Quantifying the Emergence of Symbolic Communication

Emily Cheng*, **Yen-Ling Kuo***, **Josefina Correa**†,
Ignacio Cases*, **Boris Katz***, **Andrei Barbu***
*MIT CSAIL, †MIT Brain and Cognitive Sciences Department
Cambridge, MA 02142 USA

Abstract

We quantitatively study the emergence of symbolic communication in humans with a communication game that attempts to recapitulate an essential step in the development of human language: the emergence of shared signs. In our experiment, a teacher must communicate a first order logic formula to a student through a narrow channel deprived of common shared signs: subjects cannot communicate with each other with the sole exception of car motions in a computer game. Subjects spontaneously develop a shared vocabulary of car motions including indices, icons, and symbols, spanning both task-specific and task-agnostic concepts such as "square" and "understand". We characterize the conditions under which indices, icons, and symbols arise, finding that symbols are harder to establish than icons and indices. We observe the dominant sign category being developed transition from indices to icons to symbols, and identify communicating in ambiguous game environments as a pressure for icon and symbol development.

Keywords: emergent communication; symbolic communication; cooperative games

Introduction

The development of language is quite recent given the timeline of our species. Though modern Homo Sapiens have existed for at least 300,000 years, speech is thought to have arisen only 100,000 years ago (Perreault & Mathew, 2012; Richter et al., 2017). This lengthy process to establish language suggests that developing a shared symbolic vocabulary is a delicate task. Moreover, the precise conditions of this process remain uncertain, and while animals can be taught to use symbols, they are not observed to naturally employ them (Grouchy, D'Eleuterio, Christiansen, & Lipson, 2016). Therefore, a true, spontaneous emergence of symbolic communication in nature has not yet been documented (Deacon, 1997; Grouchy et al., 2016; Schilhab, Stjernfelt, & Deacon, 2012).

We posit a novel symbolic communication game and observe the emergence and evolution of symbolic communication in human subjects over the course of gameplay. We witness the dominant *sign category* being developed transition from indices to icons to symbols, where the development of further sign categories enables better performance.

Prior work in the origins of symbolic communication has considered language emergence in a controlled environment. We note two broad approaches in designing this environment. One approach computationally simulates the emergence of communication protocols in a multi-agent setting (Baroni, 2020; Grouchy et al., 2016; Havrylov & Titov, 2017; Lazaridou & Baroni, 2020; Lazaridou, Peysakhovich, & Baroni, 2016; Lewis, 1969; Lotito, Custode, & Iacca, 2021; Mordatch & Abbeel, 2018). The second approach directly uses human subjects to study the spontaneous development of a shared code after removing pre-existing communicative conventions (De Ruiter et al., 2010; Galantucci, 2005; Galantucci & Garrod, 2011; Scott-Phillips, Kirby, & Ritchie, 2009). Our experiment, a cooperative two-player game (fig. 1), similarly probes how humans spontaneously develop a shared sign vocabulary and reveals the emergence of a symbolic sign system. Like previous studies, we remove from the game familiar modes of symbolic communication such as writing or speaking. A teacher must communicate a first-order logic (FOL) task, encoded in natural language, to a student through continuous spatial movements in a teaching environment. The student then carries out that task in novel test maps. Similar to Scott-Phillips et al. (2009), the players' communication is embodied in their movements in the game world, which tasks players to lift their own communication channel and detect communicative intent.

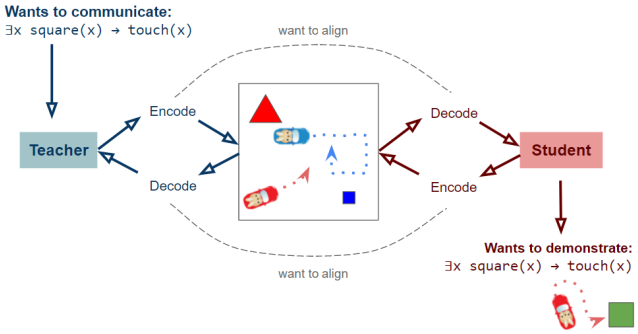


Figure 1: An overview of the game. The teacher is shown a first-order logic task (*top left*) that they must communicate to the student. To do so, the teacher and student interact using their car avatars via a communication channel of spatial movements (*middle*) that are grounded in a map containing several objects. Throughout this interaction, the players encode and decode their internal understandings of the task with the goal of aligning them. Lastly, the student demonstrates their understanding of the task (*bottom right*) to the teacher. In the figure, movement traces (dashed lines) are shown for effect – players do not see them persist when playing the game.

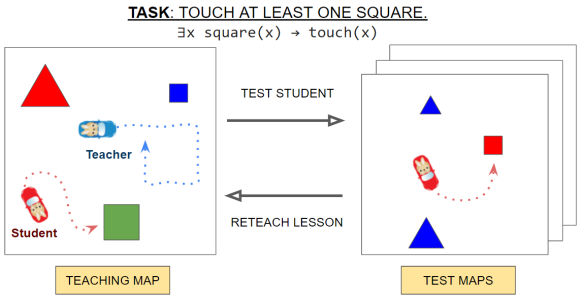


Figure 2: The game mechanics for one task. The players start in a teaching map (*left*), where the teacher communicates the task (unknown to the student) via car motions. When the teacher is done teaching, they advance the student to a set of test maps to perform the task alone (*right*) while the teacher watches. At any point during testing, the teacher may restart the task in order to reteach. At the last test map, the teacher may advance the players to the next task.

Our experiment is novel in four ways. (1) The space of what is communicated is greatly expanded compared to prior work. Instead of focusing exclusively on objects or actions, participants communicate statements in FOL including: objects, object attributes, actions, and logical connectives. This corresponds to communicating whole sentences rather than individual words. (2) Participants engage in multiple rounds of exchanges, determining on their own when they are satisfied with the results. This corresponds to a more natural setting where no outside entity determines when a dialog terminates. (3) We create a rich new interactive setting, driving a car in a multi-agent environment, where humans have little to no existing symbol inventory. (4) We introduce a zero-shot generalization task that probes how the emergent sign sets respond to stresses and sudden changes in the underlying task. Together, these additions intend to bring emergent communication experiments closer to real-world conditions.

Communication Game

We have designed a communication game to study the emergence of meaningful communication between players (see fig. 2). This communication is grounded in the game environment and its quality can be measured quantitatively by performance on tasks.

Game Mechanics

A teacher and a student, on separate computers and in separate rooms, are forbidden to interact with each other except through the game. Players each control a 2D car avatar that navigates a set of maps containing several objects, where they change the velocity and angle of their cars using their keyboard arrow keys. Their avatar movements are their only form of communication in the game. The game works as follows:

1. Players begin in a teaching map where the teacher communicates the task to the student via spatial movements only. The student does not know the task nor the task space.
2. Once the teacher is done teaching, they advance the student to three test maps to perform the task while the teacher watches. At any point in testing, the teacher may choose to reteach the task using the same maps. At the last test map, the teacher can advance to the next task.
3. After each task, each player submits a reflection form (fig. A.5)[1] where they draw and describe the "actions" they used to communicate and the overall communication in this round, such as strategy, level of confusion, etc. In addition, the student submits a guess of the task. Players do not have access to each others' reflection forms.

Players score one point for each successful test map. These points are reset for a task when the players re-attempt it. The teacher sees the total points earned, or *raw score*, throughout the game, and the student sees it every other task. Because players are afforded unlimited attempts per task, we evaluate them using their *weighted score*, defined as the raw score divided by number of attempts.

A game session repeats the above steps with forty tasks. The game motivates cooperative sign development in the following ways: (1) players cannot communicate with the sole exception of moving the cars on a shared 2D map, requiring new modes of communication; (2) players are equally incentivized by student performance on test maps, which implicitly rewards good communication; (3) players complete numerous tasks together, allowing iterative sign development over the game.

Task and Map Generation

A task in the game is a FOL formula expressed in natural language. In operating over FOL and continuous motions, our task space is greatly expanded from that of prior emergent communication games. A FOL task takes the form

$$\texttt{quantifier}\ x.\ \texttt{attribute}(x) \rightarrow \texttt{action}(x).$$

More complex tasks can be constructed using logical connectives **and** ($\wedge$), **or** ($\vee$), or **not** ($\neg$). Tasks are sampled from a probabilistic grammar (table A.5). The task predicate space as well as examples are given in table 1.

A game session is comprised of a forty-task *regular session*, then a four-task *zero-shot generalization test* which evaluates player performance on novel compositional tasks. Regular session tasks are described in table 1, and generalization tasks are described in Appendix A.

Given a task, we generate a teaching map and three test maps. Each object on a map is defined by its shape, size, color, and position. The space of object attributes is shown in table 1. Teaching maps uniformly sample four categories (Appendix A, fig. A.1): (1) *base*, where one can discriminate the solution set by touching objects; (2) *ambiguous in the quantifier*, where one cannot disambiguate the quantifier by touching objects; (3) *ambiguous in the attribute*, where one cannot disambiguate target attributes such as color or shape by touching; and (4) *inconvenient*, where the large number of target objects makes it hard to teach by touching objects. For each generalization

[1] Appendices are available at `https://tinyurl.com/yc6wvtez`

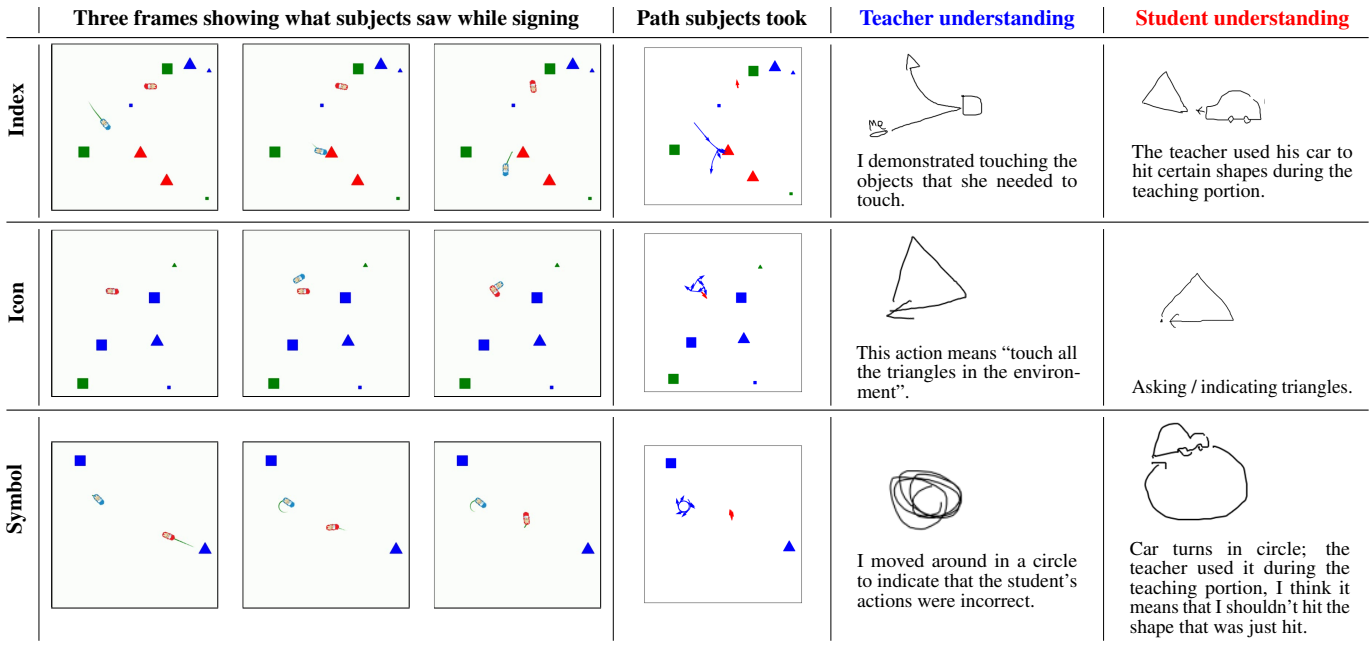


Figure 3: Several signs produced and interpreted by Pairs 1 (index & symbol) and 3 (icon). Each row is (left to right): (1) the players' view of the teaching map, shown as three frames sampled evenly from the time interval of sign production; (2) our render of the sign using player trajectories; (3) reflection form data of the sign registered by the teacher and student thereafter.

Attributes	Color: red, blue, green Shape: square, triangle Size: big, small Pattern: *partially filled*
Actions	touch, touch going forwards, touch going backwards, avoid
Quantifiers	all (∀), at least one (∃), exactly one, exactly two, *exactly three*
Logical Connectives	and (∧), or (∨), not (¬)
Example Tasks	1. Touch all objects that are small. 2. Going backwards, touch exactly one object that is not red. 3. Touch all objects that are [blue and square]. 4. Touch all objects that are not green, or touch exactly two objects that are red.

Table 1: The first order logic task space of the game, where *italicized tokens* are introduced in the generalization test. Four example tasks are shown, representing a typical spread of tasks that players encounter in the game.

task, the teaching map is blank and we hand-design three fixed test maps. Finally, all test maps sample the *base* distribution.

To ease players into the game, the first five tasks in the regular session take the form *Touch all objects that are* ____, sampling the remaining token from the attributes in table 1, and its teaching maps from the base distribution. However, to minimize the bias of task order on player sign sets, we do not set an explicit curriculum after the first five tasks and rather sample tasks and maps at random. Finally, the four generalization tasks at the end of the game session are shown to pairs of players in a random order. The five task ramp-up and the generalization test mark two shifts in task distribution.

Experiments and Analysis

We recruit thirteen pairs to play the game, categorizing signs as in previous work (De Ruiter et al., 2010; Galantucci & Garrod, 2011; Grouchy et al., 2016; Lotito et al., 2021), according to their most salient properties, as belonging to one of the following categories introduced by Peirce as his *second trichotomy*: indices, icons, and symbols (Peirce, 1965) (also cf. Deacon 1997). We quantify a *symbolic gap*, or systematic bias towards the establishment of non-symbols over symbols.

Emergence and Development of Signs

The players' shared sign sets converge over the course of the regular session (fig. 4 and fig. A.8). Pairs generally then introduce novel signs in the generalization test upon seeing new task predicates.

Sign introductions were asymmetric according to role. In particular, teachers initiated communication with their students, introducing 144 / 147 total signs (98%). Twenty-seven percent of new signs are introduced when encountering new task predicates. Systematically across pairs, such signs include `TOUCH`, introduced in the first task, as well as `BACKWARDS`. Because pairs initially use indexical `TOUCH` to refer to solution sets, non-indexical signs for object attributes and quantifiers may have delayed introductions. These signs are most often introduced to disambiguate solutions when indices fall short (see section The Symbolic Gap).

A sign undergoes a process of negotiation between players

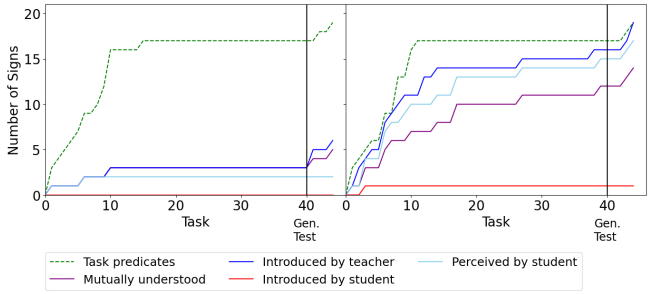


Figure 4: Evolution of the player sign sets for Pair 2 (*left*) and Pair 3 (*right*). We consider a sign to be *introduced* by task t if a player uses a sign and registers it in the reflection form at some $t' \leq t$. A sign is *mutually understood* by task t if that sign has been registered in the student's reflection form (indicating understanding) at some $t' \leq t$, and similarly for the teacher. A sign is *perceived* by the student by task t if the sign has been registered as a teacher-introduced sign (with any degree of uncertainty) in the student's reflection form at some $t' \leq t$. For reference, the number of *task predicates* encountered by the teacher in the task statements is shown as a green dashed line. We observe player sign sets saturate once task predicates stop being introduced, and desaturate upon new task predicates in the generalization test. In addition, teachers do not introduce a sign for every task predicate, students do not perceive all the teacher-introduced signs, and they do not understand all perceived signs (indicated by gaps between lines).

until it is either established or discarded (Appendix B: Sign Establishment). Of all 162 (form, meaning) combinations that have been introduced at some point in time, there are 73 symbols, 46 indices, and 43 icons. However, all updates to signs, either re-introductions of the same meaning with a new form or repurposes of an existing form for a new meaning, occur in symbols or indices regardless of whether they were already perceived or understood. In particular, there are 18 total updates to symbols (25% of symbols), 7 to indices (15% of indices), and, notably, 0 to icons (fig. A.3). This implies that icons are the most stable, then indices, and lastly symbols.

Icons, Indices, and Symbols

Across all participants, we identify 80 mutually understood signs: 18 symbols, 34 icons, and 28 indices (table A.1). We observe that a sign's meaning biases its sign category, for example, direction and shape signs are categorically icons instead of symbols (fig. 5, fig. A.6, and fig. A.7). This bias is a function of how signs are grounded in the game map and embodied in player motions, and reveals that the environment itself has a strong effect on sign category. We provide examples of sign categories in fig. 3.

Developing non-indices corresponds to a higher weighted score; the correlation between pairs' average weighted score and number of non-indices is $\rho = 0.71$. The correlations between the number of indices, icons, and symbols and weighted score are $\rho = 0.29, 0.51$, and 0.58, respectively. While all pairs developed a maximum of four indices, the further development of icons and symbols allows for a wider separation in weighted score (fig. A.9). This suggests that non-indices rather than indices allow student generalization to test maps. Indeed, while indices refer to spatially or temporally proximal objects, icons and symbols, due to their forms' uncoupling from physical environment, allow pairs to communicate spatially and temporally displaced information and succeed in ambiguous settings (Hockett, 1960).

The Symbolic Gap

We observe that pairs convey meaning non-symbolically when possible. For instance, many pairs encounter a task such as *Touch all objects that are red* within the first five tasks of the game. However, teachers choose not to introduce an abstract symbol for `RED`, instead referring to `RED` indexically. This *symbolic gap*, or a systematic bias towards the establishment of indices and icons over symbols, mirrors the overwhelming prevalence of indices and icons in animal communication systems and early human communication (fig. 7) (Deacon, 1997; Grouchy et al., 2016). Furthermore, the bias towards establishing indices and icons persists despite the fact that participants use symbols in life and that teachers introduce symbols at a constant rate (fig. 7). Why do participants primarily develop non-symbols, and under what conditions do they introduce indices, icons, and symbols?

Symbols are harder to establish than icons and indices. Initially in the game, participants likely use indices for practical reasons. Symbols are harder to perceive and understand than both icons and indices: only 62% of symbols introduced were perceived and 28% understood, in contrast to 83% of indices perceived and 68% understood, as well as 90% of icons perceived and 81% understood. This may be due to symbols' arbitrary mapping from form to meaning. Moreover, icons take more tries to perceive and understand than indices, likely due to the difficulty of tracing motions with the car (fig. 6). Then, choosing to communicate via an index over an icon or a symbol, at least initially, is to choose clarity.

Sign categories saturate at different times. The process of sign introduction may be understood in a series of plateaus. Note that pairs see all task predicates before task 20 (fig. A.8). We observe that introduced and mutually understood indices plateau at around task 5, after which their rate of introduction decreases (fig. 7). As indices plateau, non-indices continue to be introduced at a similar rate; and as icons plateau around task 20 in both introduction and understanding, symbols continue to be introduced at a similar rate. Total shared lexicon growth is first attributed to indices, then icons, then symbols, shown by the slopes of understood index, icon, and symbol curves in fig. 7.

We observe that when players develop signs, (1) each sign category will *reach saturation*, or exhaust its communicative utility in the current context, and its introduction slows, making way for the next sign category; and (2) the order of sat-

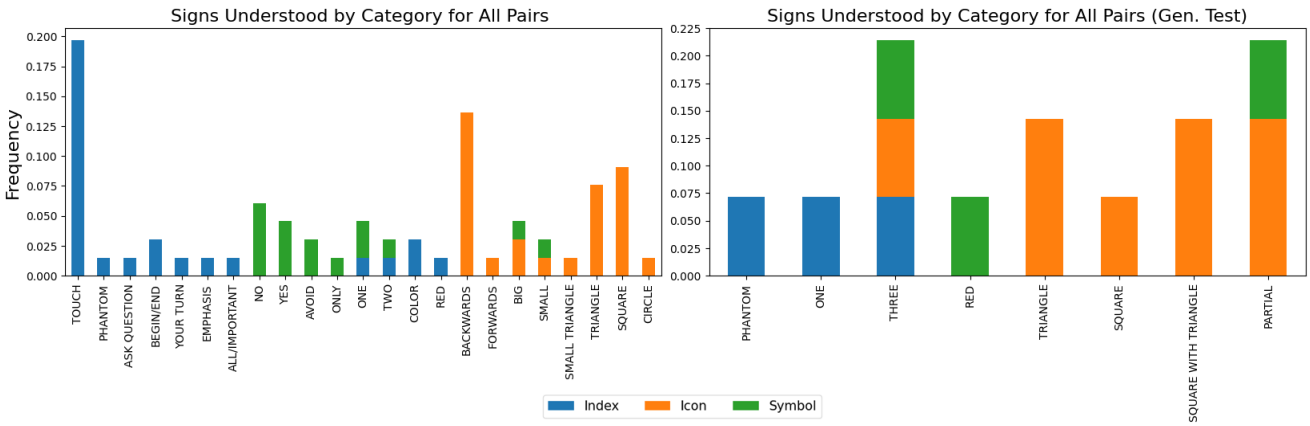


Figure 5: The frequency of signs developed by pairs in the regular session (*left*) and generalization test (*right*), where pairs cover 14/18 task predicates in the regular session. Sign meaning appears to correlate with how subjects chose to express them as icons, indices, or symbols. The same plots for introduced and perceived signs are found at fig. A.6 and fig. A.7, respectively.

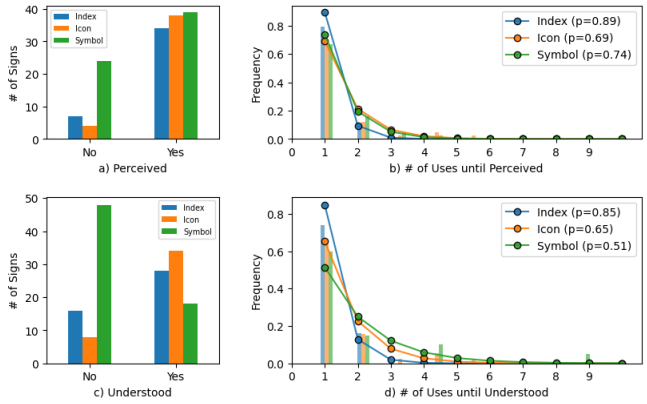


Figure 6: All signs split by whether they were perceived (a) and split by whether they were understood (c). Symbols are much harder to perceive and understand than non-symbols, given by the relative sizes of the green bars in (a) and (c). Then, in (b) and (d), a closer breakdown of those signs eventually perceived or understood– we show distributions over the number of uses of a sign until perceived or understood and fit geometric distributions, showing that indices are fastest to perceive and understand. Plots are aggregated across all pairs.

uration of sign categories is indices, then icons, then symbols (see fig. 7, fig. A.9 and fig. A.10). The effect of desaturation is seen in the uptick in icon and symbol introductions in the generalization test, where the task distribution shifts to blank teaching maps and new task predicates (fig. 7). Conversely, indices remain saturated in the generalization test as they are minimally useful in a blank teaching map.

We observe sign categories saturate concurrently with contextual shifts. The saturation of indices coincides with a shift at task 6 from the five-task ramp-up to the full distribution of tasks and maps, and is explained by what remains constant through the distribution shift: the usefulness of discriminating target sets by tapping. Where this indexical teaching strategy falls short, icons and symbols become comparatively useful.

Icon saturation also coincides with a contextual shift around task 20, when task predicate introduction plateaus. This is followed by a desaturation of icons in the generalization test upon seeing new concepts, e.g. **partial** (fig. 5). Then, icon saturation is likely tied to the rate of task predicate introduction.

Finally, it is not clear that symbols saturate during the game. Of the ten pairs who introduced symbols and the eight who developed mutually understood symbols, only half clearly reach saturation (fig. A.9). After icons saturate, symbols continue to arise likely because symbols are harder to learn.

Ambiguity drives icon and symbol production. We find that ambiguous teaching maps drive the continued introduction of icons and symbols after the saturation of indices. Icons and symbols are useful in this setting because they can communicate information decoupled from the immediate environment. For example, the teacher in Pair 1 introduces an iconic `SQUARE` in an ambiguous attribute teaching map to signal the importance of shape and not color. The effectiveness of using non-indices in these environments can be seen in table A.3, where for pairs who develop more than the median number of non-indices, the difference in weighted score between unambiguous and ambiguous teaching maps is 0.55, compared to 0.89 in pairs who develop fewer than the median non-indices.

Perhaps because indices fall short in ambiguous maps, signs introduced under ambiguity tend to be non-indices. On average, in ambiguous teaching maps non-indices comprise 77% of all sign introductions, where in ambiguous attribute maps, signs introduced tend to be symbols. Similarly, non-indices comprise 80% of sign introductions in the generalization test, where teaching maps are blank, and task predicates that existed in the regular session are re-expressed non-indexically (e.g. `RED` in fig. 5). In contrast, non-indices comprise 47% of sign introductions in base teaching maps.

Icons and symbols can communicate information more efficiently than indices in inconvenient teaching maps (imag-

ine signing ALL SQUARE instead of touching every square). Though non-indices are 81% of sign introductions on average in inconvenient maps (section 2 of table A.2), inconvenient teaching maps motivate much fewer non-index introductions, in absolute numbers, than ambiguous and even base maps. Only 12 signs are introduced in inconvenient maps compared to 23 in ambiguous attribute maps, 34 in ambiguous quantifier maps, and 45 in base maps (section 3 of table A.2). And contrary to in ambiguous maps, developing more than the median number of non-indices in inconvenient maps does not confer a large advantage over developing fewer (table A.3)– players often perform well with indexical strategies. We conclude that inconvenient maps do motivate non-index introduction over index introduction, but on a much smaller scale than ambiguous maps.

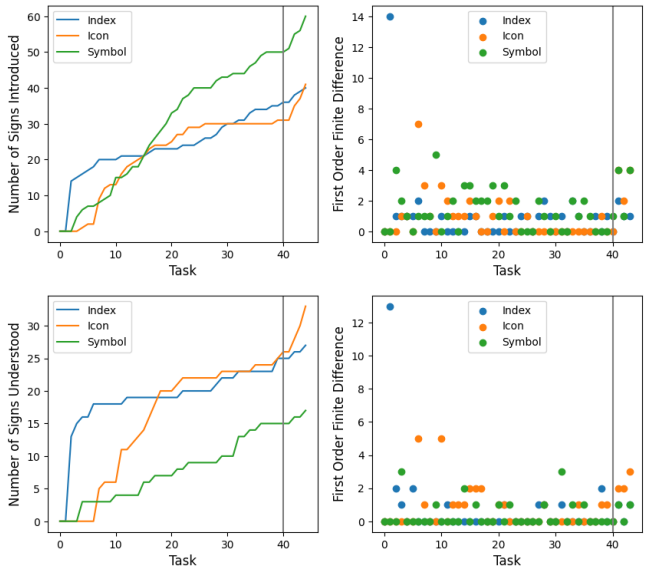


Figure 7: Cumulative number of signs introduced and understood (left), with their respective rates of change (right). The dark line at task 40 marks the beginning of the generalization test. Looking at the left figures, note that the order of saturation in introduction and understanding is indices (task ~5) then icons (task ~20), and it is unclear whether symbols saturate before task 40. Looking at the bottom right figure, note that sign introduction and understanding for indices (task ~3), then icons (task ~10), then symbols (task ~30) achieve maximum values in rate of change. Then, the increase in number of mutually understood signs is first attributed to indices, then icons, and lastly symbols.

Compositionality and Generalization

All players make use of some simple sign composition regardless of the size of sign sets, suggesting that combinational structure may emerge early on in communication systems (Galantucci, 2005; Galantucci & Garrod, 2011; De Ruiter et al., 2010). We note that sign composition carried out by all pairs of players is not only gestural but follows a *base-modification* structure (Armstrong, Stokoe, & Wilcox, 1994; Corballis, 1991).

Spatial composition develops in all pairs and involves a *primary form*, or base, on which *secondary forms*, or modifications, are superimposed. Typically, primary forms can stand alone as a sign (e.g. TOUCH), while secondary forms must modify a primary form in order to convey meaning (e.g. BACKWARDS). All signs referring to shape and actions (e.g. TOUCH) are primary. Meanwhile, 77% of signs referring to size and manner of movement are secondary.

In all nine pairs that employ temporal composition, we note a topic-comment structure (Hockett, 1960), another form of base-modification. For example, in all eight pairs that develop negation or affirmation signs, the speaker first defines the topic, or base, for example by TOUCHing or signing a target object, and then modifies it with negation or affirmation.

Number of signs developed (driven by non-indices) is a key factor in better generalization to both arbitrary test maps and generalization tasks. We proxy the pairs' ability to generalize to an arbitrary number of test maps for a given task by examining student guesses according to table A.4, finding that number of signs developed correlates to the likelihood to generalize to arbitrary test maps ($\rho = 0.64$). Likewise, developing non-indices closes the performance gap between base and blank teaching maps. The difference in average weighted score between base and blank teaching maps is 1.04 in pairs who develop higher than the median number of non-indices, while it is 1.56 in pairs who develop fewer than the median non-indices (table A.3).

Conclusion and Future Work

The advent of symbolic communication is thought to be a crucial step in the evolution of human language. We have investigated the origins of symbolic communication over a task space expanded to FOL and over a continuous, embodied communication channel. We have found quantitative evidence that symbols are the hardest to establish; that indices systematically saturate before icons, before symbols; and that icons and symbols arise to communicate displaced information under ambiguity. Finally, some form of simple sign composition appeared in emergent communication. The degree to which compositionality is required to achieve generalization requires further experiment and is left for future work.

Future work involves scaling up experiments, including expanding the task space, extending the game sequence, and recruiting more participants. This will allow us to analyze emergent sign sets over broader task predicates and time frame as well as perform statistical inferences about the population. We will also train artificial agents to play our game in a multi-agent setting, which will allow us to then compare how symbolic communication developed in human and artificial agents. By building agents that can learn to communicate symbolically, we may move one step forward understanding how human symbolic communication evolves.

Acknowledgements

This work was supported by the Center for Brains, Minds and Machines, NSF STC award 1231216, the MIT CSAIL Systems that Learn Initiative, the CBMM-Siemens Graduate Fellowship, the MIT Lemelson Minority Engineering Presidential Fellowship, the MIT-IBM Watson AI Lab, the DARPA Artificial Social Intelligence for Successful Teams (ASIST) program, the United States Air Force Research Laboratory and United States Air Force Artificial Intelligence Accelerator under Cooperative Agreement Number FA8750-19-2-1000, and the Office of Naval Research under Award Number N00014-20-1-2589 and Award Number N00014- 20-1-2643. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Armstrong, D., Stokoe, W., & Wilcox, S. (1994). Signs of the origin of syntax. *Current Anthropology*, *35*(4).

Balietti, S. (2016). nodeGame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods*, *49*(5), 1696–1715.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*.

Corballis, M. C. (1991). *The lopsided ape: Evolution of the generative mind*. Oxford University Press.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W.W. Norton & Co.

De Ruiter, J., Noordzij, M., Newman-Norlund, S., Hagoort, P., Levinson, S., & Toni, I. (2010, 03). Exploring the cognitive infrastructure of communication. *Interaction Studies*, *11*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, *5*(11).

Green, M. (2021). Speech acts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University.

Grouchy, P., D'Eleuterio, G. M. T., Christiansen, M. H., & Lipson, H. (2016). On the evolutionary origin of symbolic communication. *Scientific Reports*, *6*(1).

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems.*

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*(3), 88–96.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era.

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2016). Multi-agent cooperation and the emergence of (natural) language. *CoRR*.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

Lotito, Q. F., Custode, L. L., & Iacca, G. (2021). A signal-centric perspective on the evolution of symbolic communication. *Proceedings of the Genetic and Evolutionary Computation Conference*.

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-second aaai conference on artificial intelligence.*

Peirce, C. (1965). *Collected papers of charles sanders peirce* (C. Hartshorne, P. Weiss, & A. W. Burks, Eds.). Belknap Press of Harvard University Press.

Perreault, C., & Mathew, S. (2012). Dating the origin of language using phonemic diversity. *PLoS One*.

Richter, D., Grün, R., Joannes-Boyau, R., Steele, T. E., Amani, F., Rué, M., ... McPherron, S. P. (2017). The age of the hominin fossils from jebel irhoud, morocco, and the origins of the middle stone age. *Nature*.

Schilhab, T., Stjernfelt, F., & Deacon, T. (2012). *The symbolic species evolved*. Springer.

Scott-Phillips, T., Kirby, S., & Ritchie, G. (2009). Signalling signalhood and the emergence of communication. *Cognition*, *113*, 226–233.

Appendix A: Task and Map Generation

Zero-Shot Generalization Tasks

We hand-design a fixed set of four *generalization tasks*. The generalization tasks introduce an additional two tokens, **three** and **partially**, bringing the total number of task predicates to 20 from 18 in the regular session.

1. *Touch exactly three objects that are red*: tests induction in the quantifier from **one** and **two** to **three**.
2. *Touch all but one objects that are triangular*: tests composition of **all**, **not**, and **exactly one** in the quantifier.
3. *Touch exactly one object that is [partially but not all] blue*: tests composition of **all** and **not** to describe a new target shape.
4. *Touch exactly one object that is square on the outside and triangular on the inside*: tests composition of **triangle** and **square** to describe a new target shape.

Map Generation

So that the task is satisfiable in the generated map, we extract the set of attributes required to satisfy the formula and place N objects with these attributes on the map at random. We also place M objects of random attributes on the map as distractors. For both teaching and test maps, we sample $N \sim \mathcal{D}^{\text{required}}$ and $M \sim \mathcal{D}^{\text{distractor}}$. In the *base* case, both distributions are $\text{Unif}[1,3]$.

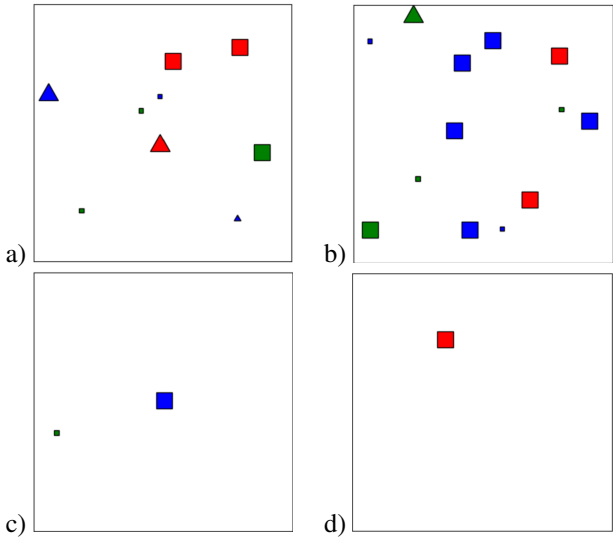


Figure A.1: Example game maps for *Touch all objects that are square*. Maps are sampled from the (a) *base*, (b) *inconvenient*, (c) *ambiguous attribute*, and (d) *ambiguous quantifier* distributions.

In addition, we introduce two pressures in the teaching map to motivate sign development by the teacher.

- *Ambiguity in the quantifier*: we place one target object ($N = 1$) so that it is difficult to teach e.g. all vs. at least one.
- *Ambiguity in the attribute*: we place no distractors ($M = 0$) so that the teacher cannot use them as negative examples to discriminate the solution set.
- *Inconvenience*: we set a large number of target objects on the map ($\mathcal{D}^{\text{required}} = \text{Unif}[9,11]$). Rather than touching all target objects, the teacher may use a sign for brevity.

Test maps sample the *base* distribution. However, for tasks composed of multiple subtasks by disjunction like *"Touch all squares, or touch all red objects"*, test maps are sampled disjointly for each subtask.

Finally, all maps measure 450×450 pixels.

Appendix B: Sign Establishment

A sign undergoes a process of negotiation between speaker and listener until established or discarded. This process consists of three broad steps: (1) the speaker internally negotiates the best form to convey the sign's meaning; (2) the listener perceives the sign as communicative; and (3) both parties reach a mutual understanding of the sign's meaning (Scott-Phillips et al., 2009). The speaker's internal negotiation of sign form depends recursively on their beliefs about steps 2 and 3 (Green, 2021; Frank & Goodman, 2012).

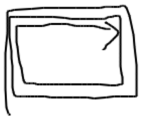
SQUARE	ALL SQUARE
	
I (the teacher) used this to mean "touch the square"	I (the teacher) used this to mean: "touch all the squares"

Figure A.2: Example of a pair of signs that the student did not perceive as distinct (Pair 3). A sign for SQUARE (left) was established; then, the teacher attempted to trace a square multiple times to indicate ALL SQUARE (right). The student was not able to perceive ALL SQUARE as distinct from SQUARE.

Whether a sign is perceived as communicative depends on whether its form departs from the range of typical car movements as well as from previously established signs. For example, pairs found it difficult to recognize direction as related to the task:

> **Student 1:** We demonstrated our understanding of the task to each other. I noticed that my partner kept on reversing into the shapes, but I wasn't sure if that was related to the task somehow.

Whether a sign is perceived also depends on whether its form departs sufficiently from that of a previously established sign. For example, in Pair 3, the teacher and student agree on an icon for SQUARE, then when the teacher introduces a compositional sign for ALL SQUARE, the student never perceives it as morphologically distinct from SQUARE (fig. A.2). In future work, it is possible to predict whether a sign is perceptible by quantifying player motions with information-theoretic metrics. We hypothesize that player movements with high Shannon information, or surprisal, correspond to a higher chance of being perceived as a sign.

Once perceived, a sign undergoes an iterative negotiation until mutually understood or discarded. That is, throughout repeated uses of the sign, the listener updates their internal understanding of its possible meanings. Simultaneously, the speaker updates their beliefs of the listener's understanding, and the sign's form and meaning accordingly. A sign is mutually understood once the listener and speaker converge upon the same (form, meaning) combination. Players' internal updates of signs are documented in reflection forms, which illustrate this iterative and reciprocal process of sign establishment. For example, in Pair 3, the teacher introduces an iconic sign for SMALL (fig. 3). The student needed four usages of the same sign and three internal updates to converge on the intended meaning. All of their descriptions for the same sign are shown below:

> **Student 3 (first usage):** Repeated small circles, indicating size, or small, i guess? teacher used it.
> **Student 3 (third usage):** I think the repeated small circles indicated that the color was important?
> **Student 3 (fourth usage):** Indicating size is important, and it's small size rather than large.

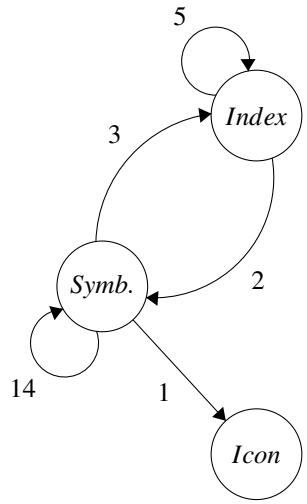


Figure A.3: All transitions between sign categories due to re-introductions and repurposes, visualized with counts on edge labels, where *symbol* is abbreviated *symb*. We compute the transition counts by determining the categories of the original and updated sign for each re-introduction and repurpose. All re-introductions rest within the original sign category, contributing 7 to the self-edge of *symbol* and 3 to the self-edge of *index*. The remaining transitions are attributed to repurposes. Note that we only consider morphological or semantic changes to signs; the vast majority of the 147 sign introductions remain in their original sign category by virtue of not being changed.

When a speaker updates a sign, they can *re-introduce* it with an updated form, or *repurpose* the same form to mean something else. Six pairs collectively attempted 10 re-introductions, or *duplicate forms*, spanning nine unique concepts: NO, AVOID, COLOR, RED, PHANTOM, EMPHASIS, and quantifiers ONE and TWO; and one pair attempted three forms for ALL. Seven out of these nine signs are symbols, and two indices EMPHASIS and PHANTOM. Concrete concepts such as shape, which are typically icons (fig. 5), are not represented. Similarly, four pairs collectively attempt 15 repurposes over 10 unique introduced forms. Eleven such changes repurpose a symbol, and the remaining four repurpose an index.

Appendix C: Game Interface

When players begin the game, they must first complete the instructions, a quiz on the instructions, and a practice map where they individually test controls. Then, they advance to the game in a synchronized step. We provide the full instructions transcript for each player below, as well as screenshots of the game user interface. The game is implemented using the NodeGame Javascript framework (Balietti, 2016).

Teacher Instructions

Welcome! In this study, you're a teacher in a communication game. You're paired with a student, and your job is to teach them to carry out a set of tasks, one task at a time. You'll control the blue avatar and the student will control the red one. You will also be asked several questions to check your understanding of the game and practice using controls before starting the game. The study is estimated to take about 4-5 hours in total. If you and your partner want to take a break, please close the window and revisit the link when you are ready to continue. You'll pick up where you left off!

Game Control You and the student will begin in a teaching stage. In the teaching stage, you will see the following:

- A **teaching area** containing different objects where you and your student will play.
- A **dashboard** displaying the current task, your total score, your score on this task, a legend showing your avatar, and a button to test the student.
- A **notification area** where you can see the latest updates from the student,
- And **instructions** describing the current step.

You will be given a task, such as "Touch at least one object that is square and blue," which the student does not know. In **only communicating with the student through spatial movements**, your job is to teach the student to perform this task.

When done teaching, you will test the student on the task in several **test maps**. When the student thinks they're done with a map, they will choose to move onto the next one, and you will get notified in the **notification area**. The student also may reset their test map if they make a mistake. Based on their performance, you can choose to bring them back to the teaching area to reteach the lesson or advance to teaching a new task. The "Reteach Lesson" option will be available in the **dashboard** throughout testing, while the "Next Lesson" option will appear at the last test map only. After each task, you will fill out a short reflection where you will be asked to describe how you and the student communicated during the task.

You will earn one point for every test map the student completes correctly. You will be able to see your total score throughout the game. Results for teacher-student pairs will be displayed on a leaderboard at the end of the session.

How to Interpret Tasks When evaluating whether a task is satisfied on a test map, please assume that actions not stated in the task description are still permitted. For example,

- "Touch at least one object that is square and blue": as long as the student touches a blue square, it is fine if the student also touches other objects.
- "Touch at least one object that is square and blue, and do not touch any objects that are not square and blue": the student must touch at least one blue square, but not other types of objects.

For tasks like "Touch exactly one object that is blue": the student must touch exactly one object, and the object must be blue. In complex tasks like "Touch at least one object that is not [blue or [square and red]]", the brackets help make reading easier (they don't contribute any extra meaning to the task).

In addition, in tasks such as "Touch all objects that are blue, or touch all objects that are square", students will be tested on both components disjointly and to maximize your score, you will need to teach both sub-tasks in the teaching map. You may also see tasks like "While going forwards, touch all objects that are square and blue". This doesn't mean that the player needs to be moving forwards at all times in the map, but rather just when they're touching objects.

Before starting, please complete the following quiz on the instructions and a practice map. If you understood the instructions correctly press the "Next" button to proceed to the quiz.

Student Instructions

Welcome! In this study, you're a student in a communication game. You're paired with a teacher, and they will teach you to carry out a set of tasks, one task at a time. (More details on the next page.) You'll control the red avatar and the teacher will control the blue one. You will also be asked several questions to check your understanding of the game and practice using controls before starting the game. The study is estimated to take about 4-5 hours in total. If you and your partner want to take a break, please close the window and revisit the link when you are ready to continue. You'll pick up where you left off!

Game Control You and the teacher will begin in the teaching stage (shown below). In the teaching stage, you will see the following:

- A teaching area containing different objects where you and your teacher will play.
- A dashboard displaying the current task and a legend showing your avatar,
- A notification area where you can see the latest updates from the teacher,
- And instructions describing the current step.

The teacher will be given a sentence describing a task which you will not know, such as “Touch all blue objects”. In only communicating with each other through spatial movements, the teacher will teach you to perform this task.

When done teaching, the teacher will test you on the task in several test maps. Now, your dashboard will also show you which test map you’re on and “I did it!” and “Reset” buttons. When you think you’re done with a test map, you can move onto the next one by pressing the “I did it!” button. If you make a mistake on a test map, you may also reset that map using the “Reset” button. Based on your performance, the teacher will choose to either bring you back to the teaching area to reteach the lesson or to advance to teaching a new task. After each task, you will fill out a short reflection where you will be asked to describe how you and the teacher communicated during the task.

You will earn one point for every map you complete correctly. You’ll be shown your total score several times throughout the game. Depending on your final score, you may be eligible for a bonus. Results for teacher-student pairs will be displayed on a leader board at the end of the session.

Before starting, please complete the following quiz on the instructions and a practice map. If you understood the instructions correctly press the “Next” button to proceed to the quiz.

User Interface

Fig. A.4 displays the interface players see when they interact with each other in teaching and testing maps. Fig. A.5 shows the reflection forms that players complete after each task.

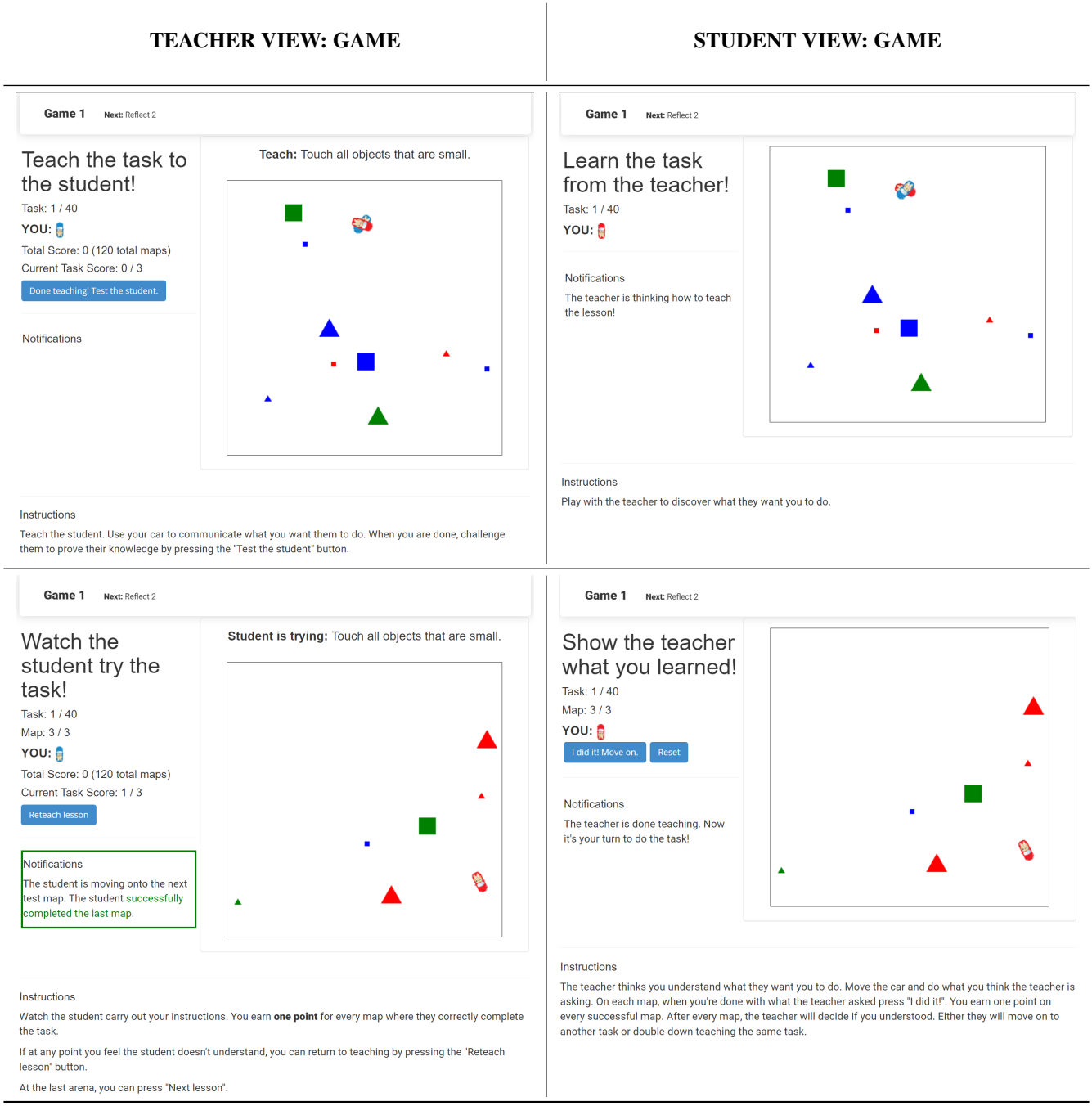


Figure A.4: Teacher and student game views (left and right, respectively) for a teaching stage (top row) and a test map (bottom row) for the task *Touch all objects that are small*.

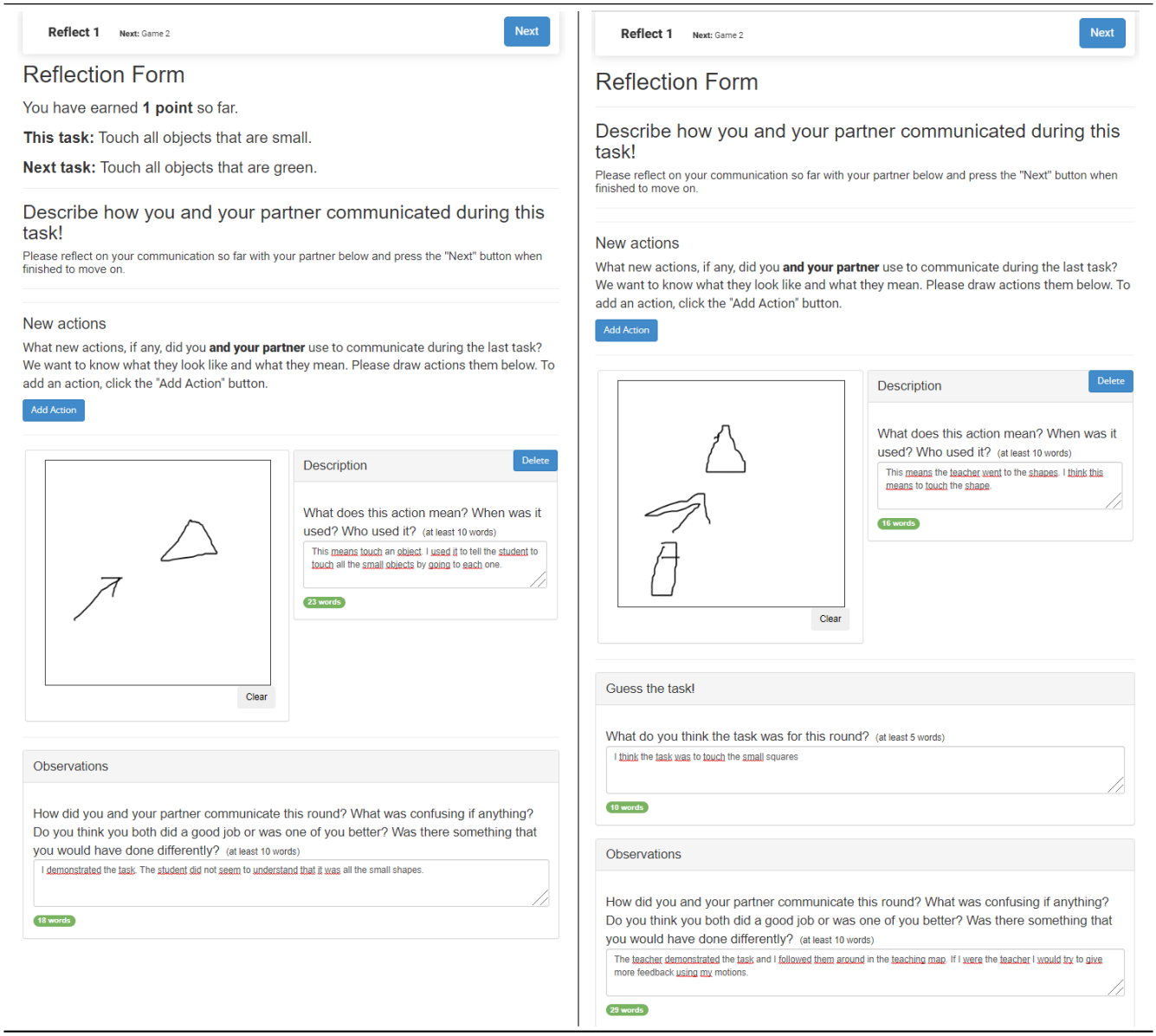


Figure A.5: Teacher and student reflection form views (left and right, respectively) for the task *Touch all objects that are small*.

Appendix D: Figures and Tables

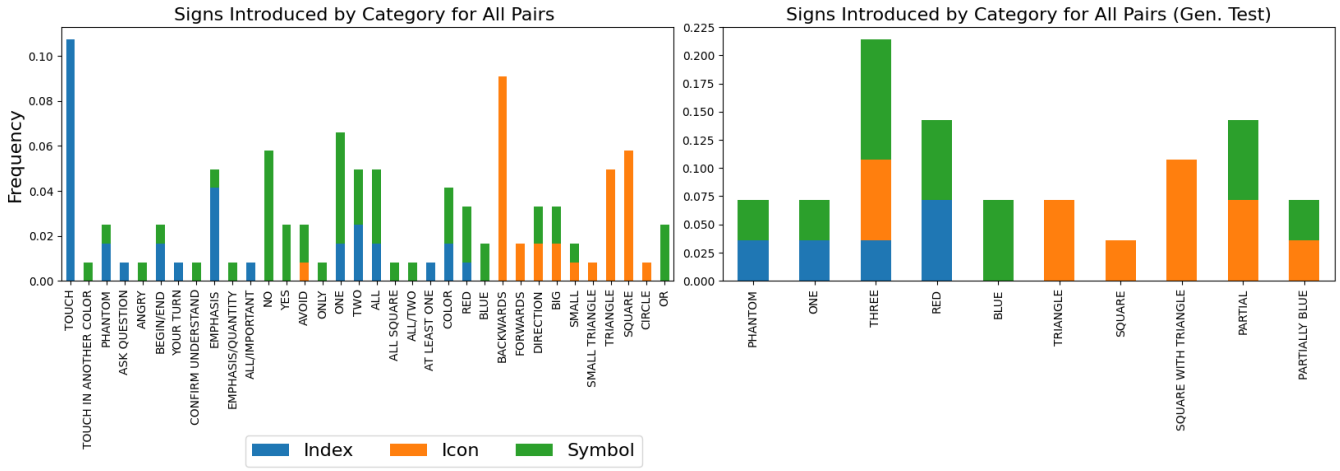


Figure A.6: The frequency of signs introduced by pairs of players in the regular session (left) and generalization test (right).

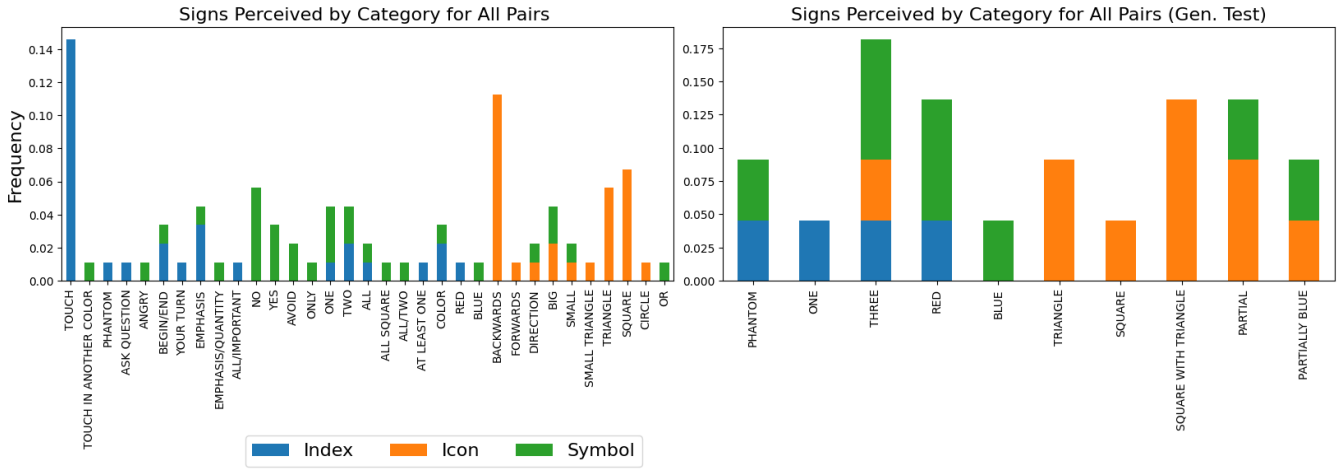


Figure A.7: The frequency of signs perceived by pairs in the regular session (left) and generalization test (right).

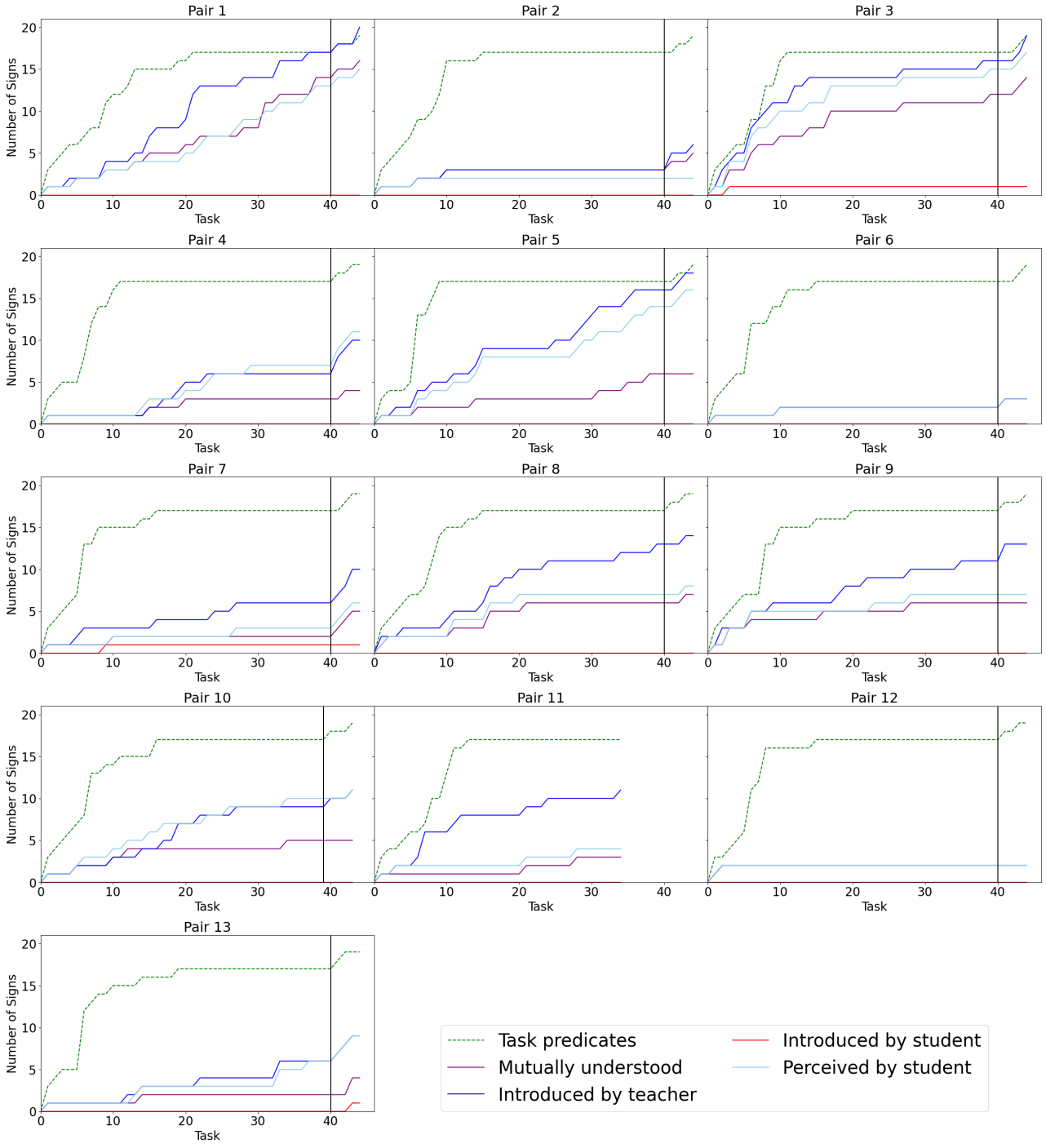


Figure A.8: Evolution of player lexicon over task index for Pairs 1-13. Pairs are numbered in descending order of weighted score. We consider a sign to be *introduced* by task t if a player uses a sign and registers it in the reflection form at some $t' \leq t$; The set of *task predicates* are those that teachers encounter in a task description; A sign is *mutually understood* by task t if that sign has been registered in the student's reflection form (indicating understanding) at some $t' \leq t$, and similarly for the teacher; A sign is *perceived* by the student by task t if the sign has been registered as a new teacher-introduced sign (with any degree of uncertainty) in the student's reflection form at some $t' \leq t$.

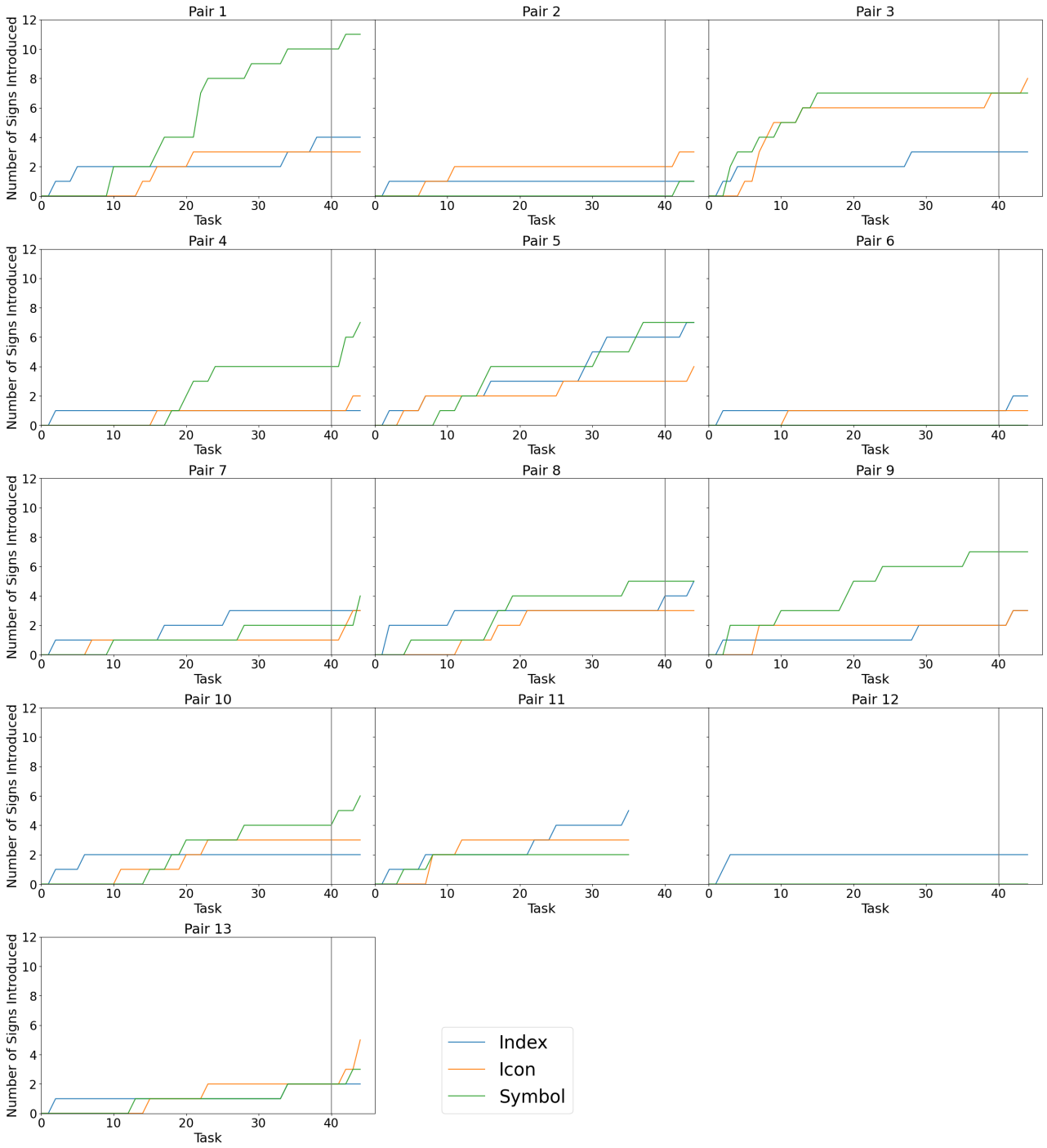


Figure A.9: Sign introductions over task index for Pairs 1-13, split by sign category. Pairs are numbered in descending order of weighted score. In each pair, the order of saturation is indices, then icons, then symbols with the exception of Pairs 7 and 11.

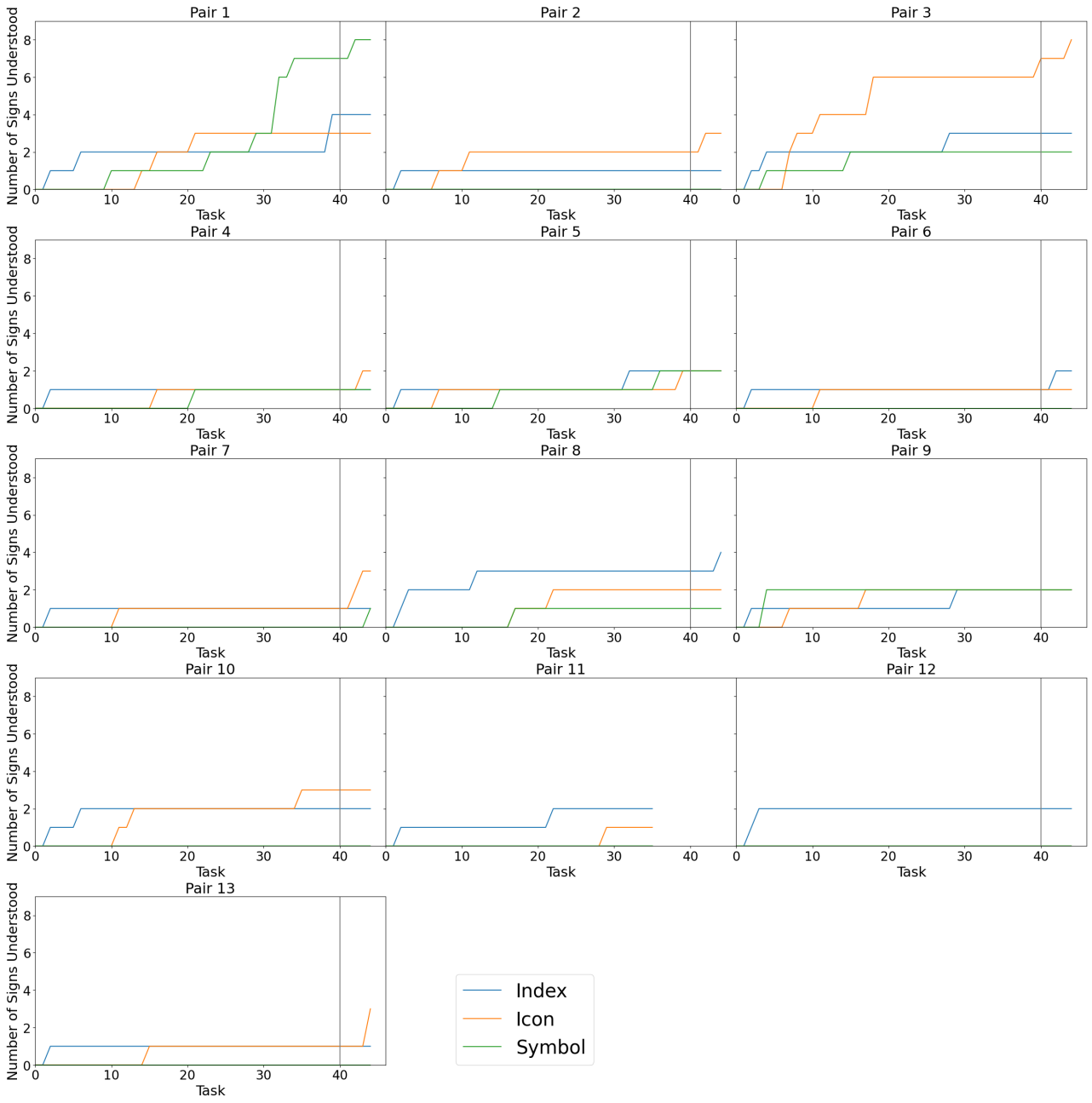


Figure A.10: Mutually understood signs over task index for Pairs 1-13, split by sign category. Pairs are numbered in descending order of weighted score. In each pair, the order of saturation is indices, then icons, then symbols with the exception of Pairs 3, 8, and 9. In all pairs, indices saturate before non-indices.

Pair Performance Data

Pair	Avg. Weighted	Avg. Raw	Avg. # Attempts	Correct Guesses	# Signs	Symbol	Icon	Index
1	**2.09**	**2.57**	1.68	**0.45**	**16**	**9**	3	4
2	**2.04**	2.48	2.0	0.36	5	0	**4**	1
3	**1.82**	**2.84**	2.68	**0.52**	**14**	2	**8**	4
4	1.69	2.29	1.93	0.43	4	1	2	1
5	1.59	1.84	1.5	0.11	6	2	2	2
6	1.50	**2.72**	3.89	N/A	3	0	1	2
7	1.41	1.80	2.11	0.34	5	1	3	1
8	1.37	2.32	3.73	**0.55**	**7**	1	2	4
9	1.26	1.63	1.73	0.18	6	2	2	2
10	1.20	1.45	1.51	0.15	5	0	3	2
11	1.08	1.38	2.09	0.00	3	0	1	2
12	1.06	1.18	1.23	0.18	2	0	0	2
13	1.03	1.23	1.41	0.18	4	0	3	1

Table A.1: Breakdown across pairs of average weighted score, average raw score, average number of attempts per task, percent correct student guesses, and mutually understood signs across the entire game (regular and generalization sessions). A student guess is *correct* if its denotation is the same as that of the task, given the universe of attributes in table 1. Pair 6 is labelled N/A because the student did not understand the question prompt in the reflection form and submitted irrelevant answers.

Conditions For Teacher Sign Introduction

	Base	Inconvenient	Ambig. Attribute	Ambig. Quantifier	Blank
Avg. # Signs	0.27 (0.04)	0.12 (0.03)	0.20 (0.04)	**0.30** (0.01)	**0.59** (0.1)
Avg. # Immediate	**0.10** (0.02)	0.01 (0.001)	0.04 (0.02)	0.02 (0.01)	**0.29** (0.06)
Avg. # Delayed	0.17 (0.04)	0.11 (0.03)	0.16 (0.04)	**0.28** (0.05)	**0.31** (0.07)
Avg. % Indices	**0.53** (0.08)	0.19 (0.10)	0.23 (0.09)	0.23 (0.09)	**0.36** (0.12)
Avg. % Icons	0.17 (0.05)	**0.56** (0.13)	0.15 (0.08)	**0.44** (0.09)	0.29 (0.08)
Avg. % Symbols	0.30 (0.07)	0.25 (0.11)	**0.62** (0.11)	0.33 (0.06)	**0.35** (0.11)
Total # Indices	19	3	6	6	6
Total # Icons	8	5	3	15	11
Total # Symbols	18	4	14	13	13
Total Signs	45	12	23	34	30

Table A.2: We analyze the teaching map conditions under which signs are introduced by the teacher (note that sampling a base, inconvenient, ambiguous attribute, or ambiguous quantifier map is equally likely). Introduced signs are either *immediate* or *delayed*, with immediate signs representing 27% and delayed signs representing 73% of all introductions. Delayed signs are more likely to be introduced in the ambiguous quantifier or blank maps. In the first section, the average number of signs introduced per task is shown with the standard error. Comparing across rows, we see that in the regular session, the highest proportion of immediate signs is introduced in base maps, while the highest proportion of delayed signs is introduced in ambiguous quantifier maps. In the second section, for each pair and type of map we compute the proportion of introduced signs that are indices, icons, and symbols, then report the averages and standard error across all pairs. We observe that base maps have the highest average proportion of index introduction, while inconvenient maps have the highest proportion of icon introduction and ambiguous attribute maps the highest proportion of symbol introduction. In the third section, we report the total number of indices, icons, and symbols introduced, aggregated across pairs. Much fewer signs are introduced in inconvenient maps compared to other maps, despite their equal representation in the task sequence.

% of Mutually Understood Non-indices vs. Weighted Score

% Non-indices	Base	Inconvenient	Ambig. Attribute	Ambig. Quantifier	Blank
$> 67\%$	1.97	1.93	1.35	1.46	0.93
$\leq 67\%$	1.83	1.80	0.99	0.87	0.27

Table A.3: The median proportion of mutually understood non-indices in a sign set is 67%. We report the average weighted score split by teaching map distribution for the pairs whose proportion of non-indices is greater than the median and whose proportion is less than or equal to the median. The average difference in weighted score between unambiguous (base and inconvenient) and ambiguous (attribute and quantifier) maps is 0.55 for the first group and 0.89 for the second group. To compute the average difference, and noting that the number of base, inconvenient, ambiguous attribute, and ambiguous quantifier maps are the same, we average the average weighted scores of the unambiguous maps, e.g. (1.93 + 1.97)0.5 = 1.95, and the average weighted scores of the ambiguous maps, e.g. (1.35 + 1.46)0.5 = 1.40, and subtract them. There are 6 pairs in the first category and 7 pairs in the second.

Rubric for scoring student guess

1	2	3	4	5
$G \cap T = \emptyset$	$G \subset T$	$G = T$	$G \supset T$	$\exists\, i \neq j \; s.t. \; (G_i \subset T_i) \wedge (G_j \supset T_j)$
No relation	G less general	Equivalent	G more general	G both more and less general

Table A.4: Rubric for scoring student guess G with respect to original task T. Let $A \subset B$ mean that task A is *less general* than task B, that is, to satisfy A is to satisfy B, given the universe described in table 1. Let T be the target task and G be the student guess. Note that, since **action**, **quantifier**, **color**, **size**, and **shape** in tasks are disentangled, we may define each task T and guess G as a 5-dimensional vector indexable as T_i, G_i, respectively. We grade each student guess, where a score of 2 or 3 implies the pair will generalize to satisfy arbitrary test maps for that task, a score of 4 or 5 implies the pair will sometimes generalize, and a score of 1 implies the pair will not generalize.

Probabilistic Task Grammar

S	→ Filter then Action	0.7	Con	→ or	0.5
	→ S Con S	0.2		→ and	0.5
	→ S and NegS	0.1	Color	→ red	0.34
NegS	→ Filter then NegAction	0.5		→ blue	0.33
Filter	→ for Quantifier x, Attribute' (x)	1		→ green	0.33
NegAttribute	→ Attribute	0.5	Size	→ big	0.5
	→ not Attribute'	0.25		→ small	0.5
	→ NegAttribute or NegAttribute	0.25	Shape	→ square	0.5
Attribute'	→ Size'	0.6		→ triangle	0.5
	→ Attribute' or Attribute'	0.4	Action	→ touch	0.34
Size'	→ Color'	0.5		→ touch going forwards	0.33
	→ Size (SizeArg)	0.5		→ touch going backwards	0.33
Color'	→ Shape(x)	0.5	NegAction	→ avoid	1.0
	→ Color (ColorArg)	0.5	Quantifier	→ all	0.25
Shape'	→ Shape(x)	1.0		→ exactly one	0.25
SizeArg	→ x	0.7		→ exactly two	0.25
	→ Color'	0.3		→ at least one but not all	0.25
ColorArg	→ x	0.7			
	→ Shape'	0.3			

Table A.5: Probabilistic grammar for task generation, where nonterminals are capitalized and terminals are lowercase. As the grammar is recursive, we impose a negative bias on sample depth so tasks are simpler. We do so by first reweighting the highest daughter node probability p as $p \leftarrow \min(\alpha p, 1.0)$, $\alpha = 1.1$ after each sample, then re-balancing the remaining daughter nodes' probabilities proportionally. Furthermore, at runtime, we cull object attributes from the grammar according to context so that sampled tasks are nontrivial. For example, in tasks composed of two subtasks and a conjunction, if we sample "red" in one subtask, we remove it from the grammar in sampling the other subtask. This avoids producing tasks like *Touch all objects that are red, and avoid all objects that are red.*