Published in final edited form as:
Nat Hum Behav. 2021 June 01; 5(6): 774–786. doi:10.1038/s41562-020-01034-z.

Reverse-Engineering the Cortical Architecture for Controlled Semantic Cognition

Rebecca L. Jackson[1,*], **Timothy T. Rogers**[2], **Matthew A. Lambon Ralph**[1,*]

[1]MRC Cognition & Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF

[2]524 WJ Brogden Hall, Department of Psychology, University of Wisconsin-Madison Madison, Wisconsin, 53706

Abstract

We employ a 'reverse-engineering' approach to illuminate the neurocomputational building blocks that combine to support controlled semantic cognition: the storage and context-appropriate use of conceptual knowledge. By systematically varying the structure of a computational model and assessing the functional consequences, we identified the architectural properties that best promote some core functions of the semantic system. Semantic cognition presents a challenging test case as the brain must achieve two seemingly contradictory functions: abstracting context-invariant conceptual representations across time and modalities, whilst producing specific context-sensitive behaviours appropriate for the immediate task. These functions were best achieved in models possessing a single, deep multimodal hub with sparse connections from modality-specific regions, and control systems acting on peripheral rather than deep network layers. The reverse-engineered model provides a unifying account of core findings in the cognitive neuroscience of controlled semantic cognition, including evidence from anatomy, neuropsychology, and functional brain imaging.

At heart, cognitive neuroscience is an effort to understand how mental representations and processes arise from, and relate to, underlying neural mechanisms. Toward this goal, researchers typically begin by seeking relationships between neural data (neurophysiological activity and/or neuropathology) and patterns of behaviour in various tasks. Here we employ an alternative 'reverse-engineering' approach that first considers the functions that a given cognitive system must support and then evaluates what neuro-computational machinery best achieves those functions. By systematically varying structure within a computer simulation and assessing the functional consequences, one can establish the architectural elements

---

*Correspondence to: Dr. Rebecca Jackson, MRC Cognition & Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge, CB2 7EF, Rebecca.Jackson@mrc-cbu.cam.ac.uk, ttrogers@wisc.edu, Matt.Lambon-Ralph@mrc-cbu.cam.ac.uk, Tel: +44 (0)1223 769452.

Author Contributions: RLJ, TTR and MALR made substantial contributions to the conception and design of the work, the interpretation of data and the manuscript revisions. RLJ acquired the results and drafted the manuscript.

Competing Interests: The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

critical to the targeted functions, potentially explaining why the system is organised in a particular way.

We apply this approach to understand the cortical network underlying semantic cognition, the controlled access to and manipulation of conceptual knowledge or meaning[1,2]. Semantic cognition provides a challenging test case because the system must concurrently achieve two functions that appear diametrically opposed. First, it must abstract over episodes and across time to acquire context-independent representations that express conceptual similarity structure and thereby promote knowledge generalisation across items and contexts[1,3,4]. This 'conceptual abstraction' supports the ability to discern conceptual similarity amongst items denoted by images, words, or other attributes, despite sometimes dramatic variability in their surface properties[4–6]. For instance, if one learns that wolves are dangerous after being attacked in the woods, this knowledge should generalise to a different wolf on a farm or in the house promoting the inference 'dangerous' for all wolves. Second, the system must flexibly adapt semantic representations to suit immediate task demands[1,5,7,8] - differentiating wolves and dogs when generating an inference about safety, but treating them similarly to infer physical appearance. Thus, the subset of features governing the similarity space and consequent generalisation in the moment must also be context-sensitive.

Whilst the literature currently advances several hypotheses about the cortical architecture of the semantic system, no prior work has compared their ability to simultaneously achieve conceptual abstraction and context-sensitivity. Interestingly, the extant hypotheses are variants of a central idea dating back at least to Wernicke: that semantic knowledge arises from interactions amongst various surface representations (sensory, motor, linguistic, affective, etc.) distributed throughout cortex[9]. They differ principally in their proposals about the pathways through which these surface representations interact. Consequently, they can be contrasted using computer simulations with a family of neural network models, all learning to compute the same interactive mappings amongst various surface representations, but differing in their architecture. Illuminating the architectural elements that best support both conceptual abstraction and context sensitivity, generates a cognitively- and computationally-motivated hypothesis about cortical network structure that can be weighed against critical empirical evidence, including (1) extant data about the anatomy of the cortical semantic system, (2) differing patterns of semantic impairment arising from damage to representation[10–12] versus control[2,13] systems, and (3) key functional brain imaging results from the study of semantic cognition under conditions of control[14,15]. The rest of this paper reports such an analysis, arriving at a model of controlled semantic cognition that provides a unified account of these disparate phenomena.

Core functions for semantic cognition

Research in semantic cognition has largely focused on how we acquire representations that capture conceptual structure from sensory, motor, and linguistic inputs that do not transparently reflect such structure (i.e., conceptual abstraction). This ability is thought to arise from sensitivity to patterns of covariation in experience[16]. While birds vary wildly in appearance, they possess properties that covary: feathers, beaks, wings, flying ability, the name "bird," etc. Many otherwise competing theories agree that the human semantic system

detects and exploits such structure, representing items as conceptually similar when they share coherently-covarying properties, even if they differ in many other respects[6,17,18]. By this view, concepts reflect clusters in the high-order covariance structure of experience.

This idea becomes challenging, however, when one considers how the various properties purported to co-occur in experience are distributed across learning episodes. While birds typically possess the abilities to fly and lay eggs, those behaviours do not directly co-occur: a bird laying an egg is not flying, and vice versa[6]. Each experience provides only partial exposure to an item's properties. Moreover, such exposure can be highly context- and modality-specific; for instance, learning that birds have hollow bones only via verbal statements in science class. Thus, conceptual abstraction relies upon extracting the relevant covariances across many different episodes over time, each providing only limited, context-bound access to a subset of properties: the system must track sameness in kind across contexts and experiences to detect that the flying item observed in one episode is similar to an item labelled "bird" in another.

This requirement to form representations abstracted across items, modalities and contexts seems at odds with the second core function of semantic cognition, context-sensitivity. Context-appropriate behaviour requires flexible construction of task-relevant similarity structure, as different subsets of features and aspects of meaning are crucial in different contexts, whilst other often more dominant meanings must be inhibited[1,6]. For instance, representing a piano and a computer keyboard as similar when generating action plans, but dissimilar when generating inferences about their weight. This flexibility underlies construction of ad-hoc categories, e.g., 'things that fit in a pocket'[19], and the tendency for new learning to generalise differently depending upon the nature of the information[20]. Often the context-appropriate behaviour requires access to a particular subset of features within a modality. For instance, the features relevant for moving a piano conflict with those relevant to playing it; accessing both simultaneously may produce an action inappropriate in strength or nature[7]. Likewise, naming a piano requires a highly specific output differing from the information required to draw it or mime its action. Thus, conceptual abstraction and context-sensitivity may appear diametrically opposed by virtue of their treatment of context, yet the semantic system must achieve both concurrently.

These considerations highlight three core functions of the human semantic system that informed the design of our simulations:

(1) It must acquire representations that capture overall conceptual similarity structure and not merely the perceptual, motor, and linguistic structure apparent within various modalities.

(2) It must acquire context-independent conceptual representations from learning episodes that provide only partial context-specific information about an item's properties.

(3) It must adapt to context so as to generate only context-appropriate behaviours.

Proposals about the neural architecture of semantic representation

Hypotheses about the architecture of the semantic network abound, but few have been specified with sufficient mechanistic precision to compare and contrast their implications for understanding conceptual abstraction and context-sensitivity (though see[6]). The reverse-engineering approach we adopt, identifies a series of architectural 'building blocks' that distinguish various theories, then parametrically varies these to delineate a space of possible architectures that encompass some existing proposals, as well as hypotheses not typically considered. Formal comparison of the effects of each building block provides critical insight into hypotheses that have been articulated only verbally, and allows us to determine which possible model best supports both conceptual abstraction and context sensitivity.

To identify the building blocks, we first considered how contemporary hypotheses about the cortical architecture of semantics vary. One view derives from Wernicke's[9] proposal that semantic processing reflects direct, interactive communication amongst various associated surface representations—perceptual, motor, linguistic, affective, and so on. This perspective foreshadows modern 'embodied cognition' views[21,22] and has been explored computationally[23–26]. Other work emphasises the importance of multimodal 'hub' regions for mediating interactions between sensorimotor modalities. For instance, 'convergence zones' may connect different modalities within a network that adopts multiple hubs, such as a pathway connecting visual and linguistic representations, another connecting visual and haptic representations, and so on[27,28]. Such hubs could be the sole vehicle for crossmodal communication, or could connect via a broader multimodal region, producing a hierarchical convergence of information across modalities. Whilst neither proposal has been investigated computationally, both have received support from functional neuroimaging and neuropsychological evidence[27,29,30]. Alternatively, the various representational modalities might all communicate via a single multimodal hub, an idea supported by convergent computational[31,32], neuropsychological[1,10,11], neuroimaging[33,34], neurophysiological[35,36] and neurostimulation[37,38] evidence.

With this landscape in mind, we considered two factors governing how modality-specific "spokes" connect to multimodal "hubs". First, communication could be direct or involve one or more intermediating regions. Deeper models (possessing more layers between input and output) can acquire more complex representations and behaviours as attested by the recent explosion of research in deep neural networks[39,40], and conceptual representations arise within the deepest layers of visual and language networks[41–43]. Thus, depth seems an important factor to consider. Second, network layers may connect only to immediately adjacent areas, or may additionally send direct 'shortcut' connections to anatomically distal regions, an analogue of white matter pathways. Such connectivity would need to be sparse due to cortical metabolic and packing constraints[44,45].

From these considerations we discern five building blocks that may influence the behaviour of the cortical semantic network: a) presence of multiple hubs connecting subsets of modality-specific spokes, b) presence of a single multimodal hub, c) network depth, d) presence of shortcut connections, and e) hierarchical convergence across modalities. These building blocks combine in different ways to produce the space of candidate architectures

shown in Figure 1 (note, not all possible combinations are explored, only those relating to the theoretical literature and allowing assessment of the impact of each building block). Importantly, all architectures employ identical inputs/outputs, and possess equal numbers of hidden units and connections, differing only in the pattern of connectivity amongst units. We assess how each building block impacts conceptual abstraction, first without and then with simultaneous context sensitivity, and whether successful architectures provide insight into existing anatomical, neuropsychological and neuroimaging data.

Results

The models were fully recurrent neural networks with activity unfolding over time, trained in the same learning environment to the same performance criterion (all output units within 0.2 of their targets), differing only in their connectivity (Figure 1). We created activation patterns for 16 items in each of three 'modalities' (e.g., word, image, action) designed to capture the central challenge of conceptual abstraction: conceptual structure was latent in the relationship between unit activations across modalities but differed strongly from the structure apparent within each modality considered independently (Figure 2, see Methods). In an initial phase, models received input from a single modality and learned to reproduce the complete pattern across all three modalities (as in prior work[25,31,46]). In a second phase, a control signal was provided as an additional input to indicate different task contexts, and models learned to activate, from input in a single modality, only those output units both true of the item and relevant to the task. For context-sensitive simulations, the required output modality was designated by an additional input signal.

To measure success in conceptual abstraction we computed a 'true' conceptual similarity matrix by concatenating the vectors from all three modalities for each item and tabulating the resulting correlations for all item pairs. We used this as the target matrix for the model representations in a representational similarity analysis[47]: for each model hidden layer, we correlated the pairwise similarities in its activation patterns and the true conceptual similarities. The highest such correlation for each model indicated its overall success in abstracting conceptual structure (see Supplementary Notes 1 and 2 for further details). We considered how this 'conceptual abstraction score', and learning speed, varied with model architecture, with and without the additional requirement of context-sensitivity. We also assessed how well each model (1) captures the multimodal structure, ignoring the modality-specific structure, and (2) generalises newly-learned conceptual and modality-specific information across contexts—however since these results align fully with the conceptual abstraction scores, they are reported in Supplementary Notes 3-5. The simulations strongly favoured one architecture, which we assessed in its ability to explain core phenomena in anatomical, neuropsychological, and neuroimaging studies of controlled semantic cognition (Phase 3).

Phase 1: Conceptual representation without control

Consistent with prior work[6], phase 1 models learned to generate, from input provided to any individual modality, the full pattern associated with the concept across all modalities. The models varied dramatically in their conceptual abstraction scores ($F(6, 1273)=1168.575$,

p<.001; Figure 3, Table S1, significant contrasts had p<.001 throughout, Hedges g effect size and confidence intervals (CI) are reported). Scores were better when the architecture included some form of hub (Spokes-Only < Bimodal Hubs; t(718)=-38.763; g=3.064, CI =-.21383, -.19322) and better still with a multimodal hub (Bimodal Hubs < Shallow Multimodal Hub; t(718)=-46.634, g=2.877, CI=-.17007, -.15633). No evidence was found for an effect of depth (Shallow Multimodal Hub ~= Deep Multimodal Hub; t(105.648)=-0.645, p=.521, g=0.099, CI=-.01207, .00614) and hierarchical convergence had a significant detrimental effect compared to a single multimodal hub (Deep Multimodal Hub > Convergent Hubs; t(133.049)=22.503, g=3.558, CI=.16428, .19594; Multimodal Hub-plus-Shortcut > Convergent Hubs-plus-Shortcut; t(114.224)=11.436, g=1.808, CI=.06229, .08838). Addition of shortcut connections further improved performance for both deep architectures (Deep Multimodal Hub < Multimodal Hub-plus-Shortcut; t(138.57)=-5.444, g=0.861, CI=-.03815, -.01782; Convergent Hubs < Convergent Hubs-plus-Shortcut; t(158)=-14.748, g=2.332, CI=-.15054, -.11498). Thus, the Multimodal Hub-plus-Shortcut model performed better than all other architectures.

The number of training epochs required to reach the performance criterion varied by architecture (F(6, 98)=39.307, Figure 4). The Spokes-Only structure learned fastest, with the presence of a hub increasing the time taken (Spokes-Only < Bimodal Hubs; t(14.163)=-13.057, g=4.768, CI=-18409, -13219), yet a multimodal hub decreasing it somewhat (Bimodal Hubs > Shallow Multimodal Hub; t(15.735)=9.545, g=3.485, CI=9238, 14522). Depth slowed training dramatically (Shallow Multimodal Hub < Deep Multimodal Hub; t(14.846)=-10.408, g=3.801, CI=-22042, -14543), but this slowing diminished substantially with shortcut connections (Deep Multimodal Hub > Multimodal Hub-plus-Shortcut; t(20.653)=7.152, g=2.612, CI=9827, 17896; Convergent Hubs > Convergent Hubs-plus-Shortcut; t(15.581)=5.065, g=1.849, CI=9168, 22417). There was no evidence hierarchical convergence affected learning time (Deep Multimodal Hub ~= Convergent Hubs; t(28)=-0.622, p=.539, g=0.227, CI=-9326, 4983; Multimodal Hub-plus-Shortcut ~= Convergent Hubs-plus-Shortcut; t(28)=-0.213, p=.833, g=0.076, CI=-2557, 2076).

The efficient abstraction of context-independent conceptual structure depended critically on the presence of one multimodal hub, resulting in the largest effects of all contrasts. Depth alone did not improve representation quality and greatly increased training time, but adding shortcut connections produced the highest-quality representation whilst speeding learning somewhat. An interim conclusion from Phase 1 is that conceptual abstraction benefitted from a single multimodal hub.

Phase 2: Controlled semantic cognition

Phase 2 simulations addressed the full challenge of controlled semantic cognition - achieving context-independent conceptual abstraction when experiencing and generating a limited, context-sensitive subset of an item's properties per learning episode. Three additional context units were added, each coding the task-relevance of a modality. For instance, if modality 1 and 2 are important but modality 3 is not (e.g., picture naming requiring visual input and verbal output without action), context units 1 and 2 would be active. This Control Layer sent trainable unidirectional connections to all units, providing a

simple way of implementing control as an influence of the current context on the flow of activation through the network to generate task-appropriate representations and behaviours[6,48,49] (Figure 5.A.).

The models were trained to generate context-sensitive outputs from partial inputs for 16 items in each of 9 ‘tasks’, defined by specifying the relevant input/output modalities. Tasks could involve the same modality for both (e.g., word repetition), or one modality as input and another as output (e.g., picture naming). From the task representation and an item’s input features, the models learned to activate the item’s task-relevant features while keeping task-irrelevant features inactive. The models experienced a limited subset of an item’s features in any given training example, both in the inputs and outputs.

Conceptual abstraction score varied substantially by architecture ($F(6, 1673)=2326.016$, Figure 3, Table S2). Bimodal hubs improved performance (Spokes-Only < Bimodal Hubs; $t(611.94)=-29.141$, $g=1.719$, CI=-.07598, -.06639), but a multimodal hub performed still better (Bimodal Hubs < Shallow Multimodal Hub; $t(718)=-59.050$, $g=4.668$, CI=-.22886, -.21413). In contrast to Phase 1, depth significantly improved conceptual abstraction under conditions of control (Shallow Multimodal Hub < Deep Multimodal Hub; $t(92.161)=-14.049$, $g=2.486$, CI=-.16050, -.12074). Hierarchical convergence dramatically reduced conceptual abstraction (Deep Multimodal Hub > Convergent Hubs; $t(90.826)=37.123$, $g=5.984$, CI=.35030, .38991; Multimodal Hub-plus-Shortcut > Convergent Hubs-plus-Shortcut; $t(137.278)=15.495$, $g=4.753$, CI=.18296, .23648). Shortcut connections improved conceptual abstraction in both deep architectures (Deep Multimodal Hub < Multimodal Hub-plus-Shortcut; $t(148.947)=-11.548$, $g=1.826$, CI=-.16490, -.11671; Convergent Hubs < Convergent Hubs-plus-Shortcut; $t(87.551)=-26.027$, $g=4.638$, CI=-.32419, -.27819). Only the Multimodal Hub-plus-Shortcut architecture acquired representations significantly closer to the context-independent conceptual structure than the control structure (Supplementary Note 6).

Training time varied by architecture (Figure 4; $F(6,98)=113.036$), with effects mimicking those observed without control. The Spokes-Only architecture was fastest, with a bimodal hub leading to slowing (Spokes-Only < Bimodal Hubs; $t(14.504)=-15.720$, $g=5.752$, CI=-7530, -5371), and a multimodal hub reducing this (Bimodal Hubs > Shallow Multimodal Hub; $t(15.833)=9.424$, $g=3.441$, CI=3145, 4973). Depth significantly slowed learning (Shallow Multimodal Hub < Deep Multimodal Hub; $t(15.778)=-21.041$, $g=7.683$, CI=-10121, -8267), yet shortcut connections alleviated this effect (Deep Multimodal Hub > Multimodal Hub-plus-Shortcut; $t(28)=8.235$,$g=3.007$, CI=3937, 6545; Convergent Hubs > Convergent Hubs-plus-Shortcut; $t(28)=6.077$, $g=2.219$, CI=2759, 5565). There was no evidence that hierarchical convergence changed learning time (Deep Multimodal Hub ~= Convergent Hubs; $t(28)=1.802$, $p=.576$, $g=0.658$, CI=-146, 2280; Multimodal Hub-plus-Shortcut ~= Convergent Hubs-plus-Shortcut; $t(28)=-0.017$, $p=1$, $g=0.006$, CI=-1494, 1470).

In these simulations, context units connected to all units in the semantic network, with their influence shaped by learning. In the best-performing Multimodal Hub-plus-Shortcut architecture, learned weights from control to the hub were smaller in magnitude than those projecting to shallower hidden ($t(1517.275)=11.824$, $g=.507$, CI=.50724, .70901) and spoke

units (t(1512.273)=10.364, g=.445, CI=.43016, .63102) suggesting that control should operate on more superficial layers (Figure 5.B.). To test this, we compared models in which control connected only to Spokes, Hidden Layer 1, or Hidden Layer 2 units (Figure 5, Supplementary Note 7). Conceptual abstraction suffered when control operated on the multimodal hub compared to the spokes (t(127.713)=31.981, g=5.057, CI=.29835, .33770) or Hidden Layer 1 (t(149.191)=27.631, g=4.369, CI=.27904, .32202). There was no evidence these differed from one another (t(158)=2.089, p=.115, g=.330, CI=.00095, .03404). Moreover, control connectivity to just the spokes (t(138.78)=21.504, g=1.977, CI=.09496, .13063) or shallow hidden units (t(158)=9.493, g=1.501, CI=.07547, .11513) produced reliably better conceptual abstraction than control connecting to all layers, despite employing fewer connections; and only these models acquired internal representations significantly closer to the context-independent than context structure (Supplementary Note 6). There was no evidence that locus of control affected training time (F(2,42)=2.073, p=.139, η^2=.090, CI=-1247.957, 729.690; -1952.690, 24.957; -1693.557, 284.090). Thus, the reverse-engineering approach suggests that controlled semantic cognition, is best achieved within an architecture employing a single, deep multimodal hub and shortcut connections, with control systems acting on superficial rather than deep network components.

Phase 3: Accounting for empirical phenomena with the reverse-engineered model

The reverse-engineered model differs from other proposals in a variety of ways, raising two questions. First, how does its structure accord with existing evidence about the anatomy of the cortical semantic network? Second, does the model help to explain important behavioural and neural phenomena in the study of controlled semantic cognition? We assessed these questions in phase 3.

Anatomy

It is well known that the ventral ATL forms a multimodal conceptual hub, as demonstrated in SD[10–12], brain imaging[33,50–52], neurostimulation[37,38] and intracortical electrode recording[35,53]. Indeed, this observation motivated the original hub-and-spoke view of semantic representation[31]. Additionally, the progression from unimodal perceptual representations to multimodal conceptual representations occurs in a graded fashion across many cortical areas[51,54] corresponding to a deep network. The simulations establish that such an architecture better promotes conceptual abstraction in conditions of context-sensitivity than other possible arrangements (including popular multi-hub theories[27], see Discussion).

The reverse-engineered model suggests two additional properties that differ from prior models. First, it proposes sparse long-range “shortcut” connections connecting posterior modality-specific regions directly to the multimodal hub, in addition to region-to-region connectivity. Both varieties of white-matter connection may be seen within the temporal lobe in assessments of the inferior longitudinal fasciculus[55–57] and were highlighted within a detailed assessment of connectivity between anterior and posterior subsections of the fusiform gyrus[58]. Second, it suggests that neural systems of semantic control should connect with semantic regions primarily via more posterior regions distal to the anterior temporal

hub. Whilst the literature does not definitively answer this question, the core ventral ATL hub region does have few connections to distal regions[59,60]. A connection from control regions to shallower areas of the semantic network, is highly consistent with observations from functional neuroimaging that control demands act upon spoke representations[14]. Thus, the reverse-engineered model is in high accord with known anatomy and provides a testable hypothesis as to the structural connectivity between control regions and the anterior temporal hub.

Distinct neuropsychological syndromes

Damage to the anterior temporal hub versus frontal and temporoparietal control regions causes qualitatively distinct semantic syndromes, termed semantic dementia (SD) and semantic aphasia (SA) respectively[10,12,13]. Both produce comparably severe semantic deficits with frequent omissions in various tasks, but differ in errors of commission. Patients with SA more often generate context-inappropriate intrusions, producing associative errors ("acorn" for squirrel) and circumlocutions ("has stripes" for zebra) in naming, losing track of the target category in verbal fluency (e.g., for birds: robin, sparrow, chicken, pig), or failing to grasp a tool in a manner that affords its correct use in a given task context[2,13,61]. Patients with SD more often generate context-appropriate but semantically incorrect behaviours: committing coordinate (e.g., "horse" for zebra) or ordinacy (e.g., "animal" for squirrel) errors in naming, generating fewer but mainly correct items in verbal fluency, and grasping a tool correctly but exhibiting a semantically inappropriate use (e.g., brushing hair with a comb)[2,10,13,62]. Figure 6 shows these patterns for semantic fluency[13] and picture naming[63] from cohorts of each patient type studied in prior work. Does the reverse-engineered model explain these differences?

To answer this question, we simulated disordered control in SA by adding noise to the control unit activations, and degraded representation in SD by removing a proportion of connections to, from and within the multimodal hub[31]. We simulated increasing levels of damage for each syndrome, matched for severity (indexed by total number of errors) and compared the relative frequency of three error types: omission (inactivation of a correct features), context-appropriate (activation of an incorrect task-relevant feature), and intrusion errors (activation of a task-irrelevant feature).

Damage to control produced fewer context-appropriate errors (damage type; $F(1, 190)=1292.758$, $\eta^2=.540$, CI=-376.230, -332.070; damage level; $F(4,190)=128.784$, $\eta^2=.215$, CI=-441.830, -30.670; -207.980, -163.820; -127.180, -83.020; -33.780, 10.380; interaction; $F(4,190)=99.301$, $\eta^2=.166$) and more intrusion errors (damage type; $F(1, 190)=2194.628$, $\eta^2=.541$, CI=320.775, 378.925; damage level; $F(4,190)=168.893$, $\eta^2=.245$, CI=-54.975, 3.175; -44.425, 13.725; -38.575, 19.575; -31.825, 26.325; interaction; $F(4,190)=168.893$, $\eta^2=.167$) than damage to representation (across all damage levels, see Supplementary Note 8). There was no evidence for differences in feature omissions, with frequency reflecting damage severity (damage type; $F(1,190)=0.613$, $p=.435$, $\eta^2=.000$, CI=-21.340, 26.440; damage level; $F(4,190)=440.445$, $\eta^2=.901$, CI=-326.690, -278.910, 134.590, -86.810; 41.090, 6.690; -31.090, 16.690; interaction; $F(4,190)=0.570$, $p=0.685$, $\eta^2=.001$). The reverse-engineered architecture accounts for the qualitatively different

patterns of impaired semantic cognition arising from damage to control versus representational elements of the system identified in the patient data (Figure 6).

Functional brain imaging

In classic neuroimaging experiments, participants viewed a stimulus (word or picture) and retrieved either the item's colour or its action[14]. An identical stimulus elicited different functional activation depending on the task: engaging regions just anterior to colour perception for colour retrieval and motion perception for action retrieval (Figure 7.a., also see[64]). To see whether the reverse-engineered model explains this effect, we contrasted activation for each hidden and output unit across two tasks using the same input (e.g., a 'word' in modality 1) but differing outputs (e.g., a 'colour' in modality 2 or an 'action' in modality 3). In both the Spoke Layer and Hidden Layer 1, the 'retrieve colour' task activated 'colour' units more, while 'retrieve action' activated 'action' units more (Figure 7.b., independent-samples t-tests per unit, all Bonferroni-corrected ps<.05). Consistent with the imaging, no evidence of differential activation was observed in the hub, the input modality spoke or its associated hidden units.

Functional Connectivity

Recent evidence suggests functional connectivity between the ATL hub and modality-specific regions changes depending upon the information required for a task[15,65]. In one fMRI study participants judged social status or traits of faces[15]. Whilst ATL connectivity to the fusiform face area (i.e., the input spoke) was stable, functional connectivity to spokes associated with status- (IPL) or trait- (PCC) processing differed by task (Figure 7.c.). In the reverse-engineered model, we assessed the change in functional connectivity between the hub and spokes for two conditions with varying output requirements. Stimuli were always presented in modality 1 ('faces'), but output was either in modality 2 ('status') or modality 3 ('trait'). T-tests contrasted the correlation strengths of the time series of the hub and each spoke region between contexts. There was no evidence for differential hub connectivity with the input spoke across contexts (t(158)=-0.568, p=1, g=.090, CI=-.08856, .04900), but connectivity to the two output spokes varied significantly (status *vs.* trait; Modality 2 'status' spoke; t(158)=3.659, p=.001, g=.578, CI=.06227, .20837; Modality 3 'trait' spoke; t(158)=-3.030, p=.009, g=.479, CI=-.18139, .03819). Thus, despite stable physical connectivity, task context effects on unit activations account for dynamic functional connectivity.

Discussion

We applied a reverse-engineering approach to discover a neural network architecture capable of achieving the core, opposing functions of controlled semantic cognition: conceptual abstraction across modalities and contexts with simultaneous context-sensitivity. The optimal network had four important characteristics: (1) a multimodal hub only, (2) a deep architecture, (3) sparse shortcut connections, and (4) control operating on shallow rather than deep network components. The reverse-engineered model subsequently accounted for several disparate phenomena in controlled semantic cognition including the coarse anatomy of the temporal cortex, qualitative differences in error patterns observed in SD vs. SA,

differential functional activation to the same stimulus depending on the task, and task-dependent shifts in functional connectivity between the ATL hub and sensory regions. In this discussion we consider why these architectural elements may be critical for conceptual abstraction and context-sensitivity, and implications for theories of controlled semantic cognition.

Why a multimodal hub only?

Prior work established that feedforward neural networks exploit shared structure across modalities and contexts only when information from each gets passed through the same units and weights somewhere in the network, termed the 'convergence principle'[6,31]. Here we show that convergence remains critical for learning structure in more neurobiologically-plausible recurrent networks: architectures lacking a single multimodal hub can learn the same input/output mappings, but do not acquire internal representations reflecting the full conceptual representational structure across modalities and learning episodes. Even models possessing a multimodal hub fail to learn the desired structure if they also possess shallower and more direct pathways between modalities as there is little pressure to use the connections mediating all modalities. These findings are problematic for distributed-only[9,66] and multi-hub[27,28] theories of semantic cognition, but consistent with the hub and spoke theory[31]. The reverse-engineered model merges the hub-and-spoke model with the controlled semantic cognition framework[1,31] with the additional constraints of depth (also instantiated in[41]), shortcut connections and a shallow interface between control and representation systems.

Why should depth help?

Deep networks can acquire complex internal representations that generalise well when trained on large corpora of naturally occurring stimuli[67]. Yet, only when required to generate context-sensitive outputs did the deeper model outperform the shallow model. Thus, depth particularly facilitates the ability to discover representational structure when learning involves experience with limited, context-dependent inputs and outputs. Context-sensitive training pressures the system to represent the same item differently in different contexts, making it difficult for the system to exploit feature covariance across contexts. If the multimodal hub connects directly to unimodal representations, context strongly influences the representations. Likewise, a deep model in which the multimodal hub directly receives context inputs acquires context-bound representations. Only when the model is deep and control operates on the shallower elements is the hub sufficiently insulated from contextual information to acquire more context-invariant representations.

Why shortcut connections?

Deep networks initially learn slowly due to 'vanishing' or 'exploding' gradients[68]: with many weights intervening between input and output (and little initial differentiation between inputs), changes to earlier weights may have negligible impact, so error-driven learning produces minimal (or inordinately large) weight changes[6,69]. Even when sparse, shortcut connections significantly remediate this problem by propagating error through fewer layers to learn more quickly, increasing pattern differentiation and speeding overall learning. They also produce a concomitant improvement in conceptual abstraction, perhaps by roughly

approximating the core structure in the environment or 'warming up' the deep hub early in a trial. A similar cortical mechanism may be in play, with an early feedforward sweep bringing the hub online and generating approximately correct states, followed by continued, iterative interaction between hub and spoke regions[70,71].

Accounting for distinct semantic syndromes

Simulated damage to control versus hub representations produced qualitatively similar damage to the error pattern found across SD and SA: equivalent reductions in accuracy but more context-inappropriate intrusions following control damage and context-appropriate errors following representation damage. In the intact model, control can be viewed as "selecting" which properties matter for the task, potentiating context-relevant while suppressing context-irrelevant properties[49]. Context-sensitive responding arises from the joint influence of representational and control systems on surface properties. Distortion of the control signal incorrectly potentiates context-irrelevant units, allowing them to produce context-inappropriate behaviour. With damaged representations the intact control signal only potentiates context-appropriate features, but distorted feedback from the hub activates the wrong features within this subset, producing context-appropriate but semantically incorrect behaviours.

Why separate systems for representation and control?

A broad literature in neuropsychology, functional neuroimaging[2,72,73], and connectivity[74,75] suggests semantic representation and semantic control are supported by the interaction of anatomically and functionally segregated systems. The current work suggests why this might be. The hub-and-spoke theory has long suggested that the anatomy of the temporal lobe promotes the extraction of conceptual structure across modalities and time in the multimodal hub. The current work extends the set of anatomical features critical to support this function alongside the additional constraint of context-sensitivity. This ability is compromised when the hub region is strongly influenced by the immediate task context. Perhaps the gross segregation of systems for representation *vs.* control is evolution's way of promoting acquisition of deep conceptual representations while preserving the flexibility required to think and act as the situation demands.

Method

Model Environment & Control

Each concept consisted of 12 features in each of 3 modalities (M1, M2 and M3; see Figure 2). Concepts were constructed based on a critical aspect of conceptual structure; unimodal perceptual structures only weakly correlate with the conceptual structure which is more predictive but requires extraction across modalities. The model environment included four orthogonal structures; one distinct unimodal (based on 5 perfectly correlated or anti-correlated features within a single modality) structure per modality (unimodal M1, unimodal M2 and unimodal M3) and a multimodal (based on 12 highly correlated features spread across all three modalities) structure. In each modality the unimodal structure is greater, yet overall, the multimodal structure is stronger. Whilst the main analyses focus on the full structure, highly consistent results are displayed for the unimodal and multimodal structures

in Supplementary Notes 3 and 4. Input was always in a single modality, with the other two modalities set to 0. For simulations without control, the target was the full concept, resulting in 48 versions of the 16 examples. For simulations with control, each concept was presented in one of three modalities with a control signal designating a required output in one of three modalities (as well as the input modality), resulting in 144 versions of the 16 examples. Task-irrelevant modalities had targets of 0.

Model Architecture

Code for replicating all simulations is available in the Supplementary Materials and at https://github.com/JacksonBecky/reverse-engineered-semantics. Additional code to process the results is available online. All architectures utilised a single framework, consisting of 12 pairs of input and output units per modality (connected on a one-to-one basis with a frozen weight of 6 and a fixed bias of -3 for the output units), 60 hidden units and 3132 bidirectional connections with learnable weights. All learnable weights were initialised using the default LENS command resulting in small random weights (mean = 0, range =1). All hidden units employed a sigmoidal nonlinearity, scaling their activity between 0 and 1, in keeping with prior explorations of semantic representation[6]. Matching the number of resources allowed clear interpretation of the differences between the architectures. All architectures had connections between the three modality-specific regions of the Spokes Layer and the six subsections of Hidden Layer 1 (with two subsections connected to each modality-specific spoke region) and within each portion of Hidden Layer 1. Deep architectures had connections from Hidden layer 1 to Hidden Layer 2 and within Hidden Layer 2. Whilst it may be noted that the modality-specific input-output regions are not technically ‘spokes’ without a hub, these sensorimotor regions are referred to as such across all the architectures for consistency. To match the number of connections between architectures, some connections were sparse (see Supplementary Method 1). Two factors varied between the 7 model architectures; the hidden layer configuration (shallow; a single layer of 60 units *vs*. deep; one layer of 42 units and a deeper layer of 18 units) and the presence or absence of four types of connections (Direct Spoke Connections; connections between modality-specific output units in the Spokes Layer; Bimodal Hub Connections; connections between pairs of Hidden Layer 1 regions that receive different modalities of input, resulting in the formation of bimodal hubs; Multimodal Hub Connections; connections between hidden units to form a single multimodal hub, either within Hidden Layer 1 or Hidden Layer 2; Shortcut Connections; direct but sparse shortcut connections between the Spokes Layer and Hidden Layer 2 that bypass Hidden Layer 1). Although none of these models are of the depth typically associated with deep neural networks, employing two *vs*. one hidden layers reflects a great relative increase in depth and the term ‘deep’ is used here in the relative sense to distinguish the shallow and relatively deeper architectures. Whilst long-range connections are likely to be relatively sparse, their precise sparsity is not known. Shortcut Connections were included at a sparse but non-trivial proportion of 1 in 24 (although see Supplementary Note 9 for an assessment of systematically varying the sparsity within the *Multimodal Hub-plus-Shortcut* architecture).

Figure 1 represents each architecture. Three architectures were constructed from the shallow configuration; a ‘Spokes-Only’ architecture employing Direct Spoke Connections only, a

'Bimodal Hubs' architecture with Bimodal Hub Connections only and a 'Shallow Multimodal Hub' architecture with Multimodal hub Connections only, resulting in Hidden Layer 1 forming a single multimodal hub. All four deep architectures have Multimodal Hub Connections resulting in a multimodal hub in Hidden Layer 2. The 'Deep Multimodal Hub' architecture has no additional connections, thus three modality-specific routes connect via a deep multimodal hub. The 'Multimodal Hub-plus-Shortcut' architecture also included Shortcut Connections and the 'Convergent Hubs' architecture included additional Bimodal Hub connections, resulting in hierarchical convergence as multiple bimodal hubs connect to a single deep multimodal hub. The 'Convergent Hubs-plus-Shortcut' architecture combined the Bimodal Hub Connections, Multimodal Hub Connections and Shortcut Connections. The seven architectures allowed contrasts separating the effect of each architectural feature; the effects of a hub (Spokes-Only vs. Bimodal Hubs), a multimodal hub (Bimodal Hubs vs. Shallow Multimodal Hub), depth (Shallow Multimodal Hub vs. Deep Multimodal Hub), shortcut connections (Deep Multimodal Hub vs. Multimodal Hub-plus-Shortcut and Convergent Hubs vs. Convergent Hubs-plus-Shortcut) and hierarchical convergence (Deep Multimodal Hub vs. Convergent Hubs and Multimodal Hub-plus-Shortcut vs. Convergent Hubs-plus-Shortcut).

In Phase 2, a 'Control Layer' consisting of three units (each corresponding to one modality) was added to provide a context signal. The models had unidirectional learnable connections from the control units to the Spokes Layer, Hidden Layer 1 and Hidden Layer 2 (where present). Initially, no assumptions were made as to where control should connect, allowing a fair comparison across architectures. Following this analysis, the emergent reliance on the connections to each layer was investigated using an equal number of connections to all layers (81 per layer if shallow, 54 per layer if deep). Then, the effectiveness of this emergent pattern was verified by contrasting versions of the model where the Control Layer was connected to each single layer (with the same number of connections).

Training Parameters

The models were constructed and trained using the Light Efficient Network Simulator (LENS, version 2.63) software [76]. Each simulation employed a fully recurrent network with 24 activity updates per example (6 time intervals and 4 ticks per time interval). Inputs were presented for the first 3 time intervals. Each training batch consisted of all examples presented once in a random order. At the end of each batch, error derivatives were calculated and all weights in the model adjusted by a small amount. All simulations employed the same training parameters, found to allow learning in pilot simulations. The models were trained using gradient descent with a learning rate of 0.001 and a weight decay parameter of 0.0001 with no momentum. Training ended when all output feature units were within 0.2 of their target. Thus, all architectures were matched on accuracy. Analyses were performed using the final time step of a test trial. Each simulation was performed 80 times. No power analysis was used to determine this sample size, however, it is much higher than typical modelling simulations (e.g.[41]).

Europe PMC Funders Author Manuscripts

Assessment Metrics

Data processing was performed in MATLAB and statistics in the Statistical Package for the Social Science (SPSS, 2013). The similarity structure of the models representations were compared to the ground-truth similarity structure to determine each architectures ability to accurately discover and represent the full structure in the environment. The critical example structure used to form this conceptual abstraction score is the 'context-independent' semantic representation structure (the relationships between examples based on the full set of features regardless of the current input or output domain). For the simulations with control it is also possible to look at the similarity of the representations to the context signal (context-only) or the full structure varying by context and concept (context-sensitive), see Supplementary Note 10.

Correlation-based similarity matrices were constructed from the activity in a model region across all examples after learning. Model regions were defined as portions of the model with the same potential connections (before sparsity is taken in to account) as connectivity constrains function[44,78,79]. This resulted in 3 Hidden Layer 1 regions in the Spokes-Only, Shallow Multimodal Hub, Deep Multimodal Hub and Multimodal Hub-plus-Shortcut architectures and 6 in the Bimodal Hubs, Convergent Hubs and Convergent Hubs-plus-Shortcut architectures. The similarity between each result-based similarity matrix and the example-based similarity matrix was determined using a correlation. This resulted in a value per model run and layer subregions for statistical comparisons, although these equivalent values are averaged when reported. The values for the region with the highest similarity to the context-independent semantic representation were used to contrast the models (although for comparison of all regions and consideration of the effect of the number of units see Supplementary Note 1). Additionally, the number of epochs taken to train each architecture to criterion was determined for 15 runs of each model. For the simulations with and without control, a repeated measures ANOVA assessed the differences between the 7 architectures and *a priori* two-sided between-samples t-tests (with Levene's tests for equality of variance) were used to compare the effect of each architectural feature with Bonferroni correction for the seven multiple comparisons. All p values for significant contrasts are below .001 unless specified otherwise. As a difference of any magnitude may reach significance with a sufficient number of observations, we complement the statistical analyses by reporting an Hedges g (a measure of effect size for t-tests that is weighted by sample size) or eta squared (for ANOVAs) effect size and 95% confidence intervals.

To determine how the Control Layer should connect to the rest of the model, two assessments were used. Firstly, 40 models were ran with connectivity to each layer. The emergent preference for receiving and employing the control signal in each layer was examined by contrasting the sum absolute magnitude of the weights to each layer using 3 t-tests (Spokes vs. Hidden Layer 1, Spokes vs. Hidden Layer 2, Hidden Layer 1 vs. Hidden Layer 2) in the deep architectures and one (Spokes vs. Hidden Layer 1) in the shallow architectures. Bonferroni multiple comparison correction for 3 contrasts was applied to the deep architectures. Secondly, the effectiveness of this emergent pattern was verified by only connecting the models to one layer (either the Spokes Layer, Hidden Layer 1 or, where possible, Hidden Layer 2). These model versions were compared on their extraction of the

context-independent representation structure using two-sided between-samples t-tests and Bonferroni correction applied.

Lesioning the Model

To assess the effect of lesions to representation and control regions, the models were constructed and trained using the optimal architecture identified within Phase 2 (including connections from control to the Spokes Layer only). To damage representational processes, the connections to, from and within Hidden Layer 2 were removed as this region had the greatest conceptual abstraction score and thus, showed the greatest specialisation for representation processes. In the Control Damage simulations, Gaussian noise was added to the input to the control units (this noise was stable across a trial and varied between trials and model runs). This addition of noise was intended to simulate damage within the control system that produces this signal. Thus, a different mechanism of damage was employed to simulate a similar effect within the control system proper and the representation system (see Figure 6.A.). Three types of errors may be made; omission of a feature that is correct both for that concept and that context, commission of a feature that is in the correct context but incorrect for that concept and commission of a feature in the incorrect context. To allow comparison across damage type, controlling for the effect of damage severity, each simulation was performed at a variety of levels with the proportion of weights removed (for Representation Damage) or the amount of noise added (for Control Damage simulations) varied systematically. Then, points at which the number of errorful features (those further than 0.2 from the correct output) were matched across the damage types were identified. At the chosen levels, t-tests showed the three damage types did not have significantly different numbers of errors (each $p>.25$). This resulted in the identification of four damage levels at which the effect of damage type on the three possible error types could be assessed; Representation Damage with the removal of connections at proportions of 0, 0.1, 0.25, 0.3 and 0.35, and Control Damage with Gaussian noise added to the control signal with ranges of 0, 0.625, 1, 1.25 and 1.375. For each error type (Correct Feature-Type Commission, Incorrect Feature-Type Commission, Omission) an ANOVA was performed to assess the effects of damage type (Representation Damage, Control simulations) and damage level (No damage, Level 1, Level 2, Level 3, Level 4). Error types were compared across the damage types using two-sided independent samples t-tests at each level. As the proportion of errors of each type is highly similar across damage levels, only Level 3 is presented in Figure 6. The full pattern of results across damage levels is provided in Supplementary Note 8. The simulation data were compared to item-level error patterns in picture naming and fluency tasks. The picture naming data were previously published by Jefferies & Lambon Ralph[13] and included 10 patients with SA and 10 with SD. Intrusion errors are associative, and context-appropriate errors are all other semantic errors (including category coordinate and superordinate errors). Intrusion, context-appropriate and omission errors are provided as a proportion of these errors, excluding phonological errors and perseverations. The category fluency data were previously presented in Rogers et al.[63] (without the present split of intrusion and context-appropriate errors). The data includes responses from 7 SD and 8 SA patients to 8 basic categories (e.g., birds). Omissions are based on comparison to the average correct responses of 16 age-matched neurologically-intact control participants. Intrusions include semantic associates, responses to a prior category and unrelated words. Context-

appropriate errors include concepts from similar categories and specific-level responses. All errors are shown as a proportion of these semantic and omission errors, excluding phonological errors and repetitions (which could have a non-semantic cause).

Simulating Dynamic Changes in Activation and Functional Connectivity of the Semantic Network

Martin et al.,[14] performed H^2O_{15}-PET on 12 participants viewing word stimuli and a further 12 participants viewing line drawings. To simulate the univariate activation differences they identified, two conditions were contrasted using the same data presented in Phase 2 – a 'colour' context (modality 1 input and modality 2 output) and an 'action' context (modality 1 input and modality 3 output). An independent-samples t-test was used to compare activation in each output or hidden unit between the two conditions. The p-values presented are Bonferroni-corrected for the number of units contrasted.

The differential connectivity of hub and spoke regions based on varying output requirements in Wang et al.[15] was simulated in the reverse-engineered model. Simulations were identical to Phase 2, except for the addition of a negative bias of -4 on each hidden unit to simulate the metabolic cost of activating neurons, as in prior imaging simulations[41]. The model was ran 80 times and activity at the final time point of each trial in context 1 (modality 1 'face' input, modality 2 output, or 'status') and context 2 (modality 1 'face' input, modality 3 output, or 'trait') was concatenated in a different random order per model run to create a time series for each voxel, per context. Each run of the model is treated as a different participant. To collapse across units within a region, a PCA was performed per region for each context in each run, analogous to extracting an ROI time course for a psychophysiological interaction analysis as in Wang et al.[15]. The correlation between the time course in Hidden Layer 2 and each spoke region was calculated and (as a PCA result is equivalent to its reverse) the absolute value of this correlation taken as a measure of the functional connectivity of these regions in this context for this model run. The correlation values for each run were compared between context 1 and 2 for each pair of regions using an independent-samples t-test to assess whether there was a significant change in the connectivity of the hub and a spoke between the two contexts. The p-values were Bonferroni-corrected for the three connections assessed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by a British Academy Postdoctoral Fellowship awarded to RLJ (pf170068), a programme grant to MALR and TTR from the Medical Research Council (grant number MR/R023883/1), an Advanced Grant from the European Research Council to MALR (GAP: 670428) and Medical Research Council intramural funding (MC_UU_00005/18). The funders had no role in the conceptualisation, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Data Availability

Data is available upon request or can be generated using the code provided.

Europe PMC Funders Author Manuscripts

Code Availability

Code for replicating all simulations is available in the Supplementary Materials and online at https://github.com/JacksonBecky/reverse-engineered-semantics. Code for further analysis is available online.

References

1. Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT. The neural and computational bases of semantic cognition. Nat Rev Neurosci. 2017; 18:42–55. [PubMed: 27881854]
2. Jefferies E. The neural basis of semantic cognition: Converging evidence from neuropsychology, neuroimaging and TMS. Cortex. 2013; 49:611–625. DOI: 10.1016/j.cortex.2012.10.008 [PubMed: 23260615]
3. Abel TJ, et al. Direct physiologic evidence of a heteromodal convergence region for proper naming in human left anterior temporal lobe. J Neurosci. 2015; 35:1513–1520. [PubMed: 25632128]
4. Wittgenstein, L. Philosophical Investigations. Blackwell Publishing; 1953.
5. Lambon Ralph MA, Sage K, Jones RW, Mayberry EJ. Coherent concepts are computed in the anterior temporal lobes. Proc Natl Acad Sci U S A. 2010; 107:2717–2722. [PubMed: 20133780]
6. Rogers, TT, McClelland, JL. Semantic cognition: A parallel distributed processing approach. MIT Press; 2004.
7. Saffran EM. The organization of semantic memory: In support of a distributed model. Brain Lang. 2000; 71:204–212. [PubMed: 10716846]
8. Thompson-Schill SL, D'Esposito M, Aguirre GK, Farah MJ. Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. Proc Natl Acad Sci U S A. 1997; 94:14792–14797. [PubMed: 9405692]
9. Eggert, GH. Wernicke's works on aphasia: A sourcebook and review. Mouton; 1977.
10. Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 2007; 8:976–987. DOI: 10.1038/nrn2277 [PubMed: 18026167]
11. Acosta-Cabronero J, et al. Atrophy, hypometabolism and white matter abnormalities in semantic dementia tell a coherent story. Brain. 2011; 134:2025–2035. [PubMed: 21646331]
12. Warrington EK. Selective impairment of semantic memory. Quarterly Journal of Experimental Psychology. 1975; 27:635–657. DOI: 10.1080/14640747508400525
13. Jefferies E, Lambon Ralph MA. Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. Brain. 2006; 129:2132–2147. DOI: 10.1093/brain/awl153 [PubMed: 16815878]
14. Martin A, Haxby JV, Lalonde FM, Wiggs CL, Ungerleider LG. Discrete cortical regions associated with knowledge of color and knowledge of action. Science. 1995; 270:102–105. DOI: 10.1126/science.270.5233.102 [PubMed: 7569934]
15. Wang Y, et al. Dynamic neural architecture for social knowledge retrieval. Proc Natl Acad Sci U S A. 2017; 114:E3305–E3314. [PubMed: 28289200]
16. Rosch E, Mervis CB, Gray W, Johnson D, Boyes-Braem P. Basic objects in natural categories. Cogn Psychol. 1976; 8:382–439.
17. Murphy GL, Medin DL. The role of theories in conceptual coherence. Psychol Rev. 1985; 92:289–316. [PubMed: 4023146]
18. Keil, FC. The Jean Piaget Symposium series. The epigenesis of mind: Essays on biology and cognition. Carey, S, Gelman, R, editors. Lawrence Erlbaum Associates, Inc.; 1991. 237–256.
19. Barsalou LW. Perceptual symbol systems. Behavioral and Brain Sciences. 1999; 22:577–660.
20. Gelman SA, Leslie SJ, Was AM, Koch CM. Children's interpretations of general quantifiers, specific quantifiers and generics. Language, Cognition and Neuroscience. 2015; 30:448–461.
21. Martin A, Chao LL. Semantic memory and the brain: structure and processes. Current Opinion in Neurobiology. 2001; 11:194–201. [PubMed: 11301239]

22. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 2016; 532:453–458. [PubMed: 27121839]
23. McCrae K, de Sa VR, Seidenberg MS. On the nature and scope of featural representations of word meaning. Journal of Experimental Psychology: General. 1997; 126:99–130. [PubMed: 9163932]
24. Lambon Ralph MA, McClelland JL, Patterson K, Galton CJ, Hodges JR. No right to speak? The relationship between object naming and semantic impairment: Neuropsychological abstract evidence and a computational model. J Cogn Neurosci. 2001; 13:341–356. [PubMed: 11371312]
25. Farah MJ, McClelland JL. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. Journal of Experimental Psychology: General. 1991; 120:339–357. [PubMed: 1837294]
26. Devereux BJ, Clarke A, Tyler LK. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific Reports. 2018; 8
27. Binder JR, Desai RH. The neurobiology of semantic memory. Trends Cogn Sci. 2011; 15:527–536. DOI: 10.1016/j.tics.2011.10.001 [PubMed: 22001867]
28. Damasio H, Grabowski TJ, Tranel D, Hichwa RD, Damasio AR. A neural basis for lexical retrieval. Nature. 1996; 380:499–505. [PubMed: 8606767]
29. Damasio, AR, Damasio, H. Computational neuroscience. Large-scale neuronal theories of the brain. Koch, C, Davis, JL, editors. MIT Press; 1994. 61–74.
30. Mahon BZ, Caramazza A. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J Physiol-Paris. 2008; 102:59–70. [PubMed: 18448316]
31. Rogers TT, et al. Structure and deterioration of semantic memory: A neuropsychological and computational investigation. Psychol Rev. 2004; 111:205–235. DOI: 10.1037/0033-295x.111.1.205 [PubMed: 14756594]
32. Lambon Ralph MA, Lowe C, Rogers TT. Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. Brain. 2007; 130:1127–1137. [PubMed: 17438021]
33. Binney RJ, Embleton KV, Jefferies E, Parker GJM, Lambon Ralph MA. The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: Evidence from a novel direct comparison of distortion-corrected fMRI, rTMS, and semantic dementia. Cereb Cortex. 2010; 20:2728–2738. DOI: 10.1093/cercor/bhq019 [PubMed: 20190005]
34. Visser M, Jefferies E, Embleton KV, Lambon Ralph MA. Both the middle temporal gyrus and the ventral anterior temporal area are crucial for multimodal semantic processing: Distortion-corrected fMRI evidence for a double gradient of information convergence in the temporal lobes. J Cogn Neurosci. 2012; 24:1766–1778. [PubMed: 22621260]
35. Shimotake A, et al. Direct exploration of the ventral anterior temporal lobe in semantic memory: Cortical stimulation and local field potential evidence from subdural grid electrodes. Cereb Cortex. 2014
36. Matsumoto R, et al. Functional connectivity in the human language system: a cortico-cortical evoked potential study. Brain. 2004; 127:2316–2330. DOI: 10.1093/brain/awh246 [PubMed: 15269116]
37. Pobric G, Jefferies E, Lambon Ralph MA. Anterior temporal lobes mediate semantic representation: Mimicking semantic dementia by using rTMS in normal participants. Proc Natl Acad Sci U S A. 2007; 104:20137–20141. DOI: 10.1073/pnas.0707383104 [PubMed: 18056637]
38. Pobric G, Jefferies E, Lambon Ralph MA. Amodal semantic representations depend on both anterior temporal lobes: Evidence from repetitive transcranial magnetic stimulation. Neuropsychologia. 2010; 48:1336–1342. DOI: 10.1016/j.neuropsychologia.2009.12.036 [PubMed: 20038436]
39. Krizhevsky, A, Sutskever, I, Hinton, GE. Advances in Neural Information Processing Systems. Nevada, U.S.A: 2012. 1097–1105.
40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. 2015
41. Chen L, Lambon Ralph MA, Rogers TT. A unified model of human semantic knowledge and its disorders. Nature Human Behaviour. 2017; 1

42. Kriegeskorte N. Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science. 2015; 1:417–446.
43. Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. A task-optimized neural network replicates human auditory behaviour, predicts brain responses, and reveals a cortical processing hierarchy. Neuron. 2018; 98:630–644. [PubMed: 29681533]
44. Plaut DC. Graded modality-specific specialisation in semantics: A computational account of optic aphasia. Cogn Neuropsychol. 2002; 19:603–639. DOI: 10.1080/02643290244000112 [PubMed: 20957556]
45. Nelson ME, Bower JM. Brain maps and parallel computers. Trends in Neurosciences. 1990; 13:403–408. [PubMed: 1700511]
46. McNorgan C, Reid J, McRae K. Integrating conceptual knowledge within and across representational modalities. Cognition. 2011; 118:211–233. [PubMed: 21093853]
47. Kriegeskorte N, Mur M, Bandettini P. Representational Similarity Analysis - Connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience. 2008; 2
48. Dilkina K, Lambon Ralph MA. Conceptual structure within and between modalities. Front Hum Neurosci. 2013; 31:1–15. DOI: 10.3389/fnhum.2012.00333
49. Cohen JD, Dunbar K, McClelland JL. On the control of automatic processes: A parallel distributed processing account of the Stroop effect. Psychol Rev. 1990; 97:332–361. [PubMed: 2200075]
50. Visser M, Embleton KV, Jefferies E, Parker GJ, Lambon Ralph MA. The inferior, anterior temporal lobes and semantic memory clarified: Novel evidence from distortion-corrected fMRI. Neuropsychologia. 2010; 48(46):1689–1696. DOI: 10.1016/j.neuropsychologia.2010.02.016 [PubMed: 20176043]
51. Rice GE, Hoffman P, Lambon Ralph MA. Graded specialization within and between the anterior temporal lobes. Annals of the New York Academy of Sciences. 2015; 1359:84–97. [PubMed: 26502375]
52. Halai A, Welbourne S, Embleton KV, Parkes L. A comparison of dual-echo and spin-echo fMRI of the inferior temporal lobe. Human Brain Mapping. 2014; 35:4118–4128. [PubMed: 24677506]
53. Chen Y, et al. The 'when' and 'where' of semantic coding in the anterior temporal lobe: temporal representational similarity analysis of electrocorticogram data. Cortex. 2016; 79:1–13. [PubMed: 27085891]
54. Marinkovic K, et al. Spatiotemporal dynamics of modality-specific and supramodal word processing. Neuron. 2003; 38:487–497. [PubMed: 12741994]
55. Herbet G, Zemmoura I, Duffau H. Functional anatomy of the inferior longitudinal fasciculus: From historical reports to current hypotheses. Frontiers in Neuroanatomy. 2018; 12 doi: 10.3389/fnana.2018.00077
56. Catani M, Jones DK, Donato R, ffytche DH. Occipito-temporal connections in the human brain. Brain. 2003; 126:2093–2107. DOI: 10.1093/brain/awg203 [PubMed: 12821517]
57. Bajada CJ, Banks BA, Lambon Ralph MA, Cloutman LL. Reconnecting with Joseph and Augusta Dejerine: 100 years on. Brain. 2017; 140:2752–2759. [PubMed: 28969389]
58. Bouhali F, et al. Anatomical connections of the Visual Word Form Area. J Neurosci. 2014; 34:15402–15414. [PubMed: 25392507]
59. Binney RJ, Parker GJM, Lambon Ralph MA. Convergent connectivity and graded specialization in the rostral human temporal lobe as revealed by diffusion-weighted imaging probabilistic tractography. J Cogn Neurosci. 2012; 24:1998–2014. [PubMed: 22721379]
60. Jung J, Cloutman L, Binney RJ, Lambon Ralph MA. The structural connectivity of higher order association cortices reflects human functional brain networks. Cortex. 2016; 97:221–239. [PubMed: 27692846]
61. Morton, J, Patterson, K. Deep dyslexia. Patterson, K, Coltheart, M, Marshall, JC, editors. Routledge & Kegan Paul; 1980.
62. Bozeat S, Lambon Ralph MA, Patterson K, Garrard P, Hodges JR. Non-verbal semantic impairment in semantic dementia. Neuropsychologia. 2000; 38:1207–1215. [PubMed: 10865096]
63. Rogers TT, Patterson K, Jefferies E, Lambon Ralph MA. Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. Neuropsychologia. 2015; 76:220–239. [PubMed: 25934635]

64. Kuhnke P, Kiefer M, Hartwigsen G. Task-dependent recruitment of modality-specific and multimodal regions during conceptual processing. Cereb Cortex. 2020; 30:3938–3959. [PubMed: 32219378]

65. Chiou R, Humphreys GF, Jung J, Lambon Ralph MA. Controlled semantic cognition relies upon dynamic and flexible interactions between the executive 'semantic control' and hub-and-spoke 'semantic representation' systems. Cortex. 2018; 103:100–116. [PubMed: 29604611]

66. 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. 2016; 23:979–990. [PubMed: 25968087]

67. Bengio, Y, Delalleau, O. International Conference on Algorithmic Learning Theory. Kivinen, J, Szepesvári, C, Ukkonen, E, Zeugmann, T, editors. Springer; 18–36.

68. Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998; 06:107–116.

69. Saxe AM, McClelland JL, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv. 2014

70. Bar M. A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci. 2003; 15:600–609. [PubMed: 12803970]

71. Bar M, et al. Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A. 2006; 103:449–454. DOI: 10.1073/pnas.0507062103 [PubMed: 16407167]

72. Noonan KA, Jefferies E, Visser M, Lambon Ralph MA. Going beyond inferior prefrontal involvement in semantic control: Evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. J Cogn Neurosci. 2013; 25:1824–1850. [PubMed: 23859646]

73. McKee JL, Riesenhuber M, Miller EK, Freedman DJ. Task dependence of visual and category representations in prefrontal and inferior temporal cortices. J Neurosci. 2014; 34:16065–16075. [PubMed: 25429147]

74. Jackson RL, Cloutman L, Lambon Ralph MA. Exploring distinct default mode and semantic networks using a systematic ICA approach. Cortex. 2019; 113:279–297. [PubMed: 30716610]

75. Davey J, et al. Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. Neuroimage. 2016; 137:165–177. [PubMed: 27236083]

76. Rohde, DLT. Technical Report CMU-CS-99-164. Carnegie Mellon University, Department of Computer Science; Pittsburgh, PA: 1999. LENS: The light, efficient network simulator.

77. IBM SPSS Statistics for Windows v. 25.0. Armonk, NY: 2017.

78. Cloutman LL, Binney RJ, Drakesmith M, Parker GJM, Lambon Ralph MA. The variation of function across the human insula mirrors its pattern of structural connectivity: Evidence from in vivo probabilistic tractography. Neuroimage. 2012; 59:3514–3521. [PubMed: 22100771]

79. McIntosh AR. Mapping cognition to the brain through neural interactions. Memory. 1999; 7:523–548. DOI: 10.1080/096582199387733 [PubMed: 10659085]

Europe PMC Funders Author Manuscripts

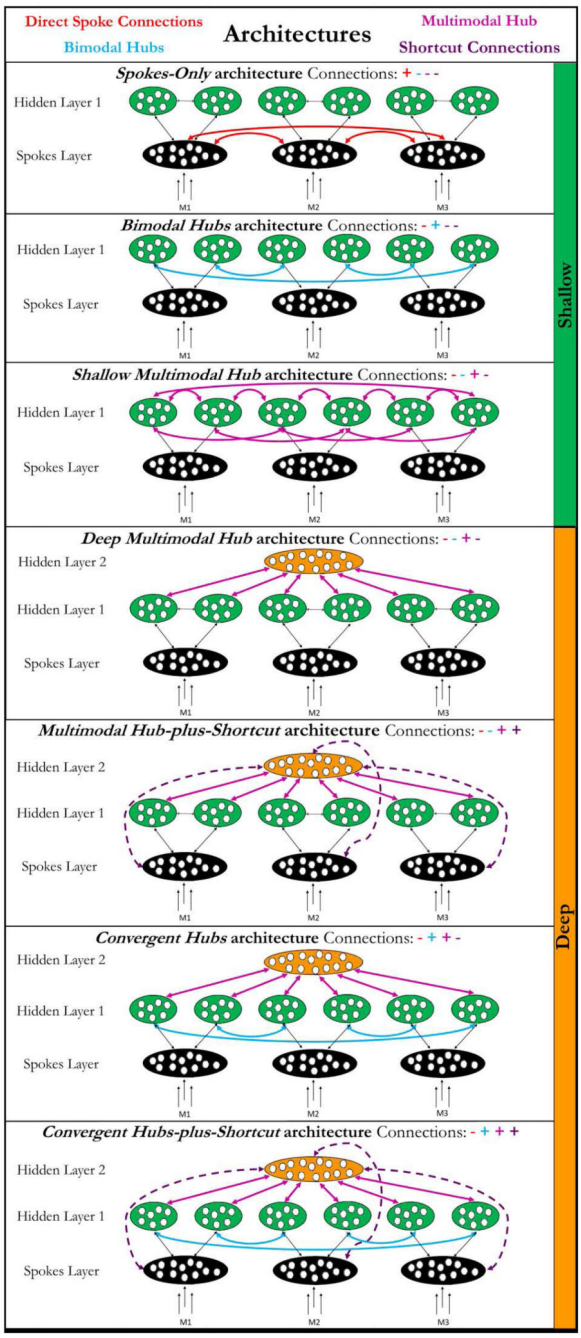

Figure 1.

The seven different architectures. Each architecture is based on the shallow or deep configuration and has one or more of the four types of variable connections (Direct Spoke Connections, Bimodal Hubs Connections, Multimodal Hub Connections and Shortcut Connections). The presence or absence of each connection type is demonstrated using + (where present) or - (where absent). The connection is also shown diagrammatically using arrows in the same colour. Black arrows represent the connections that are stable between architectures. Although only a subset of these connections may be displayed, the Shallow

Multimodal Hub architecture has connections between all Hidden Layer 1 subregions. Connections between Hidden Layer 1 regions with projections from the same modality are shown in grey; these are part of the connectivity changes needed to construct architectures without hubs in this layer and are not shown in the same colour as the other changes simply as this change is necessary for different connections and may cause confusion as to where the key change is. Many of the connections shown create coherent regions in Hidden Layer 1 - these are visualised as separate regions so that the correspondence between the architectures is apparent. The resulting 7 architectures are provided with labels (italics) for reference within the text. All employ the same total number of weights and units. M = modality.

Europe PMC Funders Author Manuscripts

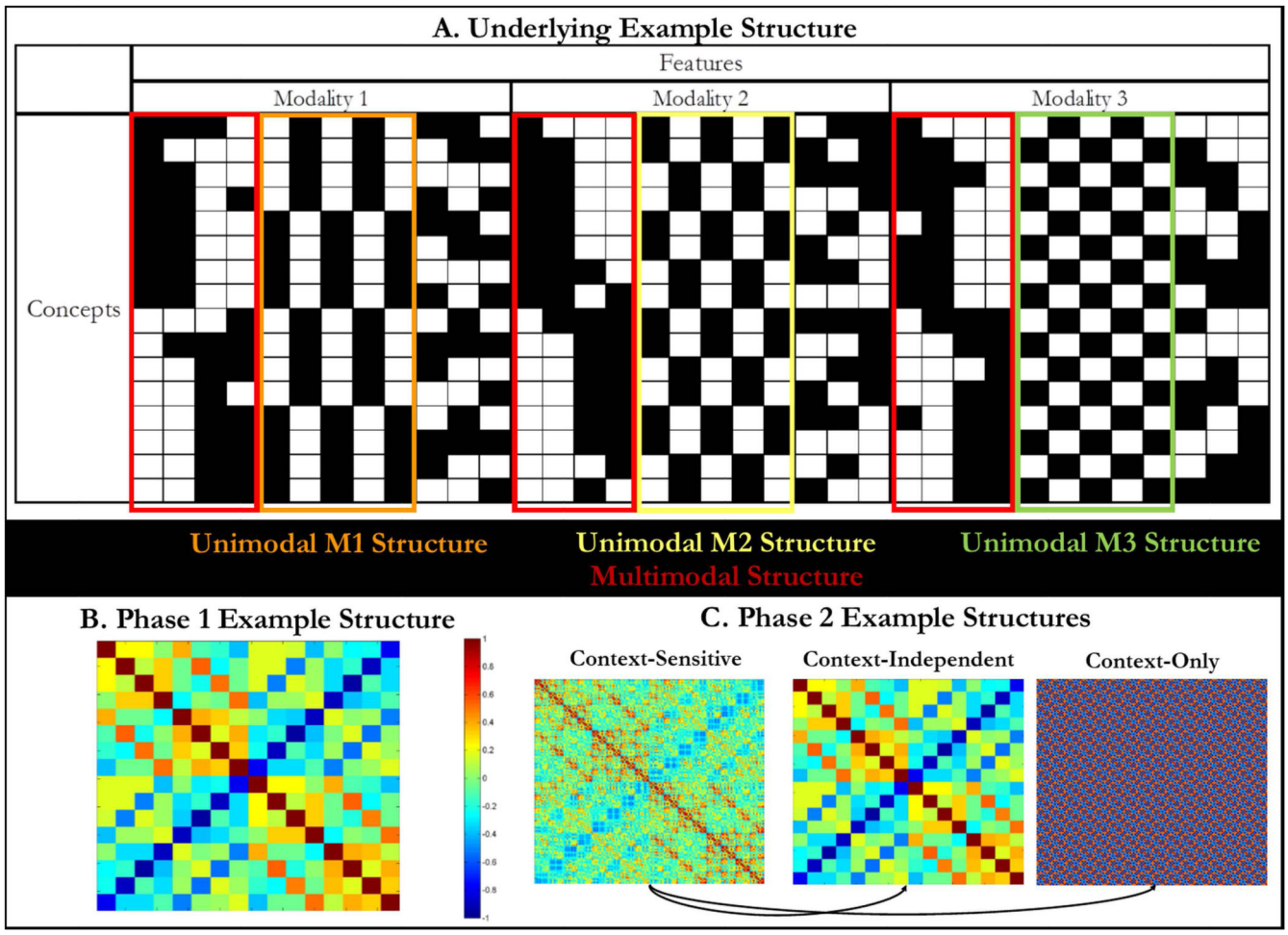

Figure 2.

The model environment. A. The full set of features for each concept. Each row is a concept and each column is a feature. A black box indicates that the feature is present and a white box that it is absent for that concept. All 16 concepts are shown here. Red boxes show features that covary strongly across modalities (multimodal structure) whilst the orange, yellow and green boxes highlight features that covary reliably within each modality (unimodal structure). The structures expressed by the multimodal and each unimodal feature set are mutually orthogonal. B. A matrix showing the context-independent conceptual similarity structure across all modalities for the 16 items for Phase 1. Colours show the correlation (ranging from -1 to 1) for all pairs of vectors based on each full row of Panel A. C. Matrices showing correlations amongst examples used in the context-sensitive simulations in Phase 2, including (left) the full context-sensitive example structure for all 144 input/output patterns, (middle) the example structure based on all features of a concept regardless of task context (same as panel B), and (right) similarities based on the control signal alone regardless of the features of a concept. The 144 patterns arise from crossing 16 items with the 9 possible task contexts. The context-sensitive example structure is a blend of the context-independent conceptual structure used to measure conceptual abstraction (middle; based on the features of each concept only) and the context-only similarity

structure (right) that indicates the appropriate input and output modalities regardless of concept.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

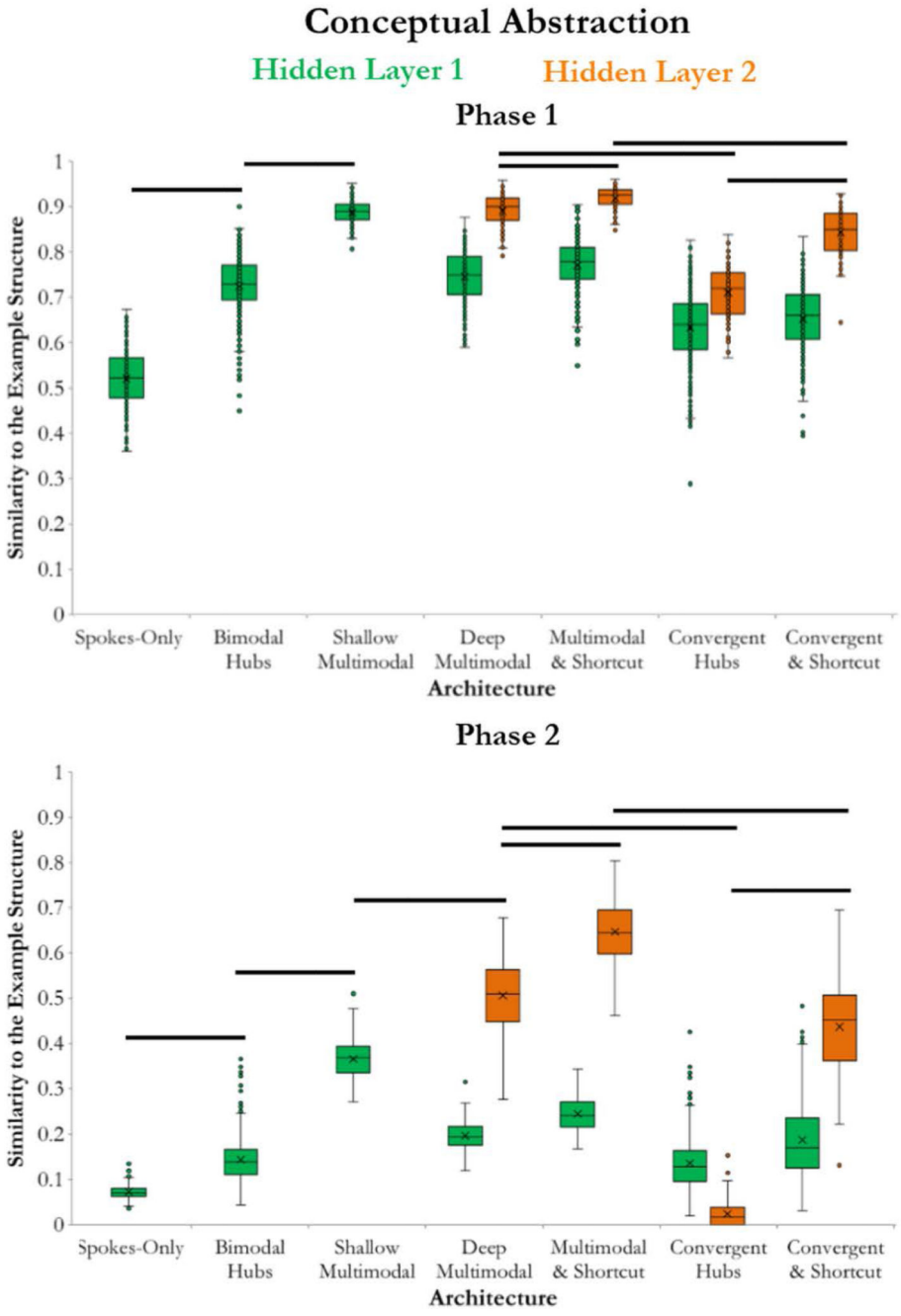

Figure 3.

Comparing the conceptual abstraction across the architectures without (in Phase 1) and with (in Phase 2) the additional demand of context-appropriate output. The similarity between the context-independent example structure and the representations in Hidden Layer 1 (green) and Hidden Layer 2 (orange) in 80 observations of each architecture are displayed. The higher box reflects the conceptual abstraction score for that model architecture. The middle bar shows the median similarity value and the cross reflects the mean across the different runs of the model (additional bars show the first and third percentile, values more than 1.5

times the interquartile range are displayed as dots, otherwise the minimum and maximum values are reflected by the whiskers). Planned contrasts with significant differences in the conceptual abstraction score are highlighted with a black line ($p<.05$).

Europe PMC Funders Author Manuscripts

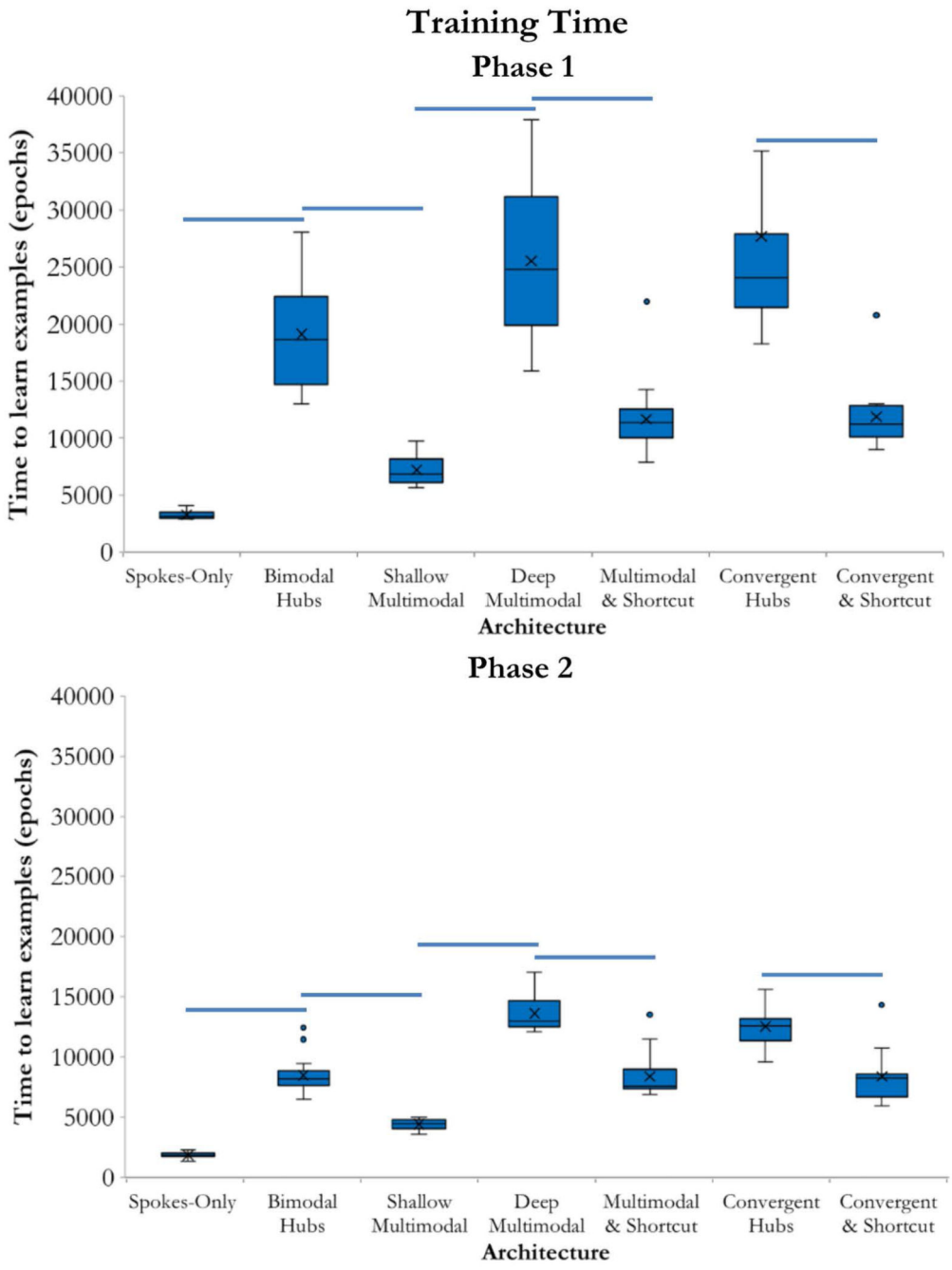

Figure 4.

Comparing the training time across the architectures without (in Phase 1) and with (in Phase 2) the additional demand of context-appropriate output. The time taken to learn the examples as the number of epochs of training is displayed for 15 observations of each architecture. The middle bar shows the median number of epochs across different runs of the model and the cross reflects the mean (additional bars show the first and third percentile, values more than 1.5 times the interquartile range are displayed as dots, otherwise the minimum and maximum values are reflected by the whiskers). A single outlier from the Convergent Hubs

architecture is not shown as it was greater than 60000 epochs. Significant differences in the planned contrasts are highlighted with a line ($p<.05$).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

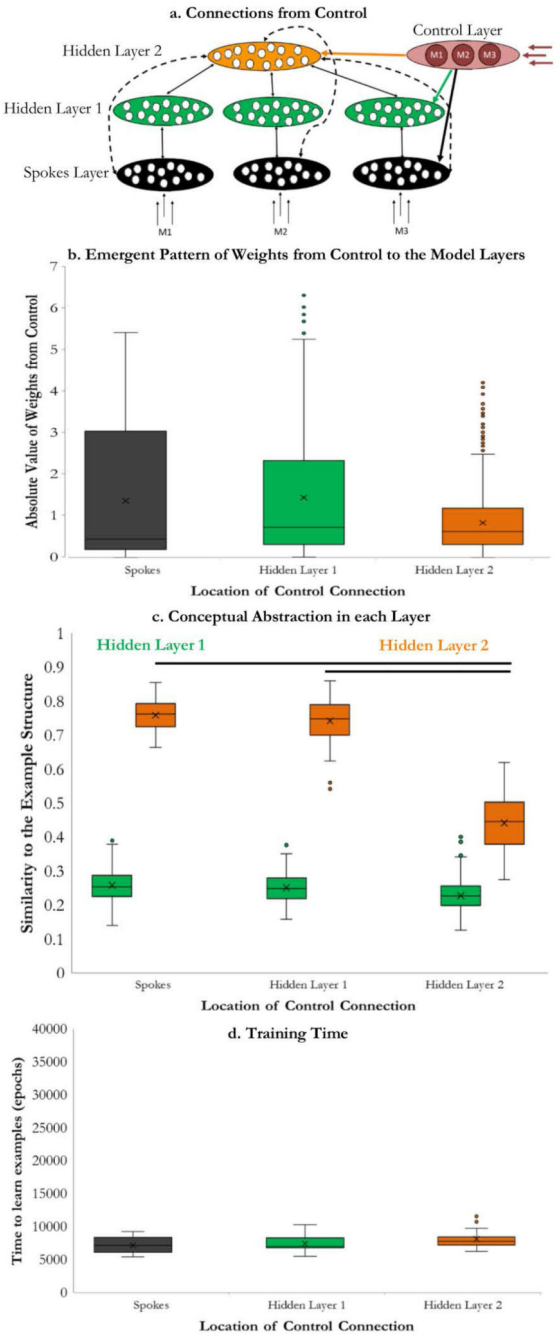

Figure 5.

Consequences of the location of the connection to control. For all plots, the middle bar shows the median number of epochs across different runs of the model and the cross reflects the mean (additional bars show the first and third percentile, values more than 1.5 times the interquartile range are displayed as dots, otherwise the minimum and maximum values are reflected by the whiskers). Significant differences in the planned contrasts are highlighted with a line ($p<.05$). A. The different ways control was connected to the Multimodal Hub-plus-Shortcut architecture. This diagram is equivalent to Figure 1, yet simplified as the

correspondence between architectures is not being highlighted. The modalities to attend to (i.e., those where an input is received or an output expected) are input to the 3 units in the Control Layer. Learnt unidirectional connections from the Control Layer allow the control signal to enter the semantic system at different points. Connections to all layers were present in initial simulations and the magnitude of weights to each layer compared. Then, the Control Layer was selectively connected to either the Spokes Layer (black arrow), Hidden Layer 1 (green arrow) or Hidden Layer 2 (orange arrow) and the results of these simulations compared. B. The emergent pattern of the absolute value of the weights from the control units to each layer in 40 observations of the Multimodal Hub-plus-Shortcut architecture. C. The effect of connecting the Control Layer to each layer of the Multimodal Hub-plus-Shortcut architecture on the similarity between the context-independent example structure and the representations in Hidden Layer 1 (green) and Hidden Layer 2 (orange) across 80 different runs of the model. The highest box reflects the conceptual abstraction score. D. The time taken to learn the examples is shown as the number of epochs of training across 15 different runs of the model when the control signal is connected to the Spokes Layer (black), Hidden Layer 1 (green) or Hidden Layer 2 (orange).

Europe PMC Funders Author Manuscripts

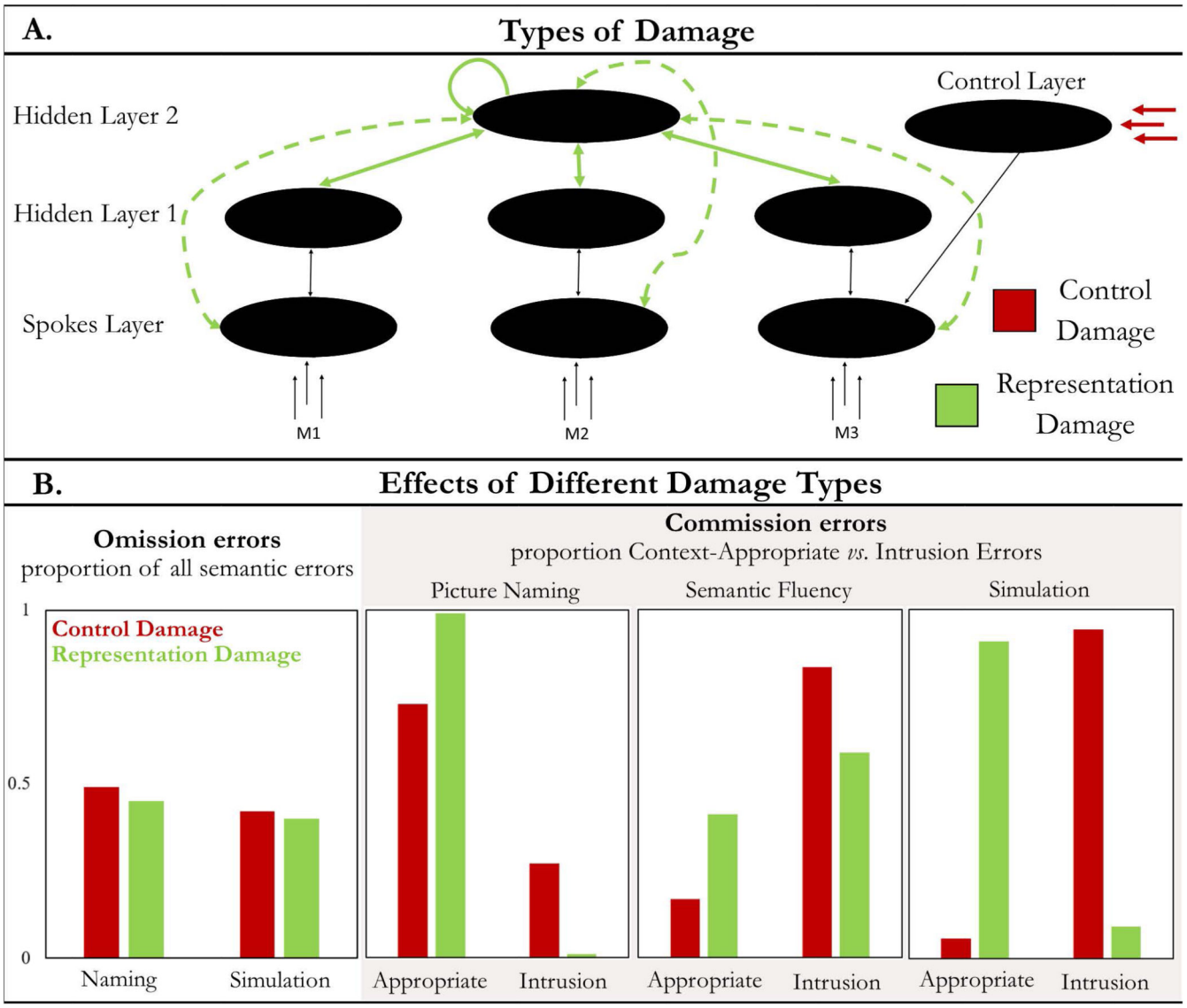

Figure 6.

Simulating different error patterns in SD vs SA. A. The connections and inputs affected in the different patterns of damage. The Representation Damage simulation involved removing a proportion of all connections within, to and from Hidden Layer 2 (shown in green). The Control Damage simulation involved adding noise to the input to the control units (shown in red). B. The left panel shows the total proportion of errors that were omissions ("don't know" or no response) for a cohort of 20 patients with damage to control regions (in SA) or representation regions (in SD) in a picture naming task, together with proportion of item-level error types that were omissions (target units that did not activate) in 80 observations of the model under control vs. representation damage. The remaining panels show the total proportion of commission errors that involved producing context-appropriate vs context-inappropriate intrusion errors, for cohort of SD and SA patients in a picture naming task (20 participants) and in a semantic fluency task (15 participants), and for the model under damage simulating these disorders (80 observations of each). As different damage levels

result in a highly similar proportion of each error type only the intermediate level is shown here (see Supplementary Note 8 for further details). The two syndromes show equal probability of omissions in naming, but differential probability of producing context-appropriate and intrusion errors in both fluency and naming. The pattern of changes following both control and representation damage are captured by the corresponding pattern of damage in the model.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

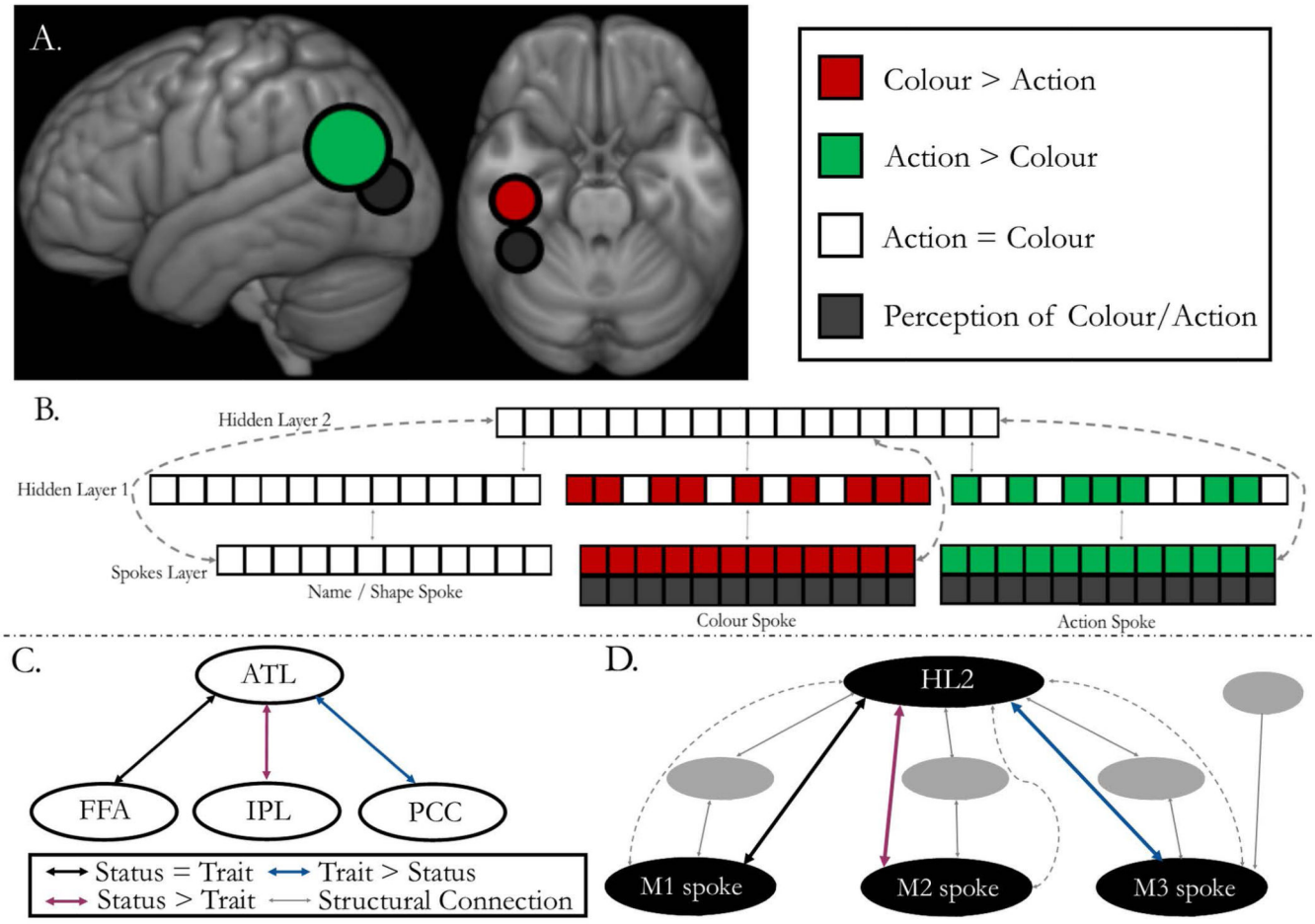

Figure 7.

Simulating dynamic changes in univariate activation and functional connectivity across contexts. A. Martin et al.,[14]found differences in and around the regions responsible for perception of colour and action (black) demonstrated greater activation when colour (red) or action knowledge (green) was required for a task. B. The reverse-engineered model simulates this effect successfully; across 80 observations of a context where the features in modality 2 are required there is greater activation of units in and around this 'colour' spoke (red), and for 80 observations where modality 3 is the required output, there is greater activation of units in and around this 'action' spoke (green). There are no changes in the involvement of the input spoke or Hidden Layer 2. C. Wang et al.[15] used a psychophysiological interaction analysis to demonstrate dynamic connectivity between the ATL hub and the spoke regions involved in trait or status processing when the required output shifted between these contexts. Functional connectivity with the input spoke did not vary. D. The reverse-engineered model demonstrated the same dynamic functional connectivity – whilst connectivity between Hidden Layer 2 and the input spoke (M1) stayed constant, the requirement to produce modality 2 features ('status') as output increased its functional connection with the hub and reduced the connectivity of the modality 3 spoke with the hub compared to the production of modality 3 features ('traits').