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

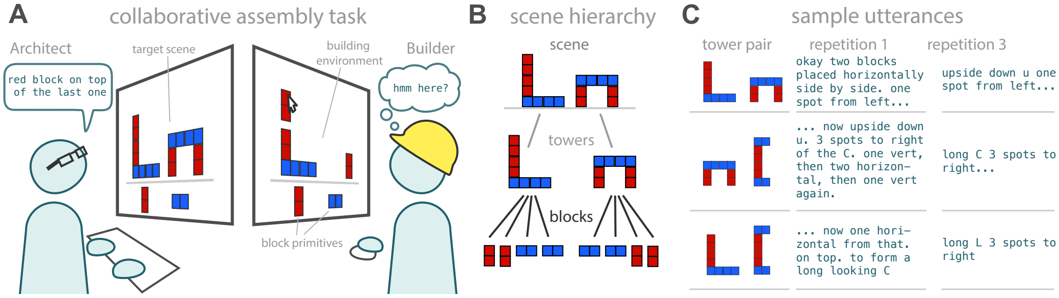
Many real-world tasks require agents to coordinate their behavior to achieve shared goals. Here we investigate how humans use natural language to collaboratively solve physical assembly problems more effectively over time. Human participants were paired up in an online environment to reconstruct scenes containing a pair of block towers. Each participant was assigned either the role of Architect or Builder: the Architect provided assembly instructions to the Builder, who aimed to reconstruct each scene as accurately as possible. We found that Architects provided increasingly concise instructions to Builders 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 succeed. In order to coordinate effectively, collaborators must share a common underlying representation of their task and goals, including basic representations of objects and actions in the environment. Often, shared representations are not supplied to agents in advance, or must be re-negotiated on the fly as each individual agent acquires new expertise about the task structure. In other words, these tasks demand sustained *ad hoc* coordination [10, 22, 27].

One promising solution to the problem of *ad hoc* coordination is the ability to explicitly communicate using natural language [23, 17, 25, 24]. Yet, communication protocols also require some degree of coordination and adaptation over the course of interaction, as emphasized in psycholinguistics [6, 11] and natural language processing [12]. How can agents simultaneously coordinate their underlying representations of objects and the way they talk about them?

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



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 horizontal, then one vert again.	long C 3 spots to right...
	... now one horizontal 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 these questions by examining how human partners successfully manage to coordinate in a physical assembly domain requiring *ad hoc* adaptation for both object representations and language. In particular, we explore the notion that humans exploit shared expectations about the hierarchical organization of physical objects to develop more abstract referring expressions that reflect this hierarchical structure. 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.

2 Related Work

Our paper builds on prior work in cognitive science and natural language processing that has used cooperative language games to investigate the emergence of shared task representations, or conventions. This work has examined how words acquire the ability to refer to objects through repeated multi-agent interactions [6, 12, 16]. A key theme in this literature concerns the importance of compositionality in emergent communication protocols [21, 15, 20]. Rather than expressing each intended meaning with a distinct word, agents may produce multi-word utterances that derive their meanings from their component parts. Compositionality may be especially important in domains where the space of possible meanings is highly structured yet variable, as in the case of providing instructions to assemble towers from blocks [28, 29]. Our study departs from this prior work by emphasizing how agents may learn to compose referring expressions at higher levels of abstraction over time as they acquire more evidence about the structure of objects and their partner's behavior.

Our paper also draws upon a large literature in both human and computer vision that has investigated constraints on the learning of hierarchical representations for objects [8, 1, 14, 19] and scenes [7, 13, 9, 5, 26, 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, 18]. Our study leverages insights from this prior work to investigate how human collaborators develop shared object-level abstractions through social interaction.

3 Task

Human participants (N=98 participants; 48 dyads) were recruited from Amazon Mechanical Turk and automatically paired up an interactive web environment to perform a collaborative assembly task (Fig. 1A). At the outset, each participant was assigned either the role of *Architect* or *Builder* and proceeded with their partner through a series of twelve target scenes containing block towers. Critically, only the Architect was shown the target scene, and only the Builder was able to place blocks in the environment. To succeed, the Architect needed to send step-by-step assembly instructions in natural language, which the Builder used to reconstruct the target scene as accurately as possible. At the beginning of each trial, the Architect was presented with a target scene and the Builder with an empty environment. They then took turns: On the Architect's turn, they could send a single set of instructions containing a maximum of 100 characters. On the Builder's turn, they could place one or more blocks in the environment before ending their turn. Blocks could be placed anywhere so long as they were supported from beneath, but could not be moved once placed. The Builder was unable to send messages to the Architect, although the Architect could see the placement of each block in real-time. Once eight blocks had been placed, both participants received feedback about the mismatch between the target scene and reconstruction before advancing to the next trial.

Each scene was composed of two block towers that appeared side by side, consisting of two horizontal and two vertical domino-shaped blocks (Fig. 1B). There were three unique towers, and each unique combination of these towers appeared four times across a series of four repetition blocks, in which each combination appeared exactly once in a randomized order. All towers appeared in both the left and right positions an equal number of times, such that there was no statistical association between the appearance of a given tower and its location within the scene nor the tower it was paired with.

4 Results

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

4.1 Successful collaborative assembly throughout interaction

Given that the focus of our study was on changes in the language produced by Architects to achieve accurate reconstructions, 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, with an average F1 score of 0.876 (95% bootstrapped CI:[0.854, 0.898]), roughly equivalent 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 model that predicted accuracy from repetition number and included random intercepts for each dyad.

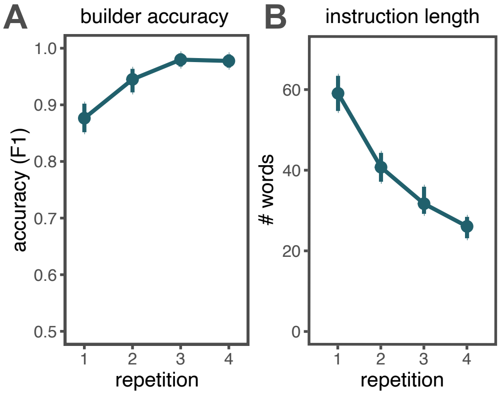


Figure 2: (A) Change in reconstruction accuracy across repetitions. B: Change in mean number of words used on each trial across repetitions.

4.2 Greater 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 linear mixed-effects models containing repetition number as a predictor, as well as random intercepts and slopes for different dyads and random intercepts for different tower combinations. 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 ($b = -0.67$, $t = -8.01$, $p < 0.001$).

4.3 Changes in words used across repetitions

What explains these changes in communicative efficiency? One possibility is that Architects increasingly omitted unnecessary words; another is that they changed which words they used. To distinguish these two 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 function words), we recruited two human annotators who were blind to the source of each utterance to manually extract referring expressions from each message. Separately for each dyad[1], we compared the word frequency distribution from the first repetition to those from the final repetition 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; Fig. 3B). These results suggest that the increasingly concise instructions at least partially reflect shifts in *which* words were used, and not only the omission of unnecessary words.

4.4 More abstract referring expressions across repetitions

A natural explanation for the shift in which words were used is that Architects had learned to produce referring expressions at a higher level of abstraction, and 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. Unsurprisingly, we found that the number of references to blocks was greater overall than to towers ($b = -7.41$, $t(2344) = -20.98$, $p < 0.001$), given that there were eight blocks in each scene and only two towers. More importantly, we found that this gap shrunk 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 and decrease in the number of block-level references (Fig. 3B).

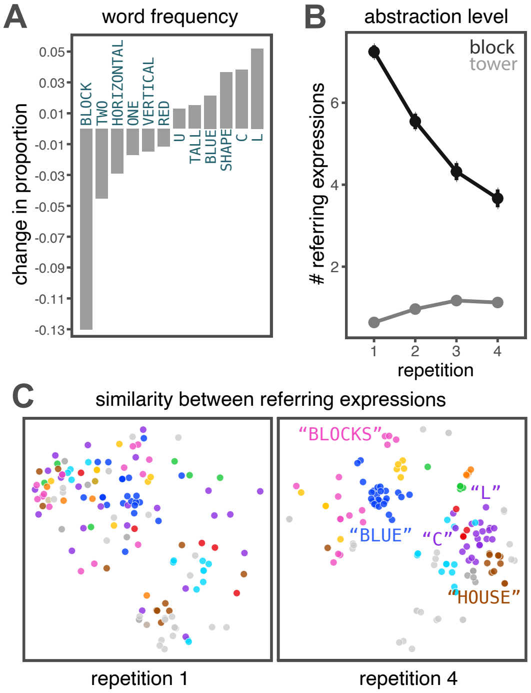


Figure 3: (A) Words with largest positive and negative changes in frequency between first and final repetitions. (B) Change in number of block-level and tower-level references across repetitions. (C) t-SNE visualization of similarity between messages from different dyads in the first and final repetitions.

4.5 Consistency and variability in referring expressions across dyads

Given the overall increase in references to "C" and "L" in the final repetition, which resemble two of the towers, the results so far suggest at least some degree of consistency between dyads with respect to the tower-level abstractions that emerged. How strong was this convergence upon a common set of labels across dyads? 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 the word frequency distributions from each pair of dyads. We noticed that although many dyads appeared to use similar referring expressions (Fig.3C), the mean JSD increased between the first and final repetitions (0.080, 95% CI:$[0.041, 0.118]$, $p = 0.004$). Taken together, these observations hint that different subsets of our sample discovered distinct solutions for mapping tokens of natural language to components of each scene — a possibility which we are exploring in ongoing work.

5 Discussion

In this paper we have shown how human dyads use compositional abstractions, expressed in natural language, to make collaboration more concise and accurate in a repeated physical assembly task. In future work, we plan to further analyze the content and structure of these linguistic conventions (e.g.,

[1] Two dyads whose language was too sparse to be represented in a contingency table were excluded from this analysis.

emergence of unique tokens for towers and scenes); analyze the emergence of these tokens in more complex, compositional scenes; and develop autonomous artificial agents who can emulate human behavior in the Architect and Builder roles. In the long term, such studies may shed light on how goal-relevant abstractions emerge from 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] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.