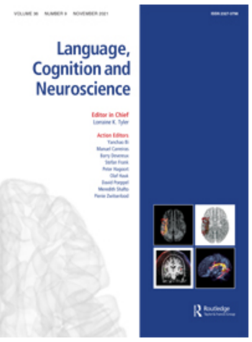

Language, Cognition and Neuroscience

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

A deep learning account of how language affects thought

Xiaoliang Luo, Nicholas J. Sexton & Bradley C. Love

To cite this article: Xiaoliang Luo, Nicholas J. Sexton & Bradley C. Love (2021): A deep learning account of how language affects thought, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2021.2001023

To link to this article: https://doi.org/10.1080/23273798.2021.2001023

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

View supplementary material

Published online: 15 Nov 2021.

Submit your article to this journal

Article views: 697

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=plcp21

LANGUAGE, COGNITION AND NEUROSCIENCE
https://doi.org/10.1080/23273798.2021.2001023

Routledge
Taylor & Francis Group

OPEN ACCESS
Check for updates

A deep learning account of how language affects thought

Xiaoliang Luo [a], Nicholas J. Sexton [a] and Bradley C. Love [a,b]

[a]Department of Experimental Psychology, University College London, London, UK; [b]The Alan Turing Institute, London, UK

ABSTRACT
How can words shape meaning? Shared labels highlight commonalities between concepts whereas contrasting labels make differences apparent. To address such findings, we propose a deep learning account that spans perception to decision (i.e. labelling). The model takes photographs as input, transforms them to semantic representations through computations that parallel the ventral visual stream, and finally determines the appropriate linguistic label. The underlying theory is that minimising error on two prediction tasks (predicting the meaning and label of a stimulus) requires a compromise in the network's semantic representations. Thus, differences in label use, whether across languages or levels of expertise, manifest in differences in the semantic representations that support label discrimination. We confirm these predictions in simulations involving fine-grained and coarse-grained labels. We hope these and allied efforts which model perception, semantics, and labelling at scale will advance developmental and neurocomputational accounts of concept and language learning.

ARTICLE HISTORY
Received 11 March 2021
Accepted 25 October 2021

KEYWORDS
deep learning; word learning; language and thought; semantic representation; Whorfian hypothesis

1. Introduction

A growing body of research suggests that linguistic labelling can affect how concepts are represented (e.g. Gleitman and Papafragou (2012), Li and Gleitman (2002), and Lupyan (2012)). Some have suggested that differences in labelling across languages can affect the acquisition of concepts, leading to differences in how the concepts are semantically represented and perceptually perceived (Lupyan et al., 2020; Özgen & Davies, 2002). In this contribution, we will briefly review evidence for this view along with computational models that offer possible mechanisms for how labelling can affect the representation of concepts. We will then offer a straightforward account of how labelling affects meaning that spans perception to decision (i.e. labelling) in the form of a error-driven, self-supervised deep convolutional neural network (DCNN) that takes photographs as input and maps them to both semantic representations and verbal labels. We propose the joint need to infer meaning and label from percepts leads to labels affecting meaning. Our model learns in an error-driven fashion in that it seeks to minimise the error on the prediction of meaning and labels.

How language may affect concepts is considered in cross-linguistic and cross-cultural studies (see Gleitman and Papafragou (2012) for a review). For example, Boroditsky and Schmidt (2000) found people's conception of the genders of objects was strongly influenced by the grammatical genders assigned to the objects in their native language. Boroditsky (2001) also found Mandarin and English speakers have different conceptions of time that may be rooted in language differences. Choi and Bowerman (1991) found that learners of English and Korean show differences in their semantic organisation of spatial meaning. Similarly, Li and Gleitman (2002) showed spatial reasoning is strongly affected by spatial lexicons used in different languages. Winawer et al. (2007) and Roberson et al. (2008) found habitual or obligatory colour distinctions made in one's language are reflected in language-specific categorical distortions in colour perception. However, it remains controversial whether language has direct effects on elementary perceptual processing (Firestone & Scholl, 2015; Klemfuss et al., 2012; Schyns et al., 1998).

A crucial function of language is to label objects, relations, properties, and events that populate everyday experience (Li & Gleitman, 2002). In category learning studies in the laboratory, artificial categories and labels are often used to investigate how verbal labelling affects meaning (Shepard et al., 1961). Verbal labels can be manipulated by experimenters to test specific

CONTACT Xiaoliang Luo ✉ xiao.luo.17@ucl.ac.uk ✉ ken.xiaoliang@gmail.com

Supplemental data for this article can be accessed at http://dx.doi.org/10.1080/23273798.2021.2001023.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

hypotheses about how people learn and represent newly acquired information. For example, research in categorical perception considers how items that straddle a category boundary can become more distinct after training on labels (Goldstone & Hendrickson, 2010).

Results from laboratory studies can be used to evaluate computational models of category learning (Wills & Pothos, 2012). Of course, one drawback of well-controlled laboratory studies of learning is that they cannot easily assess the effects of lifelong learning (Ramscar et al., 2013).

Laboratory studies of learning find a variety of effects for labelling (see Lupyan et al. (2020) for a review). Whether labels are presented before or after the perceptual stimuli can have strong effects on how learners perform and represent acquired information (Markman & Ross, 2003; Ramscar et al., 2010; Yamauchi & Markman, 1998; Yamauchi et al., 2002). Colour discrimination can be influenced by labels (Özgen & Davies, 2002). Consistent with these laboratory studies, experts, such as bird watchers, tend to have a richer feature and labelling vocabulary than novices, reflecting the fine distinctions that experts can draw (Tanaka & Taylor, 1991).

One category learning model, SUSTAIN, was developed to account for how labelling affects learning (Love et al., 2004). Unlike prototype and exemplar models that always form the same knowledge representations regardless of labelling interactions, SUSTAIN is a clustering model that forms clusters (i.e. new knowledge representations) in response to surprising failures in using labels. The mechanisms supporting these rapid learning processes have been localised to the hippocampus and medial prefrontal cortex (Davis et al., 2012a, 2012b; Love & Gureckis, 2007; Mack et al., 2016, 2018, 2020), though semantic and labelling information may subsequently be consolidated throughout cortex (Norman & O'Reilly, 2003).

The strong effect labelling can have on semantic representations was demonstrated by a brain imaging study by Mack et al. (2016) in which participants learned through trial-and-error to correctly label beetle stimuli as belonging to one of two contrasting categories. After mastering one labelling scheme, participants learned a second labelling scheme involving the same beetle stimuli. Participants' behavioural data were fit with the SUSTAIN clustering model which indicated that category knowledge was organised very differently under the two labelling schemes. Representational Similarity Analyses (RSA; Kriegeskorte et al., 2008) revealed that the model's clustering representations mirrored neural representations in the hippocampus. This study provided a strong demonstration of how labels can direct the formation of semantic representations.

This relatively simple clustering model has been able to account for how conceptual representations are affected by linguistic contrasts. For example, Davis and Love (2010) demonstrated that learning to minimise labelling errors leads to stimuli sharing a label becoming more similarly represented whereas items with contrasting labels become more dissimilar (also see Lupyan (2012)). Connectionist models also explore the effect of labelling on learning (Boroditsky & Schmidt, 2000; Ramscar et al., 2010; Rogers & McClelland, 2004; Saxe et al., 2019)

One inherent limitation of previous modelling efforts is that the stimuli and categories were artificial and hand coded in that they were constructed by the experimenters. Such approaches are unlikely to capture the richness of people's perceptual experience and the role natural language labels can play in organising the knowledge people gleam from such experiences. In laboratory studies, empirical studies and accompanying simulations are typically conducted on a small scale (e.g. two artificial categories) using low-dimensional and handcrafted stimuli that do not match the richness of the real world. In contrast, DCNNs can process photographs from a wide range natural object categories and classify them into categories denoted by natural language labels.

DCNNs are error-driven, connectionist models loosely inspired by the human visual system (Fukushima, 1980). A DCNN typically consists a sequence of layers that are responsible for extracting visual information from photographs of natural objects. DCNNs are usually trained through supervised learning to minimise classification error over millions of naturalistic images. DCNNs have achieved competitive results on benchmarks that reflect key elements of human cognition, such as object recognition (Russakovsky et al., 2015; Szegedy et al., 2015). Although DCNNs were designed to satisfy engineering and performance objectives, they have recently been shown to be good models of the mammalian visual system (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Kubilius et al., 2019; Yamins et al., 2014).

DCNNs can serve as building blocks in more encompassing models that address semantics. For example, Devereux et al. (2018) extended a DCNN model to include visuo-semantic processing to account for how visual properties of an object elicit semantic information in the visual ventral stream. We believe DCNNs hold promise for the study of how verbal labels may affect the representation of meaning.

1.1. Shared representations for labelling and meaning

Semantic meaning does not require linguistic knowledge. For example, a rabbit can determine that an approaching creature in the bushes is a danger without labelling the object as a fox. In this case, there is one prediction task, from perceptual input to meaning, which can be viewed as residing in a continuous multidimensional space. Learning can optimise for this one prediction task. In contrast, humans are faced with the additional task of inferring discrete linguistic labels.

For humans, we propose that common representations underlie these two prediction tasks, which provides a route for labelling to affect semantics (Figure 1). We model this process with a perceptual front-end in the form of a self-supervised DCNN that is pre-trained on natural images in the absence of labels (Chen et al., 2020). The choice of a self-supervised front-end is appealing in that the perceptual features are not optimised towards a labelling objective. We view this perceptual front-end as akin to the ventral visual stream. We propose that through experience, high-level perceptual representations from this DCNN are mapped into a semantic layer, which provides a high-dimensional continuous space to represent concepts. In this way, people (or rabbits) could learn to associate perceptual input with meaning. However, in species that can also label objects, there is an additional final layer in which semantic knowledge is mapped to an appropriate discrete linguistic label. During learning, the system seeks to minimise error on both prediction tasks, that is the system seeks to both predict the meaning of a percept and to appropriately label it. This joint optimisation of semantic interpretation and labelling can lead to labels affecting meaning. For instance, objects that share the same label, such as “dog”, may come to be represented in a more similar fashion in our model. Although we do not explicitly model internal categories or concepts as discontinuities within a semantic space, our approach could accommodate these theories of representation by various means, such as introducing clusters to the semantic layer.

One exciting aspect of our approach is that it can operate on pixel-level images because of its DCNN perceptual front-end. The model’s semantic representations (at the semantic layer) also reflect real-world structure. The model was trained to map from the visual images into a high-dimensional semantic space derived from word embeddings[1] (Devlin et al., 2018; Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018). Importantly, word embeddings capture real-world semantics based on word usage patterns in natural languages leading to related concepts having similar semantic vectors. For example, the concepts relating to pens, pencils and typewriters are close together in semantic space, reflecting their similar function, even though they differ in visual features. Conversely, oranges and basketballs are further apart in semantic space, despite somewhat similar visual features.

We should note that even though the semantic representations in the model are derived from word usage patterns that this layer is not intended to reflect labelling. It is only for convenience we use word embedding models to determine semantic representations as they provide rich semantic representations that do not require hand coding. In principle, we could use semantic representations derived from a non-linguistic source. For example, Hornsby et al. (2020) derive representations for objects that can be purchased in a supermarket by examining purchasing patterns as opposed to analysing word patterns. To provide another viable alternative, we could have used a semantic space derived from large-scale human similarity ratings (Hebart et al., 2020; Roads & Love, 2020).

The final, labelling layer of the model predicts the categorical target class of the stimulus in the form of verbal labels. Whereas category learning models applied to laboratory tasks typically use artificial labels and categories, we consider real-world categories and labels (Miller, 1995). For details regarding model architecture and training procedure, we refer the readers to the Appendix.

2. Experiments

One general conclusion from the literature reviewed in the Introduction is that labelling creates pressures that affect semantic and perhaps even perceptual representations. The need to label correctly can lead to differentiation of semantic information when labels differ, and convergence when there is a common label. Across two simulation studies, we evaluated our DCNN account of how labelling affects semantic representations (Figure 1) in light of these observations.

In both experiments, we systematically evaluated the importance or pressure to correctly label images relative to the importance of inferring semantic information in a veridical manner. The relative importance of labelling was controlled by the parameter β. In our simulations, we examined how this emphasis on label learning affects semantic representations in our model. Because activity passes through the semantic layer when predicting labels (i.e, from percept to meaning to label), the labels learned will affect the network’s semantic representations. For implementational

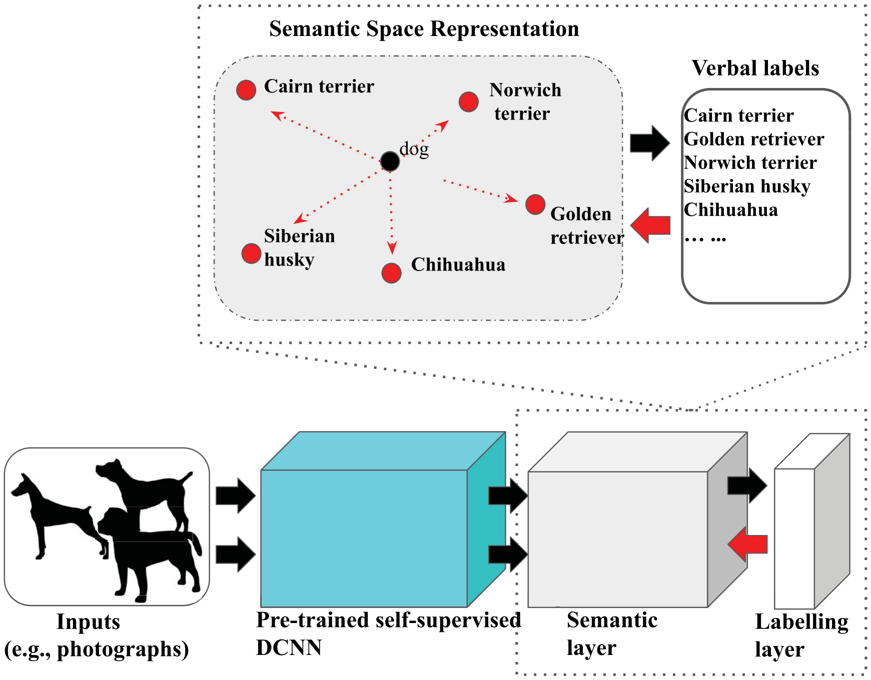

Figure 1. A proposal for how labelling can affect meaning. When viewing an object, we propose people both attempt to infer its meaning and its label. A common internal representation is learned to satisfy these two prediction tasks, which provides a route for labels to affect meaning. We implement this account in a DCNN that takes photographic images as input, processes this visual input through a series of computations intended to parallel the visual ventral stream, and associates higher-level visual representations with a semantic vector that reflects the meaning of the object. Finally, from this semantic vector, the linguistic label for the object is predicted. Prediction errors from both the semantic and label layers will affect the network's internal representations. In the above example, we demonstrate how the semantic representation of dog breeds change when the labelling task shifts from coarse-grained labelling (i.e. all dogs share the same label) to fine-grained labelling (i.e. label dog by their breed). When there is no pressure to distinguish different dogs, the model develops more homogeneous representations of dogs (the black dot). When there is pressure for fine-grained distinctions (bold red arrow), representations become more distinct (the red dots) to reduce label confusions.

details of the semantic aspects of the model, see Appendix.

We evaluated the alignment of this resulting space with the canonical semantic space the model was trained on (in addition to the linguistic labels) using representational similarity analysis (RSA). RSA evaluates how similar two similarity spaces are. In our case, we can calculate pairwise similarities for concepts in our network's semantic layer and Spearman correlate them with pairwise similarities from the semantic embedding space that served as training targets for our network's semantic layer (see Appendix).

We hypothesise that semantic distortion, operationalised as a decreased Spearman correlation between representational similarity matrices, will be greatest when the relative importance of labelling is high. We also predict that the type of distortion observed will depend on whether the labels used are fine-grained (e.g. "Chihuahua") vs. coarse-grained (e.g. "dog"). For fine-grained labels (Experiment 1), we hypothesise that semantic representations with different labels should move apart from one another to increase semantic discriminability. In contrast, coarse-grained labels (Experiment 2) shared by multiple concepts should lead to concepts sharing a label converging with one another in semantic space.

We trained and evaluated our proposed model on the ImageNet-2012 dataset (Russakovsky et al., 2014) for both meaning prediction and label prediction tasks at varying degrees of labelling pressure, β (Table 1). We examine a wide range of labelling pressure. We start off from $\beta = 0$ where there is no labelling pressure and the model completely focuses on meaning prediction to $\beta = 10$ where the pressure to use labels correctly is high. ImageNet-2012 is a large-scale dataset of naturalistic images drawn from 1000 categories based on the WordNet ontology (Miller, 1995). Full details of model architecture and training are presented in Appendix.

2.1. Experiment 1: fine-grained labels

In Experiment 1, we evaluated the effect of the weight given to labelling, β, on the resulting semantic space of the model. Pressure to correctly use fine-grained labels should lead to representations in the semantic layer differentiating from one another to reduce label confusions, which suggested two empirical predictions. First, we anticipated that increasing levels of β will lead to increasing distortions of the model's semantic space as evaluated by RSA. Second, the basis for this distortion should be semantic representations moving further apart from one another (akin to a caricature effect) to support label use.

2.1.1. Method

A model was independently trained at each level of β (labelling pressure). The labelling task was a 1000-way image classification (i.e. there were 1000 labels for the model to master). Images were shown to the DCNN and mapped to 1 of a 1000 ImageNet classes (i.e. the labels) after passing through the semantic layer (Figure 1). Further details on how the model was trained and evaluated are available in the Appendix.

2.1.2. Results & Discussion

Both predictions were confirmed. First, we evaluated whether increasing labelling pressure (i.e. increasing β) led to greater representational distortion, which should manifest in a lower Spearman correlation in an RSA comparing model representations in the semantic layer with the canonical reference embedding space. As β increased, the semantic similarity structure did become less aligned with the reference embedding space (Figure 2(A)). The second prediction was that the basis of this distortion would be representations in the semantic space moving apart from one another to support label discrimination. We evaluated the Euclidean between-class distances in the semantic layer and found that distances increased with increasing β (Figure 2(B)). In summary, the pressure to use fine-grained labels correctly led to semantic distortions in which representations moved apart from one another to reduce label confusion.

2.2. Experiment 2: coarse-grained labels

In Experiment 1, we found fine-grained labelling led to distortions in semantic space because representations moved apart to reduce label confusions. In Experiment 2, we considered how coarse-grained labels affect semantic representations. People often use labels that are broad and encompass subcategories. For instance, superordinate labels, such as "mammal", by definition refer to a collection of distinct categories. The specificity of labelling can also differ as a function of expertise within a domain. For example, a child may refer to all dogs as "dog" whereas a veterinarian may refer to specific breeds. We predicted that items from different categories sharing a label, such as different species of birds all being referred to as "birds", will develop semantic representations that are relatively more similar to one another, which would complement the differentiation result in Experiment 1 with fine-grained labels. To test our hypothesis, we paralleled Experiment 1 but with some categories sharing a coarse-grained (i.e. higher-level) linguistic label.

Table 1. Different levels of labelling pressure.

β	0	0.1	1	2	3	5	7	10

We trained and tested our model across a range of labelling pressure. When $\beta = 0$, the proposed model focused exclusively on the semantic prediction task. As β increases, the model focuses more on the labelling task. The semantic representations acquired by the model (Figure 1) should vary depending on the level of labelling pressure and the nature of labelling task (e.g. fine-grained vs. coarse-grained).

2.2.1. Method

Models were trained in a similar fashion to Experiment 1. A subset of the original 1000 classes were grouped into 1 superordinate class (hence coarse-grained) and each model was trained to classify images into either a pre-defined superordinate or the remaining individual classes. For example, when training using the 'reptile" superordinate label, standard ImageNet labels such as "thunder snake", "rock python", "African crocodile", and "mud turtle" were not used but instead the model was trained to label items from all of these reptile classes as "reptile". Individual classes not under the superordinate were trained as in Experiment 1. We explored five superordinates, "reptile", "amphibian", "primate", "bird", and "dog". In total, 40 models were trained and evaluated in Experiment 2 (8 levels of $\beta \times 5$ superordinate classes). Please see the Appendix for further details.

2.2.2. Results & Discussion

First, we replicated Experiment 1's finding that semantic space becomes increasingly distorted as pressure to label correctly increases (Figure 3(A)). Second, we considered our main prediction for Experiment 2, namely do categories sharing a higher-level label become more semantically similar as the pressure to label increases. Indeed, across all five coarse-grained labels considered, the pairwise distances between items in the semantic layer sharing a common coarse-grain label decreased relative to items in other categories as the pressure to label increased (Figure 3(B)). One caveat is that when labelling pressure became extreme at high β the integrity of the semantic space itself was weakened, leading to a minor reversal in our distance ratio. The main finding was that a common umbrella term for lower-level categories led to their semantic representations becoming relatively more similar.

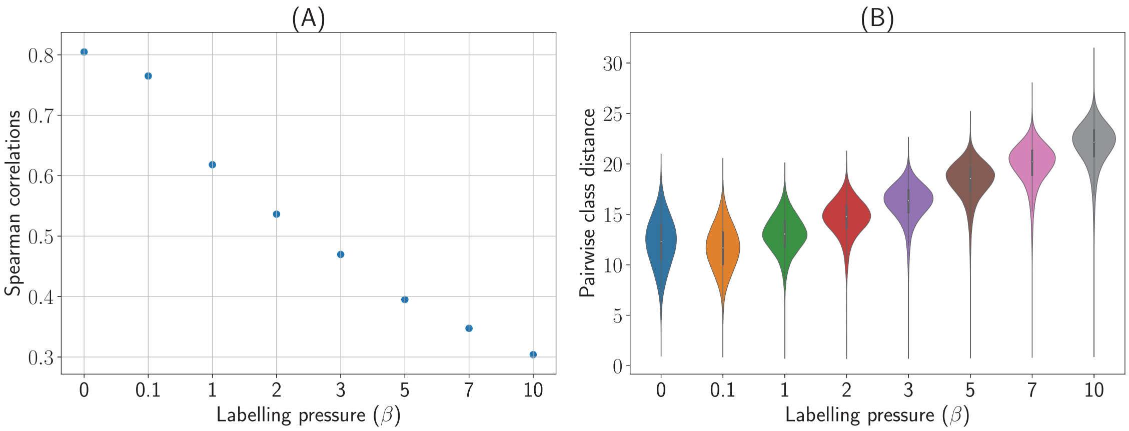

Figure 2. Main results for Experiment 1 with fine-grained labels. To evaluate the prediction that the pressure to label affects the representation of meaning, semantic representations of the same object categories were evaluated at varying degrees of labelling pressure. (A) Using representational similarity analysis (RSA), we observed a decreasing Spearman correlation between the predicted semantic space and the reference embedding space, indicating that the semantic space became more distorted as the pressure on labelling increased. (B) We confirmed that the basis for this distortion was items between classes moving apart from one another. As the pressure to label increased, the average distance between each pair of semantic concepts increased. This pattern of results indicates that to accommodate the pressure of labelling, the distance between individual concepts increased to reduce label confusability.

3. General discussion

We offered a straightforward account of how linguistic labels can shape meaning. Why do labels appear in cases to make dissimilar objects appear alike and in other cases highlight differences between perceptually similar objects? Can we understand how differences in labelling schemes across languages and cultures lead to differences in mental representation? To help answer these questions, we developed a DCNN model that spans from perception to decision, taking photographic images as input, inferring their meaning, and then finally settling on a label (see Figure 1).

Our basic theory was that performing two prediction tasks, namely predicting the meaning and label of a photographic stimulus, requires a compromise in network representations. In particular, semantic representations in later network layers end up reflecting label use. The more pressure or emphasis there is on correctly using linguistic labels, the more semantic representations will distort to support labelling. In particular, items sharing a label will tend to be

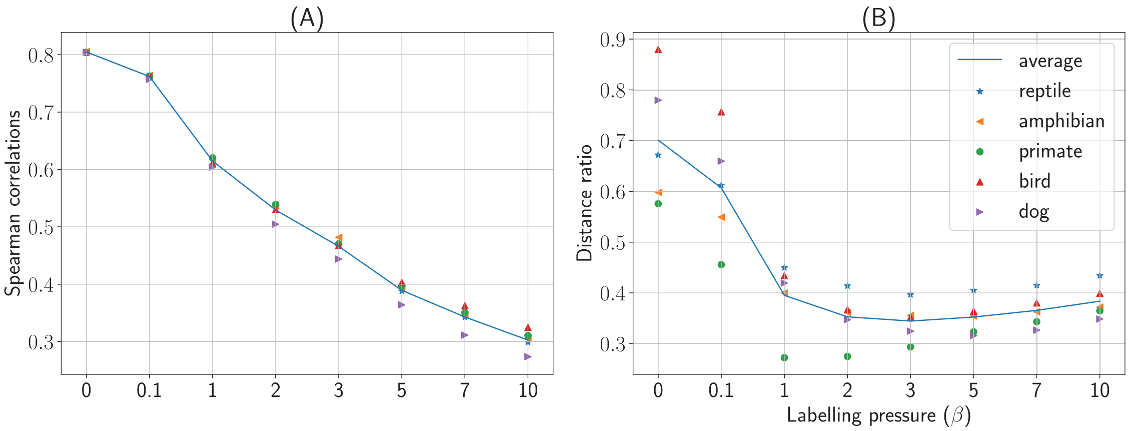

Figure 3. Main results for Experiment 2 with coarse-grained labels. To evaluate the prediction that the pressure to label affects the representation of meaning, semantic representations of the same object categories were evaluated at varying degrees of labelling pressure with a subset of categories sharing a common superodinate label (e.g. all birds labelled as "bird"). (A) As in Experiment 1, using representational similarity analysis (RSA), we observed a decreasing Spearman correlation between the predicted semantic space and the reference embedding space, indicating that the semantic space became more distorted as the pressure on labelling increased. (B) As the pressure to label increased, the relative distance between concepts sharing a superordiante label decreased relative to other concepts. Sharing a higher-level label makes objects from different lower-level categories become more semantically similar.

represented in a more similar fashion (Experiment 2), whereas those with conflicting labels will be differentiated (Experiment 1) to reduce labelling confusion. Both these effects arise from discriminative learning principles (Nixon, 2020; Ramscar et al., 2010).

Although these principles recapitulate work in the category learning literature (e.g. Love et al. (2004)), previous work relied on artificial categories from laboratory studies using hand-coded stimulus representations. Here, we demonstrated these basic principles can scale to natural categories and labelling schemes using photographic stimuli as model input. In doing so, it is straightforward to envision how differences in labelling schemes across languages could influence thought. Indeed, our model as specified could be applied to such questions to provide quantitative predictions and a candidate mechanism for how labelling affects thought.

Although we have discussed labelling as having a distortive effect on meaning, labels are of course an important signal to meaning (Waxman & Markow, 1995). The boundaries for label use in natural language are not arbitrary and can serve as useful learning signals, highlighting commonalities for items sharing a label and differences for items with conflicting labels. Although distortive, labels can be an important cue to meaning.

As we briefly reviewed in the Introduction, it remains debatable whether linguistic labelling affects lower-level perceptual experiences. Models like the one we have proposed can help clarify this issue. In our current simulations, the weights of the lower network layers, which can be viewed as corresponding to the ventral visual stream (Güçlü & van Gerven, 2015), were kept fixed during learning. This was in accord with the intuition that our visual system does not reorganise when we learn a new concept or word. However, future simulations could allow these weights in lower network layers to also change in response to error when learning new concepts or labels. A number of questions are likely to arise in such simulations. First, what constitutes a significant change in learning? Weights in the lower network layers will change, but for most learning problems the changes are likely to be minuscule as most novel concepts can be mastered by reweighting high-level features present in more advanced network layers. A related issue is whether one only considers weight changes critical that lead to interesting changes in behaviour. Second, where does one draw the line between perception and cognition? Researchers may disagree on where perception ends and cognition begins both within the brain and the corresponding network layer. Some researchers may dismiss the question entirely. Finally, changes in perception and decision can occur by top-down modulation of lower-layers absent weight changes in DCNNs by applying selective attention (Luo et al., 2021).

Some view language as key to what makes humans unique, but what is labelling in our computational account? In our account, labelling is a task pressure that shapes semantic representations. We view choosing a label as a mental action that could of course also result in a motor action, such as when communicating a label to others. Decisions not communicated through language should have similar effects. In a sense our example from the Introduction about the rabbit being constrained by only the semantic prediction task, whereas a human both predicts semantics and a label, was over simplified. The rabbit could of course need to make a number of other predictions, such as predicting the value of different actions (e.g. hide, flee, rest, etc.) in the presence of the fox. In this light, just as common labels may lead to the meaning of objects converging, so too might objects who engender the same action. For example, the semantic representations of a fox and a dog may become more similar in the rabbit's mind because in both cases the rabbit would flee. In this light, linguistic labelling is but one more discrimination task that influences the formation of semantic representations.

Interestingly, the importance of labelling in shaping representations has recently been acknowledged by the deep learning community. Most deep learning models, like our own, are trained on the ImageNet database of images and labels. ImageNet propelled the deep learning revolution, which led to the best models of human object recognition in terms of accounting for behaviour and brain response. However, closer inspection of ImageNet reveals that the frequency of categories does not closely correspond to human experience. A more recent image and label database, ecoset, attempts to remedy the situation by creating a more representative training set for models (Mehrer et al., 2021). Deep learning models trained on these images and labels prove to be better models of humans (Mehrer et al., 2021), which highlights the importance of labelling in shaping representation.

We hope that our work, exploring how linguistic labelling can affect semantic meaning, serves as a starting point for others to further consider a wide array of issues. For example, DCNN models of language and thought can be useful in understanding the interplay between concept and word learning in children during both typical and atypical development. Because aspects of DCNNs can be put in correspondence with brain regions, neurocomputational accounts of language and learning can be proposed and evaluated. As the science progresses, we expect network architectures to diverge from those used here, which were chiefly motivated by engineering concerns, to better reflect implementational aspects of the brain, such as long-range recurrence. We encourage other researchers to take advantage of DCNNs and large databases to model mental processes from perception to decision at scale, which can complement well-controlled laboratory studies and, hopefully, add value to those studies by informing their design and interpretation.

Note

1. For details of deriving word embeddings for image concepts, we refer the readers to the Appendix.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Royal Society, [Royal Society Wolfson Fellowship 183029] National Institutes of Health, [NIH Grant 1P01HD080679] Wellcome Trust [Wellcome Trust Investigator Award WT106931MA].

ORCID

Xiaoliang Luo http://orcid.org/0000-0002-5297-2114
Nicholas J. Sexton http://orcid.org/0000-0003-1236-1711
Bradley C. Love http://orcid.org/0000-0002-7883-7076

References

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43* (1), 1–22. https://doi.org/10.1006/cogp.2001.0748

Boroditsky, L., & L. A. Schmidt (2000). Sex, syntax, and semantics. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *22*. Retrieved from https://escholarship.org/uc/item/0jt9w8zf

Cadieu, C. F., H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, & J. J. DiCarlo (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10* (12), 1003963. https://doi.org/10.1371/journal.pcbi.1003963www.ploscompbiol.org.

Chen, T., S. Kornblith, M. Norouzi, & G. Hinton (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, PMLR* (pp. 1597–1607). http://arxiv.org/abs/2002.05709.

Choi, S., & M. Bowerman (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition*, *41*(1-3), 83–121. https://doi.org/10.1016/0010-0277(91)90033-Z

Davis, T., & B. C. Love (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. https://doi.org/10.1177/0956797609357712

Davis, T., B. C. Love, & A. R. Preston (2012a). Learning the exception to the rule: Model-based FMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*(2), 260–273. https://doi.org/10.1093/cercor/bhr036

Davis, T., B. C. Love, & A. R. Preston (2012b). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 821–839. https://doi.org/10.1037/a0027865

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

Devlin, J., M. W. Chang, K. Lee, & K. Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference* (Vol. 1, pp. 4171–4186). http://arxiv.org/abs/1810.04805

Firestone, C., & B. J. Scholl (2015). Cognition does not affect perception: Evaluating the evidence for top-down effects. *Behavioral and Brain Sciences*, *39*, 1–77. https://doi.org/10.1017/S0140525X15000965, e229

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Gleitman, L., & A. Papafragou (2012). New perspectives on language and thought. In *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press. doi: 10.1093/oxfordhb/9780199734689.013.0028.

Goldstone, R. L., & A. T. Hendrickson (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78. https://doi.org/10.1002/wcs.26

Güçlü, U., & M. A. van Gerven (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hebart, M. N., C. Y. Zheng, F. Pereira, & C. I. Baker (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hornsby, A. N., T. Evans, P. S. Riefer, R. Prior, & B. C. Love (2020). Conceptual organization is revealed by consumer activity patterns. *Computational Brain & Behavior*, *3*(2), 162–173. https://doi.org/10.1007/s42113-019-00064-9

Klemfuss, N., W. Prinzmetal, & R. B. Ivry (2012). How does language change perception: A cautionary note. *Frontiers in Psychology*, *3*, 78. https://doi.org/10.3389/fpsyg.2012.00078

Kriegeskorte, N., M. Mur, & P. Bandettini (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.01.016.2008

Kubilius, J., M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. J. Majaj, E. B. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, A. Nayebi, D. Bear, D. L. Yamins, & J. J. DiCarlo (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. https://arxiv.org/abs/1909.06161v2.

Li, P., & L. Gleitman (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*(3), 265–294. https://doi.org/10.1016/S0010-0277(02)00009-4

Love, B. C., & T. M. Gureckis (2007). Models in search of a brain. *Cognitive, Affective & Behavioral Neuroscience*, *7*(2), 90–108. https://doi.org/10.3758/CABN.7.2.90

Love, B. C., D. L. Medin, & T. M. Gureckis (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Luo, X., B. D. Roads, & B. C. Love (2021). The costs and benefits of goal-directed attention in deep convolutional neural networks. *Computational Brain & Behavior*: 1–18. https://doi.org/10.1007/s42113-021-00098-y.

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. *Psychology of Learning and Motivation*, *57*, 255–297. https://doi.org/10.1016/B978-0-12-394293-7.00007-8

Lupyan, G., R. Abdel Rahman, L. Boroditsky, & A. Clark (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, *24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005

Mack, M. L., B. C. Love, & A. R. Preston (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mack, M. L., B. C. Love, & A. R. Preston (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*(44), 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Mack, M. L., A. R. Preston, & B. C. Love (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46. https://doi.org/10.1038/s41467-019-13930-8

Markman, A. B., & B. H. Ross (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592–613. https://doi.org/10.1037/0033-2909.129.4.592

Mehrer, J., C. J. Spoerer, E. C. Jones, N. Kriegeskorte, & T. C. Kietzmann (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mikolov, T., K. Chen, G. Corrado, & J. Dean (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 – Workshop Track Proceedings, International Conference on Learning Representations (ICLR)*.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*: 104081. https://doi.org/10.1016/j.cognition.2019.104081

Norman, K., & R. O'Reilly (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, *110*(4), 6–11. https://doi.org/10.1037/0033-295X.110.4.611

Özgen, E., & I. R. Davies (2002). Acquisition of categorical color perception: a perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General*, *131*(4), 477–493. https://doi.org/10.1037/0096-3445.131.4.477

Pennington, J., R. Socher, & C. D. Manning. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP})*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, & L. Zettlemoyer. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–2237. https://doi.org/10.18653/v1/N18-1202

Ramscar, M., P. Hendrix, B. Love, & R. H. Baayen (2013). Learning is not decline: the mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, *8*(3), 450–481. https://doi.org/10.1075/ml

Ramscar, M., D. Yarlett, M. Dye, K. Denny, & K. Thorpe (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957. https://doi.org/10.1111/(ISSN)1551-6709

Roads, B. D., & B. C. Love (2020). Enriching imagenet with human similarity judgments and psychological embeddings. ArXiv http://arxiv.org/abs/2011.11015.

Roberson, D., H. Pak, & J. R. Hanley (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition*, *107*(2), 752–762. https://doi.org/10.1016/j.cognition.2007.09.001

Rogers, T. T., & J. L. McClelland (2004). *A parallel distributed processing approach*. MIT Press.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, & L. Fei-Fei (2014). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, & L. Fei-Fei (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Saxe, A. M., J. L. McClelland, & S. Ganguli (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546. https://doi.org/10.1073/pnas.1820226116

Schyns, P. G., R. L. Goldstone, & J. P. Thibaut (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*(1), 1–17. https://doi.org/10.1017/S0140525X98000107

Shepard, R. N., C. L. Hovland, & H. M. Jenkins (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*(13), 1. https://doi.org/10.1037/h0093825

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, & A. Rabinovich (2015). Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 7, pp. 1–9). IEEE Computer Society.

Tanaka, J. W., & M. Taylor (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. https://doi.org/10.1016/0010-0285(91)90016-H

Waxman, S. R., & D. B. Markow (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

Wills, A. J., & E. M. Pothos (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125. https://doi.org/10.1037/a0025715

Winawer, J., N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, & L. Boroditsky (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Yamauchi, T., B. C. Love, & A. B. Markman (2002). Learning non-linearly separable categories by inference and classification. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *28*(3), 585–593. https://doi.org/10.1037/0278-7393.28.3.585

Yamauchi, T., & A. B. Markman (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*, 124–149. https://doi.org/10.1006/jmla.1998.2566

Yamins, D. L. K., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, & J. J. DiCarlo (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111