlin, J. T., Gonnerman, L. M., Andersen, E. S. & Seidenberg, M. S. Category-Specific Semantic Deficits in Focal and Widespread Brain Damage: A Computational Account. *Journal of Cognitive Neuroscience* **10**, 77-94 (1998). https://doi.org:10.1162/089892998563798

37 Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R. & Levy, J. P. Conceptual Structure and the Structure of Concepts: A Distributed Account of Category-Specific Deficits. *Brain and Language* **75**, 195-231 (2000). https://doi.org:10.1006/brln.2000.2353

38 Garrard, P., Lambon Ralph, M. A., Hodges, J. R. & Patterson, K. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology* **18**, 125-174 (2001). https://doi.org:10.1080/02643290125857

39 Devlin, J. T. *et al.* Is there an anatomical basis for category-specificity? Semantic memory studies in PET and fMRI. *Neuropsychologia* **40**, 54-75 (2002). https://doi.org:10.1016/S0028-3932(01)00066-5

40 Rogers, T. T. *et al.* Anterior temporal cortex and semantic memory: reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience* **6**, 201-213 (2006). https://doi.org:10.3758/CABN.6.3.201

41 Rogers, T. T. & McClelland, J. L. *Semantic cognition: A parallel distributed processing approach*. (MIT press, 2004).

42 Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J. & Binder, J. R. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences* **119**, e2108091119 (2022). https://doi.org:doi:10.1073/pnas.2108091119

43 Chen, L. & Rogers, T. T. Revisiting domain-general accounts of category specificity in mind and brain. *WIREs Cognitive Science* **5**, 327-344 (2014). https://doi.org:10.1002/wcs.1283

44 Matlab v. 9.5 (R2018b) (The MathWorks, Natick, MA, 2018).

45 Whole-brain Imaging with Sparse Correlations (Github, doi:10.5281/zenodo.6975270 2021).

46 Figueiredo, M. & Nowak, R. in *Artificial Intelligence and Statistics.* 930-938.

47 Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. & Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research* **18**, 1-52 (2018).

48 Zhou, Y., Vales, M. I., Wang, A. & Zhang, Z. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. *Briefings in Bioinformatics* **18**, 1093-1093 (2017). https://doi.org:10.1093/bib/bbx133

49 Edgington, E. S. *Randomization tests. [electronic resource]*. 3rd ed., rev. and expanded. edn, (M. Dekker, 1995).

50 Phipson, B. & Smyth, G. K. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology* **9** (2010). https://doi.org:10.2202/1544-6115.1585

51 Alberton, B. A. V., Nichols, T. E., Gamba, H. R. & Winkler, A. M. Multiple testing correction over contrasts for brain imaging. *NeuroImage* **216**, 116760 (2020). https://doi.org:10.1016/j.neuroimage.2020.116760

52 Westfall, P. H., Young, S. S. & Wright, S. P. On Adjusting P-Values for Multiplicity. *Biometrics* **49**, 941-945 (1993). https://doi.org:10.2307/2532216

53 Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49-67 (2006). https://doi.org:10.1111/j.1467-9868.2005.00532.x

Supplementary materials

S1. Formal description of the RSL approach.

Let $\boldsymbol{S}$ be a symmetric $n \times n$ target matrix of approximate rank r describing pairwise cognitive similarities amongst *n* items, which is decomposed via singular-value decomposition (SVD) into an $n \times r$ matrix $\boldsymbol{U}$ of r orthonormal singular vectors and an $r \times r$ diagonal matrix $\boldsymbol{D}$ containing the corresponding r singular values (**Figure 2A**). The matrix product $\boldsymbol{U}\sqrt{\boldsymbol{D}}$ then contains root-weighted singular vectors that can be viewed as encoding the coordinates of each stimulus within an r-dimensional latent semantic space, while the product of this matrix with its transpose provides a rank r approximation of the original similarity matrix. That is:

$$S_{n\times n} \cong U_{n\times r} D_{r\times r} U^T = U\sqrt{D}\,(U\sqrt{D})^T$$

Let $\boldsymbol{X}_{n\times m}$ be a data matrix recording neural responses to each of n stimuli over m neural features (e.g., voxels in fMRI, electrode voltages at each timepoint in ECoG, etc.). To assess whether some weighted combination of these features jointly encode the target similarities, we seek a *decoding matrix* $\boldsymbol{\beta}_{m\times r}$ such that the matrix product $\boldsymbol{X\beta}$ produces *r* coordinates for each item, deviating as little as possible from those in target matrix $\boldsymbol{U}\sqrt{\boldsymbol{D}}$. This search can be formulated as the convex optimization:

$$argmin_{\beta} \left\| U\sqrt{D} - X\beta \right\|_F$$

That is, we seek $\boldsymbol{\beta}$ values that minimize the Froebenius norm (i.e., sum of squared differences) between the true root-weighted singular vectors $\boldsymbol{U}\sqrt{\boldsymbol{D}}$ and the predicted values $\boldsymbol{X\beta}$. The optimization essentially computes a simultaneous regression predicting all *r* coordinates for each stimulus from weighted combinations of the neural responses they evoke. Each column of the resulting decoding matrix contains one coefficient for each neural feature and predicts an

item's coordinate along one component of the target matrix $\boldsymbol{U}\sqrt{\boldsymbol{D}}$ (ie, one dimension of the semantic space).

As noted earlier, there are typically many more neural measurements (predictors) than stimuli (i.e., target data points)—that is, $m \gg n$. Thus, the optimization just defined is under-constrained and must be regularized with an additional loss H, itself a function of the coefficient matrix $\boldsymbol{\beta}$:

$$argmin_{\beta}\left(\left\|U\sqrt{D} - X\beta\right\|_F + \alpha H(\beta)\right)$$

…which is also convex so that a weighted sum of this function and the original reconstruction loss together yield a single unique solution that can be discovered via gradient descent. The constant α is a free weighting parameter governing the importance of the regularization loss relative to the model error.

For RSL we wish to use a regularization function that promotes solutions respecting the sparsity, redundancy, and spanning assumptions described earlier. Oswal et al. (2016) formulated a regularizer that accomplishes these objectives by constraining the structure of the decoding matrix $\boldsymbol{\beta}$ in a single convex loss called the *group ordered weighted LASSO* or *grOWL.* The approach generalizes both the group LASSO [53] and the ordered-weighted LASSO [46] to favor solutions that are (1) row-sparse (rows of $\boldsymbol{\beta}$ are either all zero or all non-zero), implementing both the sparsity and spanning constraints, but (2) with subgroups of selected features sharing similar coefficients if their activations covary strongly together (implementing the redundancy constraint). The grOWL loss is specified as:

$$H(\beta) = \sum_{i=1}^{m} w_i \left\|\beta_{[i]\cdot}\right\|_2$$

Here m is the number of neural features (rows of $\boldsymbol{\beta}$), $\boldsymbol{\beta}_{[i]\cdot}$ is the row of $\boldsymbol{\beta}$ with the i-th largest 2-norm (i.e., the ith longest coefficient vector), and w_i is a vector of non-negative, non-increasing weights (for instance a vector of positive values that decay linearly from an initial constant). Intuitively the loss works as follows. The first element of the weighting vector w is stipulated to be as large or larger than all other elements, and this weight is applied to whichever neural feature has the largest coefficients in the decoding matrix (i.e., the 2-norm or length of the corresponding row). Thus, the feature with the largest coefficients exerts the largest regularization cost. The optimization must therefore put the largest weights on the most informative feature to make it "worth" the high regularization cost. If the most informative feature is also correlated with many other features, their total cost can be further minimized by placing similar weights across all such features, since the cost for each such weight declines with the weight sequence in w. Thus, grOWL spontaneously clusters correlated signal-carrying neural features and gives these a similar weight, effectively averaging them. Finally, the weighting vector ensures that coefficients on features that do not reduce error beyond those already selected will be pressured to zero, ensuring overall sparsity. Formal analysis of the grOWL regularizer and mathematical proofs of its row-sparse and clustering properties were provided by Oswal et al. (2016).

S2. Anatomical distribution of decoding model weights across vATL

The decoding results reported in the main paper demonstrate that multi-dimensional semantic structure can be decoded from LFPs recorded from the vATL, and that decoding the second semantic dimension depends on regularization via the grOWL. This implies that local neural populations do not independently encode each semantic dimension, but simultaneously express multiple orthogonal components of the target semantic matrix. To evaluate whether this

is so, we examined the spatial distribution of weights across electrodes in the decoding models predicting variation on the first via the second semantic dimensions for opening windows of size 200, 400, 600, 800, and 1000 ms in the eight patients with electrodes implanted in the left hemisphere. To obtain a single number at each electrode for each model (each component is associated with a distinct model), the weight magnitudes (absolute values) were summed over time. Because the number of electrodes and their placement varied across subjects, we projected these sums onto a standard MNI cortical surface and smoothed along the surface with a 4 mm Gaussian kernel before averaging across subjects. Points on the surface that are associated with data from fewer than six subjects are excluded from analysis. To determine which sums are larger than expected by chance, the data for each subject within each window was fit to 1,000 random row-wise permutations of the target structure. Each of these solutions were projected onto the MNI surface as described above, resulting in 1,000 surface maps per subject. Then we sampled a map from each subject to simulate one “experiment”, and the data were summed and averaged as above. This was repeated 10,000 times to obtain a group-level null distribution [34].

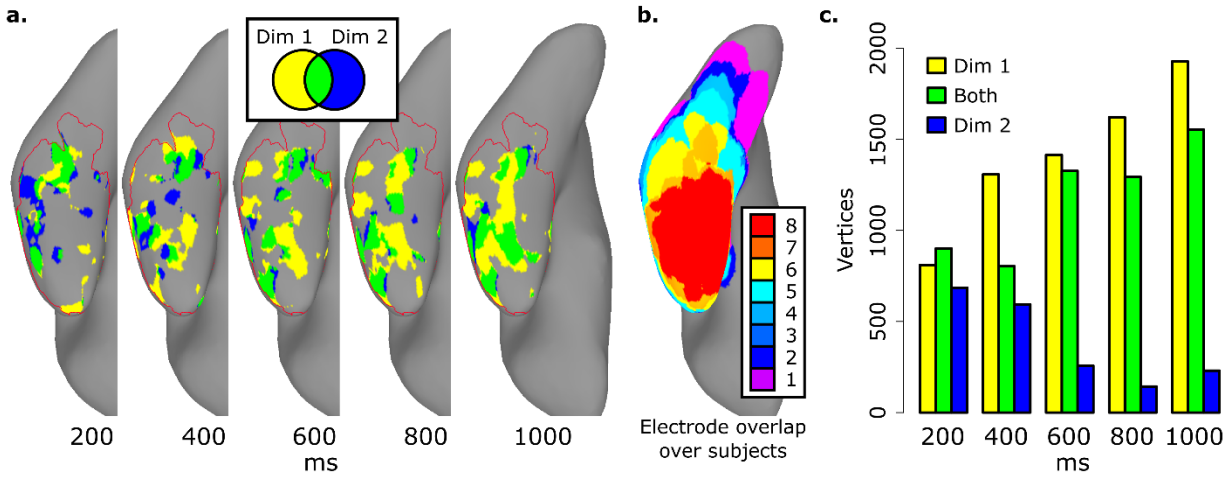

Figure S1. Significant vertices on the standard MNI N27 surface reconstruction displayed with AFNI SUMA for eight patients with left lateralized language (confirmed with the WADA test). a) RSL model coefficients (multi-task linear regression with grOWL regularization) associated with points in MNI space were projected onto and smoothed along the surface with a 4 mm Gaussian kernel, separately for dimensions 1 and 2. This was repeated for 10,000 group-level maps simulating the null distribution at each vertex by permutation. Only points on the surface where data from 6 or more patients is available are considered; vertices meeting this criterion are contained within the red contour. Vertices with average weight magnitude larger than 9,500 of the 10,000 group-level values in the permutation distribution for that vertex are shown: points that meet these criteria for dimension 1 are shown in yellow, for dimension 2 are shown in blue, and for both are shown in green. b) Number of electrodes contributing to the data at each vertex after applying a 4 mm Gaussian kernel. c) Tabulation of significant vertices.

Figure S1a displays the points on the cortical surface where the real average is larger than 9,500 of the values in this null distribution. Points on the surface that are associated with large model weights for the first component (but not the second) are shaded yellow; points that are associated with the second component (but not the first) are shown in blue. Points on the surface that are associated with large weights for both components in the same model are shown with green. Substantial overlap is observed in all windows, with overlap increasing with window length (**Figure S1c**). Furthermore, both semantic dimensions are encoded across the vATL with little functional localization.

S3. Reconstructing item similarity from SVD coordinates

The main analysis uses the first three root-weighted singular vectors as the target for fitting decoding models with RSL. Predicted coordinates are then used to generate predicted semantic distances between pairs of held-out stimuli. To understand the possible causes of

domain differences in prediction accuracy, we considered how well pairwise distances can be reconstructed from the first three singular vectors and values computed from the true matrix with a subset of items excluded. The goal was to assess whether the first three components capture inter-items distances equally well for both domains or whether such distances are better approximated for one domain than another.

To this end we conducted the following analysis. We generated 10,000 random splits of the semantic norming data, each consisting of 90 training items and ten test items (five animate and five inanimate) as in the ECoG study. For each, we computed the SVD on the semantic distance matrix for only the items in the training set. We then projected test-set items into the resulting space by matrix multiplying each item's similarity with the training set items:

$$\widehat{\boldsymbol{C}}_{\boldsymbol{q}\times\boldsymbol{r}} = \boldsymbol{S}'_{\boldsymbol{q}\times\boldsymbol{p}}\boldsymbol{U}_{\boldsymbol{p}\times\boldsymbol{r}}\boldsymbol{D}_{\boldsymbol{r}\times\boldsymbol{r}}; \boldsymbol{S}' = \left(\boldsymbol{S}_{\boldsymbol{ij}}\right)_{\substack{i\in test \\ j\in training}}$$

Here $q = 10$ is the number of test set items, $p = 90$ is the number of training-set items, $r = 3$ is the rank of the embedding, $\widehat{\boldsymbol{C}}_{\boldsymbol{q}\times\boldsymbol{r}}$ is a matrix containing estimates coordinates for the test items, $\boldsymbol{S}'$ is a matrix containing the true distances from test items to training items, and $\boldsymbol{U}$ and $\boldsymbol{D}$ are matrices containing the singular vectors and values, respectively, computed the training items. From the estimates coordinates in $\widehat{\boldsymbol{C}}_{\boldsymbol{q}\times\boldsymbol{r}}$, the predicted similarity matrix can be obtained as $\widehat{\boldsymbol{S}}_{\boldsymbol{q}\times\boldsymbol{q}} = \widehat{\boldsymbol{C}}\boldsymbol{D}\widehat{\boldsymbol{C}}^{\boldsymbol{T}}$. To measure how well the true distances are captures by the SVD, we computed correlations of lower-triangular values between $\boldsymbol{S}$ and $\widehat{\boldsymbol{S}}$, either over all values at once or correlating the animate-animate, inanimate-inanimate, and animate-inanimate similarity values separately.

Figure S2 shows the average correlations for each selection of similarities over the 10,000 splits as bars with error bars reflecting the standard deviation over splits. The mean over the 10 cross-validation splits used in the ECoG study are overlaid as points. Predicted similarities

that include or isolate between-category structure are most correlated with the true similarity structure. Predicted within-category similarities are less correlated, with predictions for inanimate items being less accurate than for animate items. In other words, a low-rank approximation of the semantic similarity matrix best captures between-domain distances and provides a better approximation of within-animate than within-inanimate semantic structure. Thus, decoding models face an intrinsic disadvantage at capturing within-domain structure generally, and within-inanimate structure specifically.

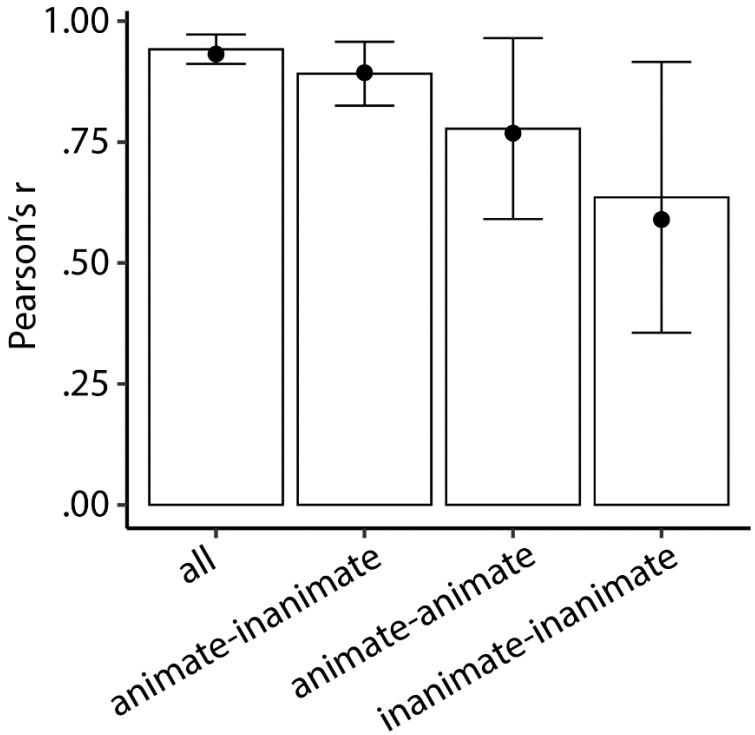

Figure S2. Correlation between true and reconstructed similarity matrices. Reconstructed similarity matrices are based on a SVD of a subset of the full similarity matrix expressing the pairwise similarities for 90 items, which is then used to reconstruct the similarities for the remaining 10 items as described in the text. This process was repeated 10,000 times with different 90/10 splits and each reconstruction was correlated with the corresponding true pairwise similarities. Bars show the average correlations over all similarities, just the between-category similarities (animate-inanimate), and just the within-category similarities (animate-animate and inanimate-inanimate). Error bars report the standard deviation of the means. The results for the specific 90/10 split used in the analysis of ECoG data reported in the main document are shown as black dots.

S4. Correlation between predicted and full-rank semantic distances

In evaluating how well the RSL decoding models predict pairwise semantic similarities, the main paper reports the correlation between predicted similarities and similarities reconstructed directly from the three singular vectors computed from the "true" full-rank matrix. We will refer to this as the *low-rank* similarity matrix. The low-rank similarity matrix represents the best possible prediction, given that only three components of the true matrix were used to fit the decoding model. Here we report correlations with the *full-rank* similarity matrix—the actual pairwise similarities amongst feature vectors as described in the main text. The low-rank and full-rank similarity matrices are very similar, $r = .94$, such that the low-rank similarity matrix accounts for ~90% of the variance in the full-rank matrix. As one would expect given the high correlation, **Figure S3** and **Figure 5** from the main document show the same pattern of results. Filled circles indicate significance at Westfall-Young FWER corrected $\alpha = .05$.

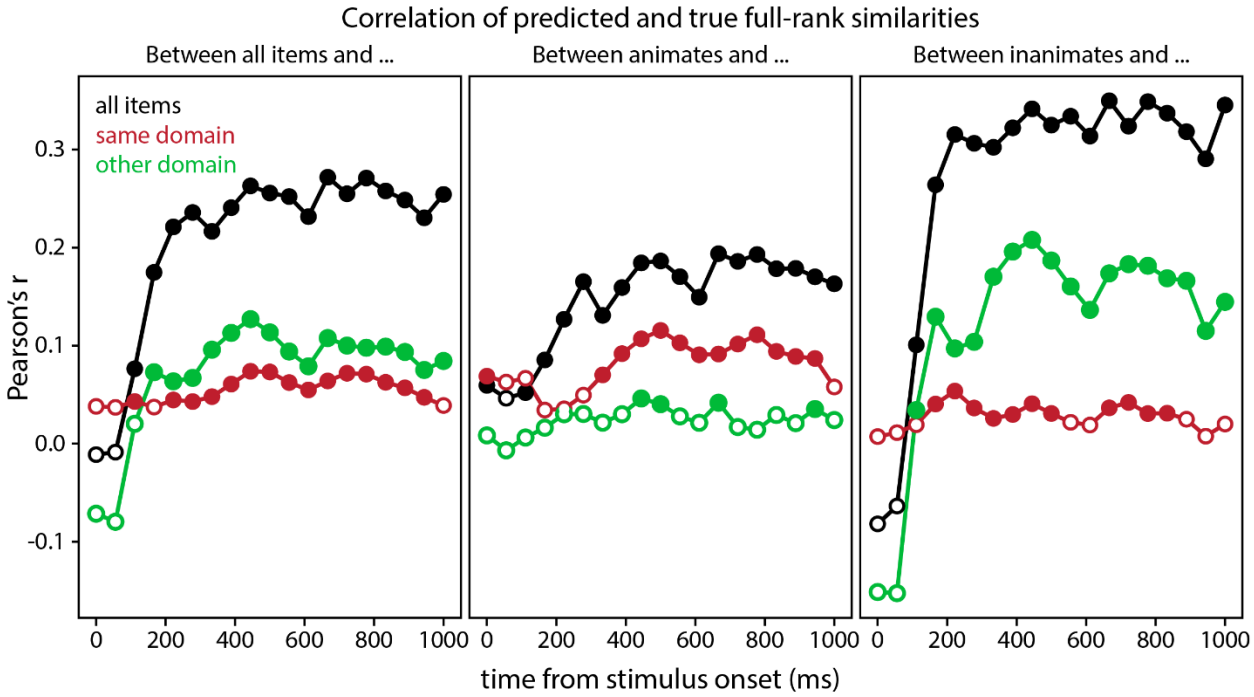

Figure S3 Correlation between predicted semantic similarities and *full-rank* cosine similarities as computed from the feature vectors. These results are very similar to **Figure 5** in the main document which involves the *low-rank* similarity matrix, as reconstructed from the three semantic dimensions modeled in the primary analysis. The results are so similar because the first

three components account for ~90% of the variance in the full-rank cosine similarity matrix. The interpretation is the same as **Figure 5**.