JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles | Behavioral/Cognitive

Causal influence of linguistic learning on perceptual and conceptual processing: A brain-constrained deep neural network study of proper names and category terms

https://doi.org/10.1523/JNEUROSCI.1048-23.2023

Received: 6 June 2023
Revised: 1 December 2023
Accepted: 6 December 2023

Causal influence of linguistic learning on perceptual and conceptual processing: A brain-constrained deep neural network study of proper names and category terms

Phuc T. U. Nguyen[1]*, Malte R. Henningsen-Schomers[1,4], and Friedemann Pulvermüller[1,2,3,4]

[1]Dept. Of Philosophy and Humanities, Brain Language Laboratory, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany

[2]Berlin School of Mind and Brain, Berlin, Germany

[3]Einstein Center for Neurosciences, Berlin, Germany

[4]Cluster of Excellence "Matters of Activity. Image Space Material", Humboldt-Universität zu Berlin, Berlin, Germany

* Corresponding authors: Phuc T. U. Nguyen (phuc.thu.uyen.nguyen@gmail.com) and Friedemann Pulvermüller (friedemann.pulvermuller@fu-berlin.de)

Number of pages: 51

Number of Figures: 6

Number of tables: 3

Number of words

Abstract: 250/ 250

Significance statement: 120/120

Introduction: 646/650

Discussion: 1521/1500

Conflict of interest: The authors declare no competing financial interests.

Acknowledgements. We would like to thank Thomas Wennekers, Rosario Tomasello, Luigi Grisoni, Laura Ciaccio, Maxime Carrière, Fynn Dobler and the other members of the MatCo research group for their insightful opinions and brilliant discussion on theoretical and practical questions and for their help and suggestions at different stages of this work. In addition, we would like to thank the high-performance computing service of the Freie Universität Berlin and Martin Freyer and Philip Krause for technical support. We would like to thank Ngan Nguyen for support in image creation and visualization.

This work was supported by the European Research Council (ERC) through Advanced Grant 'Material constraints enabling human cognition, MatCo' (ERC-2019-ADG 883811), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy through the Cluster of Excellence 'Matters of Activity. Image Space Material' (DFG EXC 2025/1-390648296).

Author contributions. M.H.-S. and F.P. conceived the study. M.H.-S., P.T.U.N and F. P. designed the study, M.H.-S., and P.T.U.N implemented research. P.T.U.N analyzed data. P.T.U.N, M.H.-S., and F.P. wrote the manuscript.

Abstract

Language influences cognitive and conceptual processing, but the mechanisms through which such causal effects are realized in the human brain remain unknown. Here, we use a brain-constrained deep neural network model of category formation and symbol learning and analyze the emergent model-internal mechanisms at the neural circuit level. In one set of simulations, the network was presented with similar patterns of neural activity indexing instances of objects and actions belonging to the same categories. Biologically realistic Hebbian learning led to the formation of instance-specific neurons distributed across multiple areas of the network, and, in addition, to cell assembly circuits of 'shared' neurons responding to all category instances – the network correlates of conceptual categories. In two separate sets of simulations, the network learned the same patterns together with symbols for individual instances ('*proper names*') or symbols related to classes of instances sharing common features ('*category terms*'). Learning category terms remarkably increased the number of shared neurons in the network, thereby making category representations more robust while reducing the number of neurons of instance-specific ones. In contrast, proper-name learning prevented substantial reduction of instance-specific neurons and blocked the overgrowth of category-general cells. Representational Similarity Analysis further confirmed that the neural activity patterns of category instances became more similar to each other after category-term learning, relative to both learning with proper names and without any symbols. These network-based mechanisms for concepts, proper names and category terms explain why and how symbol learning changes object perception and memory, as revealed by experimental studies.

Significance Statement

How do verbal symbols for specific individuals (*Micky Mouse*) and object categories (*house mouse*) causally influence conceptual representation and processing? Category terms and proper names have been shown to respectively promote category formation and instance learning, potentially by respectively directing attention to category-critical and object-specific features. Yet the mechanisms underlying these observations at the neural circuit level remained unknown. Using a mathematically precise deep neural network model constrained by properties of the human brain, we show category-term learning strengthens and solidifies conceptual representations, whereas proper names support object-specific mechanisms. Based on network-internal mechanisms and unsupervised correlation-based learning, this work offers neurobiological explanations for causal effects of symbol learning on concept formation, category building and instance representation in the human brain.

Introduction

Most signs and symbols are used to speak about objects and actions. This led philosophers and logicians to propose that the referential link between symbol and world is most essential for meaning and semantics (Wittgenstein, 1922; Frege, 1948). Yet there are quite different relationships between symbols and their related real-world entities. One most essential difference exists between ‘proper names’ used to speak about a single object or individual (e.g., “Mickey Mouse”) and ‘category terms,’ which can refer to members of an entire class or conceptual category (e.g., “house mouse”). Such differences between referential symbols are well-described at the semantic level, but not understood in terms of their underlying mechanisms in mind and brain.

The need for mechanistic neurobiological models of symbols and their meaning comes from reports about causal influences of language on perception, attention, and memory. It had long been speculated and recently been confirmed that, when human subjects learn words for objects, language may help humans to attend to and distinguish between them (Majid et al., 2004; Whorf and Carroll, 2007; Miller et al., 2018; Vanek et al., 2021). Experimental research in infants showed that learning ‘labels’ for objects increases their attention to these objects (Baldwin and Markman, 1989), which further establishes an attention-catching function of language. However, this general insight requires further specification to capture the different effects of category terms and proper names. In particular, learning a new symbol for a category of objects makes infants attend to the shared features of these objects and facilitates their learning of the conceptual category (Gelman and Markman, 1986, 1987; Plunkett et al., 2008); the latter even holds if the objects show little perceptual similarity (Graham et al., 2013). On the other hand, the category building function of language is absent when object-specific proper names are learned. In this case, the infant’s attention is directed not towards the common category features of objects, but to idiosyncratic and object-specific features instead (Scott and Monesson, 2009; LaTourrette and Waxman, 2020). In summary, category-term learning directs attention to shared features of objects (Waxman and Booth, 2001; Dewar and Xu, 2007; Althaus and Mareschal, 2014; Althaus and Plunkett, 2016), whereas unique proper-name learning highlight idiosyncratic and object-specific features (Best et al., 2010; Barnhart et al., 2018; Pickron et al., 2018; LaTourrette and Waxman, 2020). These specific and replicable effects of proper names and category terms on perception and attention have been explained in terms of different ‘strategies’ applied by the learner. A neurobiological explanation of why these specific effects occur is still missing.

Why and how can proper names and category terms direct attention to specific versus shared features of category members? To develop a mechanistic explanation, we used a brain-constrained deep neural network designed according to the area structure and connectivity of major areas relevant for language and conceptual processing (Garagnani et al., 2007; Tomasello et al., 2018; Pulvermüller et al., 2021). Six “areas” of the model simulated processes in superior temporal and inferior frontal perisylvian language areas and six extrasylvian model areas simulated inferior temporo-occipital visual ‘where’ processing stream and dorsolateral prefrontal and motor cortices (Figure 1A). In the No symbol (NoS) condition, the model learned activity patterns each representing 1 of 60 instances of objects or actions belonging to 10 different categories. In learning-with-symbols conditions, the model learned additional activity patterns representing word forms of proper names (PN) or category terms (CT) (Figure 1B-C, 2A). After learning, the model was tested by activating previously trained instance patterns of each category and, in addition, new patterns for novel instances belonging to the same categories (Figure 2B). We documented the neural and cognitive effects of proper names and category terms on instance and category learning in the model. In-depth analyses of the emerging activation patterns and representations were provided by using Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) and by classifying neurons into instance-specific and category-general ones.

Materials and Methods

Participants

The current work does not contain experiments with human participants or animal subjects.

Neurobiological constraints

In contrast to many neural network models, the brain-constrained model applied aimed at biological plausibility by applying a range of structural and functional constraints (Tomasello et al., 2018; Henningsen-Schomers and Pulvermüller, 2022; for review, see Pulvermüller et al., 2021) realizing:

(1) neurophysiological dynamics of spiking pyramidal cells (Connors et al., 1982; Matthews, 2001),
(2) synaptic weights under the modification of unsupervised Hebbian-type learning (i.e., synaptic plasticity and learning were modified according to the biologically plausible unsupervised Hebbian principles that incorporated both long-term potentiation and long-term depression) (Artola and Singer, 1993),
(3) local and global activity regulation (Braitenberg, 1978; Yuille and Geiger, 1995) based on local and area-specific inhibition mechanisms (Knoblauch and Palm, 2002),
(4) excitatory and inhibitory within-area local connectivity (including sparse, random, and initially weak excitatory links whose probability falls off with distance) (Kaas, 1997; Braitenberg and Schüz, 1998),
(5) between-area global connectivity built on neuroanatomical evidence, and
(6) built-in uncorrelated white noise in neurons of (a) all areas during training and testing mimicked spontaneous baseline neuronal firing and (b) additional noise in neurons of areas not stimulated by patterns during training, which simulated uncorrelated sensory or motor activity unrelated to instances or symbols (Rolls and Deco, 2010).

Table 2 supplies the model specifications and parameters chosen in this current work.

Model description

We applied a brain-constrained deep neural network model including spiking model neurons and twelve model areas to model sensorimotor, conceptual and linguistic mechanisms in the left-hemispheric language-dominant fronto-temporo-occipital regions of the human brain, as described in previous studies by Tomasello et al., 2018; Henningsen-Schomers and Pulvermüller, 2022.

Anatomical architecture and connectivity

To distinguish between sub-parts of neural networks from their target cortical structures of the real human brain, all model areas are marked by an asterisk before (e.g., *A1, *V1). The architecture modelled three areas representing the ventral visual system (i.e., primary visual cortex (*V1), temporo-occipital area (*TO), anterior-temporal area (*AT)) and three areas representing the dorsolateral action system (i.e., dorsolateral fronto-central motor (*M1$_L$), premotor cortex (*PM$_L$), prefrontal cortex (*PF$_L$)). These formed the extrasylvian region for sensorimotor processing where semantic information was stored. Another 6 areas of the perisylvian region for word-form processing housed articulatory-phonological and acoustic-phonological information. These areas involved the three areas of the auditory system (i.e., primary auditory cortex (*A1), auditory belt (*AB), parabelt areas (*PB)) and three inferior frontal articulatory and prefrontal areas (i.e., inferior primary motor cortex (*M1$_i$), premotor cortex (*PM$_i$), prefrontal cortex (*PF$_i$)), respectively. Between-area connections were reciprocal and connected next neighbor areas, second-next neighbors (see Schomers, 2017) and long-distance corticocortical links supported by neuroanatomical evidence in the literature (Table 1).

In the current neural network model, the fundamental information processing units are artificial neuron-like elements, or cells. Each modelled area comprised two layers of 625 e-cells and 625 i-cells that mimicked an (excitatory) pyramidal spiking neuron and a cluster of (inhibitory) interneurons hosted within the same cortical column in the cortical area. A more elaborate description of the firing behavior of such neurons could be found in Garagnani et al. (2017), Tomasello et al. (2018), Henningsen-Schomers and Pulvermüller (2022).

Activity patterns applied to the networks

60 'grounding patterns' were defined as sensorimotor activation patterns thought to represent specific sensory-motor experiences of 60 different objects or "instances". Groups of 6 instances overlapped in their neuronal grounding patterns and were taken as representations of different instances of the same concept (e.g., different robots). Note that the images of robots and cat faces for category members are to be taken purely for illustrative purposes here – the actual training patterns of the models consisted of sets of activated neurons with no systematic relationship to images of robots or cat faces. A category comprised three trained instances and three novel instances not presented during training; all six instance patterns were used for network testing (Figure 2A-B). Each category instance was neuronally coded as a set of perceptual and motor neuron activations in the primary visual and hand-motor areas of the brain-constrained network. These instance-related grounding patterns were activated either on their own or together with additional patterns of neuronal activation in the network's articulatory and auditory cortices, which were thought to implement symbol forms, that is, verbal labels or spoken word forms. These "word form patterns" were used either as proper names and therefore specifically with only one grounding pattern, or as category terms and therefore the same word form pattern co-occurred with all 3 trained grounding patterns of one category. To control the effect of non-linguistic factors, a third class of trained grounding patterns was learnt without concordant auditory-articulatory activation. Thus, we generated three classes of simulated stimulation patterns: (i) instance-related grounding patterns applied to *V1/*M1$_L$ (Figure 1B-left), (ii) category term patterns to *A1/*M1$_i$, (Figure 1B-middle) and (iii) proper name patterns to *A1/*M1$_i$ (Figure 1B-right). Sensorimotor experiences of instances were simulated with conceptual grounding patterns, (i), and symbol-related auditory-articulatory activity were simulated using word form patterns, (ii) and (iii).

For visualization and better conceptual understanding of the use of activity patterns, see Figure 1B-C. Instances belonging to the same category were simulated by similar grounding patterns, following Henningsen-Schomers and Pulvermüller (2022).: within-category instances had grounding patterns that shared 50% of their feature neurons and differed from each other in the other half; grounding patterns simulating instances from different categories had no neuronal overlap. For each grounding pattern (i), a subset of twelve out of 625 potential cells per area were randomly chosen, consisting of six unique neurons and six shared neurons. Shared neurons simulated features characterizing all instances patterns of a category; they simulated shared conceptual features of all category members (category-critical feature, e.g., members of the first category are robots in the same height and are equipped with one camera, one speaker, two antennae, a power button, two metal legs, and a pair of shoes; members of the second category are cats and have round-shaped head, eyes, nose, mouth, ears, and whiskers (Figure 1B-left). Unique neurons simulated the 'idiosyncratic', fully instance-specific visuomotor features; each of the corresponding feature neurons was only available in one instance pattern (e.g., robots vary in the body shape and color, orientation of antennas, leg forms, position of power button, shoes color). In sum, each category possessed 36 unique neurons from its six exemplars and six shared neurons. For word form patterns, category term patterns (ii) of within-category instances consisted of the same twelve neurons, which were co-activated with each of the 3 learnt grounding patterns of a category (e.g., to simulate the artificial words *fos* for all instances of the robot category, and *coxt* for all instances of the cat category) (Figure 1B-middle); each proper name pattern (iii) comprised twelve neurons, which were co-activated with one specific grounding pattern (e.g., *xub*, *vit*, *hek* for the three instances of the robot category, respectively) (Figure 1B-right). The choice of cells for pattern generation was pseudorandomized and constrained by the following criteria: First, within-category neurons had to be non-adjacent to each other. This prevented coactivation merely due to close distance. Second, no grounding patterns from two different categories shared any neuron. Last, for each instance, the grounding patterns in *V1 and *M1$_L$ followed the same principles but were not identical. The same rules applied for the grounding patterns in *A1 and *M1$_i$.

Experiment design

The current simulations involved three phases: model initialization, training phase, and testing phase, which were carried out on the high-performance computing system of Freie Universität Berlin (Bennett et al., 2020). During training, there were 3 different stimulation conditions, (1) where grounding patterns were learnt without symbol (No symbol or control condition), (2) where all grounding patterns of each category were presented together with the same word form pattern (Category term condition), and (3) where each grounding pattern was co-presented with its own specific word form pattern (Proper name condition). Thus, during learning, a stimulation pattern included two activation patterns (to *V1 and *PF$_L$) when it was learnt outside symbol context (Figure 1C-top) or, a quadruplet including the two instance-related patterns plus two word form-related ones (to A1 and PF$_i$) when learnt in symbol context (Figure 1C-bottom). Each test trial began with the presentation of a grounding pattern of an instance (projected to the two sensorimotor model areas V1 and M1$_L$).

Model initialization

One crucial step prior to training was model initialization, which randomized all synaptic links (and their corresponding weights) between within-area cells and between cells from connected areas. Twelve sets of such synaptic links and weights (i.e., 12 different instantiations of the randomly initialized neural network) were chosen, each set was then triplicated (cf. Schomers et al., 2017), and each of these three copies entered one of the three training conditions – either No symbol, Category term or Proper name. The use of distinct model instantiations can be seen as analogous to a within-subject study design with 12 subjects. We chose to implement 3 separate sets of simulations for the 3 conditions to avoid any possible interference effects between concepts and symbols that may emerge during training. Note, for example, that the relatively large representations that formed for category terms might have interfered with further learning or may even have suppressed the activation of conceptual representations without symbols. This configuration yielded a controlled ‘within-subject’ design with training condition being a three-level repeated-measures factor (*No symbol*, *Category term*, and *Proper name*). For the additional simulations performed to balance the number of word form presentations, there were 4 levels.

Training phase

The neural network model was repeatedly presented with 30 instances from ten categories. To mimic visuo-motor percepts associated with an instance, the extrasylvian primary sensorimotor areas, *V1 and *M1$_L$, were each presented with its grounding pattern (i) for 16 time steps. Following the experiment by LaTourrette and Waxman (2020) where instances were called either by a consistent label or by distinct labels each, our within-category trained instances were either paired with the same category term, by their distinct proper names, or they were not labeled at all. To mimic symbols in the Category term and Proper name conditions, we presented to the primary perisylvian areas *A1 and *M1$_i$ word form pattern (ii) and (iii), respectively, for 16 time steps (Figure 1C-bottom, 2A). Hence, in different ‘learning trials’, the word form patterns of category terms were co-presented with one of three different grounding patterns from one category, whereas those of proper names co-occurred with only one specific grounding pattern. There were no word form patterns presented in the baseline No symbol condition to control for the effect of either type of linguistic labels compared to learning without ones (Figure 1C-top, 2A).

Because activity at the end of a trial might affect learning in the next trial, the network was allowed to deactivate after each stimulated learning trial. To this end, we separated every two consecutive pattern stimulations by a waiting interval during which only the uncorrelated white noise mimicking spontaneous baseline neuronal firing was supplied to all areas (see principle 6 in Model description – Neurobiological constraints). The goal was to reset the global network (i.e., all excitatory and inhibitory cells displayed a membrane potential of zero) before a new grounding pattern was inputted to the neural network model. This interstimulus interval (ISI) was terminated only after the network activity had returned to its baseline value (thresh $= 0.18$, see table 2). As a result, the training order was not influential in this experiment.

To balance learning conditions (NoS, CT, PN), each experiential grounding pattern representing an instance was presented 2,000 times in one set of simulations. However, because each category term pattern was co-presented with 3 different instance patterns, whereas proper name patterns co-occurred with only one, this design leads to an imbalance of the number of learning trials during which individual word form patterns were presented (3 times higher for category term than for proper name presentations) (Figure 2C-top). Therefore, a second evaluation of learning trials was performed and analyzed for which the number of word form pattern activations was balanced. In this case, there were 1,000 learning trials in the Category term condition (CL_1x; each instance was presented together with a category term in 1,000 training trials, resulting in a total of 3,000 training trials per category terms) and 3,000 trials in the Proper name condition (PN_3x; each instance was presented together with a proper name in 3,000 training trials, resulting in a total of 3,000 training trials per proper name). For the control No symbol conditions, two comparison values were calculated, after 1,000 (NoS_1x) and 3,000 (NoS_3x) trials (i.e., each instance was presented without symbol in 1,000 and 3,000 training trials, respectively) (Figure 2C-bottom). These different sub-designs are summarized graphically in Figure 2C.

Testing phase

In the current experiment, we implemented a version of an old-new recognition task with the use of new instances. For each of the ten categories we presented to the neural network six testing instances: three trained instances and three novel instances (Figure 2B). In total, we used 30 previously learnt instances and 30 new instances. However, no actual old-new pairing took place because we presented trained and novel instances to the neural network in separate test trials.

Memory performance of the network model was assessed in the absence of linguistic cues, i.e., without stimulating the perisylvian primary areas *A1 or *M1i. To stimulate the experience of individual instances, the extrasylvian primary areas *V1 and *M1$_L$ were activated for 2 time steps with pure (i.e., free of any white noise) grounding patterns (i) and subsequentially deactivated towards the baseline for 28 time steps. We recorded network responses 30 time steps from the onset of this stimulation. Global resetting between two consecutive trials was conducted in the same manner as the training phase. Hence, the test order was not of interest.

Data analysis

Grounding pattern production, data processing, and data analysis were performed using Python 3.9.7, matplotlib 3.4.3 (Hunter, 2007), NumPy 1.20.3 (Harris et al., 2020), pandas 1.3.4 (Reback et al., 2022), SciPy 1.7.1 (Virtanen et al., 2020), seaborn 0.11.2 (Waskom, 2021). In the current work, statistical significances were based on a conservative p-value threshold of $.005$ suggested by Di Leo and Sardanelli (2020). We used rstatix 0.7.0 (Kassambara, 2021) in the R software environment (R Core Team, 2021) for statistical analyses.

When testing stimuli were presented to the primary sensorimotor areas, some of the 625 excitatory neurons per area fired in response to their conceptual grounding patterns. As described in Procedure, we recorded all their responses during 30 time steps from stimulation. Let $\phi(e,t)$ denote the output of an excitatory cell e at time t, such that ϕ only takes up the value 0 or 1 and t only allows discrete values up to 30 (corresponding to thirty possible simulation time steps); let $\tau_{Favg} = 5$ be a time constant, the estimated instantaneous firing rate $\omega_E(e,t)$ of cell e at time t can be calculated based on the following equation:

$$\tau_{Favg} \cdot \frac{d\omega_E(e,t)}{dt} = -\omega_E(e,t) + \phi(e,t) \quad \text{Eq. (1)}$$

Solving Eq. (1) for $\omega_E(e,t)$ returns the cell's latest spiking activity (firing rate). We estimated the mean firing rate based on $t = t_{30}$ and used this value for the subsequent RSAs. For details about relevant calculation steps, see *Appendix* of Henningsen-Schomers and Pulvermüller, (2022).

Previous research found that several of the extra-sylvian areas targeted by the deep neural model (including, for example, *V1 and *AT) are important for processing instance- and concept-related information (see, for example, Binder et al., 2005; Martin, 2007; Ralph et al., 2017; Henningsen-Schomers et al., 2023). Therefore, the current data analyses and statistical testing focused on the extrasylvian region of the deep neural network. This decision was motivated by the main aim of addressing possible causal influences of symbol learning on the perceptual processing of instances of concepts and on conceptual processing itself.

Representational similarity analysis

The estimated mean firing rate of 625 neurons in response to a testing instance reflected how this instance was represented in neural network. To understand how differently the neural network represented within- and between-category instances, we calculated the dissimilarity in firing patterns for every pair of the 60 instances. Pairwise dissimilarities computed in terms of Euclidean distance were organized in a 60×60 representational dissimilarity matrix (RDM) (Figure 3A): Each cell in the matrix reflected the dissimilarity between the firing patterns of two instances. In total, there were 36 RDMs across three training conditions and twelve areas.

We defined two classes of pairwise dissimilarities, including between-category dissimilarity ($Dissim_B$) and within-category dissimilarity ($Dissim_W$). A second way to define similarity types is based on the type of instances under study, that is, dissimilarity between two trained instances ($Dissim_{TT}$), between two novel instances ($Dissim_{NN}$), and between a trained and a novel instance ($Dissim_{TN}$). For example, within-category dissimilarity could be classified as either dissimilarity among trained instances 1 – 3 ($Dissim_{W-TT}$), among novel instances 4 – 6 ($Dissim_{W-NN}$) or between trained and novel instances ($Dissim_{W-TN}$) (Figure 3A).

Category learning. Category learning was ~~defined as~~ evaluated through the ability to (1) distinguish differences between categories and (2) group together category members. We assessed how different types of symbols impacted upon category learning performance based on (1) the dissimilarity between two between-category trained instances ($Dissim_{B-TT}$), and (2) the dissimilarity between two within-category trained instances ($Dissim_{W-TT}$) (Figure 3A). Successful category learning occurred when two instances from two distinct categories were considered as dissimilar (high $Dissim_{B-TT}$) and/or when two within-category instances were considered as similar (low $Dissim_{W-TT}$). If, as previously claimed, applying category terms invites one to encode the commonalities among instances and thereby facilitates categorization, the deep neural network should represent within-category instances similarly while highlighting the dissimilarities between instances of different categories. In the Category term condition, we expected between-category dissimilarities to be greater than within-category dissimilarities $Dissim_{B-TT_{CT}} > Dissim_{W-TT_{CT}}$. By contrast, we proposed two scenarios for the Proper name condition. In the first scenario, if proper names focus the neural network models on encoding only unique features and inhibit the encoding of category-critical features, no traces of category learning will be observable, and the representations of individual instances will be highly dissimilar regardless of their categorical membership ($Dissim_{B-TT_{PN}} \approx Dissim_{W-TT_{PN}}$). However, because within-category instances shared 50% of their activated neurons in the extrasylvian primary areas *V1 and *M1$_L$, the neural network could base on such similarities to form category representation. In this second scenario, proper names are not sufficient to override category learning; the neural network would house not only the unique representations of the instances but also the commonalities of those belonging to the same category. Like the Category term condition, the test data would as well show signs of category learning ($Dissim_{B-TT_{PN}} > Dissim_{W-TT_{PN}}$). Taking into account such intrinsic perceptuomotor similarities among instances from the same category, category learning was evaluated not only across symbol (i.e., category term or proper name) learning conditions but also in control conditions (i.e., training without symbols). For example, a superior causal influence of category terms on category learning performance would be expressed through a significantly higher $Dissim_{B-TT_{CT}}$ and lower $Dissim_{W-TT_{CT}}$ relative to training with proper names and also relative to training without symbols.

Generalization. Assuming the neural network had encoded the commonalities between within-category trained instances and formed category knowledge with the help of these shared features, they might have as well represented novel instances as members of that category when exposed to the category-critical features in these novel instances. Generalization performance would then be reflected by how similarly within-category trained instances and within-category novel instances stimulated the deep neural network. To evaluate the generalization performance of the neural network on novel instances, pairwise dissimilarities between two trained instances ($Dissim_{W-TT}$) as well as between a trained and a novel instance ($Dissim_{W-TN}$) were extracted. In the testing phase the chance was low that the neural network readily applied category knowledge earned from thousands of training trials onto a novel instance in the first and only exposure. In the case of poor generalization performance, the activation pattern of within-category novel instances would be dissimilar from that of the within-category trained instances (i.e., increasing $Dissim_{W-TN}$). Our criterion for a successful generalization after learning with symbols was that $Dissim_{W-TN}$ should be as low as $Dissim_{W-TT}$ ($Dissim_{W-TN} \approx Dissim_{W-TT}$). In other words, their absolute dissimilarity difference $DissimDiff = |Dissim_{W-TN} - Dissim_{W-TT}|$ must remain lower than when the deep neural network was trained without symbols.

Cell assembly analysis

Motivated by the notion of cell assemblies (see, e.g., Hebb, 1949; Braitenberg, 1978; Fuster, 2005), that is, strongly interlinked sets of neurons forming as a consequence of correlated neuronal activity and potentially carrying a main role in cognitive brain processing, we conducted cell assembly analyses to discover possible neuronal correlates of grounding instances, concepts and symbols along with instance-specific and category-critical neurons after repeated exposure to instances and their category terms or proper names. We extracted cell assemblies (CAs) activated by each of the 60 grounding patterns used as testing instances based on the criterion described in previous work (Garagnani and Pulvermüller, 2016; Henningsen-Schomers and Pulvermüller, 2022). Grounding patterns in the testing phase tended to coactivate several excitatory neurons (e-cells) in an area, with at least one being maximally responsive (non-response was under the threshold 0.01). To be part of a CA, the firing rate of a given e-cell had to exceed 75% of the firing rate of the maximally responsive cell of the same area. We then computed the number of unique, instance-specific and overlapping, conceptual neurons among CAs for trained instances of the same category: neurons were classified according to whether they were activated by just one grounding patterns or whether they responded to two or three instances (thus being pair or triple-shared between the learnt instances of a concept). Unique neurons were conceptualized as neurons which encoded specific, 'idiosyncratic' features of an instance; shared neurons could be understood as those that encoded common features shared by at least two instances and thus characteristic of their category. The specialized encoding of category-critical features could be indicated by a higher proportion of shared neurons per area, while traces of instance-specific features would be reflected by a larger proportion of unique neurons.

Representations are transformed through different levels of processing, i.e., from the primary areas to secondary areas, and the central "connector hub" areas of the model. We quantified such transformation as the change (i.e., gain/loss) in the number of unique and shared CA-cells in the extrasylvian central areas (AT, PF_L) comparative to the extrasylvian primary areas (V1, $M1_L$). Gains in a type of neuron, for example, shared neurons, are indicative of intensive encoding of concept related commonalities on the course of processing, while loss of shared neurons in the central areas implies that reduced encoding of idiosyncratic features and hence instance-related information. Percentage gain was calculated as the difference between the number of neurons in the central and primary areas, as a percentage with respect to the number of neurons in the primary areas:

$$Gain = \frac{n_{central} - n_{primary}}{n_{primary}} \times 100$$

Representations of category-critical features. A range of previous neurocomputational studies show, that, when brain-like networks learn concepts and word meanings, they form cell assemblies that are spread out across sensorimotor and more central areas of the network. The density of shared semantic neurons in the most central connector hubs is greatest due to their high connectivity degree and thus ample convergence of activity on these areas, resulting in especially strong activation, in particular for shared semantic neurons (for discussion, see Garagnani et al., 2017; Tomasello et al., 2018). Relative to instance-specific neurons, shared semantic neurons are activated more frequently during semantic learning, which predicts that these will recruit the largest number of additional cell assembly; these would therefore be semantic, too, and primarily located in the central hub regions. If a labeling condition specifically invites the neural network to encode category-relevant features, we expect (1) more shared neurons than unique neurons in the extrasylvian areas and (2) a greater gain in shared neurons in the central semantic areas compared to the primary areas. Category learning might still occur even in the presence of proper names because within-category similarities also characterize sensorimotor experiences. If such information is sufficient, there should be traces of shared neurons in the central, multimodal areas as well. Additionally, category terms should activate shared neurons more than proper names.

Representations of instance-specific features. When a neural network represents instances as unique entities, it shall reveal specific traces of each instance in the extrasylvian areas, especially in the semantic hubs. In an extreme case where category learning is hindered and the neural network only encodes the uniqueness of instances, there should be (1) more unique than shared neurons in the extrasylvian areas and (2) a gain only in unique neurons in the central areas with respect to the primary areas. Importantly, instances with proper names are expected to activate significantly more unique neurons than categorically labelled instances.

We gather from all twelve model instantiations the CAs in response to all 30 trained instances of 10 categories and classify CA-cells by their uniqueness to each instance (versus sharedness). To facilitate readers' understanding about the results, we offer an interactive illustration of these CAs on our web application at (https://phucthuun.shinyapps.io/CL_PN/). This web application enables one to compare the differential effects of category terms versus proper names in representing category-critical and instance-specific features of within-category and across-category instances.

Results

Representational similarity analysis

Figure 3B gives a first impression of the instance and category learning performance after 2,000 training trials. In the Category term condition, instances from the same category activated the neural network similarly, whereas instances from different categories led to substantially more dissimilar activation patterns across the different areas of the network (i.e., firing patterns were highly dissimilar, as color-coded by dark blue and pink). Category knowledge was reflected in a relatively reduced dissimilarity (light blues), which appears as homogenous within each category, contrasting with those between categories, especially in the central areas (semantic hubs). Training the deep neural network without the aid of symbols or with proper names reduced the networks' ability to distinguish instances between categories: activity pattern dissimilarities between instances from different categories were much more substantial in the Category term condition than in the Proper name condition (color-coded with shades of intermediate blue). In contrast, within-category similarities and generalization performance in the Category term condition were superior, as indicated by the more homogeneous (light) blue shade across all 6 instances (trained and not-trained) from the same category, relative to the other two conditions, where different shades of light blue are clearly visible.

Category learning

To evaluate category learning performance after 2,000 learning trials, within-category dissimilarity ($Dissim_{W-TT}$) and between-category dissimilarity between activity patterns elicited by grounding patterns of trained instances ($Dissim_{B-TT}$) were used. Figure 4A describes a global tendency of the deep neural network, across its twelve areas and three training conditions, to identify within-category instances as more similar and between-category instances as more dissimilar to each other. This feature is explained by the grounding patterns presented, which were similar across category instances, but not between. However, between-category dissimilarity is relatively enhanced in central areas, a feature not explained by the stimulations. In the next step, dissimilarity values were averaged for the six extrasylvian areas. The two-factorial repeated measures (3 × 2) – ANOVA with training condition (No symbol/Category term/Proper name) and dissimilarity type ($Dissim_{W-TT}/Dissim_{B-TT}$) confirmed the main effect of both factors ($F(2,22) = 2777.647, p < .001, \eta^2 = 0.982$ and $F(1,11) = 11155.611, p < .001, \eta^2 = 0.996$, respectively) as well as their interaction effect ($F(2,22) = 6113.987, p < .001, \eta^2 = 0.986$) on the dissimilarity between instances within these extrasylvian areas. Figure 4B illustrates category-related activation performance of the deep neural network in the extrasylvian areas of the three learning conditions: the neural network successfully grouped together instances from the same category while distinguishing between instances from the same vs. from two different categories. Pairwise comparisons with Bonferroni correction were computed to observe the effect of training condition on each level of dissimilarity type and vice versa. The results showed that $Dissim_{W-TT}$ was significantly lower than $Dissim_{B-TT}$ in all three conditions ($ps < .001$); same category-membership was thus manifest as relatively enhanced activation similarity in all conditions and across areas. The $Dissim_{W-TT}$ in the Category term condition (M=0.229, *SD*=0.005) and the Proper name condition (M=0.264, *SD*=0.004) was significantly smaller (i.e., greater similarity) than that in the control No symbol condition (M=0.29, *SD*=0.006), and they were also significantly different from each other, with greatest similarities after category term labelling ($ps < .001$). Relative to the control No symbol condition, the deep neural network responded similarly to trained instances coming from the same category when it was trained with symbols and such performance was above baseline. Importantly, the benefit of category terms was superior to both training without symbols and with proper names. Likewise, the deep neural network returned the highest $Dissim_{B-TT}$ (M=1.48, *SD*=0.018) for the Category term condition ($ps < .001$), while $Dissim_{B-TT}$ in the Proper name condition (M=0.706, *SD*=0.01) was not significantly different from that in the No symbol condition (M=0.749, *SD*=0.045) ($p = 0.01$), after application of the Bonferroni-corrected significance threshold of $.005$. Compared to the No symbol condition, training with proper names only gradually hindered the discrimination of between-category instances but left the separation of within-category instances unaffected. By contrast, both aspects of category learning were present with the aid of category terms, reduced within- and enhanced between-category similarities.

The simulations performed to control for the number of word form presentations during learning were evaluated using a two-factorial repeated measures (4 × 2) – ANOVA with training condition (now 4 levels, NoS_1x/NoS_3x/CT_1x/PN_3x) and dissimilarity type ($Dissim_{W-TT}$/$Dissim_{B-TT}$). This confirmed the main effect of both factors ($F(1.67,18.35) = 1113.758, p < .001, \eta^2 = 0.964$ and $F(1,11) = 7485.295, p < .001, \eta^2 = 0.993$, respectively) as well as their interaction effect ($F(1.65,18.10) = 1961.497, p < .001, \eta^2 = 0.973$) on the dissimilarity between instances within extra-sylvian areas. Pairwise comparisons with Bonferroni correction were computed to observe the effect of training condition on each level of dissimilarity type and vice versa. In essence, $Dissim_{B-TT}$ in the Category term condition was significantly higher than that in the Proper name and both No symbol control conditions ($ps < .001$) (Figure 4C); category-term learning increased the dissimilarity across conceptual categories relative to no-symbol learning and proper-name learning. The reverse effect, greater dissimilarity values for proper names than category terms, was found within categories. These observations were therefore valid even when proper names were ‘shown’ to the model three times more than category terms during learning.

Generalization

To evaluate the generalization performance of the deep neural network on novel instances, pairwise dissimilarities between two trained instances ($Dissim_{W-TT}$) as well as between a trained and a novel instance ($Dissim_{W-TN}$) were used. Figure 5A illustrates the tendency of the deep neural network to represent two trained instances of the same category as more dissimilar, whereas the representations of a novel and a trained instance from the same category were less dissimilar (lighter-shaded columns were mostly higher than darker-shaded columns). In the six extrasylvian areas, a 3 × 2 – ANOVA was computed with training condition (No symbol/Category term/Proper name) and type of within-category dissimilarity ($Dissim_{W-TT}$/$Dissim_{W-TN}$) as repeated measures factors. Both the main effects of training condition ($F(2,22) = 465.217, p < .001,\ \eta^2 = 0.956$) and dissimilarity type ($F(1,11) = 7711.618, p < .001, \eta^2 = 0.939$) were significant. For these two factors, there was also a significant interaction ($F(2,22) = 635.788, p < .001, \eta^2 = 0.707$) (Figure 5B). The Greenhouse-Geisser sphericity correction to the violated sphericity assumption ($p = .024$) for training conditions ($p[GG] = 2.38 \times 10^{-11}$) confirmed this result. Two-sided pairwise comparisons with Bonferroni correction showed that $Dissim_{W-TN}$ in the Category term ($M = 0.214, SD = 0.004$) and in the Proper name conditions ($M = 0.220, SD = 0.003$) were significantly lower than that in the control No symbol condition ($M = 0.249, SD = 0.004$) ($ps < .001$), but they did not differ significantly from each other ($p = .01$) (Figure 5B). $Dissim_{W-TN}$ was significantly lower than $Dissim_{W-TT}$ in all three conditions ($ps < .001$) (Figure 5B), which means that within-category trained instances were represented as less similar to each other than when each of them was compared with a novel instance from the same category. In other words, trained instances resulted in neuronal response patterns that were more similar to those caused by novel instances than those caused by trained instances from the same category, a finding easily explained by the lack of learning of the idiosyncratic features of novel instances. A further set of pairwise comparisons using Bonferroni correction revealed that the absolute $DissimDiff$ in the No symbol condition ($M = 0.041, SD = 0.016$) was significantly higher than $DissimDiff$ in the Category term condition ($M = 0.016, SD = 0.012$) ($p < .001$) but not significantly different from that in the Proper name condition ($M = 0.044, SD = 0.02$) ($p = .009$). In other words, category-term learning resulted in the most similar processing of learnt and not-learnt instances and thus to the greatest degree of generalization.

Results from the additional simulations controlling for the number of word form presentations during learning (i.e., four training conditions NoS_1x, NoS_3x, CT_1x, PN_3x, see Methods) also confirmed that generalization was maximal for novel members of categories for which category term had been learned (Figure 5C). The mere exposure to instances or learning proper names showed little generalization relative to category learning.

These results investigating brain-constrained neural network correlates of conceptual generalization sit well with well-known observations that language-learning children often generalize – or even overcategorize – category terms to novel items. In case of overgeneralization to an item, subsequent learning may establish a novel category to which the item belongs. While our results offer a mechanistic perspective on generalization, a detailed simulation of overgeneralization and reclassification learning is left for future study.

Cell assembly analysis

Figure 6A illustrates the tendency of the deep neural network to encode fewer unique neurons (U-shaped function across areas) and more shared neurons (inverted U-shaped function) in the extra-sylvian central areas than in the extra-sylvian primary areas. In the first step, the number of unique neurons and shared neurons activated by each instance were calculated and averaged across two training conditions. The repeated measures 3×2 – ANOVA with training condition (No symbol/Category term/Proper name) and neuron type (unique/shared) confirmed the significant main effects ($F(2,22) = 902.098, p < .001, \eta^2 = 0.926$ and $F(1,11) = 13966.410, p < .001, \eta^2 = 0.998$, respectively), and a significant interaction involving both factors ($F(2,22) = 5027.907, p < .001, \eta^2 = 0.985$). The supplementary 2×2 – ANOVA with training condition with symbols (Category term/Proper name) and neuron type (unique/shared) returned comparable results with 2 significant main effects ($F(1,11) = 1009.255, p < .001, \eta^2 = 0.951$ and $F(1,11) = 23994.328, p < .001, \eta^2 = 0.998$, respectively), and a significant interaction involving both factors ($F(1,11) = 4593.789, p < .001, \eta^2 = 0.986$). Pairwise comparisons with Bonferroni correction revealed that category terms made the neural network reactivate more shared neurons ($M = 11.242, SD = 0.127$) than unique neurons ($M = 2.861, SD = 0.051$) ($p < .001$). This also applied for training with proper names (shared neurons: $M = 7.963, SD = 0.222$; unique neurons: $M = 3.89, SD = 0.064$) and training without symbol (shared neurons: $M = 8.029, SD = 0.194$; unique neurons: $M = 4.493, SD = 0.08$) ($ps < .001$) (Figure 6B). Compared to this control condition, the number of unique instance-specific neurons was moderately reduced by proper names, but radically so by category terms ($p < .001$), whereas the number of shared, conceptual-category neurons remained unchanged after proper-name learning ($p = .447$), but increased dramatically with category term acquisition ($p < .001$). The latter is clear evidence for a facilitatory effect of language, more specifically, of category-term learning, on conceptual category formation in brain-constrained deep neural networks.

With respect to the gain/loss of neurons in the extrasylvian central areas relative to the primary ones, our repeated-measure 3×2 – ANOVA with two factors training condition (No symbol/Category term/Proper name) and neuron type (unique/shared) confirmed both main effects on the percentage change of neurons and their interaction to be significant ($F(2,22) = 55.17837, p < .001, \eta^2 = 0.5519424$, $F(1,11) = 6471.54090, p < .001, \eta^2 = 0.9954$, and $F(2,22) = 1484.43893, p < .001, \eta^2 = 0.966$, respectively). According to the subsequent pairwise t-tests, the deep neural networks gained shared neurons but lost unique neurons in the central areas, which held true for all conditions ($ps < .001$) (Figure 6D, upward dotted lines represent positive gains in shared neurons and downward solid lines mean negative gains in unique neurons). On the three levels of training condition, the gain in shared neurons and the loss in unique neurons in the Category term condition were significantly larger than that in the Proper name and No symbol conditions ($ps < .001$) (Figure 6D). Proper names did not significantly increase the gain in shared neurons ($p = .1$) but led only to a moderate loss of unique neurons, as compared to the control training condition ($ps < .001$). These results further confirm that training with category terms magnified both the gain in shared semantic neurons in central areas and the loss of unique instance-specific neurons there. The simulations performed for balancing the number of word form presentations during proper-name and category-term learning also confirmed these observations (Figure 6C, E). Therefore, the overgrowth of shared neurons in category-term learning does not depend on an abundant number of word form presentations and cannot be explained by adding word form information to instance-related information.

Both RSA and CA-analyses were also conducted for the whole model architecture (6 extrasylvian and 6 perisylvian model areas). The data replicated the results indicating category learning (Figure 4-1, Table 4-1), generalization (Figure 5-1, Table 5-1), and representations of category-critical as well as instance-specific features (Figure 6-1, Table 6-1).

Discussion

When sensorimotor patterns simulating the processing of similar objects or actions from different categories were presented, the brain-constrained network applied in the current study showed successful conceptual category learning. Category learning outside symbol context was manifest in greater similarities of activity patterns elicited by different instances of the same category as compared with between-category pattern similarities. Importantly, compared with training of instances per se, concurrent learning of category instances and symbols had a substantial effect on both categorial and instance-specific processes. Category-term learning led to an additional increase in dissimilarities between activity patterns across conceptual categories, while making category members substantially more similar to each other. In contrast, proper-name learning did not change between-category similarities and led to a relatively minor similarity increase between members of the same category. The model gave evidence of generalization to novel members of learned categories and showed that such generalization was maximal for novel members of categories for which category terms had been learned. Meticulous analyses of neuronal activity patterns suggest that the enhancement of within-category similarities and between-category dissimilarities in context of category symbols is due to an increase in the number of cells responding to all category members. Likewise, relative persistence of instance-specific neurons with proper-name learning underlies the maintained activation differences between category instances observed in this case. All observed effects regarding pattern dissimilarities and neuronal microstructure were greatly pronounced in the central 'connector hub' areas of the brain-constrained model applied, as compared with primary areas. Table 3 summarizes major observations in the current data and the corresponding learning aspects these observations reflect.

Relationship to experimental and neurocomputational research

Our results can be used to address observations delivered by neurocognitive and neurobehavioral experiments. Neuropsychological evidence highlights the role of the prefrontal cortex in categorical representation (for review see Kéri, 2003). Prefrontal areas (PF_i and PF_l) are part of the four central areas of our model, where conceptual neurons constituting category representations emerged most numerously. This is explained by the high degree of convergence of neural activity on these areas, which are not only located in the centre of the model architecture but also show the highest connectivity degrees. Due to ample activity converging on these connector hub areas, their frequently activated shared semantic neurons can most efficiently recruit other neurons, which therefore take on similar response properties (Doursat and Bienenstock, 2006). This mechanism may contribute to why these areas act as 'semantic hubs' and house neurons reflecting category membership (e.g., PF and AT, see Miller et al., 2002; Seger and Miller, 2010; Garagnani and Pulvermüller, 2016; Tomasello et al., 2017). On the other hand, the higher density of instance-specific neurons in the primary visual/motor model area relative to the centre is evidence for exemplar learning in the sensorimotor cortices (Bowman et al., 2020; Kéri, 2003) – a type of category learning that is based on the representations of specific category instances (Nosofsky, 1988) and should be independent of signs and symbols. Here, solid evidence for category formation was obtained even in the control condition where only sensorimotor patterns were presented to the model without symbols. In line with neural data (Freedman et al., 2001; Seger and Miller, 2010), experimental evidence shows that perceptuomotor similarities among category members are sufficient to trigger category learning in preverbal infants (Sloutsky and Fisher, 2004; de Heering and Rossion, 2015) and animals (Güntürkün et al., 2018; Pusch et al., 2023).

When learning conceptual instances in context of category terms, infants show most pronounced category building and an attention bias towards shared features of category members (Waxman and Markow, 1995; Dewar and Xu, 2007; Althaus and Mareschal, 2014). In contrast, encountering proper names for individual instances focuses their attention relatively more to object-specific features (Barnhart et al., 2018; Pickron et al., 2018; La Tourette & W, 2020). In the current network model, symbol association raises the number of neurons involving in the processing of a given sensorimotor pattern. This can be interpreted as biased attention to the object or action for which the pattern codes and thus explains why label learning generally increases attention to object features.

Furthermore, as category-term learning increases the number of category-critical shared semantic neurons in the network, at the cost of reducing the number of instance-specific ones, the pre-observed greater attention to shared features has a direct model correlate, along with the label-related tendency to build stronger category representations. Infants' attentional focusing on instance-specific features of objects is in line with the relative preservation of instance-specific neurons in the model of proper-name learning. Thus, the opposing effects of proper name and category-term learning, which, respectively, drive attention towards instance-specific and category general features of objects, are captured by the current model.

A range of neurocomputational studies previously explored the putative brain basis of cognitive processes (e.g., Deco and Rolls, 2005; Rolls and Deco, 2015; Palm, 2016), including conceptual category learning and the influence of language on object perception (Rogers and McClelland, 2014; Henningsen-Schomers and Pulvermüller, 2022). For example, Westermann and Mareschal (2014) demonstrated, using a fully distributed parallel processing model, that learning a category label made the neural patterns of category members more similar to each other, whereas different categories moved away from each other in representational space. Our RSA in models mimicking cortical area structure and connectivity, along with within-area excitatory and inhibitory connectivity, achieved the same result. In addition, we determined the neuron-level mechanisms and contributions of different model areas to this result and, in particular, revealed the model-central connector hub areas as the loci where the differences between categorical and instance-specific mechanisms as well as those between the shared- vs. specific-feature promoting roles of instance-specific and category labels are most pronounced. As to our knowledge, the contrast between activity patterns and neuronal correlates of proper names and category terms has not been addressed by previous computational work.

Model explanation

The present simulations offer explanations of the observed phenomena based on neuroscience principles. Of special relevance here are the biological learning mechanisms applied, which include unsupervised Hebbian synaptic strengthening of connections between co-activated neurons and weakening of links between cells firing independently of each other. This principle explains why category labels primarily interlink with the shared neurons of instance representations belonging to the same category. The reason lies in the highest correlation values, as instance-specific neurons are silent when the category term is used together with other category instances. This implies some weakening of connections between the category terms' and the instance-specific neurons, based on the 'anti-Hebbian' "neurons out-of-sync delink" rule. The opposite difference applies to proper names, whose neural correlates strongly connect to instance-specific neurons but weaken their links with the category-critical shared neurons whenever a different category member co-occurs with its own and thus different name. Effects are most clearly present in the central areas of the network where the neural correlates of words and entities are equally manifest so that their correlation structure can easily be mapped.

Limitations and future direction

The current simulations use idealized instance and category learning conditions. The activation patterns representing conceptual instances and word forms were chosen to be non-overlapping, except for the neurons coding for shared features. These are idealizations considering both the features of word forms and those of objects and actions could be shared across categories (cf. phonological e.g., "cat"-"hat" or perceptual color/shape similarities). Such similarities are irrelevant to category membership and hence were omitted to keep the simulation well-controlled. Secondly, only a small number of conceptual features were realized and a small set of shared features determined concept membership. This situation may hold for some concrete terms but not for others and certainly not for abstract concepts (Henningsen-Schomers et al., 2022). Furthermore, proper names and category terms were acquired by different networks to allow straightforward separation and evaluation of the mechanistic side of different label types – although label types are normally co-present in the same mind and brain. In future, it is desirable to complement this work by simulations of more realistic conceptual categories and to build one model in which interaction/interference effects between different learning conditions are possible.

Conclusion

The current study strived to meet the need for a mechanistic model of symbols and their meaning within a neurobiological computational framework by addressing specific features of proper names (Mickey Mouse) and category symbols (house mouse). Developmentalists and linguists have long been proposing that category terms and proper names distinctively impact infants’ locus of attention towards category-shared and instance-specific object and action features, respectively. By simulating concept and instance learning in a deep neural network with neurobiologically realistic architecture and brain-like connectivity, we demonstrate that learning these two different symbol types had opposing effects on the emergent neuronal cell assemblies representing and processing instances of a category and the shared conceptual features of that category, which can explain pre-observed differences in perceptual, attentive and memory processes related to the specific and shared features of category instances. These explanations were based on unsupervised Hebbian associative learning mechanism binding neurons involved in correlated processing of instance-specific category-general information. The current work could thus not only replicate but also offer underlying neuronal mechanisms and causal neurobiological explanations for well-established observations in cognitive science.

References

Althaus N, Mareschal D (2014) Labels Direct Infants' Attention to Commonalities during Novel Category Learning. PLOS ONE 9:e99670.

Althaus N, Plunkett K (2016) Categorization in infancy: labeling induces a persisting focus on commonalities. Developmental Science 19:770–780.

Arikuni T, Watanabe K, Kubota K (1988) Connections of area 8 with area 6 in the brain of the macaque monkey. Journal of Comparative Neurology 277:21–40.

Artola A, Singer W (1993) Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. Trends in Neurosciences 16:480–487.

Baldwin DA, Markman EM (1989) Establishing World-Object Relations: A First Step. Child Development 60:381–398.

Barnhart WR, Rivera S, Robinson CW (2018) Effects of Linguistic Labels on Visual Attention in Children and Young Adults. Front Psychol 9:358.

Bauer RH, Fuster JM (1978) The effect of ambient illumination on delayed-matching and delayed-response deficits from cooling dorsolateral prefrontal cortex. Behavioral Biology 22:60–66.

Bauer RH, Jones CN (1976) Feedback training of 36 – 44 Hz EEG activity in the visual cortex and hippocampus of cats: Evidence for sensory and motor involvement. Physiology & Behavior 17:885–890.

Bennett L, Melchers B, Proppe B (2020) Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin. :5 S.

Best C, Robinson C, Sloutsky V (2010) The Effect of Labels on Visual Attention: An Eye Tracking Study. Proceedings of the Annual Meeting of the Cognitive Science Society 32 Available at: https://escholarship.org/uc/item/0wn1j6px [Accessed October 9, 2022].

Binder JR, Westbury CF, McKiernan KA, Possing ET, Medler DA (2005) Distinct Brain Systems for Processing Concrete and Abstract Concepts. Journal of Cognitive Neuroscience 17:905–917.

Bowman CR, Iwashita T, Zeithamova D (2020) Tracking prototype and exemplar representations in the brain across learning Behrens TE, Barense M, Barense M, Tompary A, eds. eLife 9:e59360.

Braitenberg V (1978) Cell Assemblies in the Cerebral Cortex. In: Theoretical Approaches to Complex Systems (Heim R, Palm G, eds), pp 171–188 Lecture Notes in Biomathematics. Berlin, Heidelberg: Springer.

Braitenberg V, Schüz A (1998) Cortex: Statistics and Geometry of Neuronal Connectivity. Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-662-03733-1.

Bressler SL, Coppola R, Nakamura R (1993) Episodic multiregional cortical coherence at multiple frequencies during visual task performance. Nature 366:153–156.

Catani M, Jones DK, Donato R, ffytche DH (2003) Occipito-temporal connections in the human brain. Brain 126:2093–2107.

Catani M, Jones DK, Ffytche DH (2005) Perisylvian language networks of the human brain. Annals of Neurology 57:8–16.

Chafee MV, Goldman-Rakic PS (2000) Inactivation of Parietal and Prefrontal Cortex Reveals Interdependence of Neural Activity During Memory-Guided Saccades. Journal of Neurophysiology 83:1550–1566.

Connors BW, Gutnick MJ, Prince DA (1982) Electrophysiological properties of neocortical neurons in vitro. Journal of Neurophysiology 48:1302–1320.

de Heering A, Rossion B (2015) Rapid categorization of natural face images in the infant right hemisphere Culham JC, ed. eLife 4:e06564.

Deacon TW (1992) Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. Brain Research 573:8–26.

Deco G, Rolls ET (2005) Neurodynamics of Biased Competition and Cooperation for Attention: A Model With Spiking Neurons. Journal of Neurophysiology 94:295–313.

Dewar K, Xu F (2007) Do 9-month-old infants expect distinct words to refer to kinds? Developmental Psychology 43:1227–1238.

Di Leo G, Sardanelli F (2020) Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. Eur Radiol Exp 4:18.

Distler C, Boussaoud D, Desimone R, Ungerleider LG (1993) Cortical connections of inferior temporal area TEO in macaque monkeys. J Comp Neurol 334:125–150.

Doursat R, Bienenstock E (2006) Neocortical Self-Structuration as a Basis for Learning. 5th International Conference on Development and Learning (ICDL 2006).

Dum RP, Strick PL (2002) Motor areas in the frontal lobe of the primate. Physiol Behav 77:677–682.

Dum RP, Strick PL (2005) Frontal Lobe Inputs to the Digit Representations of the Motor Areas on the Lateral Surface of the Hemisphere. J Neurosci 25:1375–1386.

Eacott MJ, Gaffan D (1992) Inferotemporal-frontal Disconnection: The Uncinate Fascicle and Visual Associative Learning in Monkeys. European Journal of Neuroscience 4:1320–1332.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. Science 291:312–316.

Frege G (1948) Sense and Reference. The Philosophical Review 57:209.

Fuster JM (2005) Cortex and Mind. Oxford University Press. Available at: https://academic.oup.com/book/1939 [Accessed April 10, 2023].

Fuster JM, Bauer RH, Jervey JP (1985) Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. Brain Research 330:299–307.

Fuster JM, Jervey JP (1981) Inferotemporal Neurons Distinguish and Retain Behaviorally Relevant Features of Visual Stimuli. Science 212:952–955.

Garagnani M, Lucchese G, Tomasello R, Wennekers T, Pulvermüller F (2017) A Spiking Neurocomputational Model of High-Frequency Oscillatory Brain Responses to Words and

Pseudowords. Frontiers in Computational Neuroscience 10 Available at: https://www.frontiersin.org/article/10.3389/fncom.2016.00145 [Accessed March 5, 2022].

Garagnani M, Pulvermüller F (2016) Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs Barbas H, ed. Eur J Neurosci 43:721–737.

Garagnani M, Wennekers T, Pulvermüller F (2007) A neuronal model of the language cortex. Neurocomputing 70:1914–1919.

Gelman SA, Markman EM (1986) Categories and induction in young children. Cognition 23:183–209.

Gelman SA, Markman EM (1987) Young Children's Inductions from Natural Kinds: The Role of Categories and Appearances. Child Development 58:1532–1541.

Gierhan SME (2013) Connections for auditory language in the human brain. Brain and Language 127:205–221.

Graham S, Keates J, Vukatana E, Khu M (2013) Distinct Labels Attenuate 15-Month-Olds' Attention to Shape in an Inductive Inference Task. Frontiers in Psychology 3 Available at: https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00586 [Accessed July 24, 2022].

Güntürkün O, Koenen C, Iovine F, Garland A, Pusch R (2018) The neuroscience of perceptual categorization in pigeons: A mechanistic hypothesis. Learn Behav 46:229–241.

Guye M, Parker GJM, Symms M, Boulby P, Wheeler-Kingshott CAM, Salek-Haddadi A, Barker GJ, Duncan JS (2003) Combined functional MRI and tractography to demonstrate the connectivity of the human primary motor cortex in vivo. NeuroImage 19:1349–1360.

Harris CR et al. (2020) Array programming with NumPy. Nature 585:357–362.

Hebb DO (1949) The Organization of Behavior: A Neuropsychological Theory. Oxford, England: Wiley.

Henningsen-Schomers MR, Garagnani M, Pulvermüller F (2022) Influence of language on perception and concept formation in a brain-constrained deep neural network model. Philosophical Transactions of the Royal Society B: Biological Sciences 378:20210373.

Henningsen-Schomers MR, Garagnani M, Pulvermüller F (2023) Influence of language on perception and concept formation in a brain-constrained deep neural network model. Phil Trans R Soc B 378:20210373.

Henningsen-Schomers MR, Pulvermüller F (2022) Modelling concrete and abstract concepts using brain-constrained deep neural networks. Psychological Research 86:2533–2559.

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. Comput Sci Eng 9:90–95.

Kaas JH (1997) Topographic Maps are Fundamental to Sensory Processing. Brain Research Bulletin 44:107–112.

Kaas JH, Hackett TA (2000) Subdivisions of auditory cortex and processing streams in primates. Proceedings of the National Academy of Sciences 97:11793–11799.

Kassambara A (2021) rstatix: Pipe-Friendly Framework for Basic Statistical Tests. Available at: https://CRAN.R-project.org/package=rstatix [Accessed July 25, 2022].

Kéri S (2003) The cognitive neuroscience of category learning. Brain Research Reviews 43:85–109.

Knoblauch A, Palm G (2002) Scene segmentation by spike synchronization in reciprocally connected visual areas. I. Local effects of cortical feedback. Biol Cybern 87:151–167.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2 Available at: https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008 [Accessed April 8, 2023].

LaTourrette AS, Waxman SR (2020) Naming guides how 12-month-old infants encode and remember objects. PNAS 117:21230–21234.

Lu M-T, Preston JB, Strick PL (1994) Interconnections between the prefrontal cortex and the premotor areas in the frontal lobe. Journal of Comparative Neurology 341:375–392.

Majid A, Bowerman M, Kita S, Haun DBM, Levinson SC (2004) Can language restructure cognition? The case for space. Trends in Cognitive Sciences 8:108–114.

Makris N, Pandya DN (2009) The extreme capsule in humans and rethinking of the language circuitry. Brain Struct Funct 213:343–358.

Malcolm P. Young, Jack W. Scanneil, Gully A. P. C. Burns, Colin Blakemore (1994) Analysis of Connectivity: Neural Systems in the Cerebral Cortex. Reviews in the Neurosciences 5:227–250.

Martin A (2007) The Representation of Object Concepts in the Brain. Annu Rev Psychol 58:25–45.

Matthews GG (2001) Neurobiology: molecules, cells, and systems, 2nd ed. Malden, MA: Blackwell Science.

Meyer JW, Makris N, Bates JF, Caviness VS, Kennedy DN (1999) MRI-Based topographic parcellation of human cerebral white matter. Neuroimage 9:1–17.

Miller EK, Freedman DJ, Wallis JD (2002) The prefrontal cortex: categories, concepts and cognition Parker A, Derrington A, Blakemore C, eds. Phil Trans R Soc Lond B 357:1123–1136.

Miller TM, Schmidt TT, Blankenburg F, Pulvermüller F (2018) Verbal labels facilitate tactile perception. Cognition 171:172–179.

Nosofsky RM (1988) Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition 14:700–708.

Palm G (2016) Neural Information Processing in Cognition: We Start to Understand the Orchestra, but Where is the Conductor? Frontiers in Computational Neuroscience 10 Available at: https://www.frontiersin.org/articles/10.3389/fncom.2016.00003 [Accessed May 14, 2023].

Pandya DN (1995) Anatomy of the auditory cortex. Rev Neurol (Paris) 151:486–494.

Pandya DN, Barnes CL (1987) Architecture and connections of the frontal lobe. In: The frontal lobes revisited, pp 41–72. New York, NY, US: The IRBN Press.

Pandya DN, Yeterian EH (1985) Architecture and Connections of Cortical Association Areas. In: Association and Auditory Cortices (Peters A, Jones EG, eds), pp 3–61 Cerebral Cortex. Boston, MA: Springer US. Available at: https://doi.org/10.1007/978-1-4757-9619-3_1 [Accessed October 25, 2022].

Parker A, Gaffan D (1998) Interaction of frontal and perirhinal cortices in visual object recognition memory in monkeys. European Journal of Neuroscience 10:3044–3057.

Parker GJM, Luzzi S, Alexander DC, Wheeler-Kingshott CAM, Ciccarelli O, Lambon Ralph MA (2005) Lateralization of ventral and dorsal auditory-language pathways in the human brain. NeuroImage 24:656–666.

Paus T, Castro-Alamancos MA, Petrides M (2001) Cortico-cortical connectivity of the human mid-dorsolateral frontal cortex and its modulation by repetitive transcranial magnetic stimulation. European Journal of Neuroscience 14:1405–1411.

Petrides M, Pandya DN (2009) Distinct Parietal and Temporal Pathways to the Homologues of Broca's Area in the Monkey Ungerleider L, ed. PLoS Biol 7:e1000170.

Pickron CB, Iyer A, Fava E, Scott LS (2018) Learning to Individuate: The Specificity of Labels Differentially Impacts Infant Visual Attention. Child Dev 89:698–710.

Plunkett K, Hu J-F, Cohen LB (2008) Labels can override perceptual categories in early infancy. Cognition 106:665–681.

Pulvermüller F, Tomasello R, Henningsen-Schomers MR, Wennekers T (2021) Biological constraints on neural network models of cognitive function. Nat Rev Neurosci 22:488–502.

Pusch R, Clark W, Rose J, Güntürkün O (2023) Visual categories and concepts in the avian brain. Anim Cogn 26:153–173.

R Core Team (2021) R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/ [Accessed July 25, 2022].

Ralph MAL, Jefferies E, Patterson K, Rogers TT (2017) The neural and computational bases of semantic cognition. Nat Rev Neurosci 18:42–55.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724.

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proceedings of the National Academy of Sciences 97:11800–11806.

Reback J et al. (2022) pandas-dev/pandas: Pandas 1.4.3. Available at: https://zenodo.org/record/3509134 [Accessed July 25, 2022].

Rilling JK (2014) Comparative primate neuroimaging: insights into human brain evolution. Trends in Cognitive Sciences 18:46–55.

Rilling JK, Glasser MF, Jbabdi S, Andersson J, Preuss TM (2012) Continuity, Divergence, and the Evolution of Brain Language Pathways. Front Evol Neurosci 3:11.

Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, Behrens TEJ (2008) The evolution of the arcuate fasciculus revealed with comparative DTI. Nat Neurosci 11:426–428.

Rilling JK, van den Heuvel MP (2018) Comparative Primate Connectomics. BBE 91:170–179.

Rizzolatti G, Luppino G (2001) The Cortical Motor System. Neuron 31:889–901.

Rogers TT, McClelland JL (2014) Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. Cogn Sci 38:1024–1077.

Rolls ET, Deco G (2010) The Noisy BrainStochastic Dynamics as a Principle of Brain Function. Oxford University Press. Available at: https://academic.oup.com/book/7413 [Accessed October 9, 2022].

Rolls ET, Deco G (2015) Networks for memory, perception, and decision-making, and beyond to how the syntax for language might be implemented in the brain. Brain Research 1621:316–334.

Romanski L m., Bates J f., Goldman-Rakic P s. (1999a) Auditory belt and parabelt projections to the prefrontal cortex in the Rhesus monkey. Journal of Comparative Neurology 403:141–157.

Romanski LM (2007) Representation and Integration of Auditory and Visual Stimuli in the Primate Ventral Lateral Prefrontal Cortex. Cerebral Cortex 17:i61–i69.

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999b) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 2:1131–1136.

Saur D, Kreher BW, Schnell S, Kümmerer D, Kellmeyer P, Vry M-S, Umarova R, Musso M, Glauche V, Abel S, Huber W, Rijntjes M, Hennig J, Weiller C (2008) Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences 105:18035–18040.

Schomers MR (2017) Establishing action-perception circuits as a neural basis for meaning-carrying linguistic symbols – the role of frontal speech motor areas and fronto-temporal connectivity. Available at: http://doi.org/.

Schomers MR, Garagnani M, Pulvermüller F (2017) Neurocomputational Consequences of Evolutionary Connectivity Changes in Perisylvian Language Cortex. J Neurosci 37:3045–3055.

Scott LS, Monesson A (2009) The Origin of Biases in Face Perception. Psychol Sci 20:676–680.

Seger CA, Miller EK (2010) Category Learning in the Brain. Annu Rev Neurosci 33:203–219.

Seltzer B, Pandya DN (1989) Intrinsic connections and architectonics of the superior temporal sulcus in the rhesus monkey. Journal of Comparative Neurology 290:451–471.

Sloutsky VM, Fisher AV (2004) Induction and Categorization in Young Children: A Similarity-Based Model. Journal of Experimental Psychology: General 133:166–188.

Thiebaut de Schotten M, Dell'Acqua F, Valabregue R, Catani M (2012) Monkey to human comparative anatomy of the frontal lobe association tracts. Cortex 48:82–96.

Tomasello R, Garagnani M, Wennekers T, Pulvermüller F (2017) Brain connections of words, perceptions and actions: A neurobiological model of spatio-temporal semantic activation in the human cortex. Neuropsychologia 98:111–129.

Tomasello R, Garagnani M, Wennekers T, Pulvermüller F (2018) A Neurobiologically Constrained Cortex Model of Semantic Grounding with Spiking Neurons and Brain-Like Connectivity. Frontiers in Computational Neuroscience 12 Available at: https://www.frontiersin.org/article/10.3389/fncom.2018.00088 [Accessed February 12, 2022].

Ungerleider LG, Gaffan D, Pelak VS (1989) Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. Exp Brain Res 76:473–484.

Vanek N, Sóskuthy M, Majid A (2021) Consistent verbal labels promote odor category learning. Cognition 206:104485.

Virtanen P et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272.

Wakana S, Jiang H, Nagae-Poetscher LM, van Zijl PCM, Mori S (2004) Fiber Tract–based Atlas of Human White Matter Anatomy. Radiology 230:77–87.

Waskom M (2021) seaborn: statistical data visualization. JOSS 6:3021.

Waxman SR, Booth AE (2001) Seeing Pink Elephants: Fourteen-Month-Olds' Interpretations of Novel Nouns and Adjectives. Cognitive Psychology 43:217–242.

Waxman SR, Markow DB (1995) Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants. Cognitive Psychology 29:257–302.

Webster MJ, Bachevalier J, Ungerleider LG (1994) Connections of Inferior Temporal Areas TEO and TE with Parietal and Frontal Cortex in Macaque Monkeys. Cerebral Cortex 4:470–483.

Westermann G, Mareschal D (2014) From perceptual to language-mediated categorization. Philosophical Transactions of the Royal Society B: Biological Sciences 369:20120391.

Whorf BL, Carroll JB (2007) Language, thought, and reality: selected writings, 28. print. Cambridge, Mass: The MIT Press.

Wittgenstein L (1922) Tractatus logico-philosophicus. London: Routledge & Kegan Paul.

Yeterian EH, Pandya DN, Tomaiuolo F, Petrides M (2012) The cortical connectivity of the prefrontal cortex in the monkey brain. Cortex 48:58–81.

Young M, Scannell JW, Burns G (1995a) The analysis of cortical connectivity. New York : Austin: Springer ; R.G. Landes.

Young MP, Scannell JW, Burns G (1995b) The Analysis of Cortical Connectivity, 1st ed. Springer Berlin, Heidelberg. Available at: https://link.springer.com/book/9783540604013 [Accessed October 25, 2022].

Yuille AL, Geiger D (1995) Winner-Take-All Mechanisms. In: Handbook of Brain Theory and Neural Networks (Arbib MA, ed), pp 1–1056. MIT Press.

Tables

Table 1. Connectivity structure of the modelled cortical areas with neuroanatomical evidence. Table taken from Tomasello et al. (2018).

Modelled areas	**References**
Between-area connectivity (black arrows)	
Perisylvian system	
A1, AB, PB	Pandya and Yeterian, 1985; Pandya, 1995; Rauschecker and Tian, 2000
PF$_i$, PM$_i$, M1$_i$	Pandya and Yeterian, 1985; Young et al., 1995
Extra-sylvian system	
V1, TO, AT	Bressler et al., 1993; Distler et al., 1993
PF$_L$, PM$_L$, M1$_L$	Pandya and Yeterian, 1985; Arikuni et al., 1988; Lu et al., 1994; Rizzolatti and Luppino, 2001; Dum and Strick, 2002, 2005
Between system	
AT, PB	Gierhan, 2013
PF$_i$, PF$_L$	Yeterian et al., 2012
Long distance cortico-cortical connections (purple arrows)	
Perisylvian system	
PF$_i$, PB	Meyer et al., 1999 p.19; Romanski et al., 1999a, 1999b; Paus et al., 2001; Catani et al., 2005 p.200; Parker et al., 2005; Rilling et al., 2008; Makris and Pandya, 2009
PB, PM$_i$	Rilling et al., 2008; Saur et al., 2008
AB, PF$_i$	Romanski et al., 1999a, 1999b; Kaas and Hackett, 2000; Petrides and Pandya, 2009; Rauschecker and Scott, 2009
Extra-sylvian system	
AT, PF$_L$	Bauer and Jones, 1976 p.197; Fuster et al., 1985 p.198; Ungerleider et al., 1989; Eacott and Gaffan, 1992; Webster et al., 1994; Parker and Gaffan, 1998; Chafee and Goldman-Rakic, 2000
AT, PM$_L$	Bauer and Fuster, 1978; Fuster et al., 1985; Pandya and Barnes, 1987; Seltzer and Pandya, 1989; Chafee and Goldman-Rakic, 2000
TO, PF$_L$	Bauer and Jones, 1976; Fuster and Jervey, 1981; Fuster et al., 1985; Seltzer and Pandya, 1989; Makris and Pandya, 2009
Between systems	
PB, PF$_L$	Pandya and Barnes, 1987; Romanski et al., 1999a, 1999b
AT, PF$_i$	Pandya and Barnes, 1987; Ungerleider et al., 1989; Webster et al., 1994; Romanski, 2007; Petrides and Pandya, 2009; Rilling, 2014
Second-next neighbor "jumping" links (blue arrows)	
Perisylvian system (Rilling et al., 2008, 2012; Thiebaut de Schotten et al., 2012; Rilling and van den Heuvel, 2018)	
A1, PB	Pandya and Yeterian, 1985; Malcolm P. Young et al., 1994
PF$_i$, M1$_i$	Deacon, 1992; Young et al., 1995b; Guye et al., 2003
Extra-sylvian system (Thiebaut de Schotten et al., 2012)	
V1, AT	Catani et al., 2003; Wakana et al., 2004
PF$_L$, M1$_L$	Deacon, 1992; Young et al., 1995a; Guye et al., 2003

Table 2. Parameter values used in the simulations. For details and more elaborate discussion on the corresponding equations as well as their mathematical implementations, please see Henningsen-Schomers et al. (2022).

Eq. (1)	Time constant (excitatory cells)	$\tau = 2.5$ (time steps)
	Time constant (inhibitory cells)	$\tau = 5$ (time steps)
	Total input rescaling factor	$k_1 = 0.01$
	Noise amplitude	$k_2 = 7\sqrt{(24/\Delta t)}$ ($\Delta t = 0.5$ ms)
	Global inhibition strength	$k_G = 0.80$ (time steps)
Eq. (3)	Spiking threshold	thresh $= 0.18$
	Adaptation strength	$\alpha = 8.0$
Eq. (4)	Adaption time constant	$\tau_{ADAPT} = 10$ (time steps)
Eq. (5)	Rate-estimate time constant	$\tau_{Favg} = 30$ (time steps, training)
		$\tau_{Favg} = 5$ (time steps, testing)
Eq. (6)	Global inhibition time constant	$\tau_{FGLOB} = 12$ (time steps)
Eq. (7)	Postsynaptic potential thresholds	$\vartheta_+ = 0.15$ (LTP)
		$\vartheta_- = 0.14$ (LTD)
	Presynaptic output activity required for any synaptic change	$\vartheta_{pre} = 0.05$ (LTP)
	Learning rate	$\Delta w = 0.0008$

Table 3. Critical and significant observations and the corresponding aspects of learning.

Analysis	Learning aspect	Observation
RSA	Category learning	Successful category learning in all learning conditions $Dissim_{B-TT} > Dissim_{W-TT}$
		Interaction effect of Symbol type and within/between categories $Dissim_{B-TT_{CT}} > Dissim_{B-TT_{PN}}; Dissim_{B-TT_{CT}} > Dissim_{B-TT_{PN}}$ $Dissim_{W-TT_{CT}} < Dissim_{W-TT_{PN}}; Dissim_{W-TT_{CT}} < Dissim_{W-TT_{NoL}}$
	Generalization	Symbol effect on dissimilarity differences within category $DissimDiff_{CT} < DissimDiff_{NoS}$ $DissimDiff_{CT} < DissimDiff_{PN}$
CA Analysis		Tendency to encode shared features in all learning conditions $n_S > n_U$
	Representations of category-critical features	Symbol effect on the number of shared neurons $n_{S_{CT}} > n_{S_{PN}}; n_{S_{CT}} > n_{S_{NoL}}$
		Gain in shared neurons in central area in all learning conditions $n_{S-central} > n_{S-primary}$
		Symbol effect on across-area gain of shared neurons $Gain_{S_{CT}} > Gain_{S_{PN}}; Gain_{S_{CT}} > Gain_{S_{NoL}}$
	Representations of instance-specific features	Symbol effect on the number of unique neurons $n_{U_{PN}} > n_{U_{CT}}; n_{U_{NoL}} > n_{U_{CT}}$
		Loss in unique neurons in central areas in all learning conditions $n_{S-central} > n_{S-primary}$
		Symbol effect on across-area loss of unique neurons $Loss_{S_{PN}} < Loss_{S_{CT}}$

Abbreviation: $Dissim_{W-TT}/Dissim_{W-TN}$ = Dissimilarity between a trained instance and another trained instance/novel instance of the same category; $Dissim_{B-TT}$ = Dissimilarity between two trained instances from different categories; $DissimDiff = |Dissim_{W-TN} - Dissim_{W-TT}|$; n_S = number of shared neuron; n_U = number of unique neuron; CT = Category term; PN = Proper name; NoS = No symbol.

Figure legends

Figure 1. A) Area structure and between-area connectivity of the neural network model. **Left:** The network model's 12 cortical areas in the left fronto-temporo-occipital lobes: inferior-frontal articulatory (red) and superior temporal auditory systems (blue) of the perisylvian areas, and the lateral frontal hand-motor system (yellow/orange/brown) and visual "what" stream (green) in the extrasylvian cortex. **Right:** Connections among the 12 modelled brain areas: direct connections between adjacent areas (black arrows), second nearest-neighbor areas (blue arrows), and long-distant links (purple arrows). Figure modified from Tomasello et al. (2018). **B) Schematic illustrations of activity patterns for instances of two categories.** The categories are illustrated with images of robots and cat faces, but note that this is for illustrative purposes. The actual input to the model was not images, but grounding patterns consisting of sets of activated neurons (see main text for details). Active neurons of given activity patterns were either shared among instances of the same category (black) or unique to each instance (color). Each model area included 25×25 excitatory neurons, i.e., 625 cells. **Left:** In grounding patterns (i) presented to *V1/*M1$_L$, 6 shared active neurons (black) code for the common perceptual-semantic features of the category "a" and 6 unique neurons (color) represent instance-specific perceptuomotor features from each of the category members. Member instances of one category activated the same six shared neurons while the instance from another category activated a different set of six shared neurons; each instance also activated six unique neurons. **Middle:** 12 neurons (black) make up word form pattern for the category term; in the Category term condition, member instances co-activated with the same word form pattern (ii) in *A1/*M1$_i$. **Right:** 12 unique neurons (color) represent each proper name of an individual instance, which are activated 1-to-1 with these instances in the Proper name condition. Instances were co-activated with distinct different word form patterns (iii) in *A1/*M1$_i$ regardless of category. **C) Simulating no-symbol learning** (top), **category-term learning** (bottom-left), and **proper-name learning** (bottom-right) where no word form pattern, word form patterns (ii), and word form pattern (iii) were presented to *A1/*M1$_i$, respectively.

Figure 2. Experiment design used for instance learning and conceptual grounding. **A) Training phase** with 30 object instances from ten categories. The categories are illustrated with images of robots and cat faces, but note that this is for illustrative purposes. The actual input to the model was not images, but grounding patterns consisting of sets of activated neurons (see main text for details). For each trained instance, the grounding pattern (i) was either presented to the network on its own (No symbol) or combined with a 'word form pattern' of type (ii) (Category term) or type (iii) (Proper name). **B) Testing phase** with a collection of the initially trained 30 instances and 30 novel instances from the 10 original categories, resulting in 60 testing instances (i.e., 6 per category). **C) Training conditions in the main simulations** (top) **and control simulations** (bottom) differ in the number of training trials (tt) to match the number of instance representations and the number of word form representations, respectively.

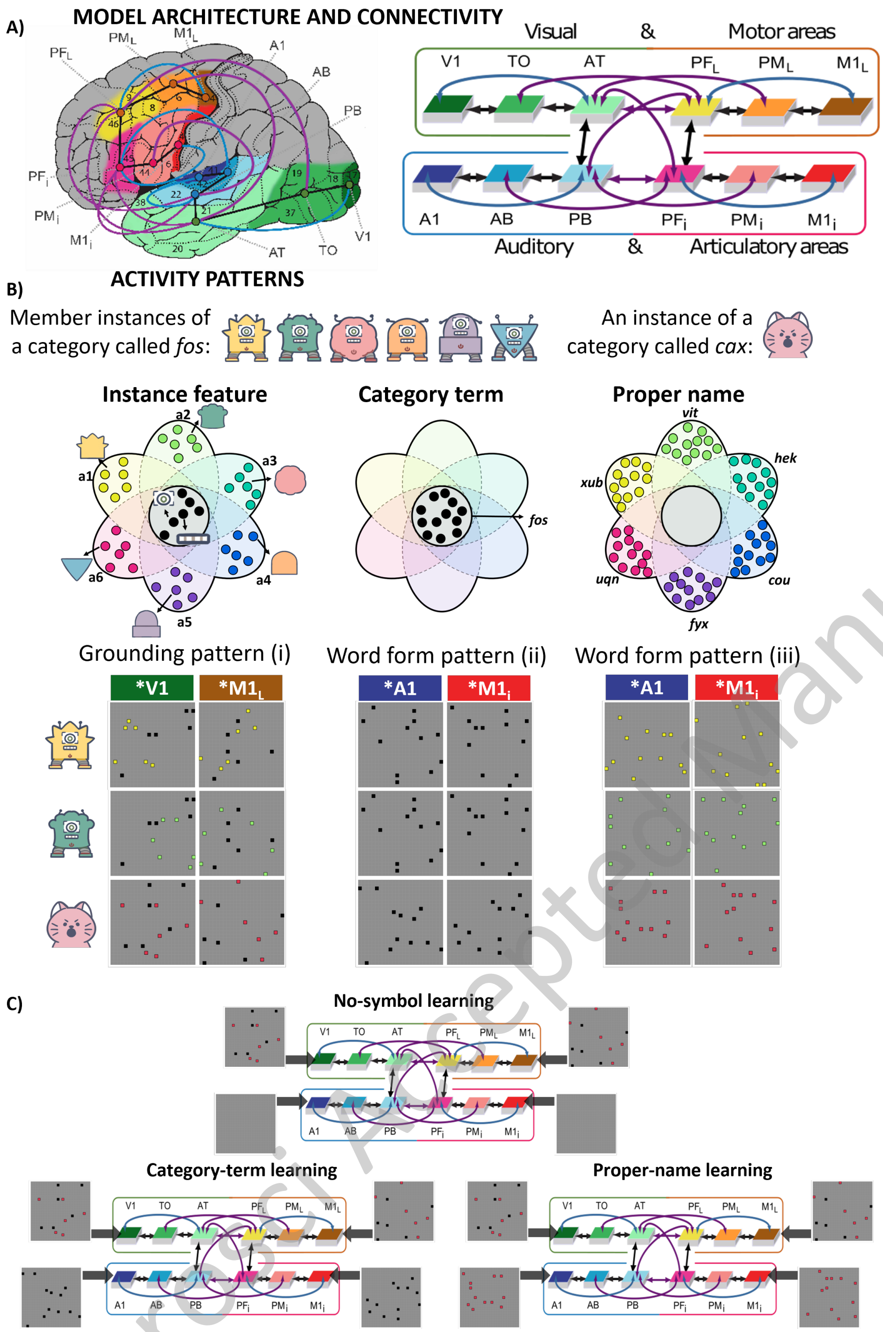
Figure 3. A) Schematic extraction of a 60×60 Representational Dissimilarity Matrix (RDM) which represents 12 instances from two different categories and the similarities between any instance pair. For illustration, we once again use the categories of robots and cat faces. The schematic dissimilarity matrix illustrates how between-category (cells outside the red boundaries) within-category dissimilarities (cells within the red boundaries) were calculated. Of interest are the (1) within-category dissimilarity among trained instances ($Dissim_{W-TT}$, lightest blue shade), (2) within-category dissimilarity between a trained and a novel instance ($Dissim_{W-TN}$, intermediate blue shade), and (3) between-category dissimilarity of two trained instances ($Dissim_{B-TT}$, darkest blue shade). The RDM is symmetric about its diagonal (grey) of zeros (representing the non-dissimilarity of each of the instances to itself). Only the upper half of the RDM is used for analysis and the lower half could be abandoned (black). **B) RDMs for each of the twelve model areas in three main simulations**: No symbol (top row), Category term (middle row), and Proper name (bottom row). Squares indicate the degrees to which network activity in the 12 network areas elicited by (12 out of 60) grounding patterns in the three learning conditions differed between each other within and between categories and are

color-coded from turquoise (no dissimilarity, $Dissim = 0$), blue, pink, and to dark red (high dissimilarity, $Dissim > 3$).

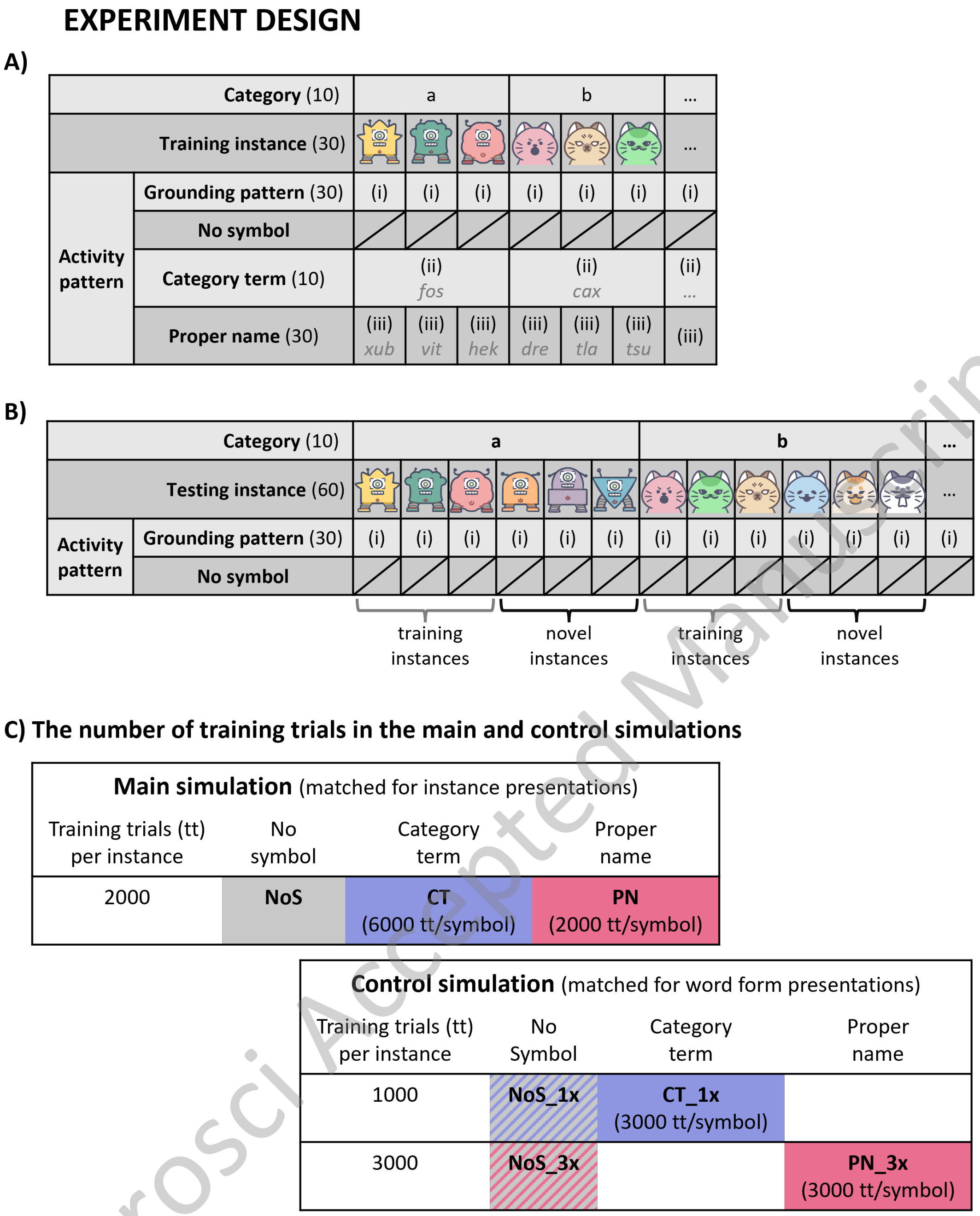
Figure 4. Bar charts depicting dissimilarities between network activity elicited by trained grounding patterns after learning for each of the three training conditions. **A)** Main simulation: Within-category (W-TT) and between-category (B-TT) dissimilarity values across all 30 trained activity patterns were averaged for each of the twelve model areas. **B-C)** Within-category (W-TT) and between-category (B-TT) dissimilarities across the 30 trained items were averaged for extrasylvian model areas. The three training conditions of the main simulations (B) were No symbol (NoS, grey), Category term (CT, blue) and Proper name (PN, pink). The four training conditions of the control simulation (C) were No symbol with each instance presented over 1,000 (NoS_1x, blue-striped grey) or 3,000 trials (NoS_3x, pink-striped grey), Category term where each instance presented over 1,000 trials (CT_1x, blue) and Proper name where each instance presented over 3,000 trials (PN_3x, pink). Error bars represent 95% confidence intervals of the mean. Circles above the bars represent post hoc pairwise comparisons between a reference (circles with filled colored) and a corresponding mean (unfilled circles) after Bonferroni correction (critical p value = 0.005). 10 comparisons relevant to the main effects of training condition and dissimilarity type and for their interaction are illustrated. Asterisks represent two-tailed p values: ** $p < .005$, *** $p < .001$, ns: not significant. The results were replicated in the whole model architecture (6 extrasylvian and 6 perisylvian model areas); see Figure 4-1 and Table 4-1.

Figure 5. Bar charts depicting dissimilarities between network activity elicited by trained novel grounding patterns after learning for each of the three training conditions. **A)** Main simulation: Within-category dissimilarity values between any two trained instances (W-TT) and between trained and novel instances were averaged for each of the twelve model areas. **B&C)** Within-category dissimilarities between any two trained instances (W-TT) and between trained and novel instances (W-TN) were averaged for extrasylvian model areas. The three training conditions of the main simulations (B) were No symbol (NoS, grey), Category term (CT, blue) and Proper name (PN, pink). The four training conditions of the control simulation (C) were NoS_1x (blue-striped grey) or NoS_3x (pink-striped grey), CT_1x (blue) and PN_3x (pink). For further explanation, see Figure 4. The results were replicated in the whole model architecture (6 extrasylvian and 6 perisylvian model areas); see Figure 5-1 and Table 5-1.

Figure 6. Bar charts depicting average numbers of instance specific ('unique') and category-general ("shared") neurons activated by grounding patterns of instances learnt in the three training conditions, No symbol (grey), Category term (blue) and Proper name (pink). **A)** Main simulation: The number of activated unique (U) and shared (S) neurons in response to each of the 30 trained instances was averaged across all twelve model areas. **B &C)** The number of activated neurons in response to the 30 trained grounding patterns was averaged for each of the six extrasylvian areas. **D&E)** Changes in neuronal activation seen between extrasylvian primary areas, where stimulation was given, and the 'higher' more central connector hub areas central to the architecture. Changes in the number of activated neurons in response to trained grounding patterns are shown for the three training conditions. Unique neurons are shown by solid lines with crossed ends, shared ones by broken lines with triangular ends. The three training conditions of the main simulations (B, D) were No symbol (NoS, grey), Category term (CT, blue) and Proper name (PN, pink). The four training conditions of the control simulation (C) were NoS_1x (blue-striped grey) or NoS_3x (pink-striped grey), CT_1x (blue) and PN_3x (pink). For further explanations see Figure 4. The results were replicated in the whole model architecture (6 extrasylvian and 6 perisylvian model areas); see Figure 6-1 and Table 6-1.


A)
MODEL ARCHITECTURE AND CONNECTIVITY
PM$_L$
M1$_L$
PF$_L$
A1
AB
PB
PF$_i$
PM$_i$
M1$_i$
AT
TO
V1
Visual & Motor areas
V1 TO AT PF$_L$ PM$_L$ M1$_L$
A1 AB PB PF$_i$ PM$_i$ M1$_i$
Auditory & Articulatory areas
B)
ACTIVITY PATTERNS
Member instances of a category called *fos*:
An instance of a category called *cax*:
Instance feature
Category term
Proper name
a1
a2
a3
a4
a5
a6
fos
vit
hek
xub
uqn
cou
fyx
Grounding pattern (i)
Word form pattern (ii)
Word form pattern (iii)
*V1
*M1$_L$
*A1
*M1$_i$
*A1
*M1$_i$
C)
No-symbol learning
Category-term learning
Proper-name learning

EXPERIMENT DESIGN



A)

	Category (10)	a			b			...
	Training instance (30)							...
Activity pattern	**Grounding pattern** (30)	(i)	(i)	(i)	(i)	(i)	(i)	(i)
	No symbol							
	Category term (10)	(ii) *fos*			(ii) *cax*			(ii) ...
	Proper name (30)	(iii) *xub*	(iii) *vit*	(iii) *hek*	(iii) *dre*	(iii) *tla*	(iii) *tsu*	(iii)

B)

	Category (10)	**a**						**b**						**...**
	Testing instance (60)													...
Activity pattern	**Grounding pattern** (30)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)	(i)
	No symbol													

training instances novel instances training instances novel instances

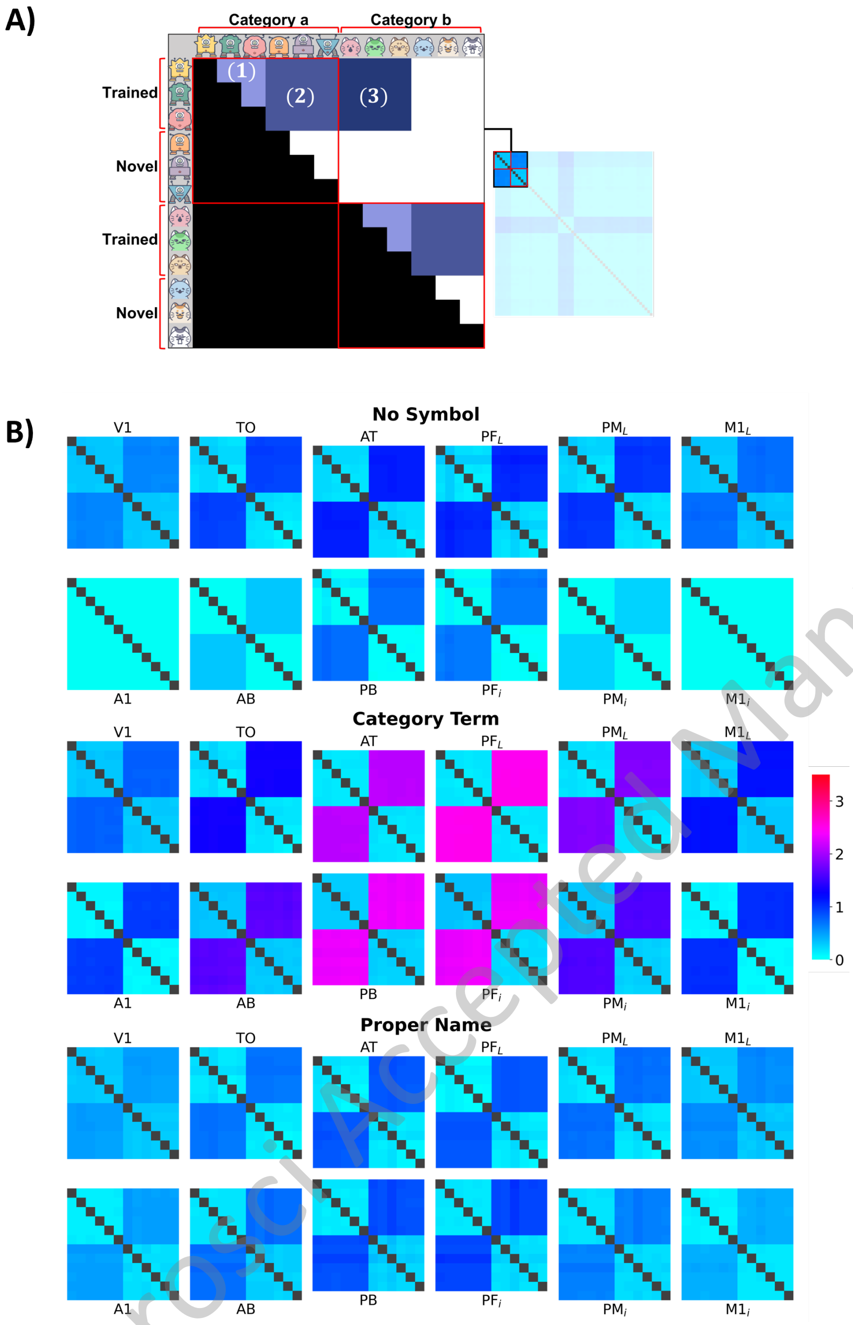
C) The number of training trials in the main and control simulations

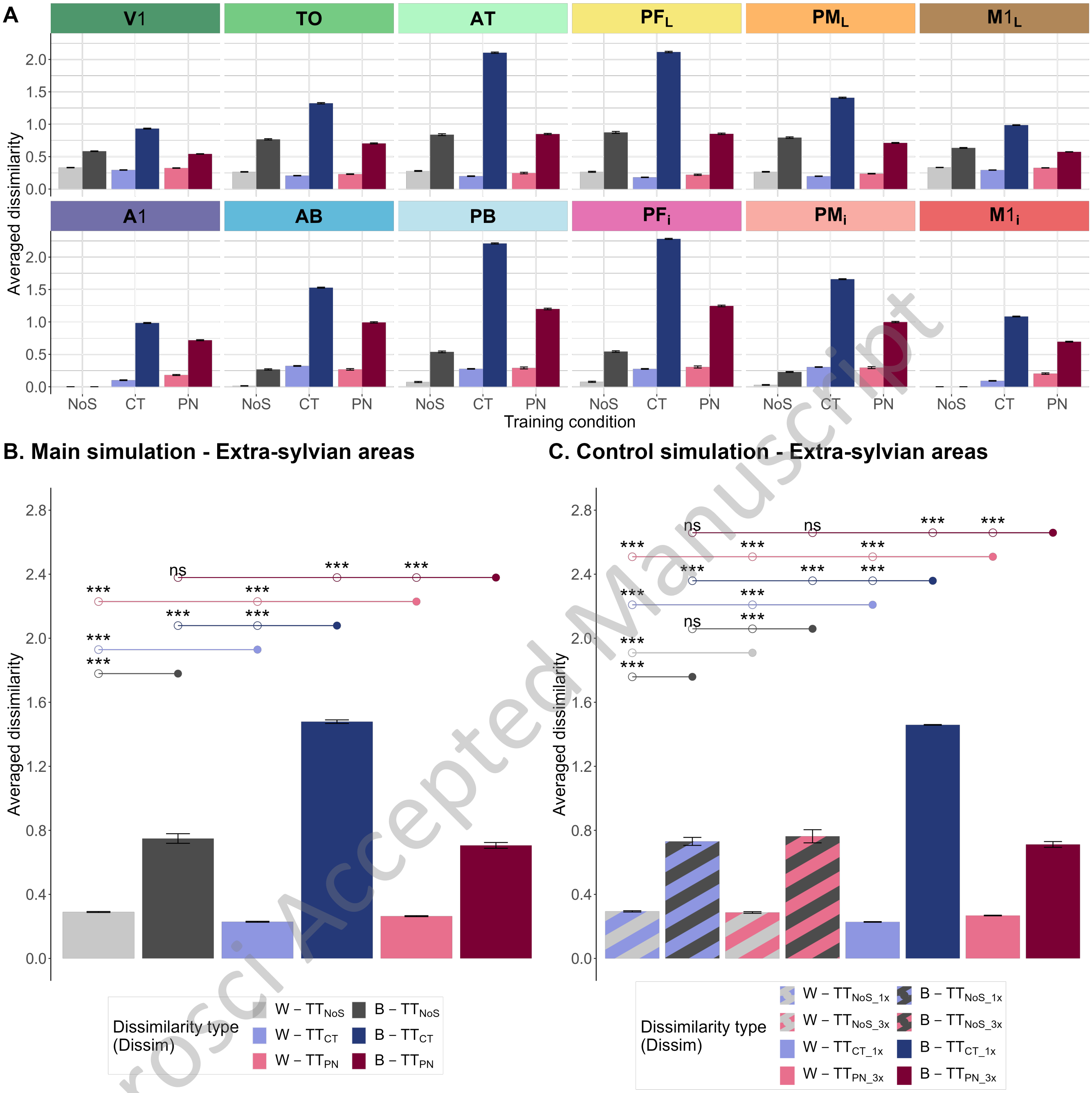
Main simulation (matched for instance presentations)

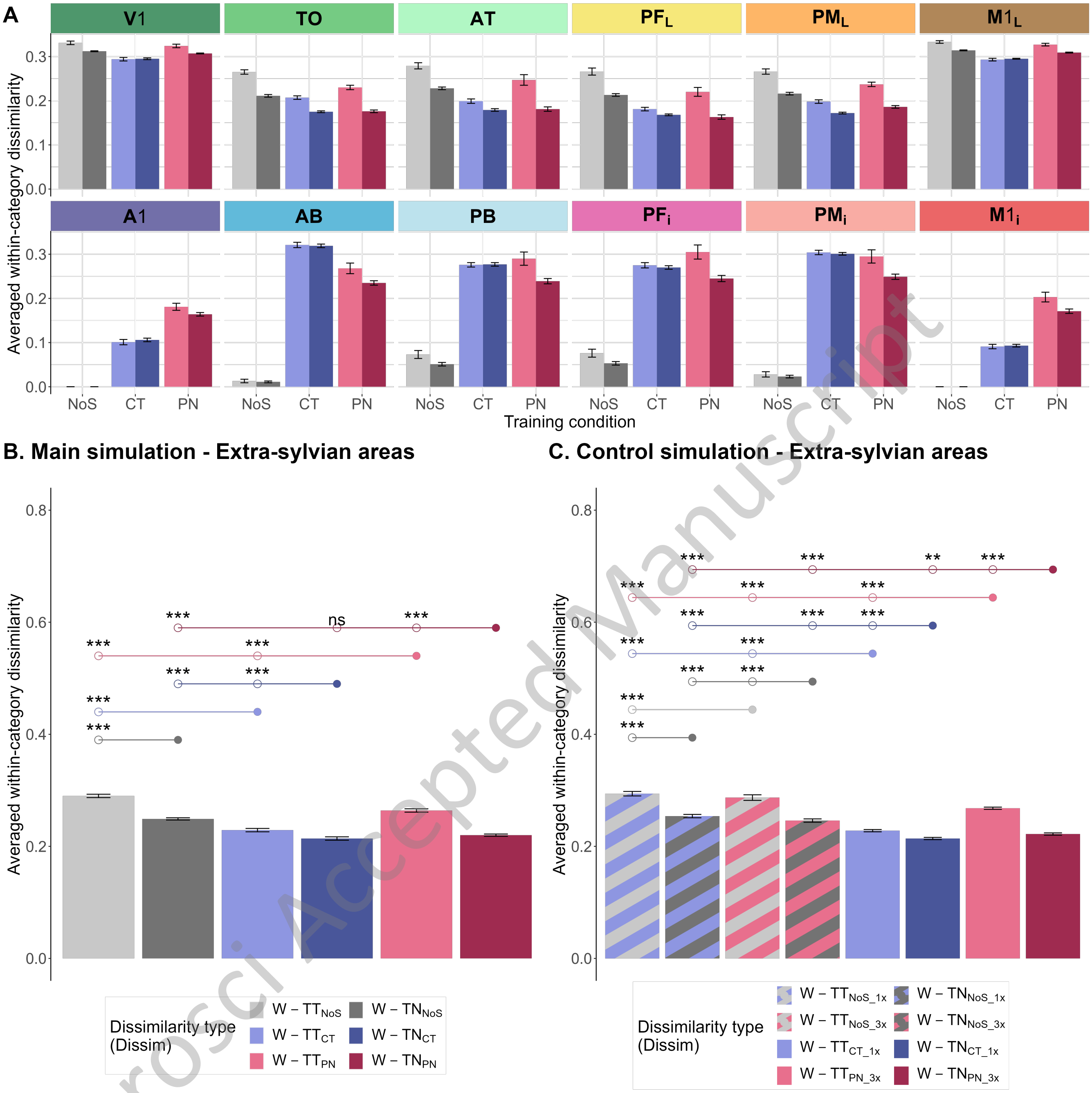
Training trials (tt) per instance	No symbol	Category term	Proper name
2000	**NoS**	**CT** (6000 tt/symbol)	**PN** (2000 tt/symbol)

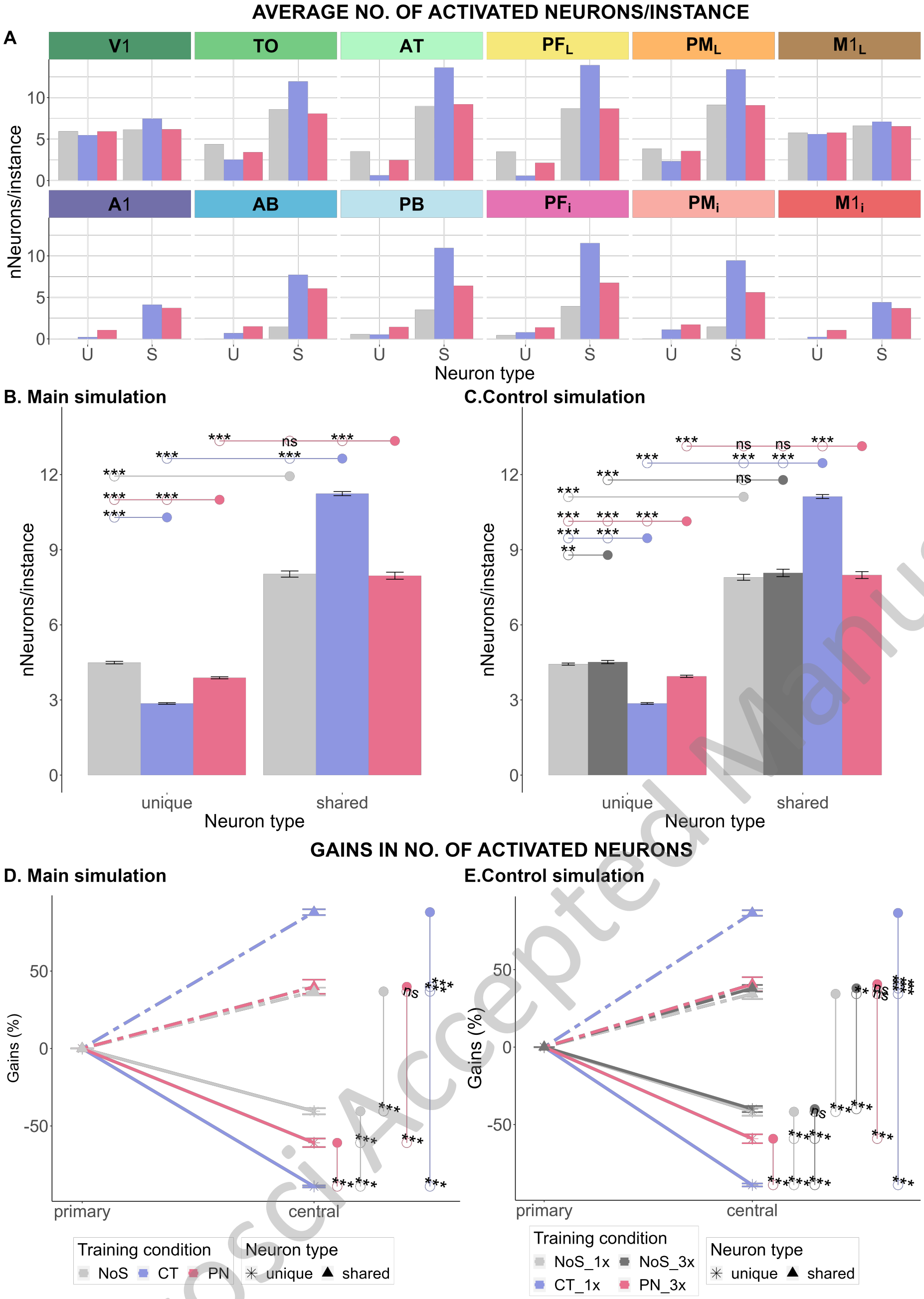
Control simulation (matched for word form presentations)

Training trials (tt) per instance	No Symbol	Category term	Proper name
1000	**NoS_1x**	**CT_1x** (3000 tt/symbol)	
3000	**NoS_3x**		**PN_3x** (3000 tt/symbol)


A)
Category a
Category b
Trained
Novel
Trained
Novel
(1)
(2)
(3)
B)
No Symbol
V1
TO
AT
PF$_L$
PM$_L$
M1$_L$
A1
AB
PB
PF$_i$
PM$_i$
M1$_i$
Category Term
V1
TO
AT
PF$_L$
PM$_L$
M1$_L$
A1
AB
PB
PF$_i$
PM$_i$
M1$_i$
Proper Name
V1
TO
AT
PF$_L$
PM$_L$
M1$_L$
A1
AB
PB
PF$_i$
PM$_i$
M1$_i$
3
2
1
0


A
V1
TO
AT
PF_L
PM_L
$M1_L$
A1
AB
PB
PF_i
PM_i
$M1_i$
Averaged dissimilarity
0.0
0.5
1.0
1.5
2.0
NoS
CT
PN
Training condition
B. Main simulation - Extra-sylvian areas
C. Control simulation - Extra-sylvian areas
Averaged dissimilarity
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
***
ns
Dissimilarity type (Dissim)
$W - TT_{NoS}$
$B - TT_{NoS}$
$W - TT_{CT}$
$B - TT_{CT}$
$W - TT_{PN}$
$B - TT_{PN}$
$W - TT_{NoS\_1x}$
$B - TT_{NoS\_1x}$
$W - TT_{NoS\_3x}$
$B - TT_{NoS\_3x}$
$W - TT_{CT\_1x}$
$B - TT_{CT\_1x}$
$W - TT_{PN\_3x}$
$B - TT_{PN\_3x}$


A
V1
TO
AT
PF_L
PM_L
$M1_L$
A1
AB
PB
PF_i
PM_i
$M1_i$
Averaged within-category dissimilarity
0.0
0.1
0.2
0.3
NoS
CT
PN
Training condition
B. Main simulation - Extra-sylvian areas
C. Control simulation - Extra-sylvian areas
Averaged within-category dissimilarity
0.0
0.2
0.4
0.6
0.8
***
ns
Dissimilarity type (Dissim)
$W - TT_{NoS}$
$W - TN_{NoS}$
$W - TT_{CT}$
$W - TN_{CT}$
$W - TT_{PN}$
$W - TN_{PN}$
**
$W - TT_{NoS\_1x}$
$W - TN_{NoS\_1x}$
$W - TT_{NoS\_3x}$
$W - TN_{NoS\_3x}$
$W - TT_{CT\_1x}$
$W - TN_{CT\_1x}$
$W - TT_{PN\_3x}$
$W - TN_{PN\_3x}$


AVERAGE NO. OF ACTIVATED NEURONS/INSTANCE
A
V1
TO
AT
PF_L
PM_L
$M1_L$
A1
AB
PB
PF_i
PM_i
$M1_i$
nNeurons/instance
0
5
10
U
S
Neuron type
B. Main simulation
C.Control simulation
nNeurons/instance
0
3
6
9
12
***
ns
**
unique
shared
Neuron type
GAINS IN NO. OF ACTIVATED NEURONS
D. Main simulation
E.Control simulation
Gains (%)
50
0
-50
primary
central
Training condition
NoS
CT
PN
Neuron type
unique
shared
NoS_1x
NoS_3x
CT_1x
PN_3x