

Identifying concept libraries from language about object structure

Catherine Wong*, William P. McCarthy*², Gabriel Grand*¹, Yoni Friedman¹,
Joshua B. Tenenbaum¹, Jacob Andreas¹, Robert D. Hawkins³ and Judith E. Fan²

¹MIT, ² University of California San Diego, ³Princeton Neuroscience Institute

*denotes equal contribution, correspondence to catwong@mit.edu

Abstract

Our understanding of the visual world goes beyond naming objects, encompassing our ability to parse objects into meaningful parts, attributes, and relations. In this work, we leverage natural language descriptions for a diverse set of 2K procedurally generated objects to identify the parts people use and the principles leading these parts to be favored over others. We formalize our problem as search over a space of program libraries that contain different part concepts, using tools from machine translation to evaluate how well programs expressed in each library align to human language. By combining naturalistic language at scale with structured program representations, we discover a fundamental information-theoretic tradeoff governing the part concepts people name: people favor a lexicon that allows concise descriptions of each object, while also minimizing the size of the lexicon itself.

Keywords: abstraction; compositionality; parts; perception; programs

The world is filled with a great variety of objects, yet people have little difficulty making sense of them. Presented with a novel object, people can readily identify its parts (Schyns & Murphy, 1994), guess its function (Tversky & Hemenway, 1984), and refer to it unambiguously (Hawkins et al., 2020). These abilities rest on the capacity to robustly connect features of the external world to a rich library of mental concepts describing not just whole objects, but their parts and how they are arranged (Miller & Johnson-Laird, 1976; Landau & Jackendoff, 1993; Rosch & Mervis, 1975; Mukherjee et al., 2019).

For example, consider the bottom-most *gadget* in Fig. 1A: even though this object does not correspond to a familiar category, we might say that it contains a row of buttons or dials, and that it is topped by an antenna or a knob. But just as we do not have a pre-existing concept for every object we encounter, we do not have a concept corresponding to every part: in Fig. 1A, for example, most people do not have a concept corresponding to a row of exactly five dials. Indeed, a complex object can be decomposed in a huge number of different ways, but people are likely to favor only a tiny subset of them. What characterizes the set of part concepts that people do use? Why these, and not others?

Identifying which parts people use to parse visual objects has been a core goal for classic theories of perceptual organization (Palmer, 1977; Marr & Nishihara, 1978;

Hoffman & Richards, 1984; Biederman, 1987) and continues to pose challenges for modern vision models (Mo et al., 2019; Bear et al., 2020). But how can we tell whether any of these proposals actually explain visual object understanding? Empirical tests of these theories have generally relied upon simple discrimination tasks rather than richer behavioral readouts (Tversky, 1989; Markman & Wachtel, 1988), limiting their ability to evaluate correspondences between a candidate object representation and the full set of parts and relations that people can identify.

Natural language offers a powerful window into our conceptual representations, given abundant evidence that our vocabularies have been shaped to efficiently communicate about the concepts we find relevant (Regier et al., 2015; Kirby et al., 2015; Zaslavsky et al., 2018; Sun & Firestone, 2021). In this work, our goal is to leverage *naturalistic language production* to identify the conceptual libraries of parts and relations used for visual object understanding, using libraries of symbolic *program* components to model how these concepts are mentally represented (Fig. 1B). In this framework, each library instantiates a different hypothesis about the underlying inventory of part concepts that people are using to decompose visual objects.

In Part I, we describe our strategy for creating a diverse collection of novel objects generated using graphics programs (Fig. 1A), and for eliciting open-ended descriptions of these objects. Analyzing these descriptions reveals hallmarks of these concept libraries: people produce longer descriptions to describe more complex objects, and invoke different part concepts to describe objects from different categories. In Part II, we refine this picture with a formal library identification model that measures the correspondence between language and candidate program libraries containing part concepts of varying complexity, building on recent work in program library discovery (Ellis et al., 2020; Tian et al., 2020; Wang et al., 2021; Wong et al., 2021). This model reveals a deeper information-theoretic principle governing the part concepts people invoke in language: they reflect a fundamental trade-off between the complexity of a concept library and the complexity of objects represented using that library.

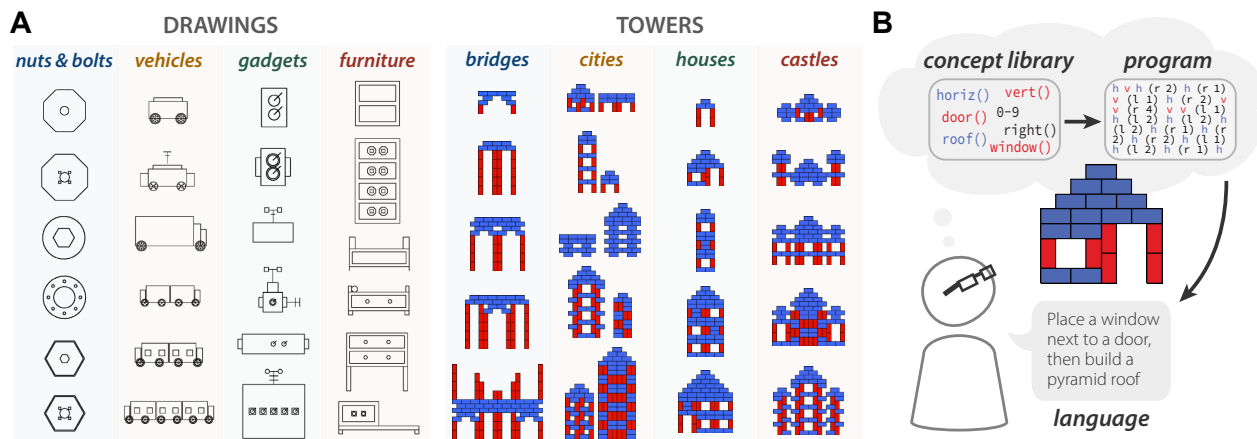


Figure 1: (A) Example objects from the *Drawings* and *Towers* domains. Each domain contains 4 subdomains of 250 novel objects. Each domain and subdomain was designed to include high variation over the type and number of base primitives (i.e., shapes, blocks). (B) This work aims to infer the latent concept library that people are using to decompose complex objects into parts, where objects are represented by executable graphics programs.

Part I: Eliciting language about object structure

Our central aim is to identify the library of part concepts that people invoke to decompose objects. Towards this end, we needed a sufficiently large and varied collection of objects, and a naturalistic task for eliciting detailed descriptions of their structure.

Methods

Participants We recruited 465 participants from Prolific to complete the task. Participants provided informed consent and were paid approximately \$15 per hour.

Stimuli To ensure that we had a sufficiently large and diverse collection of objects, we developed a hierarchical procedure for synthesizing complex configurations of shapes. Taking inspiration from recent work employing line drawings and block towers to investigate how people learn and represent the compositional structure of objects (Tian et al., 2020; McCarthy et al., 2021; Wang et al., 2021), we defined two stimulus *domains*, distinguished by the set of base shape primitives used to generate them (Fig. 1A). *Drawings* are composed of simple geometric curves (i.e., line, circle) and are evocative of familiar object categories; *Towers* are composed of rectangular blocks (i.e., horizontal and vertical dominoes) and are evocative of simple architectural models.

To investigate the degree to which people invoked category-specific part concepts to describe these objects, rather than the same set of “atomic” base primitives in all cases, we further defined four *subdomains* nested within each domain. Within *Drawings*, these were informally

designated as *nuts & bolts*, *vehicles*, *gadgets*, and *furniture*; and within *Towers*, as *bridges*, *cities*, *houses*, and *castles* (Fig. 1A). For each subdomain, we procedurally generated 250 unique examples, hierarchically composing the base primitives into increasingly complex, recursively defined parts. A dresser, for example, is composed of drawers, which are in turn composed of a panel and knobs, themselves defined by combining circles and lines. In sum, this procedure yielded a varied collection of 2000 object stimuli: 1000 *Drawings* and 1000 *Towers*, each accompanied by a graphics program that can be used to regenerate it in terms of the base primitives.

Task procedure Each participant was instructed to provide step-by-step instructions for how to “draw” or “build” 10 different “drawings” or “models” sampled from a *single* subdomain. At the start of each session, participants were first familiarized with the general characteristics of the subdomain by viewing 25 examples (none of which then appeared during the main experiment, and none of which were accompanied by any linguistic labels for the subdomain). Throughout the session, they were also shown the 7 upcoming objects they would be asked to describe, to provide concurrent information about how objects varied within the subdomain. Because we were primarily focused on interrogating which part descriptors people invoke, we designed the text-entry interface to encourage participants to describe each step by composing a *what*-phrase and a *where*-phrase, which were entered into separate text boxes. Participants could include as many instruction steps as they deemed necessary and there was no trial time limit.

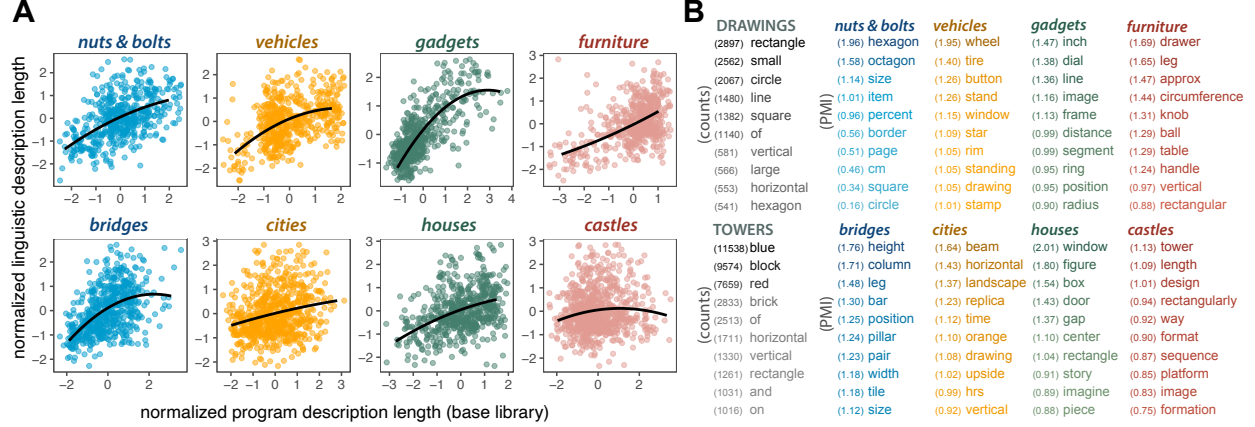


Figure 2: (A) Relationship between length of base-library programs and length of linguistic descriptions. (B) *Left*: Top-10 words that appeared most frequently in descriptions for each domain. *Right*: Top-10 words with highest pointwise mutual information (PMI) within each subdomain.

Language preprocessing To investigate the content of the instructions generated by participants, we used the spaCy NLP library to extract and lemmatize words, including part-of-speech (POS) tagging to remove determiners and punctuation. We also replaced common typos (“sqare,” “cirlce,” etc.) and spelling variations with their canonical spellings in US English.

Results

People use more words for more complex objects

The simplest way that object structure may be exposed in language is through description complexity. We consider three possibilities. First, insofar as participants decompose all objects into the same number of parts, regardless of how complex these parts are, the length of their descriptions would be predicted to remain *constant* over a wide range of objects. Second, if participants tend to decompose objects into a set of commonly recurring parts, and mention each part, the length of their descriptions would be predicted to positively correlate with object complexity: the more parts, the longer the description. A third possibility is that there is a systematic but non-linear relationship between object complexity and linguistic description length (Sun & Firestone, 2021), consistent with a compromise between the first two strategies.

We operationalize object complexity here as the length of the (base) graphics program that generated it. We measure the length of linguistic descriptions as the number of words provided in the *what* phrases (ignoring for now spatial language in the *where* phrases). To tease apart the above hypotheses, we fit a mixed-effects model to predict linguistic description length from graphics program length, including random intercepts for participants and random effects of program length at the participant level. We observed a significant main effect of program length

($t(318) = 14.8, p < 0.001$ across all subdomains), providing strong evidence against the first view. We also found that a model including an additional quadratic effect of program length, allowing for a non-linear relationship, significantly improved the fit ($\chi^2(3) = 38.6$), although the strength of this relationship varied across subdomains (Fig. 2A). These findings suggest that people generally use more words to describe more complex objects, but the strength and nature of this relationship can vary widely across object categories.

People use different parts for different subdomains

If description length scales with object complexity (expressed in the base library), a natural possibility is that speakers are simply providing descriptions at the level of those low-level primitives. For example, they may be giving block-by-block instructions for towers and line-by-line instructions for the drawings. In this case, we would not expect differences in the distribution of words used across subdomains (e.g. “bridges” and “houses” would both be described in terms of the same red and blue blocks). Alternatively, if speakers generate descriptions at higher levels of conceptual abstraction — for example, in terms of “pillars” or “windows” — we would expect their language to reflect the varying part structure of the subdomains. To assess these competing hypotheses, we computed the pointwise mutual information (PMI) for each unique word w in the language data with respect to the four subdomains d (Fig. 2B):

$$PMI(w) = \log \frac{p(w, d)}{p(w)p(d)} \quad (1)$$

Intuitively, PMI is high for words that occur more frequently in a particular subdomain (numerator) than would

be expected given the overall prevalence of the word across subdomains and the amount of language data in each subdomain (denominator). This analysis revealed highly specialized vocabularies used for particular subdomains, but not others (e.g., *drawer* and *knob* in the *furniture* subdomain), suggesting that participants did invoke subdomain-specific part concepts to some extent.

To better evaluate whether these highly diagnostic words reflected more systematic differences in word usage across subdomains, we computed the Jensen-Shannon distance (JSD) between the word frequency distributions in each set of subdomains, aggregating across all trials in that subdomain. This metric is zero when two distributions are identical and large when two distributions are far apart. We compared the mean of all pairwise JSDs to a null distribution generated by randomly permuting the subdomain group of each trial. We found that the distance between subdomains was significantly greater than expected under the null (*Drawings*: $d = 0.439$, $p < 0.001$; *Towers*: $d = 0.328$, $p < 0.001$). Taken together, these analyses indicate that people may choose distinct labels to describe visually similar parts depending on the rest of the scene (e.g. a circle may be a *knob* in one domain and a *wheel* in another domain), even when simple graphics primitives would have been sufficient.

Part II: Identifying concepts from language

The results so far suggest that people invoke subdomain-specific part concepts when describing the objects in our stimulus set, such as knobs and drawers, or windows and doors. What accounts for observed preferences for this lexicon — how many and which part concepts do people have names for?

In this section, we formalize this library identification problem by modeling the correspondence between people’s vocabularies and a space of candidate concept libraries, each containing part concepts at varying levels of complexity. We describe a procedure for constructing candidate libraries based on the hierarchical structure of each subdomain. We then introduce a library-to-vocabulary alignment model that measures how well programs written in each library predict people’s object descriptions (Wong et al., 2021).

Prior work suggests that people use language that efficiently compresses concepts into words (Regier et al., 2015; Kirby et al., 2015; Zaslavsky et al., 2018; Sun & Firestone, 2021). Our model allows us to derive an information-theoretic account of lexical choice in our object descriptions, which formally links language to efficient communication of an object’s underlying conceptual representation — we find that people favor a lexicon that trades off between concise descriptions of objects on average, and the size of the overall concept libraries.

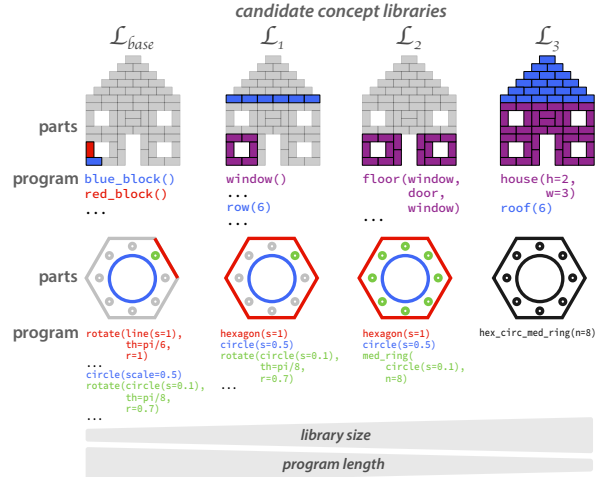


Figure 3: Graphics libraries were defined by progressively adding subroutines at higher levels of abstraction, resulting in more efficient expression of any particular program at the expense of a larger library.

Methods

Modeling a space of candidate concept libraries By design, the objects in our stimulus set are highly structured, having been generated through the hierarchical combination of increasingly complex parts. However, the corresponding graphics programs that recreate them were written using a concept library containing only the base primitives ($\mathcal{L}_{base}$): blocks and lines. As a consequence, these programs are maximally verbose: they must compose many individual blocks to represent a door, let alone an entire house; and many individual lines to represent a polygon like a hexagon, let alone a complex wheel.

To represent more complex shapes, we define higher-order graphics libraries that augment the initial set of base primitives with *program subroutines* (Fig. 3) that encapsulate part structure (e.g., a subroutine for generating an entire roof).¹ We constructed these libraries by abstracting out the nested, parametric functions used to generate each subdomain. In our experiments, we evaluate three libraries ($\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$), each containing subroutines that build recursively on those at the previous level to yield increasingly complex visual parts. For instance, $\mathcal{L}_1$ contains subroutines that abstract directly over the base library (e.g., from lines to polygons); and $\mathcal{L}_2$ contains subroutines that abstract additionally over those in $\mathcal{L}_1$ (e.g., polygons to rings of polygons). A given

¹Our approach to defining these higher-order libraries is analogous to the automated program library learning methods in (Ellis et al., 2020; Tian et al., 2020; McCarthy et al., 2021; Wang et al., 2021; Wong et al., 2021), which discover subroutines from a dataset containing programs that often correspond qualitatively to domain-relevant concepts.

program $\pi_{\mathcal{L}_{base}}$ written in the base library can therefore be expressed equivalently—and more concisely—as $\pi_{\mathcal{L}_i}$ in one of the higher-order libraries. It is worth noting that higher-order libraries are thus defined *cumulatively*: $\mathcal{L}_1$ contains the new subroutines *plus* the initial set of primitives in $\mathcal{L}_{base}$; and $\mathcal{L}_2$ contains even higher-order subroutines *plus* all of the concepts in $\mathcal{L}_1$.

Modeling alignment between libraries and vocabularies For each subdomain, the set of libraries $\{\mathcal{L}_{base}, \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3\}$ specifies a hypothesis space of alternative representations at differing levels of abstraction. We can now ask: which of these libraries best corresponds to the lexicon people use for each subdomain? We formalize this notion of lexical correspondence with a *library-to-vocabulary alignment metric* that reflects how closely the concepts in a given library co-occur with words across each subdomain. This metric is based a language-guided library learning model from the program synthesis literature (Wong et al., 2021). In brief, we leverage a standard machine translation model, IBM Model 1 (Brown et al., 1993), which can be fit to paired programs and instructions to estimate token-token translation probabilities $P(w|\rho)$ for each word $w \in W$ in the linguistic vocabulary and program component $\rho \in \mathcal{L}$ in the library. For each subdomain, we evaluate each library $\mathcal{L}_i$ using a cross-validation scheme (with batches of $n = 5$ held out stimuli). We fit the model to all but the held-out stimuli and evaluate the *mean per-word log-likelihood* for each held out instruction given its program in library $\mathcal{L}_i$. This metric varies monotonically as a function of negative perplexity (Wu et al., 2016) and normalizes for instruction lengths. As in Part I, we consider only the *what* phrases for each stimulus.

Results

Libraries produce different trade-off between concise object representation and overall library size Supposing that any of these libraries captures the part concepts that people use when describing these objects, what would lead participants to favor one over another? Our hypothesis is that this choice reflects a trade-off between the value of compressing the length of programs $|\pi_{\mathcal{L}_i}|$ that represent individual objects and the value of reducing the total number of concepts $|\mathcal{L}_i|$ stored in the library (Fig. 3)². Higher-order libraries contain concepts that compress programs to a greater degree, as each program can be written by invoking a smaller number of more abstract subroutines. However, each higher-order library is also larger than the last because it adds new concepts that must be represented along with all of the lower-level ones.

²This trade-off between program description length $|\pi_{\mathcal{L}_i}|$ and library size $|\mathcal{L}_i|$ is described in greater detail in (Ellis et al., 2020) and analogous to the formulation in (Kirby et al., 2015).

While library size increases monotonically with abstraction level, every subdomain has a non-monotonic *combined representational cost* $C_{\mathcal{L}_i} = |\mathcal{L}_i| + \frac{1}{N} \sum \pi_{\mathcal{L}_i}|$, where N is the number of programs in the subdomain. A one-way ANOVA confirms that, in every subdomain, this combined cost measure systematically varies between libraries ($ps \ll 0.001$), validating our assumption that these libraries capture different ways of negotiating the trade-off between object compression and library size. Further, as Fig. 4 reveals, $C_{\mathcal{L}_i}$ (dashed line) typically follows a U-shaped curve. At the extremes, $C_{\mathcal{L}_{base}}$ is high because programs in $\mathcal{L}_{base}$ are verbose, whereas $C_{\mathcal{L}_3}$ is high due to the large size of $\mathcal{L}_3$. In all subdomains, $C_{\mathcal{L}_1}$ and $C_{\mathcal{L}_2}$ tend to be optimal because these intermediate libraries contain a set of useful part-based abstractions that capture recurring structure across many objects.

People favor vocabularies that jointly minimize object representation and library size We can now consider the results of our *library-to-vocabulary alignment* model: which libraries best predict the words people use across each subdomain? To validate that this alignment metric is able to discriminate between libraries at all, we first conducted a one-way ANOVA on the alignment scores and found large and reliable differences between libraries in every subdomain ($ps < 0.001$; Fig. 4).

When we visualized these alignment scores (Fig. 4, solid lines), we observed that for the majority of the subdomains, the *mean log-likelihoods* follows an inverted U-shaped curve. Moreover, we generally find that the concept libraries that best predict language tend to be those containing parts of intermediate complexity—for example, part concepts (e.g., individual windows or wheels) that lie between the lowest (e.g., lines) and highest (e.g., hexagon with an inner ring of circular holes) levels of abstraction in each domain.

Finally, we observed a striking correspondence between the libraries that optimize combined representational cost ($C_{\mathcal{L}_i}$) and those that score highly on their ability to predict language. This pattern, which held for most (though not all) subdomains, suggests that people generally prefer decomposing objects into nameable parts that can be reused for many objects across the full subdomain.

Discussion

The language we use to describe the world reveals the concepts with which we represent it. In this paper, we look to natural language to investigate how people parse complex objects into meaningful parts—for example, how people decompose a whole train into its train cars and wheels, or a house into its windows, walls, and roof. We elicited descriptions for a large dataset of objects generated from *graphics programs*, and present a computational approach for linking their generative

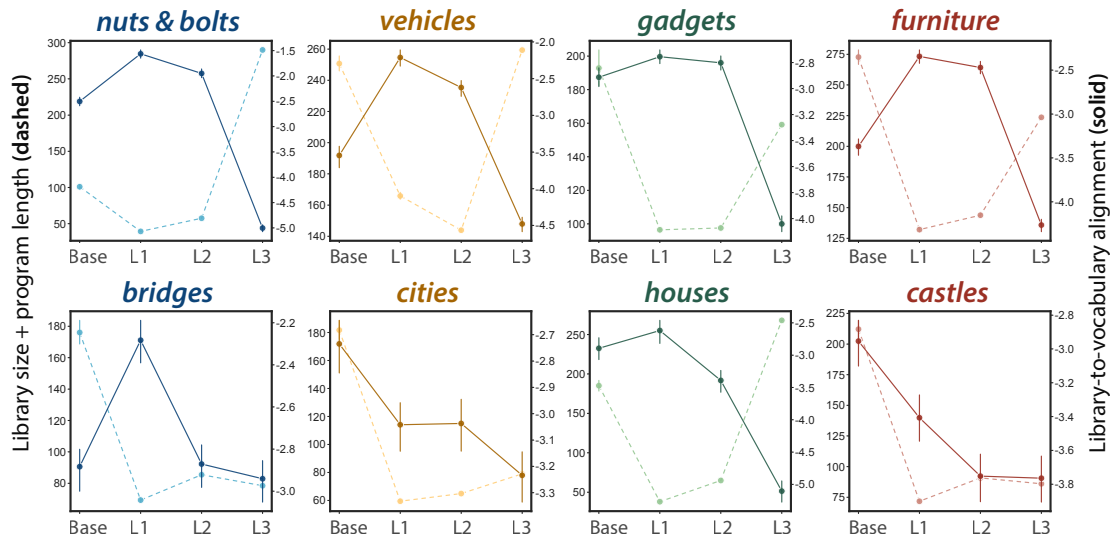


Figure 4: Relationship between concept libraries $\{L_{base}, L_1, L_2, \text{ and } L_3\}$ (x-axis); combined library size and average program length in that library (dashed); and library-to-vocabulary alignment (solid).

and hierarchical structure with human descriptions. We find that the length of people’s descriptions varies with the length of an object’s generative program, establishing a basic correspondence between language and a program representations of object structure. By constructing higher-order *concept libraries* which re-represent each object using more abstract program components, we find evidence that people’s language reflects an underlying representational trade-off – people prefer compact libraries of part concepts that efficiently capture structural motifs appearing in many objects. An intriguing implication of these findings is that there exists a “basic level” for part naming, by analogy to the well known basic level for object categories, and that can be explained by similar information-theoretic principles (Rosch et al., 1976).

While these linguistic abstraction layers enable greater compression, they may also introduce downstream challenges for communication: terms with more abstract meanings may be less interpretable and/or too lossy in some cases (e.g., pedagogical contexts where learners may not be familiar with certain concepts). To better understand how people communicate in these scenarios, it may be useful to conduct experiments manipulating what knowledge is shared between communicators to investigate the role of audience design and adaptation in interactive settings (Clark & Murphy, 1982; Krauss & Fussell, 1991; McCarthy et al., 2021).

In other settings, the level of detail contained in the descriptions we collected may not be necessary to achieve certain communicative goals, such as object identification. A promising direction is to compare our descriptions to

those produced in reference games where coarser distinctions between whole objects are sufficient, with the aim of understanding how task goals and context shape the *relevance* of different levels of abstraction (Degen et al., 2020; Bisk et al., 2020).

It is natural to expect substantial variation across descriptions in how well they support object understanding in others. To better understand why some descriptions are more informative than others, future work should also measure how well the descriptions we collected in the current study support the ability of other participants to accurately reconstruct the target objects.

Our approach and findings build on a recent and growing literature using programs (Lake et al., 2015; Goodman et al., 2014) and libraries of functional components (Tian et al., 2020; McCarthy et al., 2021; Wong et al., 2021) to model how people represent and communicate about the world. Our work generalizes previous insights into the statistical learning mechanisms that enable the rapid learning of visual regularities (Fiser & Aslin, 2001; Orbán et al., 2008; Austerweil & Griffiths, 2013) by proposing a more expressive program-like representation that can accommodate structure at multiple levels of abstraction.

More broadly, our work proposes and validates a general strategy for leveraging complex behavioral readouts (e.g., natural language descriptions) to draw rich and meaningful inferences about the content and structure of mental representations. Such approaches have tremendous promise not only to advance cognitive theory, but may contribute to the design of artificial systems that learn more human-like abstractions.

Acknowledgments

RDH is supported by NSF grant #1911835 and a Princeton C.V. Starr Fellowship. JEF is supported by NSF CAREER #2047191, an ONR Science of Autonomy award, and a Stanford Hoffman-Yee grant. GG is supported by an MIT Presidential Fellowship and an NSF Graduate Research Fellowship. CW, JBT, and JDA are supported by the MIT Quest for Intelligence, and CW and JBT have additional support from AFOSR Grant #FA9550-19-1-0269, the MIT-IBM Watson AI Lab, ONR Science of AI and DARPA Machine Common Sense.

All code and materials available at:
[https://github.com/cogtoolslab/
lax-cogsci22](https://github.com/cogtoolslab/lax-cogsci22)

References

- Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, 120(4), 817.
- Bear, D. M., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., . . . others (2020). Learning physical graph representations from visual scenes. *Advances in Neural Information Processing Systems*, 33.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2), 115.
- Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., . . . others (2020). Experience grounds language. *arXiv preprint arXiv:2004.10151*.
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311.
- Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287–299). Elsevier.
- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to “overinformative” referring expressions. *Psychological Review*, 127(4), 591.
- Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., . . . Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.
- Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6), 499–504.
- Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. In *The conceptual mind: New directions in the study of concepts*.
- Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6), e12845.
- Hoffman, D. D., & Richards, W. A. (1984). Parts of recognition. *Cognition*, 18(1-3), 65–96.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.
- Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others’ knowledge in reference. *Social Cognition*, 9(1), 2–24.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.
- Landau, B., & Jackendoff, R. (1993). “What” and “where” in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2).
- Markman, E. M., & Wachtel, G. F. (1988). Children’s use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2), 121–157.
- Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 200(1140), 269–294.
- McCarthy, W. P., Hawkins, R. D., Wang, H., Holdaway, C., & Fan, J. E. (2021). Learning to communicate about shared procedural abstractions. In *Proceedings of the annual meeting of the cognitive science society* (pp. 77–83).
- Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Harvard University Press.
- Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., & Su, H. (2019). Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 909–918).
- Mukherjee, K., Hawkins, R. D., & Fan, J. W. (2019). Communicating semantic part information in drawings. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2413–2419).
- Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105(7), 2745–2750.
- Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4), 441–474.
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings

- across languages support efficient communication. *The handbook of language emergence*, 87, 237.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3), 382–439.
- Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The Psychology of Learning and Motivation*, 31, 305–349.
- Sun, Z., & Firestone, C. (2021). Seeing and speaking: How verbal “description length” encodes visual complexity. *Journal of Experimental Psychology: General*, 151(1), 82–96.
- Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, 33.
- Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology*, 25(6), 983.
- Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, 113(2), 169.
- Wang, H., Polikarpova, N., & Fan, J. E. (2021). Learning part-based abstractions for visual object concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Wong, C., Ellis, K. M., Tenenbaum, J., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. In *International Conference on Machine Learning* (pp. 11193–11204).
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942.

Supplemental: Identifying concept libraries from language about object structure

S1. Part I Supplemental Details

This section contains additional details on the stimulus generation procedure in **Part I** of the main paper.

Stimulus generation

As discussed in the main paper, the dataset comprises two distinct stimulus domains (*drawings* and *towers*). Each domain is defined formally over an *initial library of base program primitives* $\mathcal{L}_{base}$, and a hierarchical *generative model* to procedurally construct graphics programs for rendering object stimuli withing four nested *subdomains* with varying higher-order part structure.

This section provides additional details on these domains. The full generative models for both domains, along with the generated stimuli, are released at the repository.

Drawings stimulus generation.

Initial primitives $\mathcal{L}_{base}$. The initial program primitives for the drawing domain consist of a CAD-like set of simple base shape primitives and matrix operations for transforming and combining these shapes into a final drawing:

- `line`: a unit-length horizontal line shape.
- `circle`: a unit-radius circle shape.
- `square`: a unit length square shape.
- `scaled_rect(w, h)`: a rectangle parameterized by width and height.
- `graphics_matrix(scale, theta, x, y)`: produces a transformation matrix parameterized by scale, rotation angle, and x and y translation angle.
- `apply_transform(shape, matrix)`: applies a transformation matrix to a shape.
- `repeat(shape, n, matrix)`: Applies the same transformation successively n times to a given shape and returns all of the n shapes.
- `connect(shape, shape)`: Connects two shapes into a single complex shape.
- Mathematical operations (eg. `plus`, `minus`, `sin`) over floating constants.

Drawings generative model. Drawings stimuli are defined over the base primitives as programs which com-

bine subdomain-specific parts according to varied program templates, parameterized by the number and size of each part. We generate stimuli by sampling randomized parameterizations over these templates to produce the 250 stimuli sampled for each subdomain. These subdomain are described below using semantic terms for their hierarchical part structure, but these terms simply correspond to templated graphics programs parameterized recursively by size and number of subparts.

We define the following four drawings subdomains, described with a high-level overview of their underlying part structure:

- `nuts and bolts`: combines parts for an *outer shape* (eg. a hexagonal nut) of varying size, and *inner perforation* of varying size (eg. a circular hole), and 1 or more *ring perforations* of varying size (eg. a ring of circular holes.)
- `vehicles`: combines parts for n *wheels*, *vehicle bases* of varying sizes and templated types, and varying numbers of parameterized *antenna* or *windows*.
- `gadgets`: combines parts for n *dials* or *buttons*, templated *gadget bases* of varying sizes, and varying numbers of *antenna*.
- `furniture`: combines parts for n *knobs*, *drawers*, templated *furniture bases* of varying sizes and furniture *feet*.

Towers stimulus generation. *Initial primitives* $\mathcal{L}_{base}$. The initial program primitives for the structures domain build on the towers planning domain from (Ellis et al., 2020), which consists of simple primitives for picking and placing colored blocks:

- `vertical_red`: a unit-length vertical red block.
- `horizontal_blue`: a unit-length vertical blue block.
- `left(n, canvas, block)`: moves a simulated cursor left n steps on a canvas and places a block.
- `right(n, canvas, block)`: moves a simulated cursor right n steps on a canvas and places a block.

Structures generative model. Structures stimuli are defined over the base primitives using groups of low-level

part abstractions: tiles, arches, and house-parts contained fixed arrangements of blocks; `rows(width)` and `pillars(height)` were parameterized functions. Different subsets of these were hierarchically combined in different ways to create each subdomain. Each subdomain was parameterized by numerical parameters (e.g. number of arches), as well as by part type parameters (e.g. wall tile).

Bridges contained up to 2 external arches, and up to 5 internal arches of a different type. Each low-level arch abstraction contained two red pillars, which could all be extended up to a maximum height. Each bridge could support a mid-level viaduct abstraction consisting of multiple rows. We also defined a `suspension(suspension-type)` function that placed red blocks directly above the pillars, of a uniform height, or that decreased or increased in height towards the center of the bridge.

Cities contained two skyscrapers placed a random distance apart. Each skyscraper was defined by a wall tile, which was `stacked(height)` vertically and optionally mirrored, and topped with a row or pyramid roof.

Houses were each topped with a pyramid roof, defined by the width of the house. The width of the house was determined by the width of the first floor, which could contain any permutation of `window`, `bricks`, `door`. Up to two additional floors contained permutations of `window`, `bricks`.

Castles were defined by a central mirrored wall of tiles, and two flanking stacks of tiles, each parameterized by a height (with `stack height < wall height`), and the wall additionally parameterized by width. The wall and stacks were topped with the same type of roof, providing there was space. Each roof could be a pyramid or dome- similar to a pyramid with an additional shorter row beneath.

We exhaustively enumerated items from each subdomain (with a small range for each parameter), rejecting any tower that extended beyond a 20x20 grid.

S2. Part II Supplemental Details

This section describes additional technical details for the library identification model used in **Part II** of the main paper. Code for both the defined hypothesis spaces and alignment model is also released at the provided repository.

Defining candidate program libraries

All stimulus from each subdomain are generated from the shared set of base program primitives $\mathcal{L}_0$ described in **S1**. However, as described in the main paper, our goal is to define candidate *program libraries* $\mathcal{L}_i$ that augment the initial base primitives with *additional program subroutines*

which encapsulate higher-level part structure specific to a given subdomain (such as the *wheel* components shared across the vehicles stimuli, or the *external arch* components shared across the bridge stimuli). In particular, we aim to construct *cumulatively defined libraries* such that each $\mathcal{L}_i$ contains all of the program concepts in library $\mathcal{L}_{i-1}$, along with new program subroutines at a higher level of complexity.

While we could define an enormous set of possible candidate libraries containing all possible program subroutines across our stimulus subdomains, we consider a restricted set of representative libraries constructed from the natural hierarchical structure used to generate the underlying stimuli, and which are designed to span the maximal range of candidate part concepts on our subdomains (from the common base primitives in $\mathcal{L}_0$ to extremely complex subroutines that correspond to entire categorical classes of stimuli on each subdomain.)

In particular, we construct a given library $\mathcal{L}_i$ to contain new primitives corresponding roughly to the 'next highest outer loop of parametric variation' over the part components in $\mathcal{L}_{i-1}$. On the nuts and bolts domain, for instance, $\mathcal{L}_0$ contains the original unit line and *unit circle* shape primitives, $\mathcal{L}_1$ contains additional polygons defined by parameterizing rotation over the unit lines, $\mathcal{L}_2$ contains additional **rings of shapes** (including rings of polygons) defined by parameterizing rotation over all of the shapes in $\mathcal{L}_1$, and $\mathcal{L}_3$ contains categorical stimulus concepts (like a outer shape with a ring of shape perforations parameterized by the particular shape and size of its components).

This procedure is designed to maximize discriminability of the candidate libraries while still corresponding to the underlying recursive program structure of each subdomain in a principled way. However, in future work we hope to look to *automated* program library construction procedures which can be used to generate a much larger space of candidate libraries, like those in (Ellis et al., 2020; Wang, Polikarpova, & Fan, 2021; Wong, Ellis, Tenenbaum, & Andreas, 2021).

Library-to-vocabulary alignment model

This provides further implementational details on the model used to produce the *library-to-vocabulary alignment* metric shown in **Figure 4** of the main paper.

This metric measures the *mean token log-likelihood* (which correlates negatively with perplexity) for each subdomain with respect to the set of *descriptions* for each object stimulus, *programs* that generate that object stimulus, and a given *library* in which those programs can be represented.

More specifically, each library $\mathcal{L}_i$ is comprised of $\rho_0^i, \rho_1^i \dots \in \mathcal{L}_i$. The graphics programs for each object stimulus are initially represented as a program π_{base} , but

they can be rewritten automatically with respect to any given library $\mathcal{L}_i$ into a semantically equivalent program π_i that maximally compresses the original program to use the encapsulated subroutines in the new library. We call the *tokenization* of a program π_i its left-order tree traversal (omitting variables), such that the tokenization of a program is an ordered sequence of program components $[\rho\pi\ldots]$.

The goal of our metric is to estimate the average alignment probabilities between *program components* in tokenized programs for each object stimulus (written in a given library) and the *words* in the tokenized *what* descriptions for that object stimulus. We use the IBM Model 1 translation model (Brown, Della Pietra, Della Pietra, & Mercer, 1993), which takes as input a corpus of paired (program token, description token) sequences for all object stimuli in a given subdomain, and estimates corpus-wide type-type translation probabilities $P(w|\rho)$ for each word type w across all descriptions, and each program type ρ across all programs written under a given library. This model also jointly estimates *alignments* between the individual program and word tokens in a (description, program) pair – it finds the MAP alignments $\alpha(w, \rho)$ such that $\Pi_{\alpha} P(w|\rho)$ is maximized for all aligned words and program components.

We report a *cross-validated* alignment metric over each subdomain using batches of $n=5$ heldout-stimuli: we randomly order all (description, program) pairs, fit the IBM model to estimate $P(w|\rho)$ based on all but the heldout stimuli, and then report our metric with respect to the MAP alignments on the heldout (description, program) pairs. More specifically, for each of these heldout (description, program) pairs, the mean token log-likelihood corresponds to the mean $\log P(w|\rho)$ where the mean is taken over the MAP alignments $\alpha(w, \rho)$.

References

- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311.
- Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., . . . Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.
- Wang, H., Polikarpova, N., & Fan, J. E. (2021). Learning part-based abstractions for visual object concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Wong, C., Ellis, K. M., Tenenbaum, J., & Andreas, J. (2021). Leveraging language to learn program abstractions and search heuristics. In *International Conference on Machine Learning* (pp. 11193–11204).