Linguistic Overhypotheses in Category Learning: Explaining the Label Advantage Effect

Anna A. Ivanova (annaiv@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Matthias Hofer (mhofer@mit.edu)
Department of Brain and Cognitive Sciences, MIT

Abstract

When learning to partition the world into categories, people rely on a set of assumptions (overhypotheses) about possible category structures. Here we propose that the nature of these overhypotheses depends on the presence of a verbal label associated with a given category. We describe a computational model that demonstrates how labels can either accelerate or hinder category learning, depending on whether or not the prior beliefs imposed by their presence align with the true category structure. This account provides an explanation for the phenomena described in prior experimental work (Lupyan, Rakison, & McClelland, 2007; Brojde, Porter, & Colunga, 2011) that have remained unexplained by other models. Based on these results, we argue that the overhypothesis theory of label effects provides a way to formalize and quantify the effect of language on category learning and to develop a more precise delineation between linguistic and non-linguistic thought.

Keywords: category learning; word learning; overhypotheses; shape bias; Bayesian modeling

Introduction

What is the relationship between words and concepts? According to one account (e.g., Bloom, 2002), linguistic and conceptual spaces are separate and independent, with newly learned words corresponding to mappings from a linguistic label to the corresponding concept. According to another account (e.g., Lupyan & Lewis, 2019), words have a large influence on the concepts being constructed: more than just pointers, they serve as a vehicle for abstraction and a driving force for category learning. Here, we propose a framework that aims to reconcile these two accounts by viewing linguistic labels as overhypotheses about likely category structures.

We investigate the language-concepts relationship in the context of one specific domain of human experience: category learning. When acquiring the meanings of words, the learner often needs not only to establish the mapping between the word and the category it refers to, but also to partition the underlying conceptual space itself. For example, a child needs to learn that the word "spoon" corresponds to both tablespoons and teaspoons, but not to forks. Category learning and word learning therefore happen hand-in-hand.

Experiments performed with both infants (Althaus & Mareschal, 2014) and adults (Lupyan et al., 2007) established that providing a linguistic label facilitates category learning, such that it is easier to distinguish exemplars from two different categories if these categories are accompanied by verbal labels. However, explanations for this *label advantage* effect differ. Certain theories consider labels to be "just another feature"; they posit that label-based learning is more successful simply because labels provide additional information that can be incorporated into the categorization process in a bottom-up way (the labels-as-features hypothesis; Gliozzi, Mayor, Hu, & Plunkett, 2009). Other theories claim that words have a top-down effect on category learning by enhancing attention to category-specific dimensions of the stimulus (the label feedback hypothesis; Lupyan, 2012).

We argue that neither of these hypotheses can fully account for experimental category learning data and propose a computational model that formalizes a third account: words impose specific priors on the hypothesized structure of new categories. This account differs from the labels-as-features hypothesis in that it assigns a special role to linguistic labels (they induce specific priors over hypothesized category structure); it differs from the label-feedback hypothesis in that, in contrast to the universally beneficial effect of labels proposed by that view, the prior induced by the label might either facilitate or hinder learning based on whether or not the prior-induced biases align with true category structure. We will show that both of these properties are required to account for the experimental data described below.

Experimental Evidence for the Label Effect

One line of evidence that labels play a special role during category learning in adults[1] comes from the study by Lupyan et al. (2007). The authors presented the participants with two types of "alien" stimuli (a subset of the YUFO stimuli; Gauthier, James, Curby, & Tarr, 2003), The participants had to learn to distinguish between the alien categories during a supervised learning experiment to decide whether they should approach or avoid a particular alien. The two categories differed in shape (Fig. 1).

The authors demonstrated that providing labels for the two categories ("leebish" and "grecious") facilitated category learning. Participants were receiving feedback about their choices in both label-based and label-free conditions, meaning that labels did not provide any additional information

The code for this paper is made available online at https://github.com/neuranna/labels-and-categories

[1]This effect has also been demonstrated in children, e.g. Smith, Jones, Landau, Gershkoff-Stowe, and Samuelson (2002).

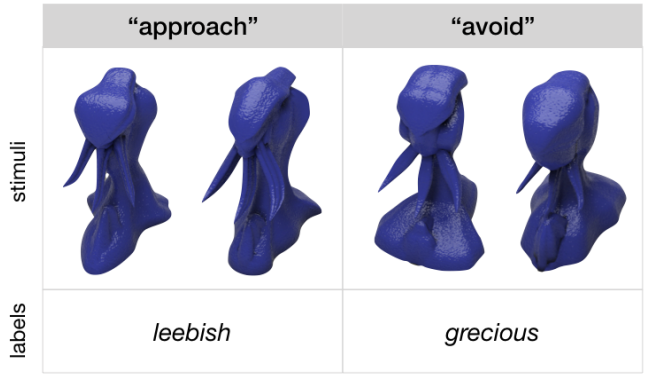

Figure 1: Sample stimuli used in Lupyan et al. (2007). Category labels were only presented to half of the participants.

about category membership, and so their influence could not be explained by the labels-as-features account (which posits that labels provide additional bottom-up information relevant for categorization). Moreover, the effect was observed for both written and spoken labels but did not generalize to non-verbal cues.

Lupyan et al. (2007) explain their results by hypothesizing that labels "highlight" dimensions where item variance aligns with category membership and thus make learning more efficient. This view was later formalized as part of the label feedback theory (Lupyan, 2012). However, a study by Brojde et al. (2011) questioned the label feedback account by providing evidence that labels are not always beneficial for category learning. Specifically, they compared learning shape-based categories with learning categories that are based on other dimensions, namely, texture and brightness. Brojde et al. found that labels indeed improved learning of shape-based categories but *hindered* learning of texture-based and brightness-based categories. They concluded that, instead of emphasizing category-relevant differences across all dimensions, labels direct the participants' attention to dimensions that have been "historically relevant" for categorization. This view is supported by other works on categorization and word learning (Smith, Jones, & Landau, 1996; Perry & Lupyan, 2014; Smith et al., 2002). However, it was never formalized as a fully fledged theory, and the label feedback model remains one of the primary ways to describe the relationship between words and categories (Lupyan & Lewis, 2019; Imai, Kanero, & Masuda, 2016).

Labels as Overhypotheses

This paper aims to provide a computational account of the effects observed by Lupyan et al. (2007) and Brojde et al. (2011). We propose that the presence of verbal labels during category learning induces a particular set of overhypotheses in the learner. Specifically, we hypothesize that the inductive constraints on the structure of categories that underlie word meanings differ from constraints on conceptual category structure more generally (e.g., Gardenfors, 2004), and that learners are sensitive to this difference. Therefore, when trying to infer the underlying structure of labeled categories, the learner will be biased toward category structures that are characteristic of word meanings (for instance, speakers of English, a language where most object noun meanings are based on shape, are more likely to classify solid objects according to shape rather than material[2]; Landau, Smith, & Jones, 1988). If this bias aligns with the true category structure, the learner will learn more efficiently compared to the "unbiased" label-free condition; if it doesn't, learning will be slower.

Our model is based on previous hierarchical Bayesian models of category learning (Anderson, 1991; Kemp, Perfors, & Tenenbaum, 2007; Xu & Tenenbaum, 2007). Those models also formalize the notion of prior knowledge of category structure as overhypotheses, which can be either fixed (Xu & Tenenbaum, 2007) or learned (Kemp et al., 2007); importantly, however, our model is the first to distinguish between overhypotheses induced by label-based and label-free category learning. In fact, Kemp et al. (2007) specifically mention that their account would predict no difference in the learning rate of labeled and non-labeled shape-based categories. We also introduce a way to model category learning based on continuous features, since previous work in this domain used binary feature representations. Modeling category learning based on multiple continuous features allows us to directly compare our model's predictions to the results obtained by Lupyan et al. (2007) and Brojde et al. (2011).

Previous models of the label effect on word learning (Gliozzi et al., 2009; Lupyan, 2012) use a connectionist rather than a Bayesian framework. Bayesian and connectionist models aim to elucidate a phenomenon on different levels of analysis and are often complementary. Bayesian models help render explicit the computational assumptions that are implicit in the solution to an inferential problem and explain why that solution works. Connectionist models provide a more mechanistic account of learning, which might be more plausible biologically but can make it more difficult to examine the models' internal representations and the computational principles underlying their behavior. Here, we aimed to demonstrate a simple computational principle that can explain human behavior during label-based vs. label-free learning without committing to implementation-level details. The Bayesian framework is well suited for that purpose. That said, given that some neural networks have been shown to have input-driven overhypotheses (e.g. shape bias; Ritter, Barrett, Santoro, & Botvinick, 2017), we expect that connectionist models would also be able to simulate the effect of linguistic overhypothesis on category learning.

Model Specification

We consider the task of learning to classify object exemplars that vary along F perceptual dimensions into C non-

[2]Unless the label occurs in a syntactic frame that suggests that it is a mass noun, in which case learners exhibit a material bias (Dickinson, 1988).

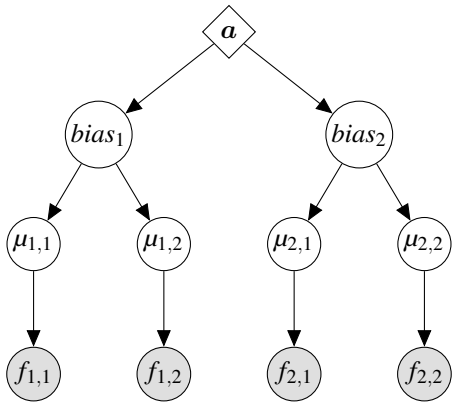

Figure 2: Graphical representation of the category learning model (for two categories and two feature dimensions). Circles indicate random variables; squares indicate fixed model parameters. Variables shaded in gray are observed during learning.

overlapping categories, with categories varying along one or more of the F dimensions.

$$i = 1,...,F;\ F := \text{Number of feature dimensions}$$
$$j = 1,...,C;\ C := \text{Number of categories}$$

The learner infers the underlying category structure by estimating the means (μ) and variances (σ^2) for each dimension of each category based on observed category exemplars f (which are also subject to perceptual noise σ_s^2):

$$f_{i,j} \sim \text{Normal}(\mu_{i,j}, \sigma_{i,j}^2 + \sigma_s^2)$$

If the dimension is not diagnostic of category membership, the inferred means for that dimension will be similar across categories. If the dimension is diagnostic, however, the means will differ. We can represent prior beliefs about which dimensions are diagnostic by constraining the relationship between the means considered by the learner:

$$\mu_{i,1},...,\mu_{i,C} := \boldsymbol{\mu}_i \sim \text{Normal}(\mathbf{0}, \Sigma_i)$$

where Σ_i is the covariance matrix that specifies the relationship between sampled category means. This relationship is best illustrated by considering the corresponding[3] correlation matrix R_i. The off-diagonal entries in R_i define the probability space over sampled pairs of category means μ_j for a given dimension i (Fig. 3A): if the correlation value is close to 1, the means sampled by the learner are almost identical; if it is 0, the means are uncorrelated. Thus, by constraining the correlation values, we can manipulate the extent to which the learner considers a given dimension as being diagnostic of category membership.

The prior over the correlation values is a function of the bias vector $\boldsymbol{k}$, which is sampled from a Dirichlet distribution with the parameter $\boldsymbol{a}$ (Fig. 3B). It is $\boldsymbol{a}$ that determines relative biases over different category dimensions and, as such, serves as an overhypothesis during learning (Figure 2).

$$k_1,...,k_F \sim \text{Dirichlet}(\boldsymbol{a})$$
$$corr(\mu_{i,v}, \mu_{i,u}) = \begin{cases} f(k_i) & \text{if } u \neq v \\ 1 & \text{otherwise} \end{cases} \qquad \forall u, v \in 1,...C$$

The correlation coefficients are a function of $\boldsymbol{k}$. The relationship between them is nonlinear because a small change in the correlation values close to 1 will result in a more radical change in the distribution of sampled means compared to a change when the values are lower (Fig. 3C). Therefore, we transform the bias values k sampled from the Dirichlet distribution using a power function with the exponent parameter $1/\gamma$, and further scale the resulting value so that it lies within the range of -1 to 1:

$$f(k_i) = (k_i^{1/\gamma} - 0.5) * 2$$

Parameter γ determines how conservative the learner is when generating hypotheses (Fig. 3D). Larger values mean that the learner is less likely to discover category structure that has been assigned low probability by the prior.

Note that, since the bias values are drawn from the Dirichlet distribution, they are interrelated such that $|\boldsymbol{k}| = 1$. If the prior is biased toward a particular dimension, the learner will be less likely to discover category distinctions based on other dimensions. Therefore, the model illustrates the general principle of label-based category learning: the presence of labels will boost learning in cases where true category structure aligns with the prior and hinder it when it doesn't.

Results

The setup We tested the model using a setup that aimed to approximate the experiments in Brojde et al. (2011). The model was trained on data from 2 categories, with 8 exemplars in each (Fig. 4). Exemplars varied along 2 dimensions, from now on referred to as "shape" and "material". The shape dimension had a high prior probability of being diagnostic during label-based learning, and material had a low prior probability; during the label-free condition, prior probabilities for the two dimensions were equal. We considered three learning scenarios: label-based learning where the label-induced prior aligns with true category structure ("right bias"), label-based learning where the prior doesn't align with true category structure ("wrong bias"), and label-free learning ("no bias").

Similarly to the experimental setup in Lupyan et al. (2007) and Brojde et al. (2011), the model was presented with all 16 stimuli in each learning block. We fitted the model to the data using the Python-based probabilistic programming language PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) and evaluated the model's performance by predicting exemplars' category membership based on the estimated means and standard

[3] $R_i = [\text{diag}(\Sigma_i)]^{-1/2} \Sigma_i [\text{diag}(\Sigma_i)]^{-1/2}$

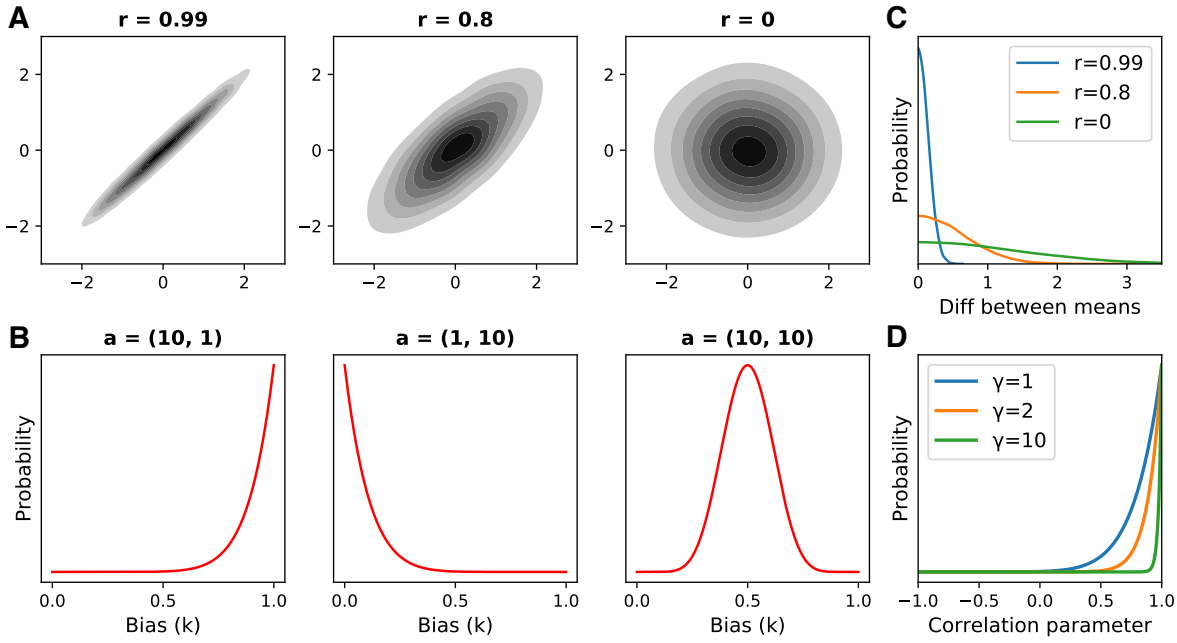

Figure 3: A. Category means for each dimension are sampled in accordance with the correlation value r. B. The vector $\boldsymbol{a}$ defines the distribution over the values of bias $\boldsymbol{k}$ (k_1 shown here; $k_2 = 1 - k_1$). C. There is a nonlinear relationship between r and the mean distance between the two sampled means. D. γ determines the distribution of the correlation values sampled under a given k.

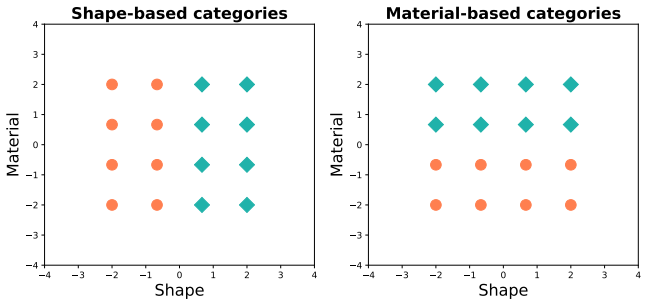

Figure 4: Category exemplars varied across two dimensions ("shape" and "material"), only one of which was diagnostic. The axis units are arbitrary.

deviations for each category. We held constant the amount of perceptual noise ($\sigma_s^2 = 1$), the nonlinear transform parameter ($\gamma = 10$) and the values of the overhypothesis parameter (label-based condition, right bias: $\boldsymbol{a} = (10, 1)$; label-based condition, wrong bias: $\boldsymbol{a} = (1, 10)$; label-free condition: $\boldsymbol{a} = (10, 10)$; see Fig. 3B for resulting probability distributions over the bias $\boldsymbol{k}$). Note that we made a simplifying assumption that, in the absence of word labels, the learner will have no bias toward either dimension; in practice, $\boldsymbol{a}$ would likely reflect both word-induced biases and general categorization biases. Further, although we keep $\boldsymbol{a}$ fixed, such overhypotheses can, in principle, be learned from data (Kemp et al., 2007).

Main results We found that the model was able to reproduce the difference between the labeled and the non-labeled conditions for both experiments (Fig. 5). In Lupyan et al. (2007), the categories were shape-based, and so the label-induced prior was beneficial for learning, as illustrated by our model; in contrast, Brojde et al. (2011) used categories that could be either facilitated or hindered by the label-induced prior, and our model successfully captured this distinction.

In order to quantitatively estimate the strength of the predicted effect, we simulated item-level predictions (with 75 participants per condition) and analyzed them using a mixed effects logistic regression model (the experimental papers report ANOVA statistics, but this is not recommended for accuracy data; Jaeger, 2008). The model we used was $accuracy \sim condition * block + (1|item)$, fitted with lme4 (Bates, Mächler, Bolker, & Walker, 2015). We found the main effect of bias in both directions (right bias: $\beta = 1.10$, $SD = 0.14$; wrong bias: $\beta = -1.01$, $SD = 0.12$), as well as the main effect of block ($\beta = 0.48$, $SD = 0.04$), indicating that model performance improved over time. We also observed an interaction between block and condition (right bias: $\beta = -0.25$, $SD = 0.05$; wrong bias: $\beta = 0.27$, $SD = 0.05$), reflecting the fact that, at the end of learning, the model performed equally well under all three conditions, despite differences in initial performance.

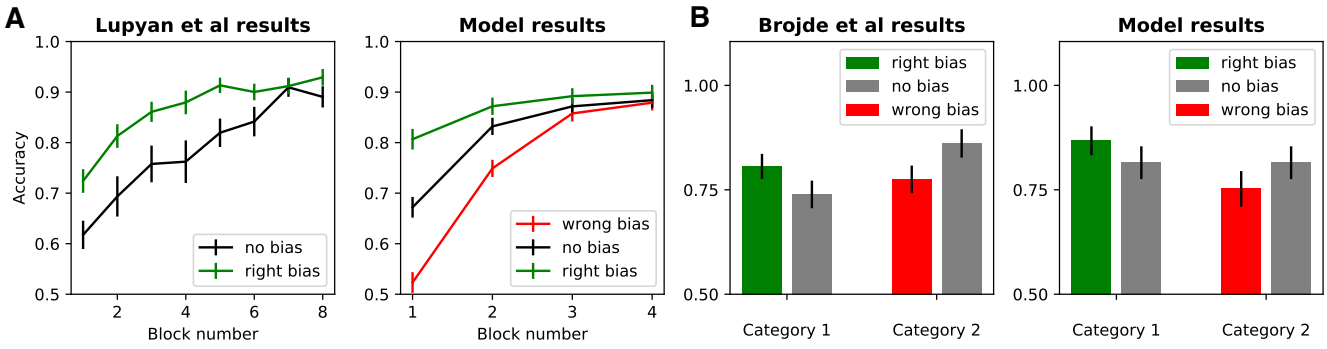

Figure 5: Modeling results: A. Learning dynamics reported by Lupyan et al. (2007; means and standard errors taken from Fig. 2) and exhibited by our model ($N_{simulated} = 75$). B. Overall accuracy reported by Brojde et al. (2011; means and standard errors taken from Fig. 3) and exhibited by our model. Note that the overall learning rate and accuracy of the model are a function of input data and noise parameters; the main result of interest here is the difference between bias conditions.

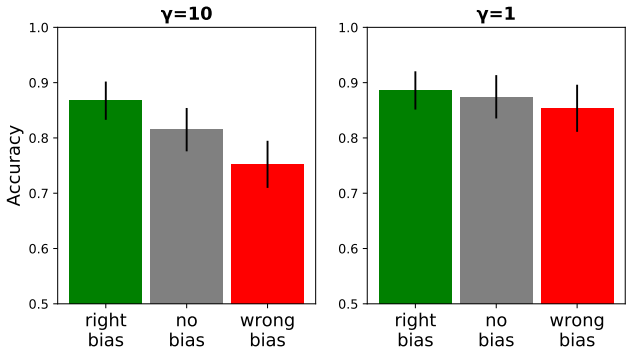

Figure 6: Effect of γ on the label advantage effect. A nonlinear tradeoff between category dimensions ($\gamma > 1$) results in the difference between all three conditions, whereas for a linear tradeoff ($\gamma = 1$), the right bias advantage disappears and the wrong label disadvantage is reduced.

The nonlinear transform effect We additionally examined a version of the model where the bias values k mapped linearly onto the correlation values ($\gamma = 1$; Fig. 6). We found that, when the mapping was linear, the difference between the right bias and the no-bias conditions was not significantly different from 0 ($\beta = 0.13$, $SD = 0.15$) and the difference between the wrong bias and the no-bias conditions was reduced ($\beta = -0.68$, $SD = 0.14$). This change in the result pattern is likely caused by the fact that a smaller value of gamma imposes a softer constraint on the correlation values of the means (Fig. 3D), which, in turn, makes the learner more likely to consider hypotheses that have lower probability under the prior and thus to discover the relevant dimension(s) even if the prior was not biased toward it. We conclude that, in order for the label advantage effect to appear, the learner needs to not only be biased toward the relevant dimension, but also to be restricted with respect to what category structures she can consider under both label-based and label-free conditions.

Discussion

We have proposed a computational account of the relationship between verbal labels and category learning. It posits the existence of overhypotheses that differ between word-based category learning and category learning in the absence of word labels. We show that an overhypothesis biased toward certain dimensions of the input stimulus can either facilitate or hinder learning depending on whether or not the privileged dimension is relevant for distinguishing these particular categories. This result explains the behavioral findings by Lupyan et al. (2007) and Brojde et al. (2011), which were not predicted by alternative models (the labels-as-features hypothesis and the label-feedback hypothesis), neither of which can account for the hindering effect of verbal labels under certain conditions.

Implications for language and thought The hierarchical Bayesian model presented here describes a set of computational principles that link learning perceptual categories and learning word meanings. We see that category learning can proceed via the same mechanism with or without verbal labels, but the learner can utilize probabilistic information about already known word meanings to infer the likely structure of a novel category. We therefore consider word meanings to be a subset of concepts more broadly (Jackendoff, 2002), with verbal labels inducing specific expectations over the structure of such concepts. Our model posits that linguistic and conceptual processes are distinct but interrelated and provides a way to examine their relationship in a principled way.

Application to other phenomena Although the experiments we aimed to model focused on shape bias, our model can be generalized to explain a number of other differences observed between label-based and label-free learning. For instance, the framework we propose can be used to model the

effect of *iconicity* on category learning (Lupyan & Casasanto, 2015) by introducing a prior defining the correspondence between the perceptual properties of the label j and the mean value of the category j along a given dimension i. Furthermore, in the current model, the prior over category dimensions defines not only which dimensions are relevant, but also how different dimensions compete for the learner's attention. A strong bias toward a particular dimension would therefore induce a *sparsity* constraint, which is another reported feature of label-based learning (e.g., Perry & Lupyan, 2014). Identifying the source of the sparsity constraint and its relationship with linguistic overhypotheses is thus a fruitful direction for future research.

Label effect in adults vs. infants Our model aims to simulate results of label-based category learning in adults; as such, we expect the participants to have fully developed priors over likely word meanings, such as shape bias. Since shape bias develops as a result of vocabulary learning (Smith et al., 2002) and becomes more pronounced with age (Landau et al., 1988), we do not expect it to affect categorization in infants. Therefore, a labels-as-features model, such as Gliozzi et al. (2009), might be more appropriate for characterizing the effects of labels on infant learning.

Overall, this paper aims to highlight the importance of formalizing computational principles that underlie the link between words and concepts. The notion of linguistic overhypotheses provides a clear and elegant way to account for the label effects observed in prior experimental work and can be leveraged further to explain other aspects of the word-concept relationship. This knowledge can later be used to develop and test mechanistic models of category learning, establish constraints on processing-level theories of concept learning and use, and probe the neural substrate of concept representations in order to elucidate when and how concepts interact with words.

Acknowledgements

We thank Roger Levy for helpful discussion at earlier stages of this project.

References

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLOS One*, *9*(7).

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bloom, P. (2002). *How children learn the meanings of words*. MIT Press.

Brojde, C. L., Porter, C., & Colunga, E. (2011). Words can slow down category learning. *Psychonomic Bulletin & Review*, *18*(4), 798–804.

Dickinson, D. K. (1988). Learning names for materials: Factors constraining and limiting hypotheses about word meaning. *Cognitive development*, *3*(1), 15–35.

Gardenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, *2*(2), 9–27.

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, *20*(3-6), 507–523.

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738.

Imai, M., Kanero, J., & Masuda, T. (2016). The relation between language, culture, and thought. *Current Opinion in Psychology*, *8*, 70–77.

Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, USA.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology*, *3*, 54.

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, *7*(2), 167–193.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.

Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language*, *135*, 66–72.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 2940–2949).

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55. doi: 10.7717/peerj-cs.55

Smith, L. B., Jones, S. S., & Landau, B. (1996). Naming in

young children: A dumb attentional mechanism? *Cognition*, *60*(2), 143–171.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245.