Function shapes form: Compositionality emerges from communicative needs, not environmental structure alone

Jess Mankewitz (mankewitz@wisc.edu)
Department of Psychology, University of Wisconsin-Madison

Robert D. Hawkins (rxdh@stanford.edu)
Department of Linguistics, Stanford University

Abstract

Human languages are compositional, combining smaller units of meaning to express more complex ideas. To explain the emergence of compositionality, researchers have appealed to functional pressures from communication. However, languages may merely inherit the component structure found in the environment. We designed a reference game to explicitly disentangle these possibilities; pairs of participants ($N = 450$) communicated about about sets of shapes that were assembled from component parts. Critically, we manipulated whether shapes that shared the same parts were competitors *within* each trial or were distributed across different trials. We found that participants successfully developed efficient conventions for referring to the shapes. However, participants who needed to distinguish shapes that shared components within the same context were more likely to develop compositional systems. When shared components appeared in separate contexts, participants favored non-compositional conventions. These results suggest compositional language structure most readily emerges from immediate communicative pressures rather than environmental structure alone.

Keywords: Compositionality; Language Evolution; Communication; Convention Formation

Introduction

Human languages have a remarkable property: they allow us to express an infinite number of ideas using a finite set of building blocks. This feature, known as compositionality, is a key driver of linguistic creativity (Hockett & Hockett, 1960; Townsend et al., 2018; Szabó, 2024). For instance, the words "house" and "boat" can combine differently to mean either a house for boats (boat house) or a boat used as a house (house boat). While languages have this capacity for compositionality, they show variation in how much they actually use it in practice, with speakers often using only a sparse subset of possible combinations (Lupyan & Dale, 2010; Sathe et al., 2023). For a communication system to be successful, both the primitive parts and the rules for how they should be combined need to be shared and conventionalized by the linguistic community (Goldberg, 2015).

The factors that determine when and how speakers use compositional structure in language remain contested. One influential explanation has pointed to the trade-off between simplicity (due to learnability pressures) and expressivity (due to communication pressures; Kirby et al., 2008, 2015). According to this view, languages must be simple enough to be reliably learned and transmitted across generations, while still being expressive enough to convey the distinctions that matter to speakers. These competing pressures have been used to explain many properties of language, from word order (Zipf, 1949) to efficient semantic categories like color words (Regier et al., 2015; Kemp & Regier, 2012), in addition to its compositional structure.

In an alternative view, compositionality may emerge from communicative pressures alone, without requiring the learnability pressures that arise from transmission. For example, Raviv et al. (2019) used a communicative reference game to demonstrate that compositional structure can emerge within a single generation. The key pressure for compositional structure came from expressivity and *generalization* – participants were more likely to create compositional systems when they needed to communicate about an expanding space of meanings. Similarly, Nölle et al. (2018) found that participants developed systematic gesture systems when communicating about open-ended rather than fixed sets of referents.

These studies highlight how the structure of the referential space itself can shape the emergence of compositional conventions. In both Nölle et al. (2018) and Raviv et al. (2019), the open, expanding nature of the referent space was shown to be critical for engendering systematic structure in the conventions. Yet many of these studies leave open another possibility: the tendency to form compositional meanings may be "baked in" from the compositional structure of the environment. Objects have composable properties like colors, shapes, and materials; events have agents, actions, and recipients. It is possible that linguistic systems may automatically mirror whatever underlying regularities are provided by the environment, regardless of specific communicative demands. This view is further supported by studies showing that languages adapt to encode features that are most relevant for disambiguation (Winters et al., 2015; Perfors & Navarro, 2014). However, these studies typically manipulate both the overall structure in the environment and the communicative context simultaneously, making it difficult to isolate which factor drives the emergence of compositional structure.

These possibilities make different predictions about when and how compositional structure should emerge. If communicative pressures are key, compositional structure should emerge most readily when the immediate communicative context requires distinguishing between referents that share components. If the mere existence of structure in the environment is sufficient, language users should spontaneously cre-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

ate compositional descriptions whenever referents have clear componential structure regardless of whether they appear as referential competitors in the same contexts.

Methods

Do speakers utilize compositionality simply because they are mirroring compositional structure in their environment, or does it arise specifically when communication demands it? Reference games provide an ideal testbed for addressing this question, as they allow us to control both the underlying statistical structure available in the global environment and the specific alternative referents in the immediate context. Using a large-scale reference game paradigm, we systematically manipulated two aspects of structure in the environment: whether the referent shapes are constructed by recombining shared components, and whether these components need to be distinguished from each other in the immediate communicative context. This study was preregistered on the Open Science Framework and pre-registration, data, and analysis sripts can be found at: https://osf.io/3t7fm/.

Stimuli

To manipulate the presence or absence of compositional structure in the environment, we constructed sets of ambiguous shape stimuli from primitive subshapes. We selected 471 unique tangrams from the KiloGram database (Ji et al., 2022) that contained at least one flat edge to serve as our subshapes. These tangrams were rotated to align their edges, resized to a 2:1 aspect ratio, and vertically stacked to generate a bank of 221,841 shapes that could be used in this experiment. We systematically manipulated the presence of shared components across two kinds of shape sets. For compositional sets, 16 shapes were generated by maximally pairing 4 top tangrams with 4 bottom tangrams. In contrast, shapes in the non-compositional sets were generated by pairing 16 unique top tangrams with 16 unique bottom tangrams, ensuring no subcomponent was repeated within a set (see Figure 1 for examples).

When the tangrams were rotated and resized, some initially distinct shapes could become visually similar. This posed two risks: the non-compositional sets might contain repeated substructures if their components were too similar, while the compositional sets may become so similar that the shapes are difficult to discriminate. To ensure that these novel shapes were sufficiently distinct, we computed pairwise similarities using embeddings from the CLIP model fine-tuned on classifications from KiloGram (Radford et al., 2021; Ji et al., 2022). We generated 1,000 candidate sets of each type and retained the 500 sets with the lowest maximum pairwise similarity scores. The final compositional sets had maximum pairwise similarities ranging from 0.465 to 0.765 ($mean = 0.604$), while the non-compositional sets ranged from 0.44 to 0.6 ($mean = 0.560$). This filtering ensured non-compositional sets contained distinct subcomponents, while the shared subcomponents in compositional sets remained discriminable.

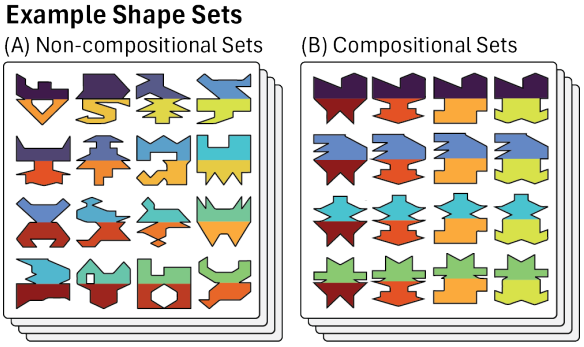

Figure 1: Two examples of the shape sets used in the experiment. (A) In non-compositional sets, 16 top components (shown in cool tones) were uniquely paired with 16 bottom components (shown in warm tones), ensuring no subcomponent appeared more than once. (B) In compositional sets, four top components were systematically combined with four bottom components to create 16 unique shapes. Colors are used here to highlight the subcomponents; in the experiment, all shapes were presented in black without color distinction.

Participants

Native English speakers from the US, UK or Canada were recruited from the online platform Prolific (www.prolific.com). Participants received a base payment of $8.25 plus a $0.03 bonus for each correct response. For quality control, participants were excluded from analysis (but still compensated) if they were missing more than 32 (50%) trials ($n = 52$) or had an accuracy rate below 75% ($n = 41$). Two additional dyads were excluded because their messages displayed clear markers of AI-generated text (e.g., unusually long, instructional descriptions like "The black box surrounding it suggests it is a target object…Let me know if you need more insights!"). The final sample includes $N = 450$ dyads distributed evenly across three experimental conditions.

Procedure

Participants were then paired into dyads to play an iterated reference game developed in the experiment developer platform Empirica (Almaatouq et al., 2021). On each trial, both participants viewed an array of 4 abstract shapes on their screen and could freely type to one another in a chat box. One participant (the director) saw one shape marked as the target with a black box and was instructed to describe the shape to their partner; their partner (the matcher) selected one of the four shapes. Both participants received immediate feedback about whether the correct shape was selected. Each dyad completed 4 blocks of 16 trials (64 trials total), with director/matcher roles alternating between trials. To examine how compositional language emerges under different environmental contexts and communicative pressures, we manipulated how shapes shared components within- and across-trials:

- **Non-Compositional Baseline:** In the baseline condition,

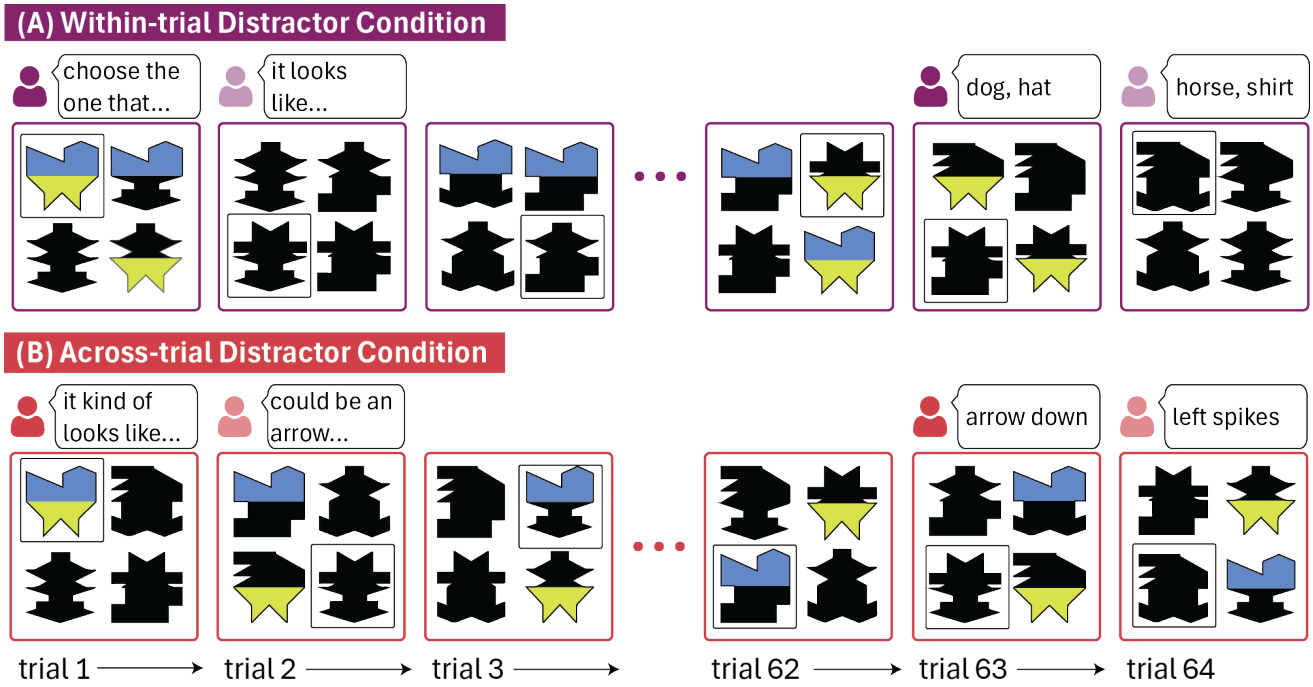

Figure 2: Example trials from the within-trial and across-trial competitor conditions. Each row shows six trials from the same dyad. (A) In the within-trial competitor condition, shapes within each display share components. (B) In the across-trial competitor condition, shapes within a single display are non-overlapping, but the same components repeat across different trials.

shapes were constructed from 32 unique components (16 tops, 16 bottoms), ensuring no components were shared across any shapes. On each trial, one shape was chosen as the target, with three competitors randomly chosen from the remaining shapes.

- **Within-trial Competitor:** Shapes were constructed by combining 4 top components with 4 bottom components to create 16 unique shapes. On each trial, after randomly selecting a target shape, one additional top component and one additional bottom component were randomly selected to construct three competitors. This ensured that shapes within each tableau systematically shared components (see Figure 2a)

- **Across-trial Competitor:** Like the within-trial competitor condition, each dyads' shapes were constructed from 4 top and 4 bottom components. After randomly selecting a target, the three competitors were randomly chosen to ensure no components were shared within the tableau. While these subcomponents repeated across different trials, and with different paired subcomponents, they never appeared together in the same display (see Figure 2b).

Across all conditions, each unique shape appeared exactly once as a target per block. In the compositional conditions, this design ensures that all subcomponents appear in the target shapes with equal frequency. This design also maintains equal exposure to the whole unique shapes across all conditions, which allows us to isolate the effects of communicative context from differences in component frequency. To control for potential biases for describing shapes in different orientations, all shapes in a given dyad's shape set were randomly rotated by a fixed amount (0°, 90°, 180°, 270°).

Text Preprocessing

Messages were preprocessed to include only referring expressions produced by directors. Multi-message descriptions within the same trial were combined into single expressions, and non-referential chat was removed through manual annotation by the first author. All messages were cleaned by removing punctuation, trailing white-space, and converting to lowercase. The processed dataset contains 28,287 unique referring expressions. For each expression, we extracted vector-space sentence embeddings using Sentence-BERT (Reimers & Gurevych, 2019). These embeddings reflect the semantic similarity between any pair of expressions using cosine distance. This approach captures meaningful relationships between descriptions that use different specific words to express similar concepts (e.g., "spiky" vs "pointy"), making it well-suited for analyzing the emergence of shared referring expressions. Recent work analyzing convention formation in reference games used similar embedding-based measures to track the development of shared descriptions (Boyce et al., 2024; Hawkins et al., 2020).

Results

We analyzed participants' behavior across the three conditions to examine (1) whether they were able to successfully coordinate to form efficient conventions in the communication game and (2) how the communicative contexts influenced the structural properties of these conventions.

Dyads form stable, efficient conventions

We first verified that participants could successfully coordinate on referring expressions across all three conditions, replicating key signatures such as an increase in accuracy and a reduction in the length of referring expression found in previous iterated reference game experiments (Hawkins et al., 2020; Boyce et al., 2024; Clark & Wilkes-Gibbs, 1986).

Accuracy A key indicator of successful convention formation is participants' increasing accuracy when identifying the correct target shape. We fit a mixed-effects logistic regression predicting correct responses from condition, trial number (scaled), and their interaction, including random intercepts for both dyad and the target shape. This analysis revealed that participants in all conditions became significantly more accurate over time ($\beta = 1.45$, $SE = .08$, $p < .001$, see Fig. 3a). While the conditions differed in their initial accuracy, all conditions reached near-ceiling performance by the final block.

Description length Another hallmark of convention formation is the reduction of referring expression length over time. We modeled word counts using a Poisson mixed-effects regression with condition, trial number (modeled with polynomials to capture non-linear trends), and their interaction as fixed effects, plus by-dyad random intercepts and slopes for trial number. The results show a significant reduction in description length over time ($\beta = -41.8$, $SE = 2.25$, $p < .001$; Fig. 3b). Conditions differed in their overall verbosity – descriptions in the within-trial condition tended to be longer, while descriptions in the across-trial condition tended to be shorter – but all conditions showed significant increases in efficiency.

Description stability To confirm that these increasingly accurate and efficient labels are actually conventional, we examined how descriptions for the same shape evolved across repeated references. We measured the stability of labels from block to block by extracting SBERT embeddings from descriptions and computing cosine similarities for the same shape across consecutive blocks. We modeled these similarities using a linear mixed-effects regression with condition, block transition, and their interaction as fixed effects, plus random intercepts for dyad (the maximal random effects structure that converged given the limited number of transitions per dyad).

The analysis revealed increasing similarity across blocks in all conditions, with descriptions becoming more similar between blocks 2-3 ($\beta = 0.078$, $SE = 0.007$, $p < .001$) and blocks 3-4 ($\beta = 0.140$, $SE = 0.007$, $p < .001$; Fig. 4). Notably, conventionalization was stronger in both compositional conditions compared to the baseline (block 3-4 increases: Within $\beta = 0.041$, Across $\beta = 0.079$, both $p < .001$). This suggests that the presence of shared visual components supported more stable convention formation, regardless of how these components were distributed in the communicative context.

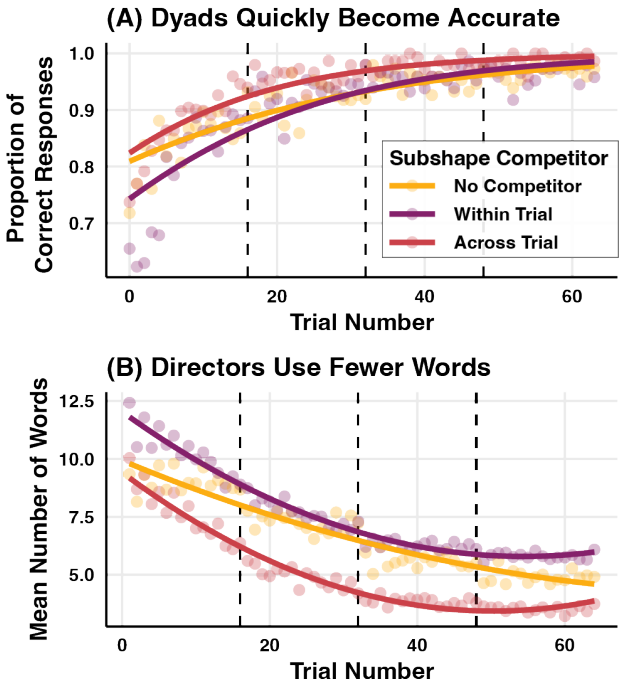

Figure 3: (A) Proportion of correct target selections across rounds by condition. Solid lines show logistic model fits. Vertical dashed lines indicate block boundaries. Participants in all conditions improved over time, with the across-trial condition. (B) Mean number of words in the directors referring expression over the course of the game.

Emergence of Compositional Structure

Having established that participants successfully formed conventions across all conditions, we next examined how these conventions differed in their structural properties. Our key question is whether compositional language structure emerges simply to mirror the presence of shared visual components in the environment, or whether it more specific communicative pressures. To address this question, we conducted a series of analyses examining how participants structured their descriptions across our three conditions.

Shared structure across descriptions We are first interested in how expressions are structured overall. To measure structure, we analyzed the similarity between descriptions for the different shapes within each round. If the descriptions are compositionally structured, systematically reusing labels for parts across multiple objects, we would expect to see higher

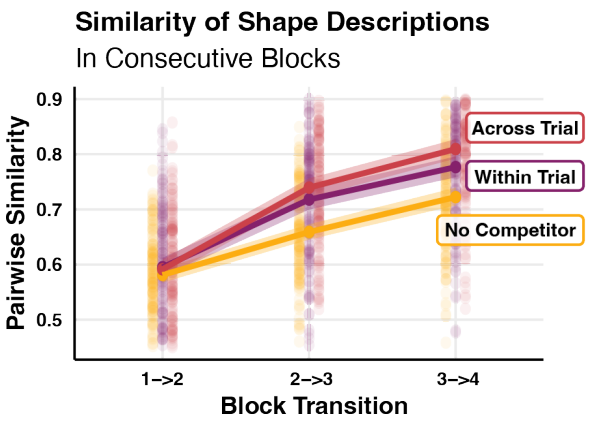

Figure 4: Average similarity between descriptions for the same shape across consecutive blocks (1-2, 2-3, and 3-4), by condition. Lines connect the mean similarities for each condition, with shaded regions indicate the bootstrapped 95% confidence intervals. Transparent points show individual dyad means.

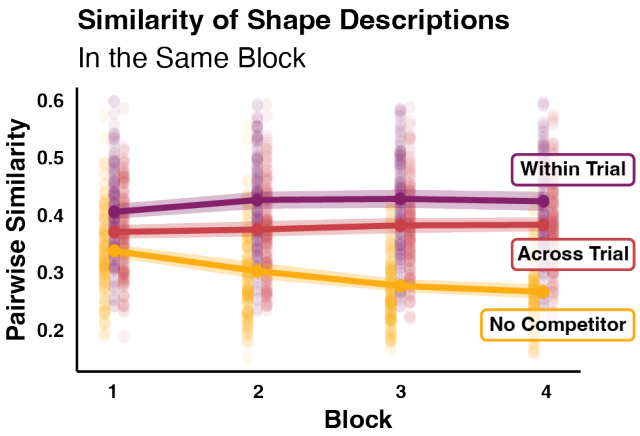

Figure 5: Average similarity between descriptions within each block, by condition. Lines connect the mean similarity between all pairs of descriptions produced in the same block. The shaded region indicate the bootstrapped 95% confidence intervals. Individual points show the mean pairwise similarities of each dyad.

similarity between descriptions, overall. However, if the expressions are less compositionally structured, with a distinct "holistic" label for each shape, we should expect little similarity between descriptions of different shapes. We should also expect any structure that emerges to emerge over the duration of the game, such that descriptions in the final block should more similar to each other than descriptions in the first block. To test this predictions, we computed the pairwise cosine similarities between the SBERT embeddings of all descriptions in each block. We then fit a linear mixed-effects model predicting similarity from condition, block number, and their interaction, with random intercepts for each dyad.

First, we found that descriptions in both conditions with compositional structure showed overall higher within-block similarity compared to the non-compositional baseline (Within-trial: $\beta = 0.087$, $SE = 0.009$, $p < .001$; Across-trial: $\beta = 0.043$, $SE = 0.009$, $p < .001$). The similarity decreased over the course of the game in the baseline condition as the descriptions for each individual shape diverged from one another ($\beta = -0.008$, $SE = 0.0004$, $p < .001$; Fig. 5). In contrast, both compositional shape conditions showed an increase in within-block similarity over time (interaction with block: $\beta = 0.026$, $SE = 0.0004$, $p < .001$ for both conditions). This suggests that participants in the compositional conditions developed conventions that maintained their overall similarity, though this pattern could also reflect the underlying greater visual similarity between shapes that share components–a distinction we examine next.

Selective use of component structure Our key question was whether participants would develop referring expressions that systematically combined labels for individual shape components. To measure how the structure of the expressions align with the component structure of the stimuli, we need to measure the way each expression refers to whole (versus component parts) of the shape. For each shape, we compared how similar the descriptions were to shapes sharing one component versus shapes sharing the other component. A small difference between these similarities would indicate a balanced reference to both components; a large difference would suggest focused reference to just one component (see Fig. 6 for an example).

To test the effect of context on the emergence of compositional language structure, we ran a linear mixed-effects model predicting this difference in similarity from condition, block number, and the iteration between the two, with by-dyad random intercepts. We found that the within-trial competitor condition maintained small differences in the similarity between shapes with one component and shapes with the other component ($\beta = 0.005$, $p = .001$). In contrast, participants in the across-trial competitor condition showed larger initial differences ($\beta = 0.100$, $p < .001$) that grew larger across the blocks ($\beta = 0.055$, $p < .001$; Fig. 7). Rather than developing systematic combinations of component labels, these participants tended to settle on conventions that referred to only one component part (e.g., "diamond bottom").

To validate our similarity measures, we coded a sample (n = 470) of final-round descriptions from each condition. Descriptions were coded by the first author based on whether they referred to only the top *or* bottom component, both components, or used holistic descriptions that didn't clearly reference distinct components. This analysis confirmed our quantitative findings: within-trial condition descriptions predominantly referenced both components (89.7%), while across-trial condition descriptions mostly referenced single components (69.2%), confirming a more metonymic strategy. The

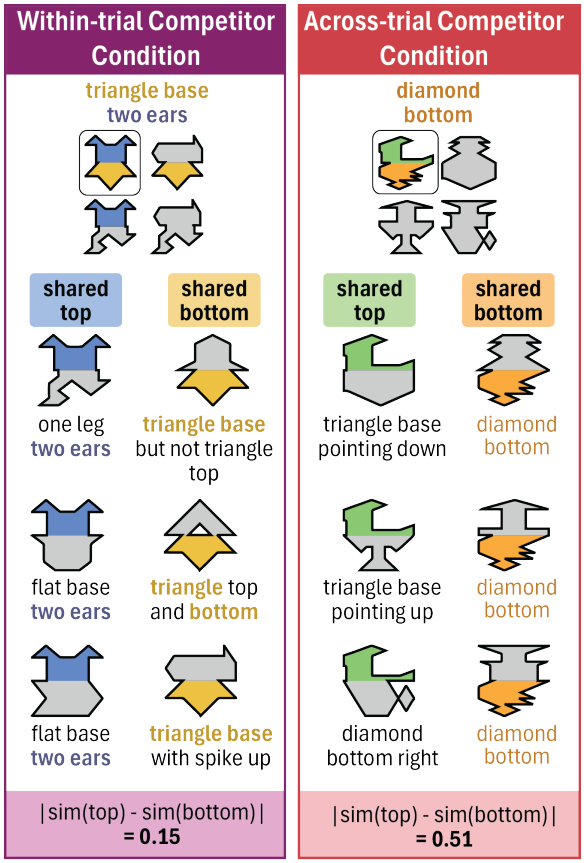

Figure 6: Examples of how the descriptions to the component parts were measured. The left and right panels show the two compositional conditions. At the top is an example trial with a target shape indicated by a box. The shapes that shared a component with the target shape, and their corresponding descriptions, are displayed below.

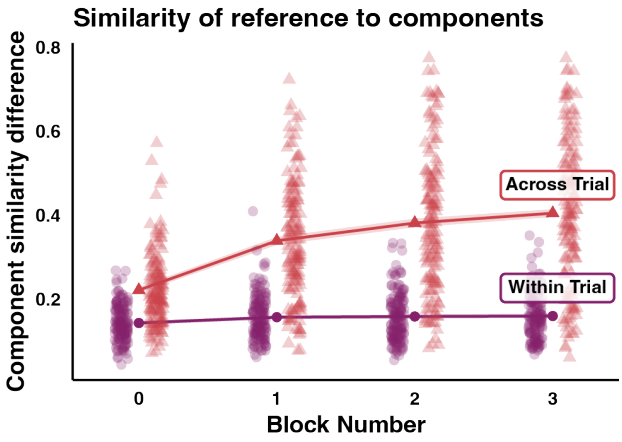

Figure 7: Mean absolute difference in similarity between descriptions sharing one component versus another across blocks. Larger differences indicate more selective reference to single components. Lines show condition means with 95% CI; points show individual dyad means.

baseline condition used both holistic descriptions (41.5%) and metonymic descriptions (42.7%).

This pattern reveals that participants developed compositional conventions only when the communicative context demanded it. When shapes could be distinguished using just one component, participants dropped reference to unnecessary components, leading to less compositional conventions..

Discussion

Previous work suggests that compositional language structure emerges from pressures for compressibility and expressivity (Kirby et al., 2008; Raviv et al., 2019; Nölle et al., 2018). Studies have shown that languages encode features that are most relevant for disambiguation in context (Winters et al., 2015, 2018; Perfors & Navarro, 2014; Müller et al., 2019). However, most studies rely on artificial learning language paradigms where participants uncover mappings between discrete meanings and messages, often using inherently compositional referent spaces. Our approach differs by examining how compositional structure emerges through naturalistic convention formation, where participants freely develop their own descriptive strategies. Furthermore, by independently manipulating environmental structure and communicative context, we isolate which factor drives compositionality.

Our results provide key evidence that the emergence of compositional structure is highly sensitive to the communicative needs engendered by the environment. Despite identical component distributions, participants in the within-trial condition developed compositional conventions while those in the across-trial condition relied on holistic or metonymic strategies. This divergence suggests that compositionality arises when distinguishing between referents with similar structures requires it, but not otherwise. Over repeated interactions, participants converged on strategies that were as efficient as possible given their condition's specific communicative demands.

This work opens several directions for future research. Future work should disentangle the role of visual similarity due to the shared component structure by manipulating these factors independently. Additionally, while participants in the across-trial condition may have developed the *capacity* for compositional descriptions, they had no reason to produce compositional descriptions. We could use held-out combinations of components as pre- and post-test items to reveal how participants generalize their conventions to novel shapes outside of the communicative context.

More broadly, these findings highlight that language is more than just a system for labeling the world – it is a tool for coordination and shared understanding. As a consequence, features of language such as compositionality may not be inevitable; they are flexible and adaptive solutions for coordinating in and communicating about a world filled with complexity.

Acknowledgments

The authors would like to thank Dr. Mike Frank, Alicia Chen, and members of the Stanford Social Interaction Lab for their comments on early versions of this work. We would also like to thank the reviewers for their helpful comments. Funding for this work was provided, in part, by the Morse Society Fellowship through the University of Wisconsin Foundation.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021, March). Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171. doi: 10.3758/s13428-020-01535-9

Boyce, V., Hawkins, R. D., Goodman, N. D., & Frank, M. C. (2024). Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, *121*(28), e2403888121. doi: 10.1073/pnas.2403888121

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Goldberg, A. E. (2015). Compositionality. In *The routledge handbook of semantics* (pp. 419–433). Routledge.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *Cognitive science*, *44*(6), e12845.

Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R. D., & Artzi, Y. (2022). Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Müller, T. F., Winters, J., & Morin, O. (2019). The influence of shared visual context on the successful emergence of conventions in a referential communication task. *Cognitive Science*, *43*(9), e12783.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104.

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775–793.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). *Learning transferable visual models from natural language supervision.*

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, 237–263.

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics.

Sathe, A., Fedorenko, E., & Zaslavsky, N. (2023). Language use is only sparsely compositional: The case of english adjective-noun phrases in humans and large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Szabó, Z. G. (2024). Compositionality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2024 ed.). Metaphysics Research Lab, Stanford University.

Townsend, S. W., Engesser, S., Stoll, S., Zuberbühler, K., & Bickel, B. (2018). Compositionality in animals and humans. *PLoS Biology*, *16*(8), e2006425.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449.

Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, *176*, 15–30.

Zipf, G. K. (1949). *The principle of least effort*. CH3.