TOWARD HUMAN-LIKE OBJECT NAMING IN ARTIFICIAL NEURAL SYSTEMS

Tiwalayo N. Eisape,[1] **Roger Levy,**[1] **Joshua B. Tenenbaum**[1,2,3] **& Noga Zaslavsky**[1,3]
[1]Department of Brain and Cognitive Sciences, [2]CSAIL, [3]Center for Brains Minds and Machines
Massachusetts Institute of Technology
Cambridge, MA 02142, USA
{eisape,rplevy,jbt,nogazs}@mit.edu

ABSTRACT

Partitioning a rich set of objects into words is a fundamental aspect of human language and a major challenge for machines. Recently, it has been argued that word meanings evolve under pressure to optimize the Information Bottleneck (IB) principle and that this framework may be used to inform AI systems with human-like semantics. However, a major challenge for invoking this approach at scale is that it assumes an underlying representation of the environment which is often unknown. Here, we address this challenge by leveraging deep learning models for specifying such underlying representations. We demonstrate our approach in the domain of containers by evaluating optimal IB container-naming systems derived from representations generated by each layer of CORnet-S, a brain-inspired deep learning image classifier. We show a gradient in success in accounting for the container-naming systems of Dutch and French, where the deeper layer of CORnet-S that roughly corresponds to a high-level object recognition area in the brain outperforms shallower layers that correspond to lower-level visual processing. This suggests that our approach may be useful for testing the relevance of various types of non-linguistic representations to the emergence of word meanings, and could potentially aid in informing artificial neural agents with human-like semantics.

1 INTRODUCTION

A key aspect of human language is the mapping of a rich, and often continuous, set of objects into a relatively small number of words (Harnad, 1990; Rosch, 1999). An understanding of the computational mechanisms that languages deploy to achieve this stands to not only elucidate the cognitive underpinnings of language, but such an understanding could potentially aid in studying how language may emerge in artificial agents. A prominent computational cognitive approach argues that word meanings, among other aspects of language, are shaped by pressure for efficient communication (for review: Kemp et al., 2018; Gibson et al., 2019). Zaslavsky et al. (2018) grounded this idea in the Information Bottleneck (IB) principle (Tishby et al., 1999) and argued that languages compress underlying representations of the environment into words by optimizing the IB complexity–accuracy tradeoff. In support of this proposal, it was shown that IB accounts for cross-linguistic data in several semantic domains, such as colors and containers (Zaslavsky et al., 2018; 2019). Furthermore, this approach can potentially inform AI with human-like semantics (Zaslavsky et al., 2017).

However, a major challenge for invoking this approach, as well as related approaches that view word meanings as partitions of some feature space (e.g. Labov, 1973; Regier et al., 2015), is the need to specify an underlying representation of the environment. While a standard perceptual space has been used in the case of color naming (Regier et al., 2007), such a space does not generally exist. For containers, non-linguistic human similarity judgments have been proposed as a proxy for such a space (Xu et al., 2016). However, this method does not scale, leaving open the question of how to specify an underlying representation in a way that is both cognitively motivated and scalable.

In the present work, we address this challenge by proposing a domain-general paradigm for leveraging deep learning models to specify underlying representations of objects. As a first step in demonstrating our approach, we consider the domain of containers and specify this domain using

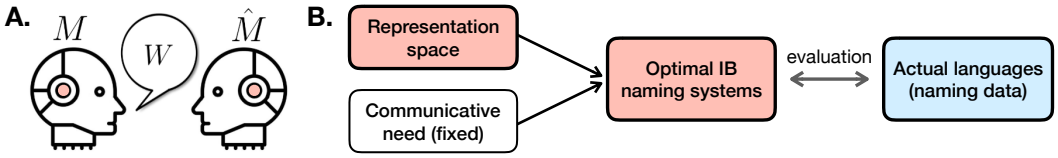

Figure 1: **A.** Communication model (see text for details). **B.** Schematic illustration of our approach. The communication model depends on the representation space $\mathcal{M}$ and the communicative need distribution $p(m)$. Here, we fix the latter and explore the influence of the representation space on the corresponding optimal IB systems and their ability to account for human naming data.

representations of container images generated by the brain-inspired deep learning image classifier CORnet-S (Kubilius et al., 2018). Each layer of CORnet-S gives an underlying representation of the domain, which in turn induces an optimal IB container-naming system. We evaluate these IB systems on container naming data from Dutch and French. We show that the deep layer that roughly corresponds to a high-level object recognition area in the brain (IT) outperforms shallower layers that correspond to lower-level visual processing areas (V1, V2, and V4). This suggests that our paradigm may be useful for testing the relevance of various types of non-linguistic representations to the emergence of human-like semantic systems. It also sets the stage for evaluating, in future work, other types of neural representations that capture more sophisticated properties of objects, such as 3D shape and intuitive physics (e.g. Kulkarni et al., 2015; Wu et al., 2015; 2018; Zhang et al., 2018; Battaglia et al., 2013), which are likely to influence semantic categories.

2 Theoretical framework

We begin by reviewing the theoretical framework of Zaslavsky et al. (2018), which the present work extends. This framework is based on the communication model shown in Figure 1A. The speaker obtains a mental representation $M \in \mathcal{M}$, sampled from a communicative need distribution $p(m)$, and maps it to a word $W \in \mathcal{W}$ using a naming distribution, or a stochastic encoder, $q(w|m)$. Given W, the listener's goal is to infer the speaker's mental representation by forming an estimator $\hat{M}$. These mental representations are formulated as beliefs over objects in the environment, and we refer to $\mathcal{M}$ as the underlying *representation space*. More specifically, each $m \in \mathcal{M}$ is defined by a probability distribution over a set $\mathcal{U}$ that characterizes the objects that the speaker and listener may communicate about. Following Regier et al. (2015), we consider similarity-based distributions over $\mathcal{U}$. That is, assuming $\mathcal{U}$ is a set of possible objects in the environment, each target object $c \in \mathcal{U}$ is represented by the distribution $m_c(u) \propto \exp(\gamma \mathrm{Sim}(c, u))$, where $u \in \mathcal{U}$, Sim is a similarity measure, and $\gamma > 0$ reflects the speaker's uncertainty. Finally, we assume a Bayesian listener whose inference $\hat{M}$ given $W = w$ is defined by $\hat{m}_w(u) = \sum_{m \in \mathcal{M}} q(m|w)m(u)$.

The IB principle predicts that efficient naming systems optimize the tradeoff between minimizing the complexity of the lexicon, measured by $I_q(M; W)$, and maximizing the accuracy of communication which is inversely related to $\mathbb{E}_q[D[M \| \hat{M}]]$, the KL-divergence between the speaker's and listener's mental representations. More formally, an optimal IB system $q_\beta(w|m)$ minimizes $\mathcal{F}_\beta[q] = I(M; W) + \beta \mathbb{E}_q[D[M \| \hat{M}]]$,[1] for some tradeoff parameter $\beta \geq 0$. The set of optimal IB systems as a function of β defines the IB theoretical limit. Zaslavsky et al. (2018; 2019) showed that color naming across languages lies near the IB theoretical limit and that this result generalizes to other semantic domains, including container naming.

In order to apply this framework across domains the underlying representation space $\mathcal{M}$ must be specified first, which is often a challenge. This work aims to address this challenge by leveraging state-of-the-art deep learning models. To this end, we explore the influence of representation spaces generated from such models, while keeping all the other components of the model fixed (see Figure 1B). Next, we describe our proposed approach for generating the representation space in cases where each object in $\mathcal{U}$ is associated with a visual stimulus.

[1] This optimization problem is equivalent to the standard IB problem $\min_q \; I_q(M; W) - \beta I_q(W; U)$. See (Zaslavsky et al., 2018) for a detailed derivation.

3 SPECIFYING THE REPRESENTATION SPACE

To specify the representation space $\mathcal{M}$ we need to define the similarity measure and the uncertainty parameter γ. While representation spaces defined from non-linguistic similarity judgments have been useful for explaining container naming systems across languages (Xu et al., 2016), this approach has several limitations. First, it requires data that may be expensive to collect. Second, it does not generalize to new objects. Third, it is unclear how to apply this to multiple domains simultaneously. Therefore, we propose another approach that may overcome these limitations.

The key idea is to generate similarity scores from deep neural networks trained on a non-linguistic task, e.g., object recognition. We assume that each object $u \in \mathcal{U}$ is associated with a visual stimulus which can be given as input to the network and represented by a latent vector h_u generated by the network. We then construct a representational similarity matrix (RSM; Kriegeskorte et al., 2008) using cosine similarity. That is, we take $\mathrm{Sim}(c, u) = 1 - \cos(h_c, h_u)$. RSMs can be generated for different networks and for different layers within a single network. Finally, to allow for a fair comparison across models, we set γ for each RSM such that the maximal possible accuracy induced by the corresponding representation spaces is constant (see Appendix A for details).

One potential concern is that these underlying representations would not necessarily be human-like. Therefore, we consider brain-inspired deep learning models. A stand-out model class in this regard is the Core Object Recognition network (CORnet; Kubilius et al., 2018) family of convolutional neural networks (CNNs). This set of architectures not only draws inspiration from the brain, but its constituents are each hierarchically organized in a manner explicitly analogous of the human ventral visual stream. As such, they are engineered with layers directly analogous to visual areas V1, V2, V4, and the inferior temporal cortex (IT). Specifically, we consider the CORnet-S model which achieves high 'Brain-Score' (Schrimpf et al., 2018) — a benchmark for quantifying the model's fit to neural and behavioral data — as well as high accuracy on ImageNet (Krizhevsky et al., 2012).

4 THE CASE OF HOUSEHOLD CONTAINERS

We demonstrate our approach in the domain of container naming for two main reasons. First, container categories are shaped not only by low-level perceptual features, such as color, but also (and perhaps primarily) by higher-level features such as spatial structure, material, and usage (Labov, 1973). Because it is a priori unclear what are the most relevant features, models that learn to extract relevant features could be particularly beneficial for this domain. While CORnet-S is limited in this sense because it is CNN-based, it may still behave as a useful proxy (Wu et al., 2015). Second, this domain has previously been studied in information-theoretic terms (Xu et al., 2016), and specifically within the IB framework (Zaslavsky et al., 2019), using human judgments to specify the domain. This prior work sets a baseline for evaluating the present approach.

Container naming data. Our analysis is based on data collected by White et al. (2017). Dutch and French monolinguals were shown container images (see examples in Figure 2) and were asked to provide a name for each container. We restrict our analysis to a subset of 77 container images with consistent image background to simplify the pre-processing of images. For each language l, we estimated a naming distribution $p_l(w|m)$ by mapping each container c to its representation m_c and averaging the naming responses across participants.

Models and predictions. We compare five IB container-naming models that differ only by their underlying representation space $\mathcal{M}$. As a baseline, we consider the model of Zaslavsky et al. (2019) in which $\mathcal{M}$ was defined from human similarity judgments, also collected by White et al. (2017). We refer to this model as SIM. In addition, we derived a model from the representation of each

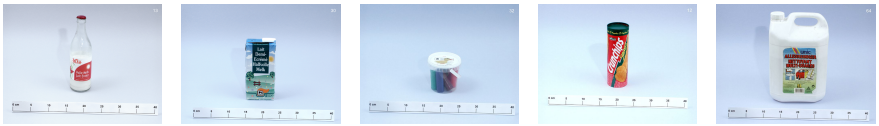

Figure 2: Examples of container images used by White et al. (2017) to elicit naming responses.

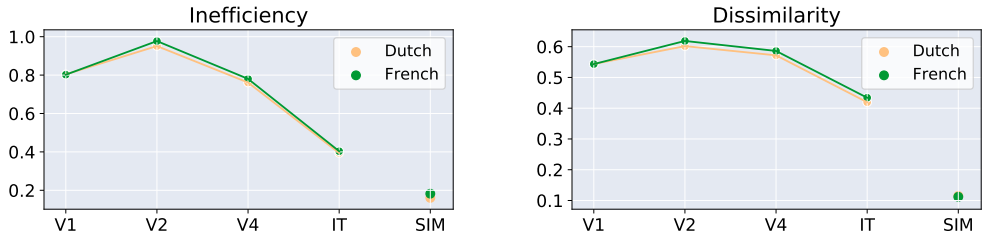

Figure 3: Inefficiency (left) and dissimilarity (right) scores for the IB container-naming models derived from CORnet-S (V1-IT) and from human similarity judgements (SIM). The IT model outperforms lower-level visual models, but does not reach the performance of the SIM model.

CORnet-S layer. That is, we presented the container images to CORnet-S and computed RSMs for the V1, V2, V4, and IT layers. For each RSM, we constructed its corresponding representation space, and computed the optimal IB systems $q_\beta(w|m)$ and theoretical limit via reverse deterministic annealing with 2000 values of $\beta \in [0, 16384]$.

We predict that optimal systems derived from low-level visual areas will not account well for naming systems across languages, because semantic categories, at least in this domain, are likely to be influenced by higher-level perceptual and conceptual features. Therefore, we expect an improvement in performance as the model's representation space is generated from deeper layers of CORnet-S.

Results. We evaluate the models using two measures (see Appendix B): (i) the inefficiency of the actual languages, i.e., their deviation from optimality; and (ii) the dissimilarity between the actual naming systems and their nearest optimal systems. Low inefficiency implies that the actual languages achieve near-optimal complexity-accuracy tradeoffs, and low dissimilarity implies that the optimal systems are human-like. Our predictions imply that inefficiency and dissimilarity should decrease with the depth of the layer.

Figure 3 shows the inefficiency and dissimilarity scores for all five models. As expected, there is a monotonic improvement from V2 to IT, and the IT representation substantially outperforms all the lower-level visual representations. This suggests that, at least in this domain, high-level visual features are more relevant for semantic categories than than low-level visual features. Unexpectedly, however, the V1 model outperforms the V2 model. One speculative explanation for this is that low-level V1 features may still have some role in shaping container categories.

Notably, the models derived from CORnet-S fail to reach the performance of the SIM model. This is likely indicative of the gap between the perceptual features captured by CORnet-S' IT layer, and more complex physical and functional features which are less likely to be captured by CORnet-S but are presumably reflected in human similarity judgments. In future work, we intend to extend our analysis to richer forms of representation that capture 3D object shapes (e.g. Kulkarni et al., 2015; Arsalan Soltani et al., 2017; Wu et al., 2017; 2018; Zhang et al., 2018) and intuitive physics of solids and fluids (e.g. Battaglia et al., 2013; Wu et al., 2015; Bates et al., 2019), in order to better capture the functional affordances of containers (or other objects) and the relationships between object form and function that are likely crucial for understanding how language picks out object categories.

5 Conclusions

We have proposed a general method for deriving efficient semantic systems from artificial neural representations of visual objects. Our initial results in the domain of container naming suggest that this method can be useful for exploring which types of underlying representations may lead to human-like semantic systems. We believe that this is noteworthy both from a cognitive perspective and from and an AI perspective. From a cognitive perspective, this work provides a potential paradigm for further studying the relation between semantic categories and perceptual or conceptual representations. From an AI perspective, this work suggests a principled method for generating efficient, and potentially human-like, semantic systems in artificial neural agents by adding a semantic IB component to existing representation learning systems.

ACKNOWLEDGMENTS

We thank Martin Schrimpf for kindly sharing with us the trained CORnet-S model and for helpful discussions. TE was supported by the GEM Consortium and the MIT-Dean of Sciences Fellowship. NZ was supported by the BCS Fellowship in Computation.

REFERENCES

Amir Arsalan Soltani, Haibin Huang, Jiajun Wu, Tejas D Kulkarni, and Joshua B Tenenbaum. Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1511–1519, 2017.

Christopher J. Bates, Ilker Yildirim, Joshua B. Tenenbaum, and Peter Battaglia. Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, 15(7): 1–29, 07 2019.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Edward Gibson, Richard Futrell, Steven T Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

Charles Kemp, Yang Xu, and Terry Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), 2018.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105. 2012.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

Tejas D Kulkarni, Pushmeet Kohli, Joshua B Tenenbaum, and Vikash Mansinghka. Picture: A probabilistic programming language for scene perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4390–4399, 2015.

William Labov. The boundaries of words and their meanings. In Charles-James Bailey and Roger W. Shuy (eds.), *New ways of analyzing variation in English*, pp. 340–371. Georgetown University Press, 1973.

Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441, 2007.

Terry Regier, Charles Kemp, and Paul Kay. Word meanings across languages support efficient communication. In B. MacWhinney and W. O'Grady (eds.), *The Handbook of Language Emergence*, pp. 237–263. Wiley-Blackwell, Hoboken, NJ, 2015.

Eleanor Rosch. Principles of categorization. In Eric Margolis and Stephen Laurence (eds.), *Concepts: Core Readings*, pp. 189–206. MIT Press, 1999.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-Score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 2018.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Anne White, Barbara C Malt, and Gert Storms. Convergence in the bilingual lexicon: A preregistered replication of previous studies. *Frontiers in Psychology*, 7:2081, 2017.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, pp. 127–135. 2015.

Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T. Freeman, and Joshua B. Tenenbaum. MarrNet: 3D shape reconstruction via 2.5D sketches. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 540–550, 2017.

Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3D completion and reconstruction. In *Proceedings of the European Conference on Computer Vision*, pp. 646–662, 2018.

Yang Xu, Terry Regier, and Barbara C Malt. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, 40(8):2081–2094, 2016.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient human-like semantic representations via the Information Bottleneck principle. In *Cognitively Informed AI workshop at the 31st Conference on Neural Information Processing Systems*, 2017.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 2019.

Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Josh Tenenbaum, Bill Freeman, and Jiajun Wu. Learning to reconstruct shapes from unseen classes. In *Advances in Neural Information Processing Systems*, pp. 2257–2268, 2018.

APPENDIX

A SETTING THE SPEAKER'S UNCERTAINTY PARAMETER

Recall that γ controls the speaker's uncertainty through $m_c(u) \propto \exp(\gamma \mathrm{Sim}(u, c))$. We estimate this parameter by assuming that the speaker has a constant memory capacity, regardless of the choice of Sim. We define the speaker's memory capacity by

$$I(M;U) = \sum_{c,u \in \mathcal{U}} p(m_c) m_c(u) \log \frac{m_c(u)}{p(u)} \,, \tag{1}$$

where $p(u) = \sum_c p(m_c) m_c(u)$, and $m_c(u)$ is treated as $p(u|c)$. It is possible to show that $I(M;U)$ is also the maximal achievable communication accuracy (Zaslavsky et al., 2018). Therefore, keeping this capacity constant across models amounts to setting the same accuracy scale for all models, which is desired for a fair model comparison.

Note that $I(M;U)$ is a function of γ for a given similarity measure. Figure 4 shows these functions for the similarity measures derived from CORnet-S' layers. While it is unclear how to determine the value of $I(M;U)$, the two extremes seem unnatural: $\gamma = 0$ corresponds to a speaker that has no memory of what they need to communicate; $\gamma \to \infty$ corresponds to a speaker with perfect memory that never confuses two objects even if they are extremely similar. Therefore, we select an intermediate value, $I(M;U) = 2$, and adjust γ accordingly for each layer (Figure 4, dashed lines). Finding better ways to estimate the speaker's memory capacity is an important direction which is left for future work.

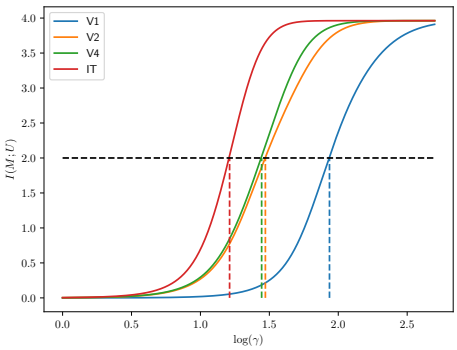

Figure 4: Speaker's memory capacity (solid lines), which is also the maximal communication accuracy, as a function of $\log(\gamma)$. The horizontal black dashed line corresponds to capacity used in this work. The vertical dashed lines correspond to values of γ used for each layer.

B MODEL EVALUATION

Our model evaluation follows the same evaluation procedure of Zaslavsky et al. (2018). Recall that a model in our setting gives a set of IB systems $q_\beta(w|m)$ as a function of β, and their corresponding optimal complexity-accuracy tradeoffs $\mathcal{F}^*_\beta = \mathcal{F}_\beta[q_\beta]$. For a given model, the tradeoff achieved by language l is computed by plugging $p_l(w|m)$ into $\mathcal{F}_\beta$. Each language is then compared to the nearest optimal system, that is, the optimal system for $\beta_l = \arg\min_\beta \{\mathcal{F}_\beta[p_l] - \mathcal{F}^*_\beta\}$.

Inefficiency is defined by $\varepsilon_l = \frac{1}{\beta_l}(\mathcal{F}_{\beta_l}[p_l] - \mathcal{F}^*_{\beta_l})$. This measures the deviation from optimality of the language's complexity-accuracy tradeoff.

The *dissimilarity* measure does not consider the complexity-accuracy tradeoffs, but rather compares directly the actual system $p_l(w|m)$ with the corresponding optimal system $q_{\beta_l}(w|m)$. This is defined by gNID (Zaslavsky et al., 2018), an information-theoretic measure for the divergence between two naming distributions.