Emergence of compositional abstractions in human collaborative assembly

William P. McCarthy
Department of Cognitive Science
UC San Diego
La Jolla, CA 92093
wmccarthy@ucsd.edu

Cameron Holdaway
Department of Psychology
UC San Diego
La Jolla, CA 92093
choldawa@ucsd.edu

Robert D. Hawkins
Department of Psychology
Princeton University
Princeton, NJ 08540
rdhawkins@princeton.edu

Judith E. Fan
Department of Psychology
UC San Diego
La Jolla, CA 92093
jefan@ucsd.edu

Abstract

Many real-world tasks require agents to coordinate their behavior to achieve shared goals. Here we investigate how humans use natural language to collaboratively solve physical assembly problems more effectively over time. Human participants were paired up in an online environment to reconstruct scenes containing a pair of block towers. One participant, who could see the target towers, sent assembly instructions to the other participant, who aimed to reconstruct them as accurately as possible. We found that participants provided increasingly concise instructions across repeated attempts on each pair of towers, reflecting the use of more abstract referring expressions that captured the hierarchical structure of each scene (i.e., tower-level expressions subsuming block-level ones). Moreover, our data suggest that different pairs of participants converged on different expressions, suggesting that multiple viable solutions exist for mapping tokens of natural language to object configurations. Taken together, our paper presents an empirical paradigm, human dataset, and set of evaluation metrics that can be used to guide the development of artificial agents that emulate human-like compositionality and abstraction.

1 Introduction

From advanced manufacturing to food preparation, many real-world tasks require multiple agents to coordinate their behavior. To coordinate effectively, collaborators benefit from sharing a common representation of the relevant objects in their environment, specified at the appropriate level of abstraction for their goals. In many cases, shared representations are not supplied to agents in advance, and thus require *ad hoc* coordination between agents as they each learn about the structure of the task [10, 23, 28].

A powerful solution to the problem of coordinating representations is the ability to communicate using natural language [24, 18, 26, 25]. Yet for communication protocols to be effective in novel task settings, these protocols must also be able to update over the course of an interaction, a phenomenon that has been explored in both psycholinguistics [6, 11] and natural language processing [12]. How do intelligent, autonomous agents simultaneously coordinate on shared object representations and language for talking about them at the appropriate level of abstraction?

NeurIPS 2020 Workshop on Object Representations for Learning and Reasoning, Vancouver, Canada.

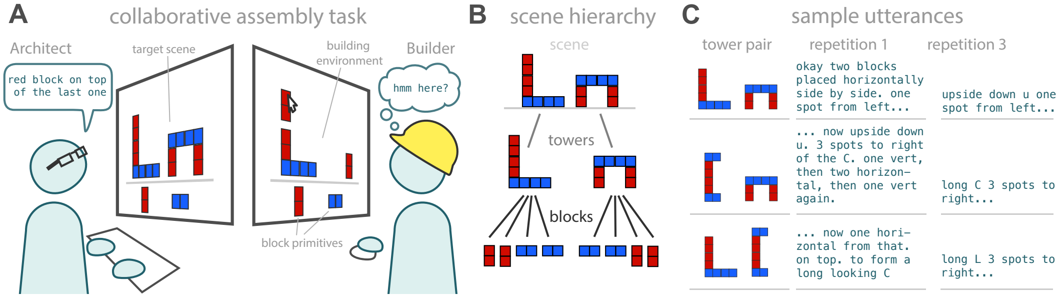
A
collaborative assembly task
Architect
target scene
building environment
Builder
red block on top of the last one
hmm here?
block primitives
B
scene hierarchy
scene
towers
blocks
C
sample utterances
tower pair
repetition 1
repetition 3
okay two blocks placed horizontally side by side. one spot from left...
upside down u one spot from left...
... now upside down u. 3 spots to right of the C. one vert, then two horizon– tal, then one vert again.
long C 3 spots to right...
... now one hori– zontal from that. on top. to form a long looking C
long L 3 spots to right...

Figure 1: Collaborative assembly task. (A) The Architect was shown a target scene and provided assembly instructions to the Builder, who aimed to reconstruct it. (B) Each scene was composed of two towers, which were each composed of four domino-shaped blocks. (C) Example messages from first and final repetitions of a tower pair, showing the emergence of expressions referring to towers.

In this paper, we approach this question by examining how humans coordinate their behavior in a physical assembly domain in which objects are hierarchically organized, and can thus be specified at different levels of abstraction. We explore the hypothesis that humans exploit shared expectations about the hierarchical organization of objects to develop more abstract referring expressions that reflect this structure. Overall, our paper presents an empirical paradigm, human dataset, and set of evaluation metrics that can be used to guide ongoing development of artificial agents that emulate human-like compositionality and abstraction.

2 Related Work

Our paper builds on prior work in both human and computer vision that has investigated how agents form hierarchical representations for objects [8, 1, 14, 20] and scenes [7, 13, 9, 5, 27, 3]. Such object-centric representations are especially valuable because of their ability to support high-level visual reasoning and planning, including the ability to compose shape primitives to form more complex objects during physical assembly [2, 19]. We leverage insights from this work to investigate how human collaborators coordinate their object representations via social interaction.

We also draw upon work in cognitive science and natural language processing that has used cooperative language games to investigate the emergence of novel linguistic conventions, and in particular how agents learn to produce informative and concise referring expressions for objects over time [6, 12, 17]. A key theme in this literature concerns the importance of compositionality in emergent communication protocols [22, 15, 21], specifically the ability to recombine language from different contexts to formulate new meanings [16]. Compositionality may be especially important in domains where the space of possible meanings is highly structured yet large, as in the case of providing instructions to assemble towers from blocks [29, 30]. Our study departs from this prior work by emphasizing how agents develop linguistic conventions for objects defined at higher levels of visual abstraction over time as they acquire more evidence about the structure of these objects and their collaborator's behavior.

3 Task

We recruited 98 human participants (N=49 dyads) from Amazon Mechanical Turk and automatically paired them up to perform a collaborative assembly task (Fig. 1A). At the outset, each participant was assigned the role of *Architect* or *Builder* and proceeded with their partner through a series of twelve trials. At the start of each trial, the Architect was presented with a target scene containing block towers. The Builder could not see the target scene, and was instead presented with an empty grid world environment in which they could place blocks. To coordinate, the Architect sent step-by-step assembly instructions, which the Builder used to reconstruct the target scene as accurately as possible.

Each scene was composed hierarchically from two block towers that appeared side by side; in turn, each tower consisted of four domino-shaped blocks– two vertical and two horizontal (Fig. 1B). To evaluate changes in behavior, we employed a *repeated* design where each tower appeared multiple times. There were three unique towers. All three pairs of these towers appeared once in each of four repetition blocks in a randomized sequence, for a total of twelve trials.All towers appeared in both the left and right positions an equal number of times, such that there was no statistical association between a given tower and its location, nor the tower it was paired with.

The Architect and Builder took as many turns as they needed to reconstruct each scene. On the Architect's turn, they sent a single message containing a maximum of 100 characters; on the Builder's turn, they placed one or more blocks before awaiting further instructions (Fig. 1C). Blocks could be placed anywhere so long as they were supported from beneath, and could not be moved once placed. The Architect could see the placement of each block in real time but the communication channel was otherwise unidirectional: the Builder was unable to send messages back to the Architect. Once all eight blocks had been placed, both participants received feedback about the mismatch between the target scene and reconstruction before advancing to the next trial.

4 Results

Although each interaction only spanned twelve trials, we hypothesized that human dyads would be able to leverage this small amount of experience to rapidly develop shared task representations, manifesting in increasingly successful and efficient collaboration over time.

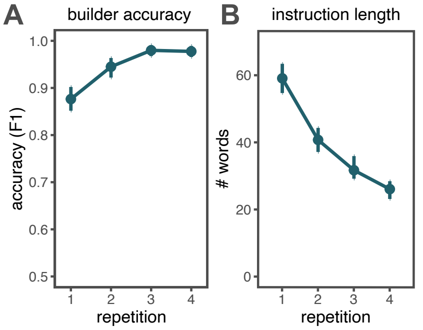

Figure 2: (A) Reconstruction accuracy improved across repetitions. (B) Mean number of words used on each trial decreased across repetitions.

Success across repetitions Given that the focus of our study was on how language produced by Architects changed over time, we sought to first verify that human dyads were able to successfully perform the assembly task. We found that even on their initial reconstructions, they were highly accurate (mean $F_1 = 0.876$; 95% CI:[0.854, 0.898]), which roughly corresponds to having just one block out of place. Even so, we found that dyads reliably improved across repetitions ($b = 3.38$, $t = 7.90$, $p < 0.001$; Fig. 2A), the magnitude of which we estimated using a linear mixed-effects (LME) model that predicted accuracy from repetition number and included random intercepts for each dyad.

Communicative efficiency across repetitions Given that the same towers recurred throughout the interaction, we hypothesized that Architects would exploit these regularities to provide more concise instructions over time. To test this hypothesis, we analyzed both changes in the total number of words used and how many messages were sent within a trial. We estimated changes using LME models containing repetition number as a predictor, as well as random intercepts and slopes for each dyad and random intercepts for each tower pair. Consistent with our hypothesis, we found that Architects sent messages containing fewer words over time ($b = -10.8$, $t = -10.9$, $p < 0.001$) (Fig. 2B), which were themselves contained in fewer messages within each trial ($b = -0.67$, $t = -8.01$, $p < 0.001$).

Changes in words used across repetitions What explains these gains in communicative efficiency? One possibility is that Architects increasingly omitted unnecessary, non-referential function words; another is that they changed which words they used to refer to objects. To distinguish these possibilities, we compared changes in the frequency of words used in the first and final repetitions. To ensure that our analyses reflected changes in the referring expression used to refer to components of each scene rather than in the use of function words, we recruited two human annotators who were blind to the source of each utterance to manually extract referring expressions from each message[1]. For each dyad, we compared the word frequency distributions between the first and final repetitions using a permutation-based χ^2 test [4], which revealed a reliable difference between the two distributions ($p < 0.001$, Bonferroni corrected for multiple comparisons). To identify the words contributing most to this shift, we calculated the overall change in proportion from the first repetition to the final repetition. We found that words such as "block" and "horizontal" were used less often while "shape," and "C" were used more often (Fig. 3A). These results suggest that increasingly concise instructions reflect shifts in *referential* words.

[1] Two dyads were excluded from this analysis because our annotators were unable to recover referring expressions from their language.

More abstract referring expressions across repetitions A natural explanation for the shift in the words were used is that Architects had learned to produce referring expressions at a higher level of abstraction, in particular ones that corresponded to entire towers rather than individual blocks. To evaluate this possibility, the same human annotators additionally tagged each referring expression with the number of references to block-level and tower-level entities they contained. Unsurprisingly, given that there were eight blocks in each scene and only two towers, we found that the number of references to blocks was greater overall than those made to towers ($b = -7.41$, $t(2344) = -20.98$, $p < 0.001$). More importantly, we found that these proportions shifted across repetitions ($b = 1.35$, $t(2344) = 10.49$, $p < 0.001$; interaction between repetition number and reference type), reflecting both an increase in the number of tower-level references (e.g. "C shape," "L shape") and corresponding decrease in the number of block-level references (e.g. "horizontal blue block," "vertical red block"; Fig. 3B).

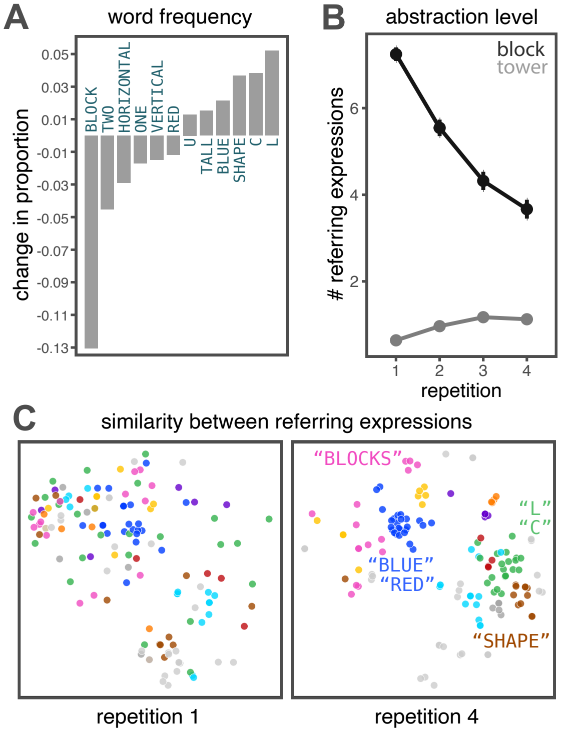

Figure 3: (A) Words with largest positive and negative changes in frequency between first and final repetitions. (B) Change in number of block-level and tower-level references across repetitions. (C) t-SNE visualization of similarity between messages from different dyads in the first and final repetitions.

Consistency and variability in referring expressions across dyads The overall increase in tokens resembling entire towers ("C" and "L" shapes) in the final repetition suggests some degree of consistency between dyads, with respect to the tower-level abstractions that emerged. To what extent did different dyads converge on the same set of labels for each tower, rather than settle on distinct, but internally consistent ways of referring to them? To explore this question, we estimated how dissimilar the language used by different dyads was within each repetition, by computing the Jensen-Shannon divergence (JSD) between their word frequency distributions, aggregating language from all trials in a repetition block. We found that the mean pairwise JSD increased significantly between the first and final repetitions ($d = 0.080$, 95% CI:$[0.041, 0.118]$, $p = 0.004$), consistent with divergence between dyads. We visualized these distances using a t-SNE embedding of word count vectors (Fig. 3C), revealing that this divergence might be attributed to the formation of distinct "clusters" (denoted by different colors shown with representative words; gray dots belong to degenerate clusters with < 4 members). Together, these findings suggest that even in this relatively simple task domain, human dyads manage to discover a diverse array of solutions for mapping tokens of natural language to components of each scene.

5 Discussion

This paper investigated how humans efficiently collaborate in a physical assembly task by developing shared abstractions for connecting language with object representations. In future work, we plan to further investigate the sources of consistency and variability in the communication protocols that emerge during collaboration, as well as constraints on generalization of these protocols to novel tasks and collaboration partners. Moreover, to probe the computational mechanisms that enable effective coordination, we plan to evaluate how well different algorithmic approaches emulate human behavior in both Architect and Builder roles (e.g., program synthesis, reinforcement learning, *seq2seq*, etc.). In the long term, such studies may shed light on the inductive biases that enable rapid coordination upon shared procedural abstractions during social interaction between intelligent, autonomous agents.

References

[1] Joseph L Austerweil and Thomas L Griffiths. A nonparametric bayesian framework for constructing flexible feature representations. *Psychological review*, 120(4):817, 2013.

[2] Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly L Stachenfeld, Pushmeet Kohli, Peter W Battaglia, and Jessica B Hamrick. Structured agents for physical construction. *arXiv preprint arXiv:1904.03177*, 2019.

[3] Daniel M Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B Tenenbaum, et al. Learning physical graph representations from visual scenes. *arXiv preprint arXiv:2006.12373*, 2020.

[4] Eric J Beh and Rosaria Lombardo. *Correspondence analysis: theory, practice and new strategies*. John Wiley & Sons, 2014.

[5] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[6] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[7] József Fiser and Richard N Aslin. Encoding multielement scenes: statistical learning of visual feature hierarchies. *Journal of Experimental Psychology: General*, 134(4):521, 2005.

[8] Robert L Goldstone. Learning to perceive while perceiving to learn. *Perceptual organization in vision: Behavioral and neural perspectives*, 233278, 2003.

[9] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019.

[10] Barbara Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 1996.

[11] Robert D Hawkins, Michael C Frank, and Noah D Goodman. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845, 2020.

[12] Robert D Hawkins, Minae Kwon, Dorsa Sadigh, and Noah D Goodman. Continual adaptation for efficient machine communication. *arXiv preprint arXiv:1911.09896*, 2019.

[13] John M Henderson and Andrew Hollingworth. High-level scene perception. *Annual review of psychology*, 50(1):243–271, 1999.

[14] R Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy J Mitra, and Daniel Ritchie. Shapeassembly: Learning to generate programs for 3d shape structure synthesis. *arXiv preprint arXiv:2009.08026*, 2020.

[15] Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

[16] Brenden M Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

[17] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.

[18] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*, 2019.

[19] W. McCarthy, D. Kirsh, and J. Fan. Learning to build physical structures better over time. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 2020.

[20] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019.

[21] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

[22] Martin A Nowak, Joshua B Plotkin, and Vincent AA Jansen. The evolution of syntactic communication. *Nature*, 404(6777):495–498, 2000.

[23] Peter Stone, Gal A Kaminka, Sarit Kraus, Jeffrey S Rosenschein, et al. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI*, page 6, 2010.

[24] Alane Suhr, Claudia Yan, Jacob Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. *arXiv preprint arXiv:1910.03655*, 2019.

[25] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[26] Stefanie Tellexll, Pratiksha Thakerll, Robin Deitsl, Dimitar Simeonovl, Thomas Kollar, and Nicholas Royl. Toward information theoretic human-robot dialog. *Robotics*, page 409, 2013.

[27] Rishi Veerapaneni, John D Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based reinforcement learning. In *Conference on Robot Learning*, pages 1439–1456. PMLR, 2020.

[28] Rose E Wang, Sarah A Wu, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. Too many cooks: Coordinating multi-agent collaboration through inverse planning. *arXiv preprint arXiv:2003.11778*, 2020.

[29] Sida I Wang, Percy Liang, and Christopher D Manning. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*, 2016.

[30] Qi Zhang, Richard Lewis, Satinder Singh, and Edmund Durfee. Learning to communicate and solve visual blocks-world tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5781–5788, 2019.