Comparative multivariate decoding adjudicates theories of semantic representation in the anterior temporal lobes and the rest of the cortex

Saskia L. Frisby[1*], Christopher R. Cox[2], Ajay D. Halai[1], Matthew A. Lambon Ralph[1], & Timothy T. Rogers[3*]

[1] MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK

[2] Department of Psychology, Louisiana State University, Baton Rouge, USA

[3] Department of Psychology, University of Wisconsin–Madison, Madison, USA

* Correspondence: saskia.frisby@mrc-cbu.cam.ac.uk

** Correspondence: ttrogers@wisc.edu

Keywords: semantic representation, 7T-fMRI, decoding, multivariate pattern analysis

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript arising from this work.

Abstract: The anterior temporal lobes (ATLs) are known to support semantic cognition, but theories about their precise representational coding vary. We collected 7T-fMRI data with a novel acquisition sequence designed to improve signal quality in the ATLs, then employed a pioneering analytical approach (comparative multivariate decoding) to adjudicate between theories. Specifically, we applied multiple decoding methods, each making different assumptions about the content, nature or location of representations within the ATLs, then used the pattern of results *across* methods to adjudicate competing hypotheses. The results suggest that the ATLs represent domain-general semantic information via a multidimensional vector-space code that is anatomically clustered within and across individuals, and that posterior temporal and occipitotemporal regions utilize a similar domain-general, vector-space code. More generally, the comparative multivariate analytical framework utilized here has the potential to reveal how the brain represents, not just semantic knowledge, but any kind of information.

1. Introduction

Humans have a remarkable ability to understand the world around them – to recognize objects and make inferences about their properties, to comprehend and produce language, and to predict and influence events. These capacities, collectively known as *semantic cognition*, rely on neural systems that remain poorly understood, with controversy surrounding two general questions. The first concerns the role of the anterior temporal lobes (ATLs) – regions known to support semantic cognition, but which are the subject of differing views about their function and the representations they subserve. The second concerns regions beyond the ATLs – which ones are also important for semantic cognition, what information they encode, and whether their functions differ from the ATL. These divergent views may arise because functional neuroimaging studies of semantics apply different analytic techniques, each adopting different assumptions about the neural code that limit the kinds of representation it can and cannot detect [1]. Because most studies adopt a single kind of analysis, each is unable to adjudicate between theories of how neural activity encodes semantic information.

The current paper instead applies a *comparative multivariate approach* to evaluate competing theories of semantic representation in the brain. The key idea is to apply multiple different decoding models to the same set of data, then to assess which hypotheses are consistent with the full pattern of results. Theories concerning the ATL in semantic representation make differing predictions about three aspects of representation:

1. Domain of representation ("what"). Building on evidence of category-specific patterns of semantic impairment and functional activation [2,3], some hypotheses propose that semantic representations in the ATL are *animal-specific* [4,5] whereas other regions support knowledge about manmade objects [6], social and emotional concepts [7], or other categories [8]. An alternative perspective proposes that seeming category-specific patterns arise mainly from confounding factors [2,9,10], and that, once such confounds are controlled or accounted for, the ATLs can be seen to support semantic knowledge for all conceptual domains [11,12].

2. Nature of representation ("how"). The literature offers three different proposals about the nature of semantic representations instantiated within the ATLs:

2A. Pointers. The ATLs may contain "pointers" (within distinct "convergence zones") for different concepts, each binding together its associated features encoded elsewhere in the cortex [13]. For

instance, the *robin* pointer may connect properties such as *has feathers, has a red breast, can fly* etc. Under this hypothesis, perception of the word "robin" activates the *robin* pointer in the ATL, which then activates the robin's shape, pattern of motion, typical color, other verbal descriptors, etc. in various remote cortical regions. Since each concept pointer is independent of every other, on this view it should not be possible to decode graded semantic similarity relations between items from the ATL pointers.

2B. Semantic features. Some theories assume that different neural populations across the cortex (including in the ATL) may each *independently* encode a distinct semantic feature dimension, such as animacy (animate/inanimate), size (large/small), movement (can fly/can't fly), and so on [14,15]. Semantic features capture graded conceptual similarity relations [16,17], so, on this view, ATL activity patterns should express such similarities within the domain represented. Because different populations encode distinct semantic dimensions, however, decoding performance should be best when *different* models are fit to decode each dimension.

2C. Semantic vector spaces. Neural populations within the ATL may *conjointly* encode a graded semantic representation space. On this view, ATL activity patterns express graded conceptual similarities between stimuli within the domain of representation, but the information encoded by each population is not straightforwardly interpretable. In other words, each dimension is encoded by patterns over multiple populations, and each population partially encodes information about multiple different semantic dimensions [18–27]. Thus decoding models should perform better when constrained to decode multiple semantic dimensions from the *same* neural populations.

3. Location of the representation ("where"). Traditionally, functional brain imaging assumes that the neural populations representing semantic information are contiguous within individual brains and that the semantic system is organized similarly across individuals. By contrast, recent applications of multivariate decoding suggest that semantic representations can be anatomically dispersed within and across individuals [15,28]. Thus a third aspect concerns the localization of representations in the ATL. If representations are anatomically clustered (by which we mean that representations are at least loosely contiguous within individual brains and located at least somewhat similarly across individuals), then decoding models constrained to prefer anatomically clustered representations should outperform models that ignore anatomical contiguity and similarity.

The combination of these three aspects leads to 12 distinct hypotheses about the nature and location of semantic representations in the ATL, listed in Table 1. Critically for comparative

multivariate decoding, one can design a family of decoding analyses across which each hypothesis predicts a *different pattern* of relative performance. For instance, pointer-based representations predict that it should not be possible to decode graded similarities between items within or outside the domain of representation, while animal-specific hypotheses predict that the ATLs only support knowledge about animals. If ATL representations are both pointer-based and animal-specific, a binary classifier should succeed at discriminating animate from inanimate stimuli (since only animals activate a pointer), but a regression model predicting between-animal semantic distances should fail (since pointers do not encode graded semantic similarities). In contrast, a domain-general pointer representation should *fail* for both decoding goals; an animal-specific feature-based representation should achieve both goals but fail at decoding between-item similarities for non-animals; and a domain-general feature-based representation should achieve all three goals (binary classification, within-animate similarities, within-inanimate similarities). Following this logic, the 12 hypotheses can be distinguished by comparing the performance of decoding models that vary three independent factors:

1. *The decoding target.* Models can be fit to categorize stimuli as animate/inanimate or to predict graded similarities among stimuli on multiple semantic dimensions within and across animate/inanimate domains. For classification we can use regularized logistic regression. For predicting graded similarities representational similarity learning (RSL), an extension of regularized linear regression, can be used to predict target coordinates on multiple dimensions simultaneously [29,30].

2. *Regularization across dimensions.* When decoding graded similarities, we can fit models that predict each semantic dimension independently (regularized with a standard loss such as LASSO, also known as the L1 norm), or that prefer solutions in which the same neural populations are used to predict all dimensions simultaneously (regularized with group ordered weighted LASSO or grOWL; [29,30].

3. Regularization across space. For binary classification, we can fit models that ignore the anatomical location of features, or that prefer solutions that are anatomically clustered within and across participants (sparse-overlapping-sets LASSO or SOS-LASSO[31]).

Table 1 indicates how each of the 12 hypotheses predicts a different pattern of comparative outcomes across these analyses:

Hyp.	**Domain**	**Rep. type**	**Location**	**ATL function**	**Animacy class**	**Animal similarity**	**Inanimate similarity**	**grOWL better**	**SOS-LASSO better**
1	Animal	Pointer	Same	Bind features for animals	Yes	No	No	No	Yes
2	Animal	Features	Same	Encode features of animals	Yes	Yes	No	No	Yes
3	Animal	Space	Same	Represent similarities among animals	Yes	Yes	No	Yes	Yes

4	Animal	Pointer	Variable	Bind features for animals	Yes	No	No	No	No
5	Animal	Features	Variable	Encode features of animals	Yes	Yes	No	No	No
6	Animal	Space	Variable	Represent similarities among animals	Yes	Yes	No	Yes	No
7	All	Pointer	Same	Bind features for all concepts	No	No	No	No	Yes
8	All	Features	Same	Encode features for all concepts	Yes	Yes	Yes	No	Yes
9	All	Space	Same	Represent similarities among all concepts	Yes	Yes	Yes	Yes	Yes
10	All	Pointer	Variable	Bind features for all concepts	No	No	No	No	No
11	All	Features	Variable	Encode features for all concepts	Yes	Yes	Yes	No	No
12	All	Space	Variable	Represent similarities among all concepts	Yes	Yes	Yes	Yes	No

Table 1. Pattern of results predicted by each of 12 hypotheses across five comparative decoding analyses. Hyp.: Hypothesis, Rep. type: representation type. **Animacy class:** Can classifiers successfully predict stimulus animacy? **Animal similarity:** Can RSL models successfully predict multidimensional semantic similarities between animate items? **Inanimate similarity:** Can RSL models successfully predict multidimensional semantic similarities among inanimate items? **grOWL better:** Do RSL models predicting graded semantic structure perform better with grOWL regularization (which prefers solutions that predict multiple target dimensions from the same neural features) than with LASSO regularization (which does not have this preference)? **SOS-LASSO better:** Do classifiers perform better with SOS-LASSO regularization (which encourages solutions where predictive features are anatomically clustered) than with LASSO regularization (which does not have this preference)?

Outside the ATL, the same comparative approach can be used to decode whole-cortex neural patterns, yielding information about where and how other cortical systems encode semantic information. For instance, by comparing classification to RSL we can assess whether any brain areas encode binary animacy information without also encoding semantic similarity structure; by comparing RSL regularized with LASSO vs. grOWL, we can assess whether distinct semantic dimensions are encoded by different or overlapping neural populations outside the ATL; and by contrasting results of individual vs. SOS-LASSO we can assess whether semantic information is anatomically clustered.

We applied this comparative approach to new data collected using 7T-fMRI with a scanning protocol designed to improve ATL signal quality [32]. We focused first on adjudicating hypotheses of ATL function, then considered implications for semantic representation in the rest of the cortex. The results illuminate the neural bases of semantic representation and suggest a new approach for resolving conflicting patterns of results in cognitive neuroscience more broadly.

2. Results

Data and approach. We provide an overview of our approach here – see Methods for full details. 7T-fMRI data were collected from 27 participants as they named 100 line drawings of familiar items, half animate and half inanimate. We fit a first-level (within-participant) general linear model (GLM) to each participant's data and extracted one whole-brain beta image per item. To adjudicate

hypotheses about ATL function, we constructed a region of interest (ROI) based on a previous electrocorticography (ECoG) study of semantic processing in the ATL. This ROI encompassed the ATL and extended into posterior temporal regions, so, for the avoidance of ambiguity, we refer to this as the ventral temporal lobe (vTL) ROI. For greater anatomical specificity, we bisected the ROI into anterior and posterior sections and analyzed the halves independently as well as the whole. For the whole-cortex analysis, we applied a cortical grey matter mask. Data from all voxels within the ROI or mask were provided as input to the decoding models.

We assessed binary classification (animate/inanimate) using logistic regression models fit using LASSO regularization (which prefers sparse solutions, i.e. only a few of the available features encode the target information) or using SOS-LASSO (which prefers solutions that are both sparse and anatomically clustered within and across participants).

Next, we used RSL to predict each item's coordinates on three orthogonal semantic dimensions derived from a set of semantic feature norms [29,33] and plotted in Figure 1. The first component separates animate from inanimate items and also somewhat differentiates animal subclasses. Significant decoding of this dimension is compatible both with a neural code representing independent features that systematically covary with animacy (e.g., *can move*, *has eyes*, etc.) or a neural code that represents a direction within a semantic vector space. The second dimension differentiates subclasses of animals, with mammals on one end and birds/invertebrates on the other; it is compatible with independent features that discriminate mammals from other animals (e.g., *has fur*, *can fly*, etc.) or a direction orthogonal to animacy within a semantic vector space. The third dimension largely differentiates manmade objects, with vehicles and instruments on one end, clothing on the other, and various miscellaneous items in between. It is compatible with independent features that differentiate kinds of artifacts (e.g., *is hard* vs. *is soft*) or as a third orthogonal direction within a semantic vector space.

RSL models were fit using LASSO regularization (which assumes that the neural code is sparse and is indifferent to whether features encode one or multiple dimensions) or using grOWL regularization (which additionally assumes that important features encode multiple dimensions simultaneously).

In all cases models were fit and evaluated using ten-fold nested cross-validation with an inner loop used to find optimal hyperparameters and with the resulting model then assessed on an outer-loop hold-out set. For classification, performance was assessed as the mean classification accuracy over hold-out folds for classification models. For RSL, the Pearson correlation between predicted and true coordinates for held-out items was calculated for each hold-out fold and the mean of these correlations was used as the performance metric. To test for significance, the mean

accuracy or correlation across participants was evaluated against an expected null distribution computed via bootstrapped permutations (see Methods; Stelzer et al., 2013).

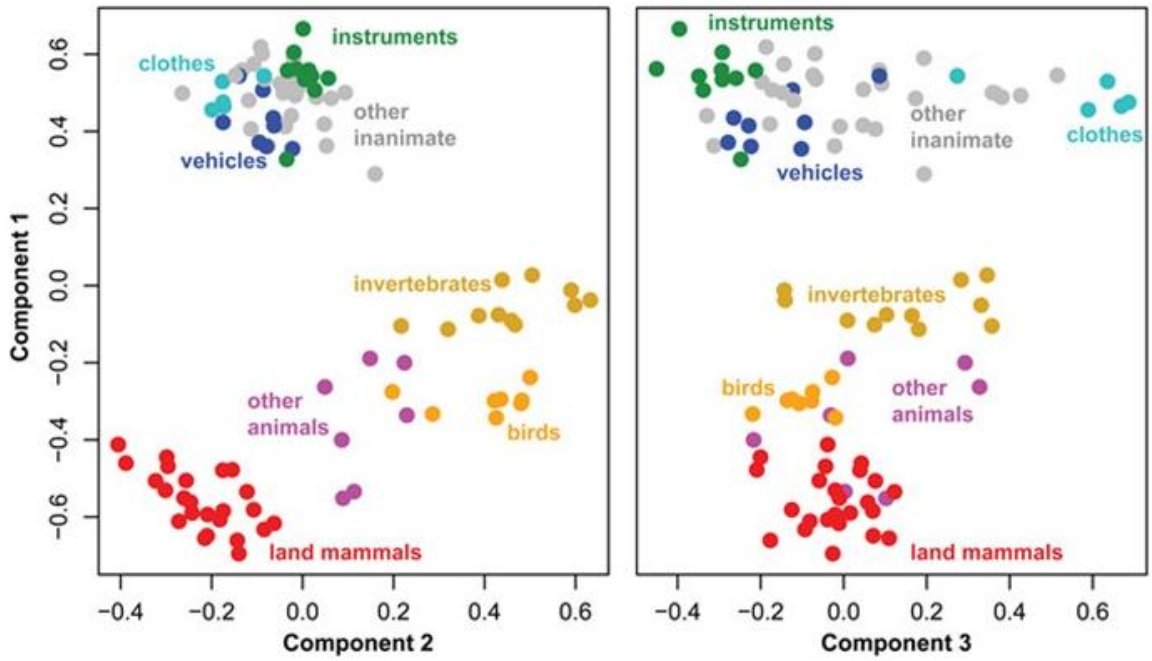

Figure 1: Coordinates of all stimuli on each target semantic dimension. Colors indicate category membership – land mammals (red), birds (orange), invertebrates (mustard), other animals (magenta), instruments (green), clothes (light blue), vehicles (dark blue), and other inanimate objects (grey). Reprinted with permission from Cox et al. [29].

<u>ATL analyses.</u> Regions of interest are shown in Figure 2 and results for binary classification are shown in Figure 2A. All models showed above-chance classification on held-out items, indicating that animacy information is encoded in both hemispheres, and in both anterior and posterior extents of the vTL — a pattern inconsistent with domain-general pointer representations. Classification with SOS-LASSO was reliably better than with LASSO for all models except right anterior vTL, suggesting that the regions encoding animacy information within vTL are anatomically clustered.

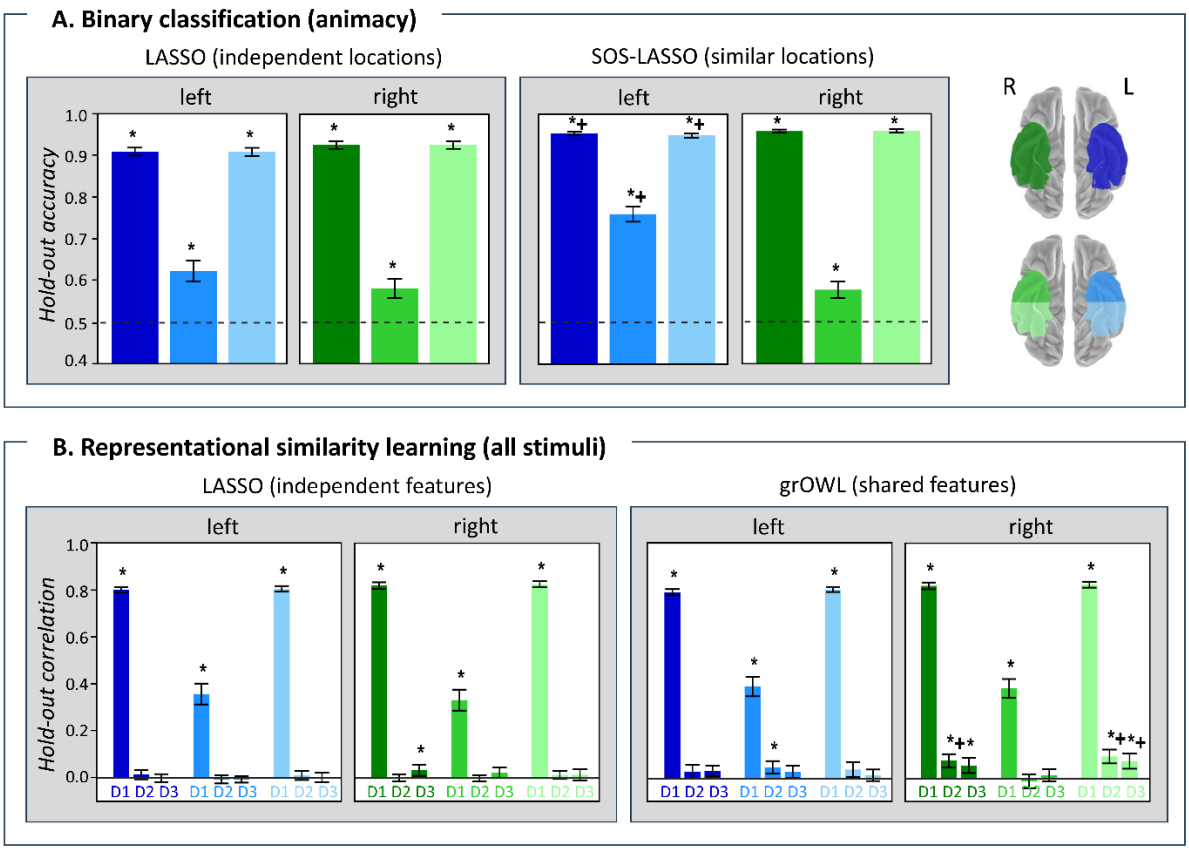

Figure 2: Decoding results in regions of interest. Surface plots (top right) indicate which bar corresponds to which ROI or half-ROI. (A) Mean and 95% confidence interval of the hold-out accuracy for regularized logistic regression classifiers trained on beta values to discriminate animate from inanimate items. Classifiers were trained using either LASSO or SOS-LASSO regularization separately for each ROI. Stars (*) indicate accuracy significantly better than expected under the null hypothesis (as determined by permutation test, probabilities adjusted to control the false-discovery rate at α = 0.05). Crosses (+) indicate a significant difference between classifier accuracy using that regularization penalty and accuracy in the same ROI using the other regularization penalty (probabilities adjusted to control the false-discovery rate at α = 0.05). (B) Mean and 95% confidence interval of the hold-out correlation for RSL models trained on beta values to predict the coordinates of held-out stimuli on three target semantic dimensions (dimension 1, D1; dimension 2, D2; and dimension 3, D3). Models were trained separately for each ROI using either LASSO or grOWL regularization fit separately to each dimension or grOWL regularization fit to all three dimensions simultaneously. Stars (*) indicate a significant correlation between predicted and target coordinates (as determined by permutation test, probabilities adjusted to control the false-discovery rate at α = 0.05). Crosses (+) indicate a significant difference between correlation on that dimension using that regularization penalty and correlation on the same dimension in the same ROI using the other regularization penalty (probabilities adjusted to control the false-discovery rate at α = 0.05).

For RSL, Figure 2B shows the mean hold-out correlation between predicted and true coordinates on each target dimension for each model, calculated across all stimuli. All models showed very good decoding of the first target dimension in both hemispheres and all ROIs. Since the first dimension separates animate and inanimate items, this is consistent with the binary

classification result. Models fit with both regularization penalties produced significant correlations on at least one other dimension that, though small relative to the first component, were significantly higher than expected by chance (dimension 3 for LASSO, and both dimensions 2 and 3 for grOWL). When fit using all vTL voxels, models regularized with grOWL showed reliable decoding of all three semantic dimensions from right-hemisphere activation, and in many cases the correlations were greater than those obtained with LASSO, which suggests that features encode multiple dimensions simultaneously.

Figure 3 shows predicted/true correlations calculated separately for animate and inanimate items in order to test for decoding of within-domain structure. For animate items, models fit with LASSO only performed above chance for dimension 1 in all ROIs. In contrast, models fit with grOWL reliably decoded both the first and second dimensions in both hemispheres, in the posterior halves of each, and in the anterior half of the left hemisphere – a result that is incompatible with representations that are both pointer-based and animal-specific.

For inanimate items, LASSO showed no reliable decoding in the left hemisphere and a somewhat mixed pattern in the right. In contrast, grOWL showed reliable decoding of dimension 3 in all three left-hemisphere ROIs, and of dimensions 1 and 3 in the right hemisphere for both the posterior half and the whole-vTL ROIs. This result indicates that representations within the vTL are not animal-specific. Again, the larger correlations obtained with grOWL suggest that features encode not one but multiple dimensions.

Note that, as shown in Figure 1, dimension 2 differentiates animals but not inanimate items, while the reverse is true for dimension 3. Accordingly, the decoding models generally showed reliable correlations within-domain on dimension 2 for animate items and on dimension 3 for inanimate items.

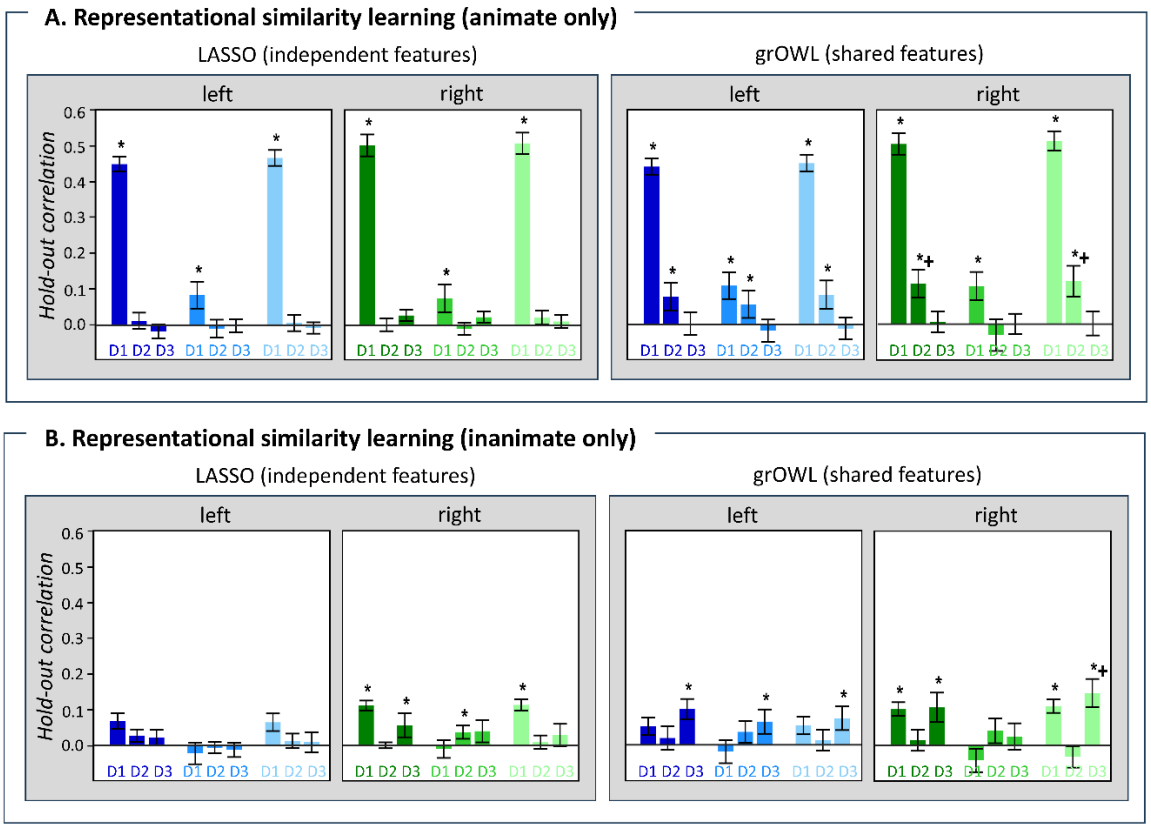

Figure 3: Hold-out correlations within-domain in regions of interest. Mean and 95% confidence interval of the hold-out correlation for RSL models trained on beta values to predict the coordinates of held-out stimuli on three target semantic dimensions (dimension 1, D1; dimension 2, D2; and dimension 3, D3). Models were trained using either LASSO or grOWL regularization separately for each ROI. Stars (*) indicate a significant correlation between predicted and target coordinates (as determined by permutation test, probabilities adjusted to control the false-discovery rate at α = 0.05). Crosses (+) indicate a significant difference between correlation on that dimension using that regularization penalty and correlation on the same dimension in the same ROI using the other regularization penalty (probabilities adjusted to control the false-discovery rate at α = 0.05). (A) Hold-out correlations for animate items only; (B) Hold-out correlations for inanimate items only.

Key observations. In summary, the reliable decoding of graded within-domain semantic structure for both animate and inanimate concepts is inconsistent with theories proposing that the ATL represents semantic information *only* within the animal domain. This graded result is also incompatible with all pointer-based theories. RSL models that preferred to predict all three dimensions from the same voxel subsets (grOWL regularization) often performed better than those with no such preference (LASSO), including the decoding of dimension 2 within animates and dimension 3 within inanimates from right-hemisphere activation. Notably, models regularized with LASSO never showed reliably better decoding than those regularized with grOWL. These findings are inconsistent with the view that separate neural populations encode independent semantic features, but consistent with the alternative view that the ATL encodes a semantic vector space in which each neural population

contributes to the representation of multiple semantic dimensions. Finally, regularization with SOS-LASSO significantly improved binary classification in almost all ROIs, suggesting that the location of animacy information within the vATL is relatively contiguous within and across individuals. Together, the only hypothesis consistent with the full pattern of results is that the ATLs use anatomically clustered neural populations to encode a domain-general semantic vector space (Hypothesis 9 in Table 1).

Whole-cortex analyses. Figure 4A shows the binary classification results. Accuracy was extremely high for classifiers fit with LASSO and SOS-LASSO (hold-out accuracy ≅ 0.95). Accuracy did not differ reliably between these approaches, possibly due to ceiling effects.

Figure 4B shows results for RSL. Models fit with LASSO and grOWL showed very good decoding of the first target dimension ($r \cong 0.8$), but only models fit with grOWL showed reliable decoding of the second and third dimensions. Note that both LASSO and grOWL prefer sparse solutions, so neither is likely to have discovered the entire semantic code. Nevertheless, this result suggests that, across the cortex, the features that best predict multidimensional semantic structure do so via a vector-space code in which neural populations conjointly of multiple semantic dimensions.

Figures 4C and 4D show hold-out correlations calculated separately for animate vs. inanimate items. Both approaches achieved reliable decoding of within-domain structure on dimensions 1 and 2 for animate items, and on dimensions 1 and 3 for inanimate items. Within-domain decoding was, however, reliably better for grOWL than LASSO on dimension 2 for animates and dimension 3 for inanimates, again suggesting a conjoint, vector-space semantic code.

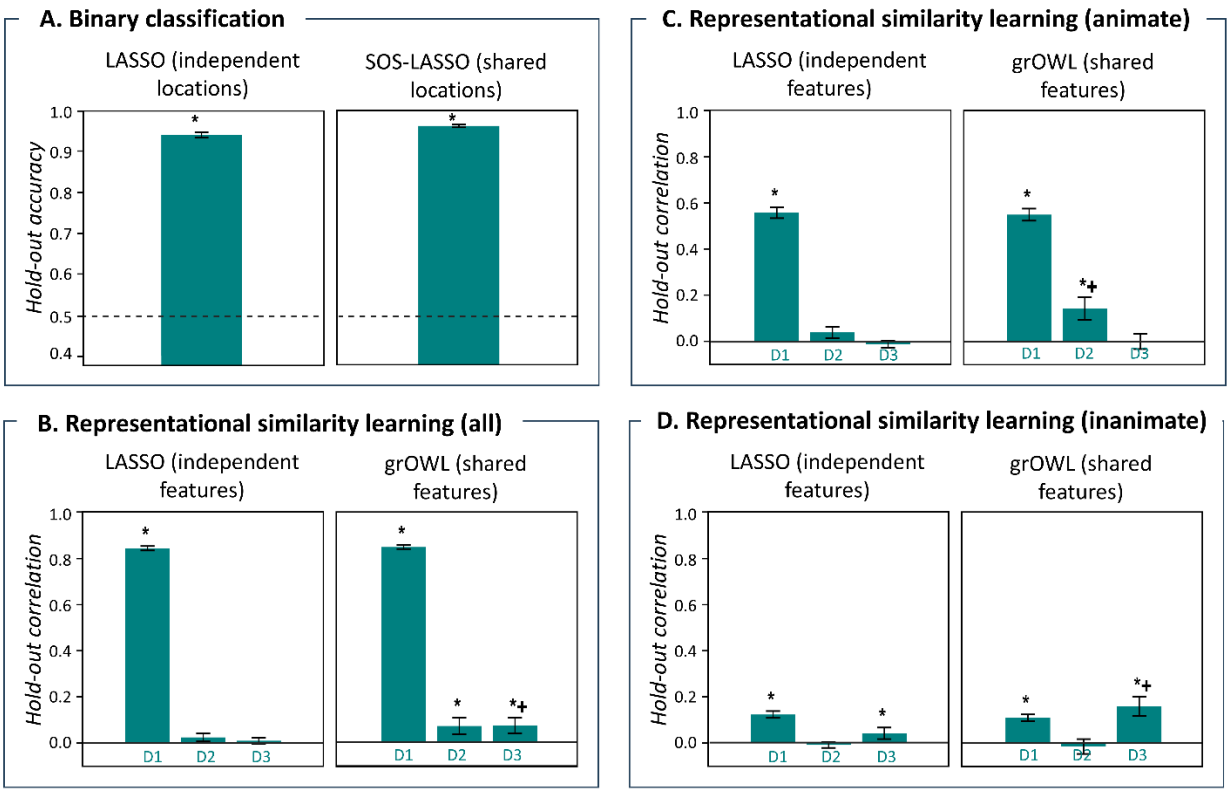

Figure 4: Decoding results at the whole-brain level. (A) Mean and 95% confidence interval of the hold-out accuracy for regularized logistic regression classifiers trained on beta values to discriminate animate from inanimate items. Classifiers are trained using either LASSO or SOS-LASSO regularization. Stars (*) indicate accuracy significantly better than expected under the null hypothesis (as determined by permutation test, probabilities adjusted to control the false-discovery rate at α = 0.05). (B) Mean and 95% confidence interval of the hold-out correlation for RSL models trained on beta values to predict the coordinates of held-out stimuli on three target semantic dimensions (dimension 1, D1; dimension 2, D2; and dimension 3, D3). Models are trained using either LASSO or grOWL regularization. Stars (*) indicate a significant correlation between predicted and target coordinates (as determined by permutation test, probabilities adjusted to control the false-discovery rate at α = 0.05). Crosses (+) indicate a significant difference between correlation on that dimension using that regularization penalty and correlation on the same dimension using the other regularization penalty (probabilities adjusted to control the false-discovery rate at α = 0.05). (C) Hold-out correlation for animate items only; (D) hold-out correlation for inanimate items only.

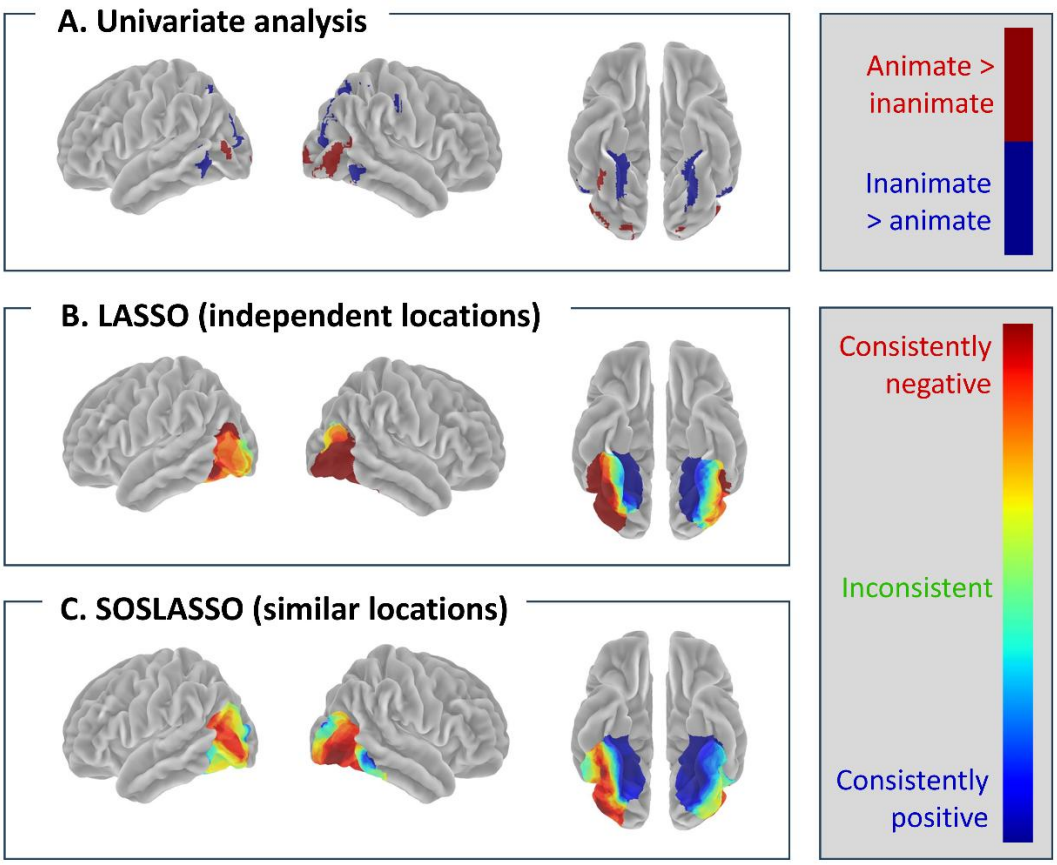

Figure 5: Coefficient maps for classification. Values are shown projected to the surface (fsaverage template). (A) Univariate effects (A>I, red; I>A, blue) are shown projected to the cortical surface and are cluster-corrected at $p < 0.001$. (B) Coefficients for logistic regression classifiers trained with LASSO regularization on beta values to discriminate animate from inanimate items. Colored areas show voxels selected more often than expected from a null distribution computed via permutation testing and controlling the false discovery rate at $\alpha = 0.05$ (see Methods). Colors indicate the proportion of coefficients that are positive across participants, with colored labels indicating what an increase in activation signifies for a given decoder. For classification, cool colors indicate areas where elevated activation systematically signifies inanimate items across participants; warm colors indicate the reverse; and green indicate regions that contribute significantly to prediction, but in different directions across participants. (B) Coefficients for logistic regression classifiers trained with SOS-LASSO regularization.

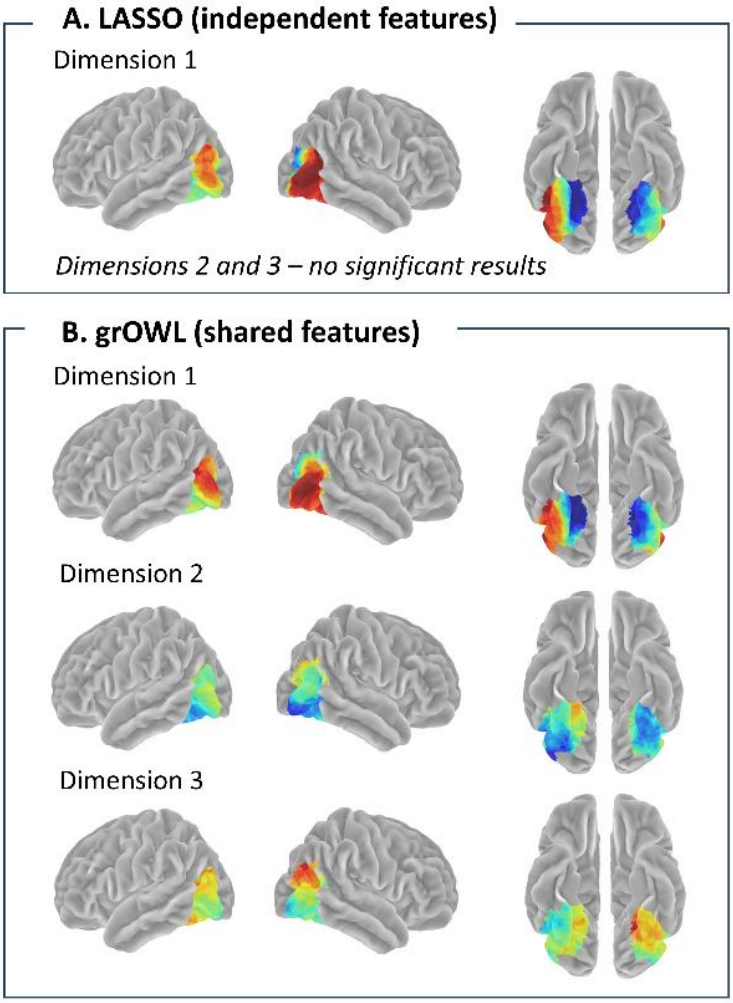

Figure 6: Coefficient maps for RSL. Values are shown projected to the surface (fsaverage template). (A) Coefficients for RSL models trained with LASSO regularization on beta values to predict the coordinates of held-out stimuli on three target semantic dimensions. Colored areas show voxels selected more often than expected from a null distribution computed via permutation testing and controlling the false discovery rate at $\alpha = 0.05$ (see Methods). Hues indicate the proportion of coefficients that are positive across participants, with colored labels indicating what an increase in activation signifies for a given decoder. For dimension 1, cool colors indicate areas where elevated activation systematically signifies inanimate items across participants; warm colors indicate regions where elevated activation signifies animate items (for dimensions 2 and 3, no coefficients reached significance). (B) Coefficients for RSL models trained with grOWL regularization. Cool colors signify regions where elevated activation reliably signifies inanimate items (on dimension 1), birds or invertebrates (on dimension 2) or clothes (on dimension 3); warm colors indicate regions where elevated activation signifies animals (on dimension 1), mammals (on dimension 2) or instruments or vehicles (on dimension 3).

To visualize the anatomical locations of voxels contributing to decoding performance, we first used permutation testing to evaluate which voxels were selected by a decoding model more often than would be expected if no relationship exists between voxel activation and outcomes (see Methods).

From these results we generated the cortical surface group maps shown in Figures 5 and 6. In these plots, colored areas show cortical regions selected by a decoding method reliably more often than expected under the permutation null distribution. The color indicates the proportion of decoding model coefficients that were positive across all participants, which illustrates how an increase in BOLD corresponds to semantic information encoding. Regions shown in blue show received mostly positive coefficients across participants; those in orange/red received mostly negative coefficients. In other words, for the classification analysis (Figure 5), increased activity in an orange/red region is associated with increased probability that the stimulus is animate and increased activity in a blue region is associated with increased probability that the stimulus is inanimate. For the RSL analysis (Figure 6), increased activity in an orange/red regions is associated with a lower coordinate and increased activity in a blue region is associated with a higher one (see Figure 1). In both analyses, regions in green were important for decoding but received mixed positive and negative coefficients across participants.

All models selected the same broad regions of ventral and lateral temporal and occipitotemporal cortex. Coefficients from binary classifiers (Figures 5B, 5C, 6A and 6B) and from the first dimension of the grOWL-based RSL model all showed a remarkably similar pattern – elevated activation in ventromedial temporal lobe bilaterally signified an inanimate stimulus, and elevated activation in more ventrolateral and lateral occipitotemporal regions signified an animate stimulus. Univariate results (Figure 5A) showed the same pattern but failed to detect regions which in which the coefficient direction differed across participants (green regions in panels B and C).

In Figure 6B, RSL coefficients for dimensions 2 and 3 occupied the same anatomical regions, consistent with grOWL's preference for solutions in which the same voxels predict all target dimensions. The RSL coefficients for dimensions 2 and 3 were mixed in direction across much of the selected regions. Areas that did show directional consistency across participants were anatomically situated somewhat differently than those for dimension 1. For dimension 2 (separating animals), elevated activation consistently signified birds or invertebrates in left posterior ventrolateral temporal cortex but signified mammals in a more medial and anterior left ventral temporal region. For dimension 3 (separating artifacts), elevated activation in right ventromedial temporal cortex and right temporo-parietal junction consistently signified a greater likelihood of instruments or vehicles than clothing. While we do not wish to attribute too much significance to these exact patterns of coefficient distributions, the results do clearly indicate that (a) the same cortical regions encode, not just animacy, but also graded information about semantic similarity along multiple orthogonal semantic dimensions, both within and between domains, and (b) that the neuroanatomical

consistency and organization of this information across the implicated regions differs for different semantic dimensions.

3. **Discussion**

Most previous multivariate studies of semantic representation employ a single kind of analysis, and so are constrained by that pipeline's assumptions about the neural code and limited in the kinds of representational coding that they can and cannot detect [1]. By adopting a comparative multivariate approach, we turned these embedded assumptions – previously seen as a limitation – to our advantage. The pattern of results across different decoding models revealed *what*, *how*, and *where* semantic information is encoded in the ATL and elsewhere. Our results suggest that the ATL encodes semantic information across conceptual domains via a multidimensional vector-space code with important neural populations anatomically clustered within and across participants (Hypothesis 9 in Table 1), and that the posterior temporal and occipitotemporal cortex encodes similar information via a similar vector-space code.

Our first question about the ATL concerned its *domain* of representation. Application of RSL revealed multidimensional structure of both animate and inanimate items, which is incompatible with hypotheses that representations within the ATL are animal-specific. This finding therefore aligns with results from neuropsychology, which demonstrate that patients with atrophy of the ATL do not show category-specific impairments [24,35]; with results from neurostimulation, which show transient domain-general semantic impairment following ATL stimulation [36]; and with univariate neuroimaging studies, which implicate the ATL in the representation of multiple domains [12,37].

Second, with regards to the *nature* of representation within the ATL, successful classification of animacy is incompatible with domain-general pointer-based representations. This implies that, rather than simply orchestrating the activity of regions elsewhere in the cortex, the ATL computes additional representations of its own [20]. Analogously, computational models in which different sensory modalities are directly connected perform poorly at semantic tasks, but models that feature an additional hidden layer, to which all modalities are connected, can represent and generalize multimodal semantic information and use it to guide task performance [19,27]. Whether this re-representation is composed of independent features, or of populations that conjointly encode a vector space, cannot be resolved simply by considering which dimensions can be decoded – for instance, the finding here of significant decoding of dimension 2 could be driven either by neural populations encoding independent features discriminating mammals from other animals (*has fur*, *can fly*, etc.) or by populations encoding a dimension that is uninterpretable, but that nevertheless distinguishes between mammals, birds/invertebrates, and others. Rather, it is improved decoding

with grOWL (which prefers solutions that predict multiple target dimensions from the same neural features) than with LASSO (which does not have this preference) that suggests that the ATL encodes semantic information via vector-space representations.

Comparison of classification with SOS-LASSO (which prefers solutions in which predictive features are anatomically clustered within and across individuals) to classification with LASSO (which does not have this preference) answered our third question, about *where* semantic information is represented. Consistently better performance with SOS-LASSO supports the conclusion that semantic representations within the ATL are at least loosely anatomically contiguous within and across individuals. To our knowledge, no extant theory of semantic cognition explicitly predicts that the ATL represents a domain-general vector-space *and* that informative neural populations are arranged in this way. The closest, a domain-general vector-space theory termed the hub-and-spoke model [21,24,26,27], argues that the both the ATL and modality-specific regions called "spokes" are consistently localized across individuals, but makes no prediction about fine-grained anatomical organization within these regions. We therefore believe that these results have the potential to spark future theoretical innovation. This quest will be aided by the development of additional regularization penalties. The current conclusion that ATL semantic representations are vector-space representations *and* anatomically clustered hinges on the assumption that it is the same neural populations that drive both better RSL with grOWL than with LASSO *and* better classification with SOS-LASSO than with LASSO. For instance, a new "SOS-grOWL" that prefers anatomically contiguous vector-space codes would be a powerful tool for further comparative multivariate analysis.

Interestingly, no advantage in binary classification was observed for SOS-LASSO relative to LASSO at the whole-brain level. Very high decoding accuracy (~0.95) suggests that this may be a ceiling effect. In a previous comparison between SOS-LASSO and LASSO, Cox and Rogers [31] also found no difference in accuracy between methods when distinguishing faces from places and objects. When inspecting coefficients, however, SOS-LASSO revealed a neural code for faces that extended across temporal, parietal, and prefrontal cortex, in contrast to a LASSO result including voxels within the fusiform face area (FFA) and nearby regions [38–40]. Splitting this extended system into voxels within canonical face/place systems and those outside, Cox and Rogers [31] found that voxels outside the canonical system were *better* at predicting the target than those within, so the result was not simply a false positive. This illustrates that, when considering whether SOS-LASSO confers an "advantage", simply inspecting accuracy may be an overly simplistic approach – if SOS-LASSO identifies additional regions that are can be demonstrated to represent the target information, this could constitute evidence that representations are anatomically clustered. In our case, however, SOS-LASSO did not highlight any regions beyond those highlighted by LASSO.

So far, our discussion of *where* semantic information is encoded has been focused on anatomical clustering. The more conventional question of localization of anatomical information also warrants attention. The posterior temporal and occipitotemporal cortex have been proposed to contain pointers for inanimate objects [13,41,42] or independent visual features [43], but successful decoding of multidimensional semantic structure is incompatible with the first of these hypotheses and superior performance of grOWL is incompatible with the second. The hub-and-spoke model implicates these regions as a visual "spoke" [21]; preliminary evidence from computational implementations of the hub-and-spoke model suggests that spokes come to represent semantic similarity structure during semantic tasks [19,44,45], but the mechanistic reasons for this are not yet understood.

Despite significant decoding from the ATL in the ROI analysis, this region was not selected in the whole-brain maps (Figures 6 and 7). Multiple methodological factors could have driven this result. First, all the regularization penalties used prefer sparse solutions and so are unlikely to have captured *every* voxel participating in representation. Second, although our acquisition sequence was designed to improve signal quality in the ATL, temporal signal-to-noise ratio was not restored to the level achieved in the rest of the cortex (Supplementary Figure S1); sparse models may prefer to select voxels with better-quality signal. Third, previous research suggests that semantic representation in the ATL is *dynamic* – the region as a whole encodes semantic information, but the contribution of individual neural populations changes over time [46]. Other regions may encode the same information in a temporally stable way and so may be preferable to a decoding model trained on fMRI data. Fourth, voxels in the ATL are selected far more frequently by models trained on *permuted* data than voxels elsewhere in the brain (Supplementary Figures S2 and S3). This means that coefficients in the ATL are unlikely to survive permutation thresholding even when they are selected by most models trained on real data. The reason for this phenomenon is unknown [45], but may reflect the fact that signal inhomogeneity increases the ratio of uncorrelated noise relative to correlated noise, increasing the probability that data will correlate with a permuted target (Frisby et al., unpublished simulations).

4. **Conclusion**

Multiple forms of neuroscientific data converge on the conclusion that the ATLs support semantic cognition [21]. This work demonstrates that, out of 12 hypotheses about the domain, nature and location of representation in the ATLs, one hypothesis best explains the data – that the ATLs represent domain-general semantic information (*what*) via a vector-space code (*how*) that is anatomically clustered within and across individuals (*where*). Posterior temporal and

occipitotemporal regions utilize a similar domain-general, vector-space code. Our results not only adjudicate competing hypotheses about semantic representation but also illustrate the power of the comparative multivariate approach – using the pattern of results across multiple decoding models to differentiate theories. The conjoining of theory with decoding analysis approach [1] has the potential to reveal how the brain represents, not just semantic knowledge, but any kind of information.

5. **Methods**

5.1. Participants

32 participants were recruited to the study. Of these, five were unable to complete the study either for medical reasons or because of technical problems during data acquisition. This left 27 participants (age range 19-50, mean age 27.96 years, 17 female, 10 male) who were all right-handed native speakers of British English, with normal or corrected-to-normal vision, and with no neurological or sensory disorders. All participants gave written informed consent and the research was approved by a local National Health Service (NHS) ethics committee.

5.2. Stimuli and task

Stimuli were 100 line drawings used in previous ECoG work [29,46–48] – 50 animals and 50 inanimate items including vehicles, clothes, musical instruments and a range of other objects [49] (https://github.com/slfrisby/7TConvergent/tree/main/stimuli/). Animate and inanimate items were matched for age of acquisition, visual complexity, familiarity, word frequency, name agreement, and non-semantic visual structure [46,48,50]. PsychoPy 2022.2.5 [51] was used to display stimuli and record speech. Stimuli were rear-projected onto a screen in the MRI scanner and observers viewed stimuli through a mirror mounted to the head coil directly above the eyes.

Participants were instructed to name each picture out loud as quickly and accurately as possible while keeping head motion to a minimum. To ensure that participants understood the instructions, they practiced the task before entering the scanner while the experimenter provided feedback about their head motion. Practice continued until the experimenter was satisfied that head motion was minimized.

Overt speech was recorded for each trial and responses were scored for accuracy with the study's aim of decoding semantic similarity structure in mind. If the participant did not respond within the 4 seconds during which the stimulus was on-screen, the trial was marked as incorrect. If the participant responded incorrectly but named an object that the rater judged to be visually and semantically similar to the stimulus and at a similar level of specificity (e.g. "wasp" for bee), the trial was marked as correct; if the object that they named was semantically dissimilar or was named at a

less specific level (e.g. “insect” for bee), the trial was marked as incorrect. It was important that neural responses did not reflect a “blend” of the semantics of multiple stimuli so, if the participant gave two synonyms (e.g. “gun, pistol”) the trial was marked as correct; if they named two different objects (“kettle, iron”) the trial was marked as incorrect even if one of the names was correct. Trials containing fillers such as “um” and “err” in addition to a correct response were marked as correct because the participants were neurologically healthy and there was no reason to believe that these utterances reflected anything other than normal speech processes. The mean and standard deviation of the accuracy scores were calculated.

The study had a fast event-related design with 4 runs per participant, collected in a single session. Each run began 16 s after the start of the MR sequence and consisted of 100 trials (each stimulus appeared once per run). Each trial consisted of a stimulus presented for 4 s followed by a fixation cross presented for 4 s. 4 sequences of animate and inanimate stimuli were created and counterbalanced across participants. The order of trials was optimized for the contrast between animate and inanimate stimuli using optseq2 (https://surfer.nmr.mgh.harvard.edu/fswiki/optseq2). For each run and each participant, a random order of animate stimuli and a random order of inanimate stimuli were created; the stimuli were then interspersed according to the sequence generated with optseq2.

5.3. Image acquisition

All images were acquired on a whole-body 7T Siemens MAGNETOM Terra MRI scanner (Siemens Healthcare, Erlangen, Germany) with an 8Tx32Rx head coil (Nova Medical, USA).
An MP2RAGE anatomical scan was acquired with the following parameters: 224 sagittal slices (interleaved acquisition), FOV 240 x 225.12 x 240 mm, voxel size 0.75 mm isotropic, TR 4300 ms, inversion times 840 and 2370 ms, TE 1.99 ms, flip angles 5° and 6°, GRAPPA acceleration factor 3, and duration 8 minutes and 50 seconds.

Next, manual B_0-shimming was performed over the volume to be used for echo-planar imaging (EPI). This was necessary because of a known bug in the scanner software which renders automatic shims inaccurate. A double-echo steady state (DESS) sequence was used to generate a field map and the radiographer adjusted shim parameters manually with the aim of reducing the water linewidth, defined as full width of the spectrum at half height, to below 40 Hz. However, for some participants the adjustments proved time-consuming and so adjustment time was capped at 30 minutes from the start of scanning after which the best shim parameters were adopted. The actual average water linewidth was 38.80 Hz (standard deviation = 5.93 Hz).

All four functional runs of EPI used the multi-echo multiband sequence designed by Frisby et al. [32] to improve ATL signal quality. The sequence had the following parameters: 48 axial slices (interleaved acquisition), FOV 210 x 210 x 210 mm to cover the whole brain in most participants (visual inspection ensured that the vATL was included in the FOV for all participants and the FOV was tilted up at the nose to avoid ghosting of the eyes into the vATL), voxel size 2.5 mm isotropic (no gap), TR = 1510 ms, 3 echoes, TE_1 = 11.80 ms, TE_2 = 27.05 ms, TE_3 = 42.30 ms, flip angle 63°, multiband factor 2, GRAPPA acceleration factor 3, phase partial Fourier 7/8, and A-P phase encoding direction. Immediately *before* each run, 5 additional volumes were acquired with the phase encoding direction changed to P-A to facilitate distortion correction during preprocessing.

5.4. Data analysis

All analysis code is available at: https://github.com/slfrisby/7TConvergent/.

5.4.1. Preprocessing

For ease of data sharing we converted all DICOMs to BIDS format [52] using *heudiconv* (v1.1.1; [53]).

Standard reproducible preprocessing pipelines designed for 3T-fMRI, such as *fMRIprep* [54–56] perform poorly on 7T-fMRI EPI data. Therefore, the analysis pipeline was split into stages using different software packages.

Since it is notoriously difficult to perform a good-quality brain extraction on MP2RAGE data (because of salt-and-pepper noise in the background and cavities), two pipelines were used. The MP2RAGE T1w (combined) image first had its background noise removed using O'Brien regularization [57] and was then submitted to CAT12 for segmentation [58]; in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/). The bias-corrected and global-intensity-corrected T1w (combined) image produced was provided to as input to the anatomy pipeline in *fMRIPrep* 21.0.1. The T1w (combined) image was corrected for intensity non-uniformity with *N4BiasFieldCorrection* [59] and skull-stripped with a Nipype implementation of the *antsBrainExtraction.sh* workflow (ANTs 2.3.3, https://github.com/ANTsX/ANTs; [60,61]. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM) and grey matter (GM) was performed on the brain-extracted T1w (combined) image using *fast* (FSL 6.0.5.1, [62]). Volume-based spatial normalization to standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with *antsRegistration.sh* (ANTs).

Functional preprocessing was performed using in-house code, composed of functions from AFNI (v.18.3.03; [63,64]; https://afni.nimh.nih.gov/), FSL (v.5.0.10; [65–68]; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki), tedana (v.24.0.0; [69–72];

https://tedana.readthedocs.io/en/stable/index.html) and ANTS (v. 2.2.0; [60,61]; https://github.com/ANTsX/ANTs). Slice timing was corrected to the middle slice using *3dTshift* (AFNI), motion was corrected with *3dvolreg* and *3dAllineate* (AFNI), using the first volume of the first run acquired as a reference (for each volume, the TE_1 image was aligned and the resulting transform was applied to the TE_2 and TE_3 images), and skull-stripped using *BET* (FSL [67]) to create a participant-specific brain mask. Then *tedana* was used to combine data from all echoes optimally based on T2* weighting [73] and to denoise the data using ICA. *tedana* conducts denoising by decomposing data using PCA and ICA, classifying components according to whether the signal scales linearly with TE (as BOLD signal does), and reconstruct the data using only BOLD-like components. The brain mask created with *BET* was used as the mask for this stage.

Next, unwarping was conducted using *topup* and *applytopup* (FSL; [74]). Field displacement maps were calculated using ten volumes (five with A-P phase encoding direction, extracted from the start of each functional run, and five with P-A phase encoding direction, collected separately before each run) and the resulting correction was applied to all images. Finally, the mean EPI for each run was coregistered to the skull-stripped native structural image using a rigid-body registration in *AntsRegistrationSyN.sh* (ANTS). This produced transforms from native EPI space to native T1 space. For all decoding analyses, images remained in native space and were not smoothed. However, the transforms from native EPI to native T1 (for the first run) and the transforms from native T1 to standard space (MNI152NLin2009cAsym), with *antsApplyTransforms* (ANTS), were used to calculate the MNI-space coordinates that each voxel *would* have *were* images to be transformed into MNI space. These coordinates were used for visualization and were provided to classifiers trained with SOS-LASSO regularization, which take anatomical information as input. The transform from native EPI to native T1 space was also used to transform the participant's grey-matter mask into native EPI space. For the univariate analysis, images were transformed into standard space (MNI152NLin2009cAsym) with *antsApplyTransforms* (ANTS) and smoothed with a 6 mm FWHM Gaussian filter in SPM12 (https://www.fil.ion.ucl.ac.uk/spm/).

5.4.2. 1st-level (within-participant) GLM

Data were processed using the general linear model (GLM) approach implemented in SPM12 in MATLAB r2023b. For decoding, each stimulus was modelled as an individual boxcar function, 1.5 seconds in length because the stimulus in our stimulus set with the longest mean naming time had a mean naming time of 1323 ms [49] (periods between stimuli were modelled implicitly). These boxcar functions were subsequently convolved with SPM's difference of gammas haemodynamic response function. The six motion parameters extracted during preprocessing were used as regressors of no

interest. Data from all four runs was concatenated, so an additional constant was included for each run. The micro-time resolution was set as the number of slices (n = 48), the micro-time onset was set as the reference slice for slice-time correction (n = 24), and the high-pass filter cutoff was 128 seconds. Each participant's grey matter mask in native EPI space was used as a mask for the analysis. The parameter estimation method was restricted maximum likelihood estimation (ReML) and serial correlations were accounted for using an autoregressive AR(1) model during estimation. This produced one beta image per stimulus in native space. For univariate analysis, the same model was run on the normalized, smoothed data with identical parameters except that a grey matter mask in standard space (MNI152NLin2009cAsym) was used, that only two task regressors were used (animate stimuli named correctly and inanimate stimuli named correctly), and that animate stimuli named incorrectly and inanimate stimuli named incorrectly were modelled as additional regressors of no interest. The univariate contrasts of interest were greater activation for animate stimuli named correctly than inanimate stimuli named correctly (A > I) and greater activation for inanimate stimuli named correctly than animate stimuli named correctly (I > A). This produced contrast images in MNI space.

5.4.3. Univariate analysis – 2nd-level (across-participant) GLM

Group-level t-tests of each contrast of interest (A > I, I > A) were conducted in SPM12. The group t-maps were assessed for significance by using a voxel-height threshold of $p < 0.001$ to define clusters and then a cluster-defining family-wise-error-corrected threshold of $p < 0.05$ for statistical inference. Group-level results were then projected to the pial surface (fsaverage template [75,76]) using *-volume-to-surface-mapping* with trilinear interpolation in Connectome Workbench 1.5.0 and the thresholded maps were plotted on the pial surface using *plot_surf_stat_map* in *nilearn* implemented in Python 3.9.

5.4.4. Multivariate decoding analysis

5.4.4.1. Input data

To prepare data for decoding, beta images from all 4 presentations of each stimulus were averaged across runs. If a participant had named a stimulus incorrectly on one to three occasions, the beta image(s) corresponding to the incorrect trial(s) were not included in the averaging. If the participant named the stimulus incorrectly on all four trials, the values for each voxel were interpolated with the median value for that voxel over all other stimuli.

For the vTL analysis, an ROI was constructed based on electrode coverage in a previous study [48]; all electrode coordinates are available at https://github.com/slfrisby/7TConvergent/). Electrodes were converted to one NIFTI volume at native EPI resolution (2.5 mm isotropic) in MNI space and smoothed at 10 mm FWHM using SPM12. In order to exclude regions that lay outside the temporal lobe, an "exclusion mask" was created by taking the Harvard-Oxford cortical and subcortical atlases, removing the temporal lobe from the atlas, and adding the cerebellum from the AAL 2 atlas [77]; since the Harvard-Oxford atlas does not include a cerebellum). Only voxels that lay *outside* this exclusion mask were included in the ROI. After masking the ROI was smoothed again at 4mm FWHM because, on inspection, this created a smooth shape that fully encompassed the temporal lobes. Since electrode coverage was far more extensive in the left than the right ROI, a mirror-image of the left ROI was used as the right ROI. The final ROIs in standard space (MNI152NLin2009cAsym) are shows projected to the surface in Figure 2 and are available at https://github.com/slfrisby/7TConvergent/tree/main/ROI/. The ROIs were backprojected into each participant's native space using the transforms from native EPI to native T1 and the transforms from native T1 to standard space generate with *antsRegistration.sh*. Only beta values from voxels that overlapped with both the ROI and the participant's native-space grey matter mask were provided as input to the decoding models.

For the whole-cortex analysis, only voxels that overlapped with the participant's native-space grey matter mask were provided as input to the decoding models.

5.4.4.2. Decoding models

We tested for semantic structure using two decoding methods. First, we used regularized logistic regression classifiers to decode binary animacy [46]. One set of classifiers was trained using LASSO regularization [78], which assumes that the neural code is sparse (only a few of the features under consideration encode the target structure) and uncorrelated (important features exhibit uncorrelated activity). These classifiers were fit using *glmnet* in MATLAB r2023b [79], using 3 – 0.2 as the range of possible lambda values. Each classifier took, as input features, beta values from multiple voxels from a single participant. A second set of classifiers was trained using SOS-LASSO regularization, which assumes that the neural code is relatively sparse but also in roughly the same anatomical location within and across participants. Full mathematical details, illustrating how SOS-LASSO incorporates these assumptions, can be found in Rao et al. [80,81] and Cox and Rogers [31]. These classifiers were fit using the WISC MVPA toolbox in MATLAB r2018b (https://github.com/crcox/WISC_MVPA/) with hyperband budget set to 25 (all other parameters were set as defaults, including a set size of 18 mm with 9 mm overlap and 3 – 0.2 as the range of possible lambda values). These classifiers took beta

743 values from multiple voxels from multiple participants as input and, though data were in native
744 space, the MNI-space coordinates that each voxel *would* have *were* images to be transformed into
745 MNI space were also provided as input to the classifier.
746 Second, we used RSL to assess whether and where graded, multidimensional semantic
747 structure was represented. Target semantic dimensions were modelled following the approach of
748 Cox et al. [29] – details are provided in full there and are summarized here. First, a representational
749 similarity matrix (RSM) was constructed that expressed the semantic similarity between each
750 possible pair of stimuli (operationalized using feature verification norms [33]). Next, singular value
751 decomposition (SVD) was applied to extract three components accounting for 89.5% of the variance
752 in the whole RSM (81.1%, 4.4%, and 4.0%, respectively). The coordinates of all 100 stimuli on each
753 dimension are shown in Figure 1.
754 RSL models were fit to predict the coordinates of each stimulus on the three target semantic
755 dimensions from beta values from multiple voxels. One set of models used LASSO regularization and
756 were trained using *glmnet* in MATLAB r2023b, using 6 – 0 as the range of possible lambda values. A
757 second set of models used grOWL regularization, which assumes that the neural code is sparse but
758 correlated – in other words, information is encoded by small groups of features that are correlated in
759 their activity. Importantly, grOWL also assumes that the neural code is interdependent – features
760 carry information about multiple dimensions simultaneously. Full mathematical details, explaining
761 how grOWL incorporates these assumptions, can be found in Oswal et al. [30] and Cox et al. [29]. grOWL
762 models were trained using the WISC MVPA toolbox in MATLAB r2018a with the hyperband budget
763 set to 25 (all other parameters were set as defaults, including 6 – 0 as the range of possible lambda
764 values). All RSL models took, as input features, beta values from multiple voxels from a single
765 participant.
766
767 5.4.4.3. Cross-validation and model evaluation
768 For both logistic regression classifiers and RSL models, model performance was assessed via ten-fold
769 nested cross-validation. Data were split into ten folds, each containing ten stimuli (five animate, five
770 inanimate). In the first instance, stimuli were assigned randomly to the ten folds, but this order was
771 then fixed so that models fit with every kind of regularization were assessed using the same fold
772 configuration. One fold was designated as the outer-loop hold-out set and a second fold was
773 designated as the inner-loop holdout set. The stimuli in the remaining eight folds were used as
774 training data to search for the best-performing hyperparameter(s). Models with these
775 hyperparameter values were assessed on the inner-loop hold-out set. The folds were then
776 reassigned so that ten stimuli previously used for training became the inner-loop hold-out set. Once

all combinations of inner-loop hold-out set and training data had been explored, the best-performing hyperparameter value(s) were used to train a final model that was assessed on the outer-loop hold-out set. The procedure was then repeated with the nine other possible outer-loop hold-out sets.

The mean classification accuracy (for classifiers) or correlation between true and predicted coordinates (for RSL models) was calculated over all hold-out folds. Statistical significance was assessed via permutation testing. This was done for two reasons. First, since we used a fast event-related design rather than a slow event-related design, the haemodynamic response overlapped between trials. This means that, if temporally adjacent stimuli were divided between training and test sets, the stimulus in the training set may contain information about the stimulus in the test set and so may produce classification hold-out accuracies above 0.5 or hold-out correlations above 0. Second, cross-validated correlation (as was calculated for RSL models) is known to produce predicted coordinates that sometimes correlate *negatively* with the target coordinates in a hold-out set – the mean correlation under the null hypothesis may be below zero [29,82]; therefore, the magnitude of the real correlation can be understood only with reference to the permutation distribution. In permutation testing, the stimulus labels were permuted before model training (the permutation orders were generated at random but then fixed so that the permutations for each model were comparable) and the procedure was repeated 100 times to produce 100 simulated accuracy or correlation values for each participant. A group-level empirical null distribution was simulated by randomly sampling one of the 100 values from each participant and computing a group average 10,000 times [34]. One-tailed *p*-values were calculated (where m is the number of values in the permutation distribution and b is the number of values in the permutation distribution larger than the true value [29,83]):

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

To test whether one model or ROI produced a higher accuracy or correlation than another, a group-level permutation distribution of differences was created using the differences between the values in both permutation distributions and using the group-level simulation procedure described above. A p-value was then calculated for the true difference. Statistical significance for all tests was defined after adjusting p-values to control the false discovery rate at α = 0.05 [84].

5.4.4.4. Coefficient visualization

Ten-fold nested cross-validation does not provide a clear picture of which voxels are informative – each of the ten models selects its own set of coefficients [31]. Therefore, to visualize coefficients, one model was trained for each participant using an inner loop of ten stimuli to identify the best hyperparameter(s) but no outer loop (i.e. all 90 stimuli that were not in the current inner loop hold-

out set were used for hyperparameter exploration; therefore, all 100 stimuli were used for training). 100 permutation models were also trained for each participant (but no random sampling was conducted at the visualization stage). Coefficients were converted to NIFTI volumes at native EPI resolution (2.5 mm isotropic) in MNI space using SPM12, separately for each participant. Coefficients both for models trained on real data and for models trained on permuted data were projected to the pial surface (fsaverage template; [75,76]) using *-volume-to-surface-mapping* with trilinear interpolation and smoothed on the surface at 6 mm FWHM using *–metric-smoothing* in Connectome Workbench 1.5.0. Two metrics were calculated – (1) the proportion of participants in which each vertex was selected and (2) the proportion of negative coefficients that each vertex received. (For classification, since animals were coded as 0 and inanimate objects as 1, a negative coefficient indicates that a positive beta value is associated with increased probability that the stimulus is animate. For RSL, although the dimensions capture more than just animacy, the same approach was taken for consistency.) The map of the proportion of negative coefficients was thresholded using the map of coefficient selection - vertices were included in the map only if they were selected significantly more frequently in the real distribution than they were in the permutation distribution, as assessed via binomial tests with p-values adjusted to control the false discovery rate at $\alpha = 0.05$ [84]. Finally, the thresholded map was plotted on the pial surface using *plot_surf_stat_map* in *nilearn* implemented in Python 3.9.

Data and code availability: Data will be made publicly available upon peer review and acceptance. Code is publicly available at: https://github.com/slfrisby/7TConvergent/.

Author contributions: Saskia L. Frisby – Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing; Christopher R. Cox – Conceptualization, Methodology, Software, Writing – Review & Editing; Ajay D. Halai – Conceptualization, Methodology, Writing – Review & Editing, Supervision; Matthew A. Lambon Ralph – Conceptualization, Methodology, Writing – Review & Editing, Supervision; Timothy T. Rogers – Conceptualization, Methodology, Writing – Review & Editing, Supervision.

Funding: This work was supported by an EPS study visit grant (G118497) to S.L.F., by an MRC Career Development Award (MR/V031481/1) to A.D.H., by an MRC programme grant (MR/R023883/1) and intramural funding (MC_UU_00005/18) to M.A.L.R., and by an NSF grant (NCS-FO 219903 DRL) to T.T.R.

Competing interests: The authors declare no conflicts of interest.

Acknowledgements: We would like to thank the participants and the Wolfson Brain Imaging Centre radiographers.

References

1. Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A. & Rogers, T. T. Decoding semantic representations in mind and brain. *Trends Cogn. Sci.* **27**, 258–281 (2023).
2. 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).
3. Noppeney, U. *et al.* Temporal lobe lesions and semantic impairment: a comparison of herpes simplex virus encephalitis and semantic dementia. *Brain* **130**, 1138–1147 (2006).
4. Tranel, D., Damasio, H. & Damasio, A. R. A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia* **35**, 1319–1327 (1997).
5. 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 Lang.* **75**, 195–231 (2000).
6. Buxbaum, L. J. & Saffran, E. M. Knowledge of object manipulation and object function: dissociations in apraxic and nonapraxic subjects. *Brain Lang.* **82**, 179–199 (2002).
7. Grafman, J. The structured event complex and the human prefrontal cortex. in *Principles of frontal lobe function* 292–310 (Oxford University Press, New York, NY, US, 2002). doi:10.1093/acprof:oso/9780195134971.003.0019.
8. Kanwisher, N. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc. Natl. Acad. Sci.* **107**, 11163–11170 (2010).
9. Clarke, A. & Tyler, L. K. Understanding What We See: How We Derive Meaning From Vision. *Trends Cogn. Sci.* **19**, 677–687 (2015).

10. 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. *Cogn. Neuropsychol.* **18**, 125–174 (2001).
11. Binney, R. J., Parker, G. J. M. & Lambon Ralph, M. A. Convergent Connectivity and Graded Specialization in the Rostral Human Temporal Lobe as Revealed by Diffusion-Weighted Imaging Probabilistic Tractography. *J. Cogn. Neurosci.* **24**, 1998–2014 (2012).
12. Chen, L. & Rogers, T. T. A Model of Emergent Category-specific Activation in the Posterior Fusiform Gyrus of Sighted and Congenitally Blind Populations. *J. Cogn. Neurosci.* **27**, 1981–1999 (2015).
13. Damasio, A. R. The brain binds entities and events by multiregional activation from convergence zones. *Neural Comput.* **1**, 123–132 (1989).
14. Binder, J. R. *et al.* Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* **33**, 130–174 (2016).
15. 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).
16. McRae, K., Seidenberg, M. S. & de Sa, V. R. On the nature and scope of featural representations of word meaning. *J. Exp. Psychol. Gen.* **126**, 99–130 (1997).
17. Ruts, W. *et al.* Dutch norm data for 13 semantic categories and 338 exemplars. *Behav. Res. Methods Instrum. Comput.* **36**, 506–515 (2004).
18. Giallanza, T., Campbell, D., Cohen, J. D. & Rogers, T. T. An integrated model of semantics and control. *Psychol. Rev.* **132**, 1128–1177 (2025).
19. Jackson, R. L., Rogers, T. T. & Lambon Ralph, M. A. Reverse-engineering the cortical architecture for controlled semantic cognition. *Nat. Hum. Behav.* **5**, 774–786 (2021).
20. Lambon Ralph, M. A., Sage, K., Jones, R. W. & Mayberry, E. J. Coherent concepts are computed in the anterior temporal lobes. *Proc. Natl. Acad. Sci.* **107**, 2717–2722 (2010).

21. Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
22. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. in *Advances in Neural Information Processing Systems* vol. 26 (Curran Associates, Inc., 2013).
24. Patterson, K., Nestor, P. J. & Rogers, T. T. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* **8**, 976–987 (2007).
25. Pereira, F., Gershman, S., Ritter, S. & Botvinick, M. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.* **33**, 175–190 (2016).
26. Rogers, T. T. *et al.* Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychol. Rev.* **111**, 205–235 (2004).
27. Rogers, T. T. & McClelland, J. L. *Semantic Cognition: A Parallel Distributed Processing Approach*. (MIT Press, Cambridge, Mass., 2004).
28. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* **9**, 1–13 (2018).
29. Cox, C. R. *et al.* Representational similarity learning reveals a graded multidimensional semantic space in the human anterior temporal cortex. *Imaging Neurosci.* **2**, 1–22 (2024).
30. Oswal, U., Cox, C. R., Lambon Ralph, M. A., Rogers, T. T. & Nowak, R. D. Representational Similarity Learning with Application to Brain Networks. in *Proceedings of The 33rd International Conference on Machine Learning* 1041–1049 (PMLR, 2016).
31. Cox, C. R. & Rogers, T. T. Finding distributed needles in neural haystacks. *J. Neurosci.* **41**, 1019–1032 (2021).

924 32. Frisby, S. L. *et al.* Optimising 7T-fMRI for imaging regions of magnetic susceptibility.
925 2025.03.17.643748 Preprint at https://doi.org/10.1101/2025.03.17.643748 (2025).
926 33. Dilkina, K. & Lambon Ralph, M. A. Conceptual structure within and between modalities. *Front.*
927 *Hum. Neurosci.* **6**, (2012).
928 34. Stelzer, J., Chen, Y. & Turner, R. Statistical inference and multiple testing correction in
929 classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size
930 control. *NeuroImage* **65**, 69–82 (2013).
931 35. Rogers, T. T., Patterson, K., Jefferies, E. & Lambon Ralph, M. A. Disorders of representation and
932 control in semantic cognition: Effects of familiarity, typicality, and specificity. *Neuropsychologia*
933 **76**, 220–239 (2015).
934 36. Pobric, G., Jefferies, E. & Lambon Ralph, M. A. Category-specific versus category-general
935 semantic impairment induced by transcranial magnetic stimulation. *Curr. Biol.* **20**, 964–968
936 (2010).
937 37. Binney, R. J., Hoffman, P. & Lambon Ralph, M. A. Mapping the Multiple Graded Contributions of
938 the Anterior Temporal Lobe Representational Hub to Abstract and Social Concepts: Evidence
939 from Distortion-corrected fMRI. *Cereb. Cortex* **26**, 4227–4241 (2016).
940 38. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally
941 defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).
942 39. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: A module in human
943 extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
944 40. Martin, A. The representation of object concepts in the brain. *Annu. Rev. Psychol.* **58**, 25–45
945 (2007).
946 41. Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends Cogn. Sci.* **15**, 527–536
947 (2011).
948 42. Damasio, H., Tranel, D., Grabowski, T., Adolphs, R. & Damasio, A. Neural systems behind word
949 and concept retrieval. *Cognition* **92**, 179–229 (2004).

43. Martin, A. GRAPES — Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychon. Bull. Rev.* **23**, 979–990 (2016).

44. Chen, L., Lambon Ralph, M. A. & Rogers, T. T. A unified model of human semantic knowledge and its disorders. *Nat. Hum. Behav.* **1**, 0039 (2017).

45. Cox, C. R. Testing neurocognitive predictions of the hub-and-spoke model of semantic memory with network representational similarity analysis. (University of Wisconsin-Madison, 2016).

46. 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).

47. 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).

48. Frisby, S. L. *et al.* All spectral frequencies of neural activity reveal semantic representation in the human anterior ventral temporal cortex. 2025.04.17.649404 Preprint at https://doi.org/10.1101/2025.04.17.649404 (2025).

49. Morrison, C. M., Chappell, T. D. & Ellis, A. W. Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Q. J. Exp. Psychol. Sect. A* **50**, 528–559 (1997).

50. Barry, C., Morrison, C. M. & Ellis, A. W. Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency, and name agreement. *Q. J. Exp. Psychol. Sect. A* **50**, 560–585 (1997).

51. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).

52. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 160044 (2016).

975 53. Halchenko, Y. O. *et al.* HeuDiConv — flexible DICOM conversion into structured directory
976 layouts. *J. Open Source Softw.* **9**, 5839 (2024).
977 54. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods*
978 **16**, 111–116 (2019).
979 55. Gorgolewski, K. J. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data
980 processing framework in python. *Front. Neuroinformatics* **5**, (2011).
981 56. Markiewicz, C. J. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. Zenodo
982 https://doi.org/10.5281/zenodo.12774221 (2024).
983 57. O'Brien, K. R. *et al.* Robust T1-weighted structural brain imaging and morphometry at 7T using
984 MP2RAGE. *PLOS ONE* **9**, e99676 (2014).
985 58. Gaser, C. *et al.* CAT – A computational anatomy toolbox for the analysis of structural MRI data.
986 2022.06.11.495736 Preprint at https://doi.org/10.1101/2022.06.11.495736 (2023).
987 59. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–
988 1320 (2010).
989 60. Avants, B. B., Tustison, N. & Johnson, H. Advanced Normalization Tools (ANTS). (2009).
990 61. Avants, B. B. *et al.* A reproducible evaluation of ANTs similarity metric performance in brain
991 image registration. *NeuroImage* **54**, 2033–2044 (2011).
992 62. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov
993 random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**,
994 45–57 (2001).
995 63. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance
996 neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
997 64. Cox, R. W. & Hyde, J. S. Software tools for analysis and visualization of fMRI data. *NMR Biomed.*
998 **10**, 171–178 (1997).
999 65. Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo
1000 echo-planar images: application to diffusion tensor imaging. *NeuroImage* **20**, 870–888 (2003).

1001 66. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL.
1002 *NeuroImage* **62**, 782–790 (2012).
1003 67. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
1004 68. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation
1005 as FSL. *NeuroImage* **23**, S208–S219 (2004).
1006 69. The tedana Community *et al.* ME-ICA/tedana: 23.0.1. Zenodo
1007 https://doi.org/10.5281/zenodo.7926293 (2023).
1008 70. DuPre, E. *et al.* TE-dependent analysis of multi-echo fMRI with tedana. *J. Open Source Softw.* **6**,
1009 3669 (2021).
1010 71. Kundu, P., Inati, S. J., Evans, J. W., Luh, W.-M. & Bandettini, P. A. Differentiating BOLD and non-
1011 BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage* **60**, 1759–1770 (2011).
1012 72. Kundu, P. *et al.* Integrated strategy for improving functional connectivity mapping using
1013 multiecho fMRI. *Proc. Natl. Acad. Sci.* **110**, 16187–16192 (2013).
1014 73. Posse, S. Multi-echo acquisition. *NeuroImage* **62**, 665–671 (2012).
1015 74. Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo
1016 echo-planar images: application to diffusion tensor imaging. *NeuroImage* **20**, 870–888 (2003).
1017 75. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and
1018 Surface Reconstruction. *NeuroImage* **9**, 179–194 (1999).
1019 76. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis: II. Inflation, flattening, and
1020 a surface-based coordinate system. *NeuroImage* **9**, 195–207 (1999).
1021 77. Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a
1022 macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–
1023 289 (2002).
1024 78. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*
1025 **58**, 267–288 (1996).

79. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, (2010).

80. Rao, N., Cox, C., Nowak, R. D. & Rogers, T. T. Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. in *Advances in Neural Information Processing Systems* vol. 26 (2013).

81. Rao, N., Nowak, R. D., Cox, C. & Rogers, T. Classification with the sparse group lasso. *IEEE Trans. Signal Process.* **64**, 448–463 (2016).

82. Zhou, Y., Vales, M. I., Wang, A. & Zhang, Z. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. *Brief. Bioinform.* **18**, 744–753 (2017).

83. Phipson, B. & Smyth, G. K. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, (2010).

84. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).