

Understanding Symbolic Communication

by

Emily Cheng

S.B., Mathematics, Computer Science and Engineering
Massachusetts Institute of Technology (2020)

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author.....
Department of Electrical Engineering and Computer Science
December 9, 2021

Certified by
Boris Katz
Principal Research Scientist
Thesis Supervisor

Certified by
Andrei Barbu
Research Scientist
Thesis Supervisor

Accepted by
Katrina LaCurts
Chair
Master of Engineering Thesis Committee

Understanding Symbolic Communication

by

Emily Cheng

Submitted to the Department of Electrical Engineering and Computer Science
on December 9, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

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 abstract symbols in order to accomplish complex tasks. In our experimental setup, a teacher must communicate an abstract notion, a formula in first order logic rendered to them in natural language, to a student. Subjects do so through a narrow channel that deprives them of common shared symbols: they cannot see or speak to one another and must only communicate via the motions of cars in a computer game. We observe that subjects spontaneously develop a shared vocabulary of car motions for task-specific concepts, such as “square” and “forall”, as well as for task-agnostic concepts such as “your turn”. We find that symbols are harder to establish than icons and indices, and that systematically, indices develop before icons, which develop before symbols. We characterize the conditions under which indices, icons, and symbols arise, and identify communicating in ambiguous game environments as the primary pressure for icon and symbol development.

Thesis Supervisor: Boris Katz
Title: Principal Research Scientist

Thesis Supervisor: Andrei Barbu
Title: Research Scientist

Acknowledgments

This project would not have been possible without the support of many people. I am grateful to my thesis advisors Boris Katz and Andrei Barbu, whose ideas led to this project, who have patiently guided me in reading my numerous drafts, and who have believed in me during my first attempt at academic research. Thanks to Ignacio Cases for helping me shape the direction and discourse of this project, for spending 5 hours piloting my game, and for his invaluable insights in linguistics, semiotics, writing, and broadly life. Finally, infinite thanks to Yen-Ling Kuo, my extremely competent PhD student advisor, for a long and ever-expanding list of things that would be twice the length of my thesis, which includes her close technical guidance (especially Javascript), spending not only 5 hours piloting my game but countless more testing it, invaluable discussions of ideas, and moral support.

Contents

1	Introduction	17
1.1	Origins of Human Language	17
1.2	Icons, Indices, and Symbols	18
1.3	Probing Emergence of Sign Communication	19
1.4	Contributions	21
2	Emergent Communication Game	23
2.1	Game Mechanics	24
2.2	Task Space	25
2.3	Game Maps	27
2.4	Solution Generation and Verification	28
3	Human Experiments	31
3.1	Experimental Setup	32
3.2	Emergence and Development of Signs	32
3.3	Establishing a Sign	36
3.4	Iconic, Indexical, and Symbolic Signs	39
3.4.1	The Symbolic Gap	42
3.5	Compositionality	46
3.6	Generalization	48
4	Conclusion and Future Work	51
4.1	Future Work	52

A Tables	53
B Figures	57
C Probabilistic FOL Task Grammar	65

List of Figures

- 2-1 The game flow for one task. The teacher (light blue circle) and the student (light red circle) start in a teaching map containing several objects (left). The teacher is given the task expressed in natural language but the student does not have access to this knowledge. The teacher must communicate the task to the student via car movements only. When the teacher is finished teaching, the student is moved to a new set of test maps to perform the task alone (right) while the teacher watches how the student performs the task. At any point during testing, the teacher may choose to move the student back to the teaching map to reteach the task. At the last test map, the teacher may advance the pair to the next task. 23
- 2-2 Example game maps for *Touch all objects that are square*. From left to right, maps are sampled from the *base*, *inconvenient*, *ambiguous predicate*, and *ambiguous quantifier* distributions. 28

3-1 Evolution of player lexicon over task index 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 $t' \leq t$; The set of *vocabulary tokens* in a task description are those in Table 2.1; 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 $t' \leq t$; A sign's meaning is *changed* at task t if in either player's reflection form for t , the player updates the sign's description. We consider two registered actions sufficiently similar to be the same sign if their natural language descriptions have the same meaning and the drawings are similar. All plots except for *meaning changed* are cumulative. 33

3-2 Comparative view of cumulative mutually understood signs over tasks for Pairs 1-13, where the pairs are sorted in decreasing order of weighted score. High scoring pairs (green) are closer to the top and low scoring pairs (red) are closer to the bottom (see table 3.2). We observe a separation of trajectories into three groups by the end of the regular session ($x = 40$), roughly corresponding to color: number of signs $y > 10$, between $4 \leq y < 10$, and $y < 4$ 35

3-3 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. 37

3-4	All transitions between sign categories due to re-introductions and repurposes, visualized with counts on edge labels, where <i>symbolic</i> is abbreviated <i>symp</i> . 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 <i>symbolic</i> and 3 to the self-edge of <i>index</i> . The remaining transitions are attributed to repurposes. Note that we only consider morphological or semantic changes to signs; the vast majority of the 149 sign introductions remain in their original sign category by virtue of not being changed.	38
3-5	The frequency of signs developed by pairs in the regular session (left) and bonus session (right), where signs are ordered by lexical category. There were no signs developed for categories <i>iconic and indexical</i> as well as <i>symbolic, iconic, and indexical</i> . Overall, pairs cover 14/18 vocabulary tokens in the regular session. The same plots for introduced and perceived signs are found at fig. B-1 and fig. B-2, respectively.	40
3-6	Several teaching map traces and corresponding signs registered by Pairs 1 (index and symbol) and 3 (icon). Each row is (left to right) a teaching map, a sign registered by the teacher immediately thereafter, and the corresponding sign registered by the student.	41
3-7	Cumulative number of signs introduced by all pairs, split by sign category. .	43
3-8	Cumulative number of signs understood by all pairs, split by sign category.	43
3-9	Distributions of the total # of uses (Z-score) of a sign until perceived and understood, split by sign category across all pairs. Symbolic signs are by far the hardest to notice and understand. A non-normalized version of this figure may be found at fig. B-3.	44

3-10	Linguistic composition across spatial and temporal dimensions. In the spatial dimension, composition occurs by adding modifications to a primary form. We observe SQUARE as a primary form and BIG superimposes morphologically to form BIG SQUARE. Along the temporal dimension, we illustrate the typical topic-comment syntax used by pairs in negation. All images are taken from the reflection forms of Pair 3.	47
B-1	The frequency of signs introduced by pairs in the regular session (left) and bonus session (right). There were no signs introduced for the category <i>symbolic, iconic, and indexical</i>	57
B-2	The frequency of signs perceived by pairs in the regular session (left) and bonus session (right). There were no signs perceived for the category <i>iconic and indexical</i> , nor <i>symbolic, iconic, and indexical</i>	58
B-3	Distributions of the total # of uses of a sign until perceived and understood, split by sign category across all pairs. Symbolic signs are by far the hardest to notice and understand.	58
B-4	Comparative view of cumulative sign introductions over tasks for Pairs 1-13, where the pairs are sorted in decreasing order of weighted score. High scoring pairs (green) are closer to the top and low scoring pairs (red) are closer to the bottom (see table 3.2). We observe a separation of trajectories into three groups by the end of the regular session ($x = 40$), roughly corresponding to color: number of signs introduced $y > 15$, between $6 < y < 13$, and $y < 3$. . .	59

B-5 Evolution of player lexicon over task index for Pairs 1-13. 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 *vocabulary tokens* in a task description are those in Table 2.1; 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 $t' \leq t$; A sign's meaning is *changed* at task t if in either player's reflection form for t , the player updates the sign's description. We consider two registered actions sufficiently similar to be the same sign if their natural language descriptions have the same meaning and the drawings are similar. All plots except for *meaning changed* are cumulative. 60

B-6 Sign introductions over task index for Pairs 1-13, split by sign category. In each pair, the order of saturation is indices, then icons, then symbolic signs with the exception of Pairs 7 and 11. 61

B-7 Mutually understood signs over task index for Pairs 1-13, split by sign category. In each pair, the order of saturation is indices, then icons, then symbolic signs with the exception of Pairs 3, 8, and 9. In all pairs, indices saturate before non-indices. 62

B-8 Several teaching map traces and corresponding signs registered by Pairs 3 (symbolic/indexical), 7 (iconic/indexical), and 1 (symbolic/iconic). Each row is (left to right) a teaching map, a sign registered by the teacher immediately thereafter, and the corresponding sign registered by the student. 63

C-1 Probabilistic grammar for task generation. As the grammar is recursive, we impose a negative bias on sample depth so tasks are simpler. We do so by first multiplying the highest daughter node probability by a multiplicative factor of $\alpha = 1.1$ after each sample, then re-balancing the remaining daughter probabilities proportionally. Furthermore, at runtime, we cull predicates 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.* 65

List of Tables

- 2.1 The first-order logic task space of the game. There are 18 vocabulary tokens in the regular session, and 2 introduced in the bonus session (*italics*). 26
- 3.1 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 (predicate 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 predicate, 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. 35
- 3.2 We report the fraction of introduced signs that were ultimately perceived and understood. Symbolic signs were significantly harder to both perceive and understand than nonsymbolic signs, and pure symbols were the hardest to perceive and understand. Recall that *symbolic* refers to a sign with symbolic properties, while *symbol* refers to a purely symbolic sign. 42

- A.1 Breakdown across pairs of average weighted score, average raw score, percent correct student guesses, and mutually understood signs across the entire game (regular and bonus sessions). A student guess is *correct* if its denotation is the same as that of the task, given the universe of predicates in Table 2.1. Pair 6 is labelled N/A because the student did not understand the question prompt in the reflection form and submitted irrelevant answers. . . . 53
- A.2 We analyze the teaching map conditions under which signs are introduced (note that sampling a base, inconvenient, ambiguous predicate, or ambiguous quantifier map is equally likely). In the first section, the average number of signs introduced per task is shown with the standard error. 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 bonus setting. In the second section, for each pair and type of map we compute the proportion of introduced signs that are indices, icons, and symbolic signs, then report the averages and standard error across all pairs. In the third section, the total count of indices, icons, and symbolic signs over all pairs is shown. 54
- A.3 For all perceived signs (row 1), we report the number of uses until perceived in z-value. For all understood signs (row 2), we report the number of uses until understood. The standard error is reported in parentheses. Evidence of perception/understanding is taken from the student’s reflection form. 54
- A.4 For all meanings for which multiple forms were developed, we report the number of forms developed across all pairs, sorted roughly by lexical category. The first 7 signs are all symbolic and the following 2 signs are indices for all iterations of the sign. Pairs 3, 5, 8, 9, 11, and 13 developed multiple forms for one meaning. For all 13 duplicate forms (forms that are not the original for a meaning), 3 are introduced when the previous form is understood, 4 are introduced when the previous form is only perceived, and 5 are introduced when the previous form is neither understood nor perceived. 55

A.5 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 2.1. 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. 55

Chapter 1

Introduction

For centuries, humans have speculated about the origins of language. Though the precise conditions of its emergence remain uncertain, human language is thought to have arisen from a broader framework of symbolic communication [5, 7, 15]. In this text, we hope to shed light on the very nascent stages of human language by recapitulating the origins of symbolic communication in a controlled setting. To do so, we posit a novel symbolic communication game and observe the emergence and evolution of symbolic communication in human subjects over the course of gameplay.

1.1 Origins of Human Language

Thought to have coincided with the speciation of Homo Sapiens around 300,000 years ago [6, 7], the precise origins of human language remain a mystery. Putting forth theories ranging from pure human invention [26] to evolving from gesture [30], to Chomsky's Universal Grammar [9], philosophers, anthropologists and linguists have long attempted to piece together explanations from a scant fossil record. Dominant explanations of the 20th century posit that language co-evolved with the brain, that is, the history of language is a history of human cognitive evolution [15, 34, 22, 35]. In turn, the evolution of human cognition shapes, and is shaped by, the evolutionary trajectory of the human cognitive niche [34, 25].

A clue to the beginnings of language may be that human language is situated in a broader

symbolic culture that is not observed in animal communication [15, 11, 32]. Somewhere along the evolutionary path from zero signs to modern-day communication, humans have spontaneously developed *symbolic signs* defined by arbitrary, one-to-many mappings from form to meaning, while animals are observed to employ *non-symbolic signs*, characterized by non-arbitrary, one-to-one mappings from form to meaning. Investigating the origins of sign communication, both non-symbolic and symbolic, is then a crucial first step to better understand the early communication from which language grew.

1.2 Icons, Indices, and Symbols

Human communication makes use of not just symbols, but a wide array of *signs*. In fact, all communication is comprised of signs, where a sign consists of a mapping from form to meaning, or from *signifier* to *signified* in the terms of Ferdinand de Saussure [14]. Following a semiotic framework defined in Peirce's 1868 "On a New List of Categories" [33], we may classify signs as *iconic*, *indexical*, or *symbolic* as follows:

1. An *iconic* sign is a sign whose signifier bears physical likeness to its signified in a one-to-one mapping. For example, photographs, drawings, or onomatopoeia physically resemble their meaning.
2. An *indexical sign* is a sign whose signifier corresponds temporally or spatially to its signified in a one-to-one mapping. For example, the distinct calls of Titi monkeys in the wild refer to different nearby predators [8].
3. A *symbolic sign* is a sign whose mapping between signifier and signified is arbitrary and may be one-to-many [21]. For example, `dog` is an arbitrary token that refers to members of *Canis lupus familiaris*.

The sign categories are not necessarily disjoint. For example, a sign may have both iconic and symbolic properties. We refer to signs as being *an icon*, *an index*, or *a symbol* when they are primarily iconic, indexical, and symbolic, respectively. We refer to a sign as being *iconic*, *indexical*, or *symbolic* when the sign has iconic, indexical, and symbolic

properties but are not necessarily pure icons, indices, or symbols. Whenever we refer to *sign categories*, we refer only to Peirce's early notion that signs be iconic, indexical, or symbolic, though Peirce eventually defined hundreds of categories of signs [2].

Human communication employs a mix of iconic, indexical, and symbolic signs. However, the expressive power of human language is specifically made possible by symbolic signs [21, 15, 11, 24], which permit complex syntax in human language [15]. Furthermore, linguistic composition in symbols allows for *digital infinity*, the notion that a small set of symbols can describe an infinite set of meanings [10]. It would simply be impossible to describe infinitely many meanings using the one-to-one mappings in non-symbolic signs.

1.3 Probing Emergence of Sign Communication

Humans today are predisposed to communicate symbolically and, while they can be taught to use symbols, animals are not observed to naturally employ symbolic communication [21]. Therefore, a true, spontaneous development of symbolic communication in nature has not yet been observed [15, 21, 36].

In practice, research in the origins of symbolic communication seeks to, in a controlled environment, approximate the conditions under which language is thought to have emerged. There are two broad approaches in designing this controlled environment. The first is to computationally simulate the emergence of communication protocols in a multi-agent setting, and the second is to perform human experiments in the lab.

For example, agent simulation at the population level may entail a population of agents that must communicate in order to reproduce. Initialized with indexical communication protocols that are subject to random mutations, the population evolves to communicate iconically and later symbolically [21]. Other approaches use reinforcement learning agents that (symbolically) communicate in order to cooperatively solve various Lewis signalling games [4, 23, 27–29, 31]. While these studies evidence the emergence of a shared symbolic communication protocol among agents, they are limited in their faithfulness to the conditions under which human symbolic communication developed. Because these studies employ toy referential games, where players communicate under information asymmetry to pick a target

image, the set of tasks operate mainly over predicate logic. Humans and animals, however, communicate a set of concepts much broader than that of referential games. Furthermore, communication channels are often pre-specified to be closed, finite, discrete and symbolic. In contrast, the development of human symbolic communication *lifted* a discrete, symbolic communication channel from continuous body movements.

The other approach to study the origins of symbolic communication directly uses human subjects. A number of studies in cognitive science have investigated the spontaneous development of a shared code in human communication after removing existing communicative conventions [13, 17, 18, 37]. In these studies, subjects play a cooperative communication game, such as a referential task, without being able to speak, see, or write to each other. Through the shared task, subjects are motivated to develop their own communicative codes, such as sequences of discrete spatial movements on a 2D grid, thereby producing novel symbolic communication systems [13].

Our experiment, a cooperative two-player game, similarly probes how humans spontaneously develop a shared sign vocabulary and reveals the emergence of a symbolic sign system. Like previous studies, we remove familiar modes of symbolic communication such as writing or speaking from the game. A teacher must communicate a task, encoded in a natural language utterance, to a student through continuous spatial movements in a teaching environment. The student then carries out that task in novel test maps. Similar to [37], the players' communicative behavior is embodied in their movements in the game world, which implicitly tasks the players to create their own communication channel and detect communicative intent.

Our experiment is novel in two ways. First, the space of communication is greatly expanded from that of previous studies: it occurs over a channel of continuous rather than discrete spatial movements, and operates over a larger and more abstract task space specified by first order logic (FOL) propositions that is unknown to the student. This allows us to explain sign development over a broad range of task and movement complexity. Second, it is believed that increased feedback and interaction between interlocutors results in more effective communication [13, 18]. Therefore, we allow unlimited teaching time per task, removing restrictions in the amount of feedback and interaction between teacher and student.

1.4 Contributions

We posit a game environment and incentives that communicate abstract notions, encoded in first order logic, through motions which have no existing shared symbolic lexicon, encouraging the development of new abstractions (chapter 2). We have conducted a human experiment with thirteen teacher-student pairs (chapter 3), where we evidence the notion of the *symbolic gap*, showing not only that symbols are much harder to establish than non-symbols, but also that indices develop first, then icons, then symbols. We illuminate the conditions under which each category of sign develops, demonstrating that new vocabulary, ambiguous environments, and shifts in task distribution are primary motivators. All pairs employ some degree of temporal or spatial compositionality, where how a sign is composed is determined by its lexical category. Broadly, the development of icons and symbols over indices correlates to better performance and generalization. We conclude with a discussion of future directions including extending to artificial agents (chapter 4).

Chapter 2

Emergent Communication Game

We have designed a communication game that motivates the emergence of meaningful communication between players (see Fig. 2-1). This communication is grounded in the game environment and its quality can be measured quantitatively by performance on tasks.

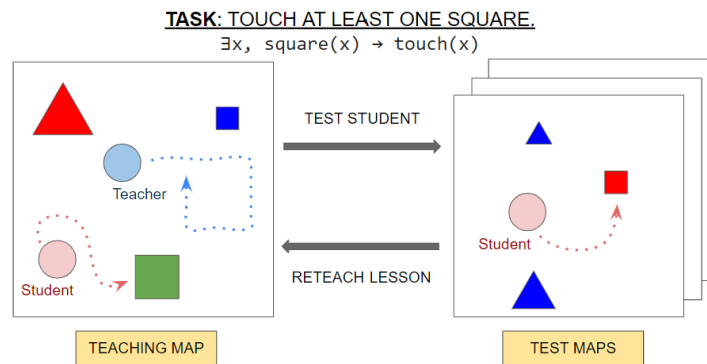


Figure 2-1: The game flow for one task. The teacher (light blue circle) and the student (light red circle) start in a teaching map containing several objects (left). The teacher is given the task expressed in natural language but the student does not have access to this knowledge. The teacher must communicate the task to the student via car movements only. When the teacher is finished teaching, the student is moved to a new set of test maps to perform the task alone (right) while the teacher watches how the student performs the task. At any point during testing, the teacher may choose to move the student back to the teaching map to reteach the task. At the last test map, the teacher may advance the pair to the next task.

2.1 Game Mechanics

On separate computers, a teacher and a student each control a 2D car avatar that navigates a set of maps containing several objects. Players 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 players need to segment the start and end of the signs and negotiate their meanings through gameplay.

The game works as follows:

1. The teacher and student begin in a teaching map containing several objects, where the teacher communicates the task to the student via spatial movements only.
2. Once the teacher is done teaching, the student goes on to three test maps for the same task to execute the actions while the teacher watches. At any point in testing, the teacher may choose to reteach, returning the pair to the teaching map to repeat the task with the same sequence of maps. At the last test map for the task, the teacher may choose between reteaching the lesson and moving onto the next task.
3. The teacher and student are rewarded a point for each successfully completed test map. The teacher sees the total score throughout the game, and the student sees the total score once per two tasks.
4. After each task (one teaching map + three test maps), each player fills out a reflection form where they draw and describe the actions they used to communicate and select whether previously registered actions are used, answering the questions “What does this action mean? When was it used? Who used it?”. To avoid biasing players, we use the word *action* instead of *sign* or *symbol* (however, we post-interpret all registered *actions* as *signs*). Both players also submit a description of the overall communication in this round, detailing e.g. whether the player was confused, or the teaching strategy if the player was a teacher. In addition, the student submits a guess of the task description they were just asked to perform. Players do not have access to each others’ reflections nor registered actions.

A game session repeats these steps with forty tasks so the players develop a shared set of signs over time. The game motivates cooperative sign development in the following ways: (i) players cannot communicate with the sole exception of moving the cars on a shared 2D map, requiring new modes of communication; (ii) players are equally incentivized by student performance on test maps, which implicitly rewards good communication; (iii) players play through numerous tasks and reflections with the same partner, which reinforces sign meaning and offers continued opportunity to evolve the shared sign sets. The players additionally play through four novel bonus tasks in which the teaching maps are blank, allowing us to measure the generalization capabilities of their sign sets.

Since players may only communicate through spatial motions, they must lift their own communication channel from the regular space of movement. This differs from prior emergent communication experiments in which the communication channel is pre-specified, and more closely approximates the conditions under which human communication originated [37]. Tasking players with negotiating their own communication channel is more difficult than providing a dedicated communication channel, as players must additionally detect whether motions are communicative and determine their start and end.

Prior to the game start, the student is not privy to the task space. The players are only given instructions about their roles– the teacher is told that they must communicate a set of natural language tasks to the student, and the student is told that they must learn a set of natural language tasks from the teacher and perform these tasks on a set of test maps. Both players are shown a practice round where they individually complete a task on one practice map in order to test game controls.

2.2 Task Space

A task in the game is a first-order logic (FOL) formula expressed in natural language. We want tasks to be easy enough for the teacher to understand, but hard enough to eventually require the development of symbols and exhibit some degree of linguistic compositionality. Our task space greatly expands the space of tasks of prior referential games by introducing

actions and quantifiers. An atomic FOL task takes the form

$$\text{for } \mathbf{quantifier} \ x, \mathbf{predicate}(x) \rightarrow \mathbf{action}(x).$$

More complex tasks can be constructed using logical connectives **and** (\wedge), **or** (\vee), or **not** (\neg). We use connectives to combine multiple predicates to describe objects or construct complex tasks from atomic formulas. The task space is described in Table 2.1.

Predicates	Color: red, blue, green Shape: square, triangle Size: big, small Pattern: <i>partially filled</i>
Actions	touch, touch going forwards, touch going backwards, avoid
Quantifiers	all (\forall), at least one (\exists), exactly one, exactly two, <i>exactly three</i>
Logical Connectives	and (\wedge), or (\vee), not (\neg)

Table 2.1: The first-order logic task space of the game. There are 18 vocabulary tokens in the regular session, and 2 introduced in the bonus session (italics).

Tasks are sampled from a probabilistic grammar (Appendix C) and are classified into four levels of complexity. Level 1 consists of touching **all** objects with single predicates; for example *Touch all objects that are red*. Level 2 adds one modification to an action or a quantifier; for example *Touch at least one object that is red*, or introduces one predicate composition such as *Touch all objects that are red and square*. Level 3 introduces multiple such modifications, and Level 4 uses logical connectives at the task level. The levels of the tasks are approximately distributed as 10% level 1, 30% level 2, 50% level 3, and 10% level 4.

Let the set of tasks in levels 1-4 be *regular tasks*. We additionally design a fixed set of four *bonus tasks* that test the zero-shot generalization of player sign sets as follows:

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.

The bonus tasks introduce an additional two tokens, **three** and **partially**, bringing the total vocabulary size to 20.

2.3 Game Maps

Given a task, we generate a teaching map and three test maps, all measuring 450×450 pixels. Each object on a map is defined by its shape, size, color, and position. The space of object attributes is the predicate space in Table 2.1.

So that the task is satisfiable in the generated map, we extract the set of predicates required to satisfy the formula and accordingly place N objects with these attributes on the map at random. We also place M objects of random attributes on the map as distractors. Distractors are important especially for test maps to measure student comprehension. For example, if the task is *Touch at least one square* and there are only squares on the test map, the student trivially succeeds and we gain no information about their level of understanding. 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 $\text{Unif}[1, 3]$.

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

- *Ambiguity:* We consider two forms of ambiguity by modifying distributions to sample objects. (1) Ambiguity in the quantifier: we place one target object ($N = 1$) so that it is more difficult to teach **all** vs. **at least one**, **exactly one**, and **exactly two**. (2) Ambiguity in the predicate: we place no distractors ($M = 0$) so that the teacher cannot use them as negative examples to discriminate the solution set.

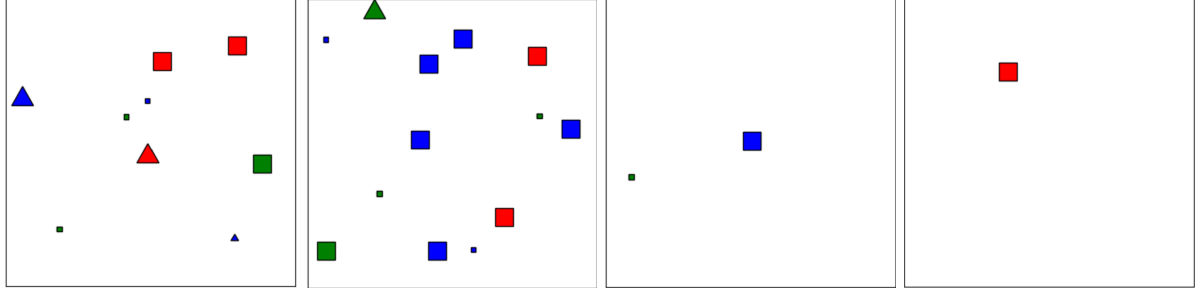


Figure 2-2: Example game maps for *Touch all objects that are square*. From left to right, maps are sampled from the *base*, *inconvenient*, *ambiguous predicate*, and *ambiguous quantifier* distributions.

- *Inconvenience*: We set $\mathcal{D}^{\text{required}} = \text{Unif}[9, 11]$ to have a large number of target objects on the map. In maps with inconveniently many target objects, rather than touch all target objects the teacher may use a sign for brevity.

For regular tasks, teaching maps are sampled from *base*, *ambiguous in the quantifier*, *ambiguous in the predicate*, and *inconvenient* distributions with equal probability (Fig. 2-2). In general, test maps are sampled from the *base* distribution only. However, for tasks that are composed of multiple subtasks by disjunction like “*Touch all squares, or touch all red objects*”, test maps are sampled disjointly for each subtask. That is, it would be impossible to achieve a perfect score on all three test maps if the teacher were to simply teach one of the subtasks. For each bonus task, the teaching map is completely blank and we hand-design three fixed test maps.

2.4 Solution Generation and Verification

For each task and map, we define the solutions and nonsolutions as follows. Let $\mathcal{S}^{\mathcal{M}, \mathcal{T}}$ be the collection of all subsets of (object, direction) pairs on map \mathcal{M} which satisfy task \mathcal{T} . Then, the *minimal solution sets* $\mathcal{S}^{\mathcal{M}, \mathcal{T}}_{\min} = \{s \mid s \in \mathcal{S}^{\mathcal{M}, \mathcal{T}}, \nexists r \in \mathcal{S}^{\mathcal{M}, \mathcal{T}} \text{ s.t. } r \subset s\}$. Likewise, $\mathcal{N}^{\mathcal{M}, \mathcal{T}}_{\min}$ is the *minimal nonsolution sets*, or the collection of minimal subsets of objects on the map that automatically fail the task. For example, for the task *Touch exactly two objects that are red*, $\mathcal{S}^{\mathcal{M}, \mathcal{T}}$ is the collection of all subsets of two red objects, and $\mathcal{N}^{\mathcal{M}, \mathcal{T}}$ is the collection of all subsets of three red objects.

Given the parse tree for \mathcal{T} , minimal solution and nonsolution sets are generated as in Algorithm 1. At runtime, there is a verifier that is given $\mathcal{S}^{\mathcal{M},\mathcal{T}}$ and $\mathcal{N}^{\mathcal{M},\mathcal{T}}$ and runs each time the student collides with a new object in \mathcal{M} as follows: (1) if the student's visited set is a superset of a minimal nonsolution, the student automatically fails; (2) if not, then the student succeeds if their visited set is a superset of a minimal solution.

Algorithm 1: Recursive Minimal Solution Generation, where the subroutine MERGE is the standard set operation over quantifiers, conjunction, negation, and disjunction.

Let T be the parse tree of task \mathcal{T} .

Let \mathcal{M}' be a collection of size-one sets $\{\{m\}\}_{m \in \mathcal{M}}$.

procedure MINSOLUTION(T, \mathcal{M}')

if T ternary branching **then**

 ▷ E.g. $\wedge(\text{red}, \text{blue})$

$\mathcal{S}_{\min}^{\mathcal{M},\mathcal{T}} \leftarrow \text{MERGE}(T_1, \text{MinSolution}(T_2, \mathcal{M}'), \text{MinSolution}(T_3, \mathcal{M}'))$

else if T binary branching **then**

 ▷ E.g. all red

$\mathcal{S}_{\min}^{\mathcal{M},\mathcal{T}} \leftarrow \text{MERGE}(T_1, \text{MinSolution}(T_2, \mathcal{M}'))$

else if T terminal **then**

$\mathcal{S}_{\min}^{\mathcal{M},\mathcal{T}} \leftarrow \{m \mid m \in \mathcal{M}', T(m) \text{ is True}\}$

 ▷ E.g. red

$\mathcal{S}_{\min}^{\mathcal{M},\mathcal{T}} \leftarrow \text{MinSolution}(T, \mathcal{M}')$

Chapter 3

Human Experiments

The evolution of sign communication is core to how humans communicate with each other. Our human experiments reveal several trends in the evolution of sign communication. We not only find that all pairs develop signs, but that a sign’s meaning biases its category. Furthermore, the number of signs and specifically non-indexical signs correlates to performance (section 3.2). The negotiative process of sign establishment, which hinges on sign stability, is strongly influenced by sign category (section 3.3). We illustrate the *symbolic gap*, or the systematic establishment of first indices, then icons, and finally symbolic signs in that order over the sequence of the game; and characterize the conditions under which indices, icons, and symbolic signs are introduced. While linguistic compositionality is a mature feature of human language, we observe early compositionality in all pairs, and describe the factors that determine how a sign is composed spatially or temporally (section 3.5). Finally, we describe the factors that predict generalization of pairs to arbitrary test maps and to novel tasks in the bonus round (section 3.6).

Player reflection forms evidence the spontaneous development of signs for all but two pairs, who communicated by drawing English words. The following data analysis will only consider the remaining 13 pairs. Our descriptions of sign meaning in this section are taken from player reflection forms, and all data annotation requiring interpretation of player signs in context is performed by expert annotators recruited from the lab.

Performance data is displayed with player sign sets by sign category (see 3.4) in Table A.1. Because pairs are allowed as many reteaches as necessary to complete a task, raw score

is not necessarily indicative of communication quality. Instead, since we observe that more reteaches corresponds to learning by trial-and-error and therefore a lesser understanding of the task, we evaluate communication by weighting the raw score by the number of attempts. Both the average weighted score and the average raw score are bounded between 0 and 3 (three test maps with one point possible per test map).

3.1 Experimental Setup

We recruited 30 adult participants (15 pairs) to play the game. Eight participants were recruited from the lab, and the remaining participants were recruited from the MIT Brain and Cognitive Sciences professional subjects email list. Among participants, 24 were native English speakers and 6 were native speakers of Spanish, Urdu, Hindi, Tagalog, Hebrew, and Bosnian/Croatian/Serbian. Participants completed the game over a period of 3.3 ± 1.1 days, playing approximately 10-15 tasks per day and taking breaks whenever they chose. When rejoining the game, participants were able to pick up where they left off. The amount of time each pair spent was 2.8 ± 0.8 hours, and participants were paid \$20/hour of active gameplay.

We sample the first five tasks in the regular session from the level I distribution, and its teaching maps from the base distribution. We do so for two reasons: (1) to ease the teachers' transition to complex tasks and ambiguous teaching maps; and (2) so that players acclimate to the controls and game flow. However, we choose not to furnish an explicit *curriculum* over the course of the game and rather sample tasks at random from the full distribution of levels to minimize the bias of task sequence on resulting player sign sets. Finally, the four bonus tasks at the end of the game session are shown to pairs in a random order. The five task ramp-up and the bonus round make pairs adapt to two shifts in task distribution.

3.2 Emergence and Development of Signs

The players' shared sign sets converge over the course of the regular session — see Fig. 3-1 for example trajectories of introduced, perceived, and understood signs for two pairs, and

Appendix B for the remaining trajectories. In the regular session in Fig. 3-1, for example, Pair 3 virtually stops introducing new signs by task 15, changing sign meaning by task 17, and perceiving new signs by task 18. Similarly, Pair 2 stops introducing new signs by task 10 and perceiving new signs by task 6. Pairs generally then introduce novel signs in the bonus round when they encounter new vocabulary tokens.

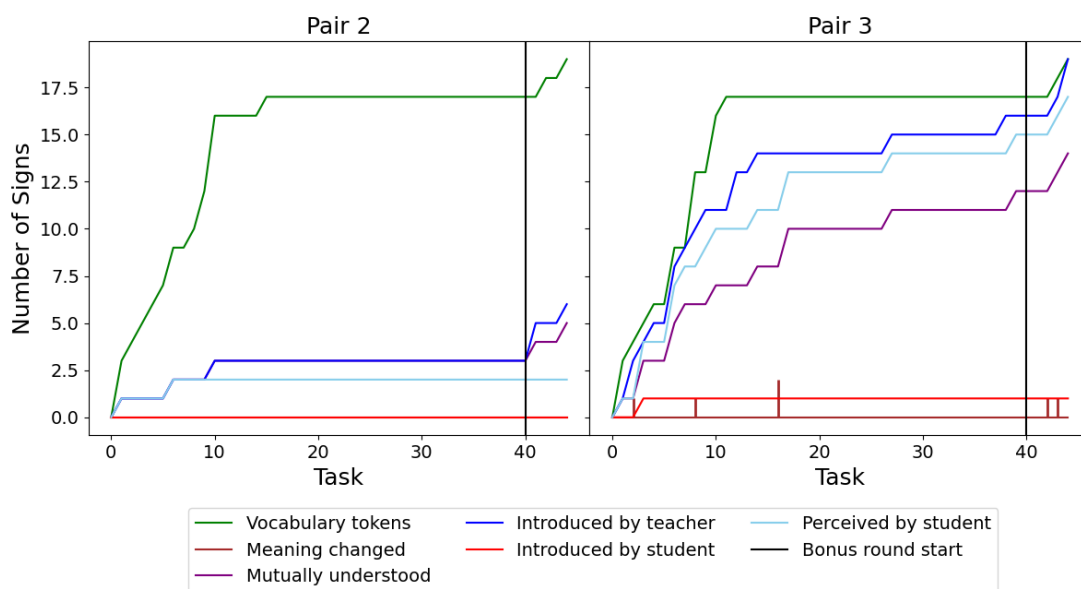


Figure 3-1: Evolution of player lexicon over task index for Pair 2 (left) and Pair 3 (right). We consider a sign to be *introduced* by task t if a player uses a sign and registers it in the reflection form at some $t' \leq t$; The set of *vocabulary tokens* in a task description are those in Table 2.1; 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 $t' \leq t$; A sign’s meaning is *changed* at task t if in either player’s reflection form for t , the player updates the sign’s description. We consider two registered actions sufficiently similar to be the same sign if their natural language descriptions have the same meaning and the drawings are similar. All plots except for *meaning changed* are cumulative.

Pairs are shown 20 vocabulary tokens, but all pairs develop fewer than 20 signs (Table A.1). They achieve this by linguistic composition (Section 3.5) and collapsing the task space. For example, in collapsing the task space, all pairs but one merged TOUCH with FORWARDS by assuming a default forwards direction, later introducing a sign for BACKWARDS (Fig. 3-5). Assuming a default also occurs for quantity; because the first five tasks concern

touching ALL objects of a certain attribute, pairs often assume ALL in their sign for TOUCH by default and later differentiate ONE and TWO. This not only illustrates that the order of vocabulary introduction biases the tokens for which signs are introduced, but also that new signs are introduced to disambiguate from established defaults.

New signs tend to be introduced under several circumstances. First, signs are introduced when encountering new vocabulary. We define a new sign as *immediate* if its meaning corresponds to a new vocabulary token seen for the first time, and all other signs are *delayed*. Twenty-seven percent of new signs are immediate. Systematically across pairs, immediate signs include TOUCH, introduced in the first task, as well as BACKWARDS. Because pairs tend to initially use TOUCH to indexically refer to solution sets, signs for object attributes and quantifiers may be delayed. Signs, especially delayed signs, are more likely to be introduced in teaching maps that are ambiguous (note that bonus maps are ambiguous in both predicate and quantifier) (Table A.2). See section 3.4.1 for discussion on the effect of sign category on performance in ambiguous maps.

Finally, the spike in sign introduction in the bonus round – which both introduces new vocabulary and where teaching maps are blank – supports our hypothesis that the pressures of new vocabulary and ambiguity are sufficient conditions for sign development (Fig. 3-1). See section 3.4.1 for further discussion of the effect of sign category on sign introduction.

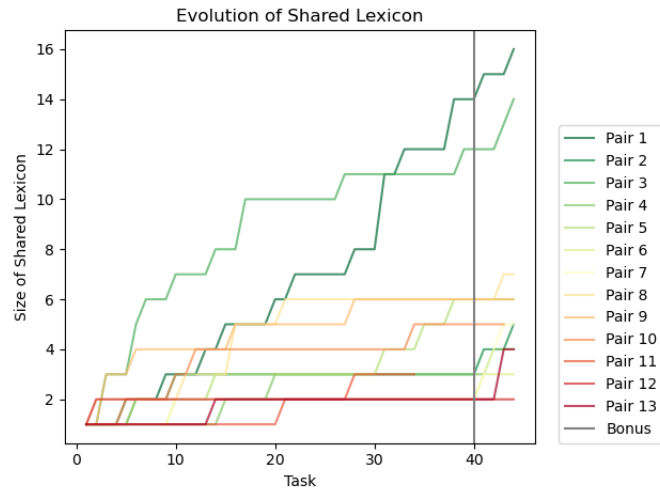
Player role also strongly influenced the type and quantity of signs introduced. Most of the time, the teacher initiated communicating task meaning to the student and the student tried to decipher task meaning from the teacher. Indeed, teachers introduced 146 / 149 total introduced signs, or 98 %. The meanings for which pairs developed signs, as well as their sign category, are shown in Fig. 3-5. Sign meanings can be classified into those describing vocabulary tokens, such as ALL or RED, and “meta-linguistic” signs such as CONFIRM UNDERSTANDING or EMPHASIS. Because the game periodically shows players their scores as reinforcement, it is not required that players develop meta-linguistic signs. The fact that five pairs organically developed these signs indicates the importance of social rewards or dense feedback for learning to communicate more efficiently.

% of Mutually Understood Non-indices vs. Weighted Score

% Non-indices	Base	Inconvenient	Ambig. Predicate	Ambig. Quantifier	Bonus
> 67%	1.97	1.93	1.35	1.46	0.93
≤ 67%	1.83	1.80	0.99	0.87	0.27

Table 3.1: 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 (predicate 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 predicate, 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.

Figure 3-2: Comparative view of cumulative mutually understood signs over tasks for Pairs 1-13, where the pairs are sorted in decreasing order of weighted score. High scoring pairs (green) are closer to the top and low scoring pairs (red) are closer to the bottom (see table 3.2). We observe a separation of trajectories into three groups by the end of the regular session ($x = 40$), roughly corresponding to color: number of signs $y > 10$, between $4 \leq y < 10$, and $y < 4$.



The trajectories of sign development across pairs empirically self-organize into several groups (figs. 3-2 and B-4): those who introduce and develop many signs early (green trajectories), those who introduce and develop few signs (red trajectories), and the remaining bulk of pairs in the middle (yellow/orange trajectories). This illustrates a positive correlation between number of signs a pair develops and its average weighted score ($\rho = 0.66$). We find that the size of sign sets, and thus the relationship between number of signs and weighted score, is driven specifically by the development of non-indices (see section 3.4.1).

3.3 Establishing a Sign

A sign undergoes a process of negotiation between speaker and listener until it is either established or discarded. Note that due to the asymmetry in player roles, the speaker is typically the teacher and the listener is typically the student. The process of negotiation 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 [37]. The speaker's internal negotiation of sign form depends recursively on their beliefs about steps 2 and 3 [19, 16]. In this section, we describe specifically the factors that contribute to the perception of a sign (step 2), the process of negotiating a mutual understanding (step 3), and, once established, what influences a sign's stability over time. The effect of sign category on all steps is discussed in section 3.4.1.



Qualitatively, whether a sign is perceived depends on whether its form departs from the range of typical car movements. In practice, for example, recognizing direction as meaningful or related to the task is difficult:

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.

In contrast, the teacher in Pair 7 later introduces a DIRECTION sign that is “turning in a full circle to emphasize the direction,” and this is perceived immediately. 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. 3-3). 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

Figure 3-3: 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.

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

understanding of its possible meanings. Simultaneously, the speaker updates their beliefs of the listener’s understanding and adjusts the sign’s form and meaning accordingly. A sign is mutually understood once the listener and speaker converge upon the same (form, meaning) pair. Players’ internal understandings of signs are documented in reflection forms, which evidence the iterative and reciprocal process of sign establishment. For example, in Pair 3, the teacher introduces an iconic sign for SMALL with the description ‘I (the teacher) used these (small) shapes to mean “touch all the small shapes”. ’ (fig. 3-6). 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.

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 13 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 (table A.4). Seven out of these nine signs are abstract and are all expressed symbolically, the exceptions being the indices EMPHASIS and PHANTOM (see section 3.4.1 for definition). Concrete concepts such as shape, which are typically icons fig. 3-5, are not represented in table A.4. Similarly, four pairs collectively attempt 16 repurposes over 10 unique introduced

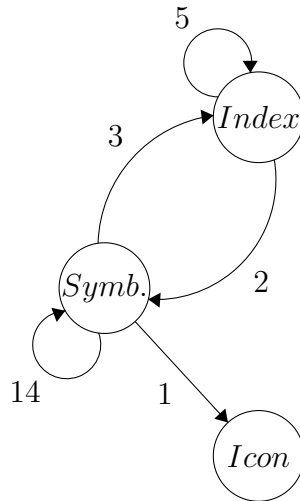


Figure 3-4: All transitions between sign categories due to re-introductions and repurposes, visualized with counts on edge labels, where *symbolic* is abbreviated *ymb.* We compute the transition counts by determining the categories of the original and updated sign for each re-introduction and repurpose. All re-introductions rest within the original sign category, contributing 7 to the self-edge of *symbolic* 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 149 sign introductions remain in their original sign category by virtue of not being changed.

forms. Nine such changes repurpose a symbolic sign, and the remaining seven repurpose an index. Again, icons are never repurposed in any pair, though Teacher 8 repurposes a symbolic sign, formed by drawing a triangle that (oddly) means COLOR, to mean an iconic TRIANGLE.

We were not able to establish whether a sign’s being perceived or understood affects its chance of re-introduction or repurpose (table A.4). Among the 13 duplicate forms, 3 are introduced when the previous form is understood, 4 are introduced when the previous form is perceived, and 5 are introduced when the previous form is neither understood nor perceived. Similarly, there are 16 repurposes, 7 of which repurpose understood forms, 3 of which repurpose perceived but not understood forms, and 6 of which repurpose unperceived forms.

The transitions between sign categories in all re-introductions and repurposes, labeled with counts, are summarized in fig. 3-4. Of all 165 (form, meaning) pairs that exist at some point in time, there are 82 symbolic signs, 49 indices, and 35 icons. However, all updates

occur in symbolic signs or indices regardless of whether they are perceived or understood.¹ In particular, there are 18 total updates to symbolic signs (22% of symbolic signs), 7 to indices (14%), and, notably, 0 to icons. This indicates that icons are the most stable, then indices, and lastly symbolic signs. In other words, the link from form to meaning in icons is the most rigid of all sign categories, and the link from form to meaning in symbolic signs the most tenuous. We expect that this is due to icons' physical resemblance to their referent, a strict requirement compared to merely a physical correspondence in indices and the arbitrary relationship in symbols.

3.4 Iconic, Indexical, and Symbolic Signs

Across all participants, we identify 80 mutually understood signs: 18 symbolic, 34 icons, and 28 indices (table 3.2). We find not only that the meaning of a sign systematically determines sign category, but also that symbolic signs are harder to learn than non-symbolic signs, and that indices, icons, and symbolic signs develop in that order.

Developing non-indices (icons and symbolic signs) corresponds to a higher weighted score; running a correlation between pairs' average weighted score and number of non-indices gives $\rho = 0.71$ ($n = 13$). The correlations between the number of indices, icons, and symbolic signs and weighted score are $\rho = 0.29, 0.51, \text{ and } 0.58$, respectively. While all pairs developed a maximum of four indices, the further development of icons and symbolic signs allows for a wider separation in weighted score (fig. B-6). This suggests that non-indices rather than indices are the primary factor in student generalization to test maps. Indeed, while indices must refer to objects that are spatially or temporally proximal, icons and symbolic signs, by virtue of their forms' uncoupling from physical environment, allow pairs to communicate information that is spatially and temporally displaced, and thus succeed even in ambiguous settings [24]. We observe a systematic connection between a sign's meaning and its sign category (figs. 3-5, B-1 and B-2). This relationship is a function of how signs are grounded in the game map and embodied in player motions. The effect of

¹165 = 149 introductions + 16 repurposes, where re-introductions are counted in sign introductions and repurposes are counted in meaning changes (fig. 3-1).

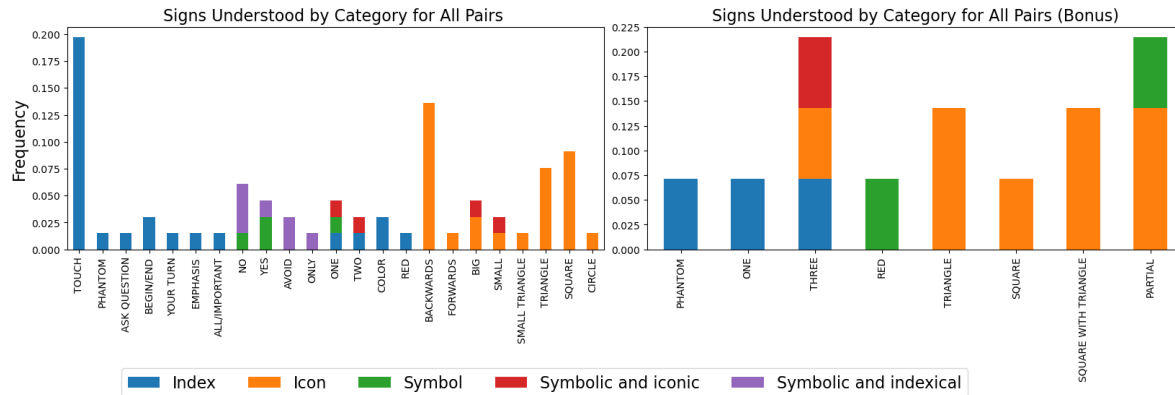


Figure 3-5: The frequency of signs developed by pairs in the regular session (left) and bonus session (right), where signs are ordered by lexical category. There were no signs developed for categories *iconic and indexical* as well as *symbolic, iconic, and indexical*. Overall, pairs cover 14/18 vocabulary tokens in the regular session. The same plots for introduced and perceived signs are found at fig. B-1 and fig. B-2, respectively.

grounding is clear, for example, in that the sign for TOUCH, where the teacher physically touches target objects in the teaching map, is indexical in all pairs and derives its meaning from *spatial proximity* to target objects. Some pairs also develop a PHANTOM index whose meaning is grounded in spatial and *temporal proximity* to target objects, where in reteaching, the teacher drives to the coordinates of target objects in the last test map. Similarly, the meaning for BACKWARDS, in which the teacher drives backwards on the teaching map, is iconic in all pairs and grounded in the *physical motions* of the teacher. When the teacher is unable to disambiguate the target set by a concrete demonstration, they develop e.g. iconic signs for shape. In shape icons, the sign's meaning is grounded in *physical likeness* to objects in the game maps, which allows the pair to communicate spatially and temporally displaced information in order to succeed in ambiguous maps. Finally, it is challenging to ground concepts that are not physically present in the game maps, such as color and conjunctions. For such abstract concepts, the sign's mapping from form to meaning may be arbitrary, thus we tend to see the teacher attempt more symbolic signs (fig. B-1). It is important to note that icons and symbolic signs are a reflection of the vocabulary space and the range of car avatar movement. For example, if our vocabulary included shapes harder to draw than triangles and squares, it is possible that shape signs would be indexical or symbolic. Likewise, if our vocabulary consisted only of concrete concepts like shape and

size, there may be fewer symbolic signs. We provide examples of several sign categories in figs. 3-6 and B-8.

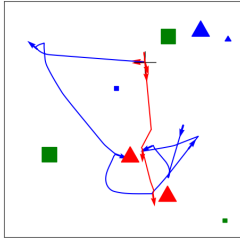
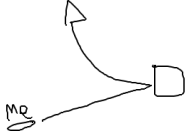
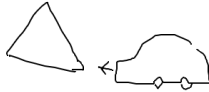
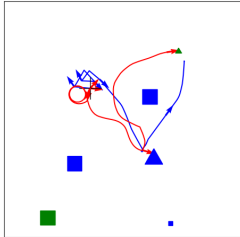


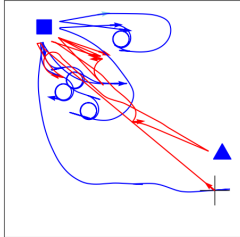


	Teaching Map	Teacher Response	Student Response
Index	 <p><i>Touch all objects that are red.</i></p>	 <p>I demonstrated touching the objects that she needed to touch.</p>	 <p>The teacher used his car to hit certain shapes during the teaching portion</p>
Icon	 <p><i>Touch all objects that are triangular.</i></p>	 <p>This action means “touch all the triangles in the environment”.</p>	 <p>Asking / indicating triangles.</p>
Symbol	 <p><i>While going backwards, touch exactly one object that is big.</i></p>	 <p>I moved around in a circle to indicate that the student’s actions were incorrect</p>	 <p>Car turns in circle; the teacher used it during the teaching portion, I think it means that I shouldn’t hit the shape that was just hit.</p>

Figure 3-6: Several teaching map traces and corresponding signs registered by Pairs 1 (index and symbol) and 3 (icon). Each row is (left to right) a teaching map, a sign registered by the teacher immediately thereafter, and the corresponding sign registered by the student.

3.4.1 The Symbolic Gap

Empirically, pairs prefer to convey meaning non-symbolically. For example, many pairs encounter a task such as *Touch all objects that are red* within the first five tasks of the game. We observe, for example, that the teacher chooses not to introduce an abstract symbol for RED, instead referring to RED indexically. Furthermore, all signs developed for SQUARE and TRIANGLE are iconic rather than symbolic. This *symbolic gap*, or a systematic bias towards the establishment of indices, and to a lesser extent, icons over symbols, mirrors the overwhelming prevalence of indexical and iconic signs in animal communication systems and early human communication (fig. 3-8)[15, 21]. Furthermore, the bias towards establishing indices and icons persists despite the fact that participants use symbols extensively in life and that teachers introduce symbolic signs at a constant rate (fig. 3-7). This motivates several questions: why do participants primarily develop non-symbols? Under what conditions do they introduce indices, icons, and symbolic signs?

Initially in the game, participants likely prefer indexical signs for pragmatic reasons. Recall that establishing an introduced sign occurs in two stages: first, the receiver needs to recognize a sign’s “signalhood”, then determine its meaning [37]. Likely due to their arbitrary mapping from form to meaning, symbolic signs are significantly *harder* to perceive and understand than both icons and indices (table 3.2), explaining the symbolic gap. Moreover, icons take significantly *longer* to perceive and understand than indices, likely due to the difficulty of tracing motions with the car (table A.3 and figs. 3-9 and B-3). Then, choosing to communicate via an index over an icon or a symbol, at least initially, is to choose clarity, a parallel to Grice’s Maxim of Manner [20].

	Nonsymbolic icon	Nonsymbolic index	Symbolic	Symbol
Fraction Perceived	0.90	0.83	0.62	0.58
Fraction Understood	0.81	0.68	0.28	0.17

Table 3.2: We report the fraction of introduced signs that were ultimately perceived and understood. Symbolic signs were significantly harder to both perceive and understand than nonsymbolic signs, and pure symbols were the hardest to perceive and understand. Recall that *symbolic* refers to a sign with symbolic properties, while *symbol* refers to a purely symbolic sign.

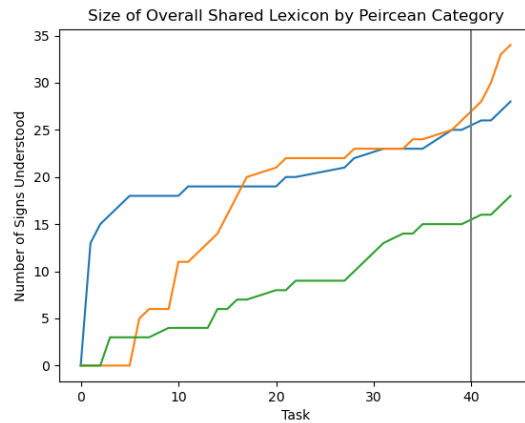
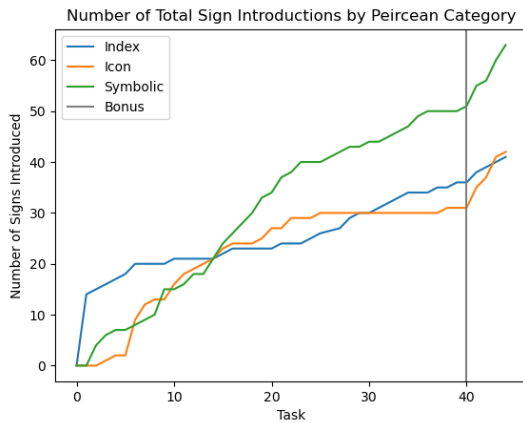


Figure 3-7: Cumulative number of signs introduced by all pairs, split by sign category. Figure 3-8: Cumulative number of signs understood by all pairs, split by sign category.

The subsequent process of sign introduction may be understood in a series of plateaus. Recall that pairs see all vocabulary tokens before task 20. We observe introduced and mutually understood indices plateau at around task 5, after which their rate of introduction is much slower (figs. 3-7 and 3-8). As indices plateau, non-indices continue to be introduced at a similar rate; and as icons subsequently plateau around task 20 in both introduction and understanding, symbolic signs continue to be introduced at a similar rate. Note the beginning of a plateau in symbolic signs around task 35, though it remains uncertain without more tasks. It is clear, however, that the growth in the size of the total shared lexicon is first attributed to indices, then icons, then symbols. We hypothesize that when pairs develop signs, (1) the fastest growing mutually understood sign category will *reach saturation*, that is, exhaust its communicative utility in the current context, and its introduction slows, making way for the next fastest growing category; and (2) the order of saturation of sign categories is first indices, then icons, then symbolic signs (see figs. B-6 and B-7). The effect of *desaturation* is seen in the uptick in introductions for icons and symbolic signs in the bonus round, where the task distribution shifts (fig. 3-7). Conversely, indices remain saturated in the bonus round as an index is minimally useful in a blank teaching map.

Sign categories saturate when the marginal utility of introducing a new sign of that category is low. We can extract why sign categories saturate by examining their response to contextual shifts. For indices, note that a distribution shift occurs at task 6; the first

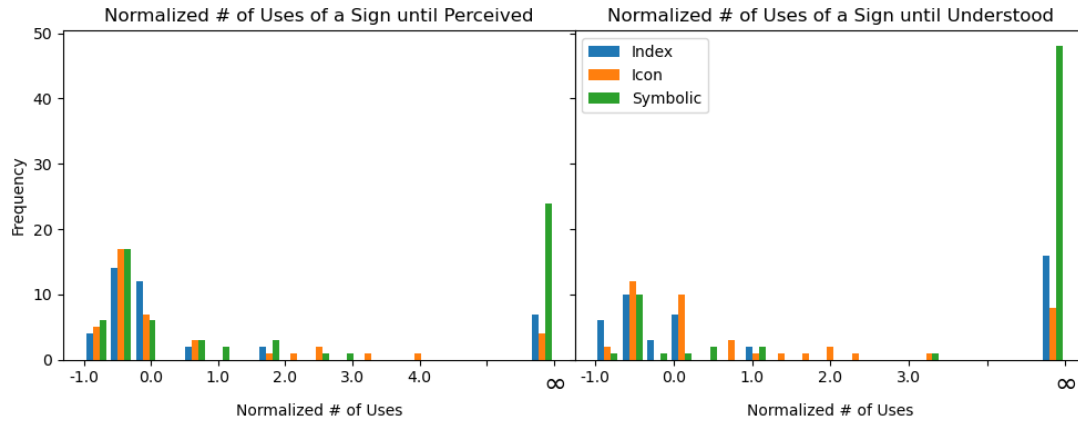


Figure 3-9: Distributions of the total # of uses (Z-score) of a sign until perceived and understood, split by sign category across all pairs. Symbolic signs are by far the hardest to notice and understand. A non-normalized version of this figure may be found at fig. B-3.

five tasks sample level I tasks and base teaching maps, where it is easy to disambiguate the target set indexically using TOUCH, and the following 35 tasks sample from harder levels and teaching maps. The saturation of indices coincides with this distribution shift, which implies additional indices would not confer additional advantage in harder levels and maps. The saturation in indices is therefore explained by what remains constant through the distribution shift, which is the usefulness of discriminating target sets by tapping or pointing. In cases where this indexical teaching strategy falls short, icons and symbolic signs become comparatively useful.

We find that ambiguous teaching maps in particular drive the continued introduction of icons and symbolic signs after the saturation of indices. In unambiguous teaching maps, quantifiers and object attributes may be disambiguated indexically using TOUCH. However, in ambiguous teaching maps, pairs can no longer perfectly discriminate the target set indexically. Instead, icons and symbolic signs are useful in this setting because they can communicate meanings that are decoupled from the immediate environment. For example, the teacher in Pair 1 introduces an iconic SQUARE in an ambiguous predicate 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 3.1, where for pairs who develop more than the median number of non-indexical signs, 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-indexical signs.

Perhaps due to their effectiveness, signs introduced in ambiguous maps tend to be non-indices rather than indices. On average, in ambiguous teaching maps non-indices comprise 77% of all sign introductions, where in ambiguous predicate maps, signs introduced tend to be symbolic in particular. In contrast, non-indices comprise 47% of sign introductions in base teaching maps. A similar phenomenon occurs in the bonus round where the teaching map is blank. For example, **red**, **square**, and **triangle** are all present in the regular session but are re-expressed iconically or symbolically in the bonus (Fig. 3-5).

Icons and symbols can communicate information more efficiently than indices in inconvenient teaching maps (imagine signing ALL SQUARE instead of touching every square). However, contrary to expectations, we find that inconvenience is not a strong motivator for icon or symbol development over index development, especially in contrast to the effect of ambiguous maps (section 3 of table A.2), and players make effective use of indices due to the abundance of diverse objects (table 3.1). For example, in an inconvenient teaching map for *Touch all objects that are triangular*, the teacher in Pair 8 “I enacted the behavior of painstakingly touching every triangle no matter color or size”. The tendency to develop non-indices in ambiguous maps over inconvenient maps shows that in our setting, necessity, rather than convenience, is the likely driving factor for icon and symbol introduction.

Similar to index saturation, icon saturation coincides with a contextual shift. Starting from task 20, about when icons saturate, pairs no longer see new vocabulary. We believe that the saturation of icons is therefore due to the plateau in vocabulary introduction, particularly vocabulary whose meanings are concrete and may be expressed iconically, such as shapes and directions. The increase in icons in the bonus round upon seeing new vocabulary for e.g. PARTIAL and SQUARE WITH TRIANGLE further evidences that, unlike indices, icon introduction is tied to vocabulary.

Finally, it is not clear that symbolic signs saturate during the game. The continued introduction of symbolic signs towards the end of the regular session indicates that teachers believe the marginal symbolic sign can still improve student performance even after all vocabulary tokens are introduced. Of the ten pairs who introduced symbolic signs and the eight who developed mutually understood symbolic signs, only half clearly reach saturation

(fig. B-6). After icons saturate, symbolic signs continue to arise likely because symbolic signs are harder to learn. In the limit, given the finite task and vocabulary space, symbolic signs would likely saturate for all pairs.

We have observed the transition from indexical signs to icons and symbolic signs under ambiguous environments. In the emergent setting, sign introductions for concrete meanings like shape lean iconic, and abstract meanings like conjunctions lean symbolic (fig. B-1). However, are there conditions under which icons, too, would eventually be re-expressed as symbols? In future work, we expect to find that, for example, in shifting from polygons with ≤ 3 sides to polygons of up to N sides, players can no longer rely on icons to depict all shapes. Using symbols rather icons would likely be helped by a large expansion in vocabulary, where it is no longer feasible to develop one icon per concept. Perhaps over a much longer timeline with a curriculum of vocabulary expansions, we will be able to see a transition from icons to symbols.

3.5 Compositionality

Compositionality, or the notion that a sign's meaning can be computed from the meanings of its constituents, is a feature of human language allowing for syntactic structure [38]. In particular, compositionality permits an unbounded number of meaningful expressions from a finite set of signs, imbuing language with infinite expressive capacity. While compositional thinking is staple in humans, the intermediate stages of compositionality in other species has not been observed; therefore, its origins remain uncertain. In this section, we examine evidence of early compositionality that arises over the course of our game and hypothesize that its precursors arise from inherently compositional gestures.

Communication protocols developed by all pairs exhibit some degree of linguistic compositionality (fig. 3-10), which manifests spatially in all pairs and temporally in nine pairs. When multiple signs are spatially composed, their meanings and form are combined simultaneously in the same physical location in order to create a new (form, meaning) pair. For example, at the most basic level, all pairs compose an indexical TOUCH and an iconic BACKWARDS, and, in reversing into a target object, spatially compose the two to mean

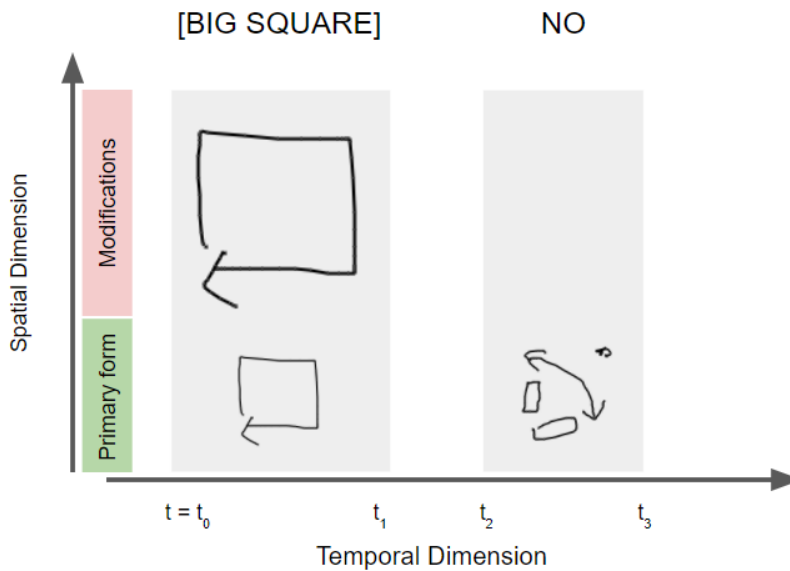


Figure 3-10: Linguistic composition across spatial and temporal dimensions. In the spatial dimension, composition occurs by adding modifications to a primary form. We observe SQUARE as a primary form and BIG superimposes morphologically to form BIG SQUARE. Along the temporal dimension, we illustrate the typical topic-comment syntax used by pairs in negation. All images are taken from the reflection forms of Pair 3.

TOUCH BACKWARDS. We observe that spatial composition involves a *primary form* on which *secondary forms* or modifications are superimposed. Primary forms tend to be *free*, or can stand alone as a separate sign, while secondary forms are typically *bound*, or must modify a primary form in order to be meaningful. For example, in fig. 3-10, SQUARE is the primary form, which may stand alone, and BIG is the secondary form, which in its given formulation must bind to a primary form. We observe that whether a component is primary or secondary in spatial composition as well as free or bound depends on its lexical category. One-hundred percent of nouns (such as shape) and verbs (such as TOUCH) for which signs were developed are primary when composed and exist as free forms. Meanwhile, adverbs and adjectives, such as quantity, size, and direction, are systematically secondary when composed and tend to not exist as free forms. Seventy-seven percent of all adverbs and adjectives for which a sign is developed are secondary, bound forms.

Signs may also be temporally composed. Temporal composition is aided by the discretization of player utterances; while player movement is continuous, they emit discrete signs separated by empty pauses as well as explicit signs for demarcation. For example, Pairs 1 and 5 develop a sign for BEGIN/END, which mark the boundaries of a communicative

motion sequence. For all nine pairs that employ temporal composition, we notice the development of a topic-comment structure in temporally composed sequences of signs, where the speaker first establishes the topic, or context, then adds a comment or modification [24]. For example, in all eight pairs that develop negation or affirmation signs, the speaker first defines the topic, for example by signing `BIG SQUARE`, then follows with negation or affirmation (e.g. figs. 3-6 and 3-10). The same occurs for quantifiers; for example, in Pair 7, the teacher first establishes the topic by `TOUCHING` the target shape, then signs `TOUCH TWICE`. Like primary and secondary forms in spatial composition, the topic-comment structure reveals another base-modification structure but at a higher level.

While linguistic compositionality is often discussed in the context of human language [38], we witness that compositionality is not relegated to complex language. The type of early compositionality that develops in our communication game instead reflects concepts and gestures that are inherently compositional, for example big shapes or driving backwards. Moreover, our observation of a topic-comment structure for negation and affirmation parallels Corballis’s proposition that “Gestures may have provided the basis for propositional language, perhaps first in the form of pointing to people or things or indicating actions” [1, 12]. We then hypothesize that gestural precursors to temporal compositionality may have followed a natural topic-comment structure.

3.6 Generalization

Generalization, or the ability to transfer knowledge to novel settings, is considered to be a hallmark of human language. In order to generalize, one must infer a set of abstractions from previous experience and then apply them across multiple situations, and by measuring this capability, we can evaluate how brittle or robust emergent communication protocols are to the immediate environment. Humans, especially children, prove to be exceptional at linguistic generalization compared to animals and neural language models [3, 4]. We analyze generalization of pairs’ sign sets along two dimensions: (1) how well pairs can generalize to arbitrary test maps given a teaching session, and (2) how well pairs generalize to novel tasks in the bonus round.

Recall that weighted score approximates the quality of communication based on only *three test maps*. We can proxy the pairs' ability to generalize to an *arbitrary number of test maps* for that task by examining the student guess for that task. We grade each student guess according to Table A.5, determining whether the pair will generalize to satisfy arbitrary test maps for that task, sometimes generalize, or never generalize.

Pairs who develop more signs and have a higher weighted score are more likely to generalize to arbitrary test maps. As expected, the correlations between number of signs and average weighted score with likelihood to always generalize are positive ($\rho = 0.64$ and 0.41 , respectively), and their correlations with likelihood to never generalize are negative ($\rho = -0.56$, -0.48 , respectively). Beyond this, we could not find a remarkable separation in correlation between number of indices, icons, and symbolic signs and likelihood of generalization.

In addition to generalization to arbitrary test maps, we test the zero-shot generalization of pairs to the bonus round, which requires that pairs not only generalize to blank teaching maps but also to novel tasks. When encountering bonus tasks, pairs introduce a total of 30 new signs. We have shown in section 3.4.1 that in ambiguous settings, indexical teaching strategies fall short, which necessitates the development of icons and symbolic signs. In blank teaching maps, this effect is similar to ambiguous teaching maps in the regular session: of the total 30 signs introduced in the bonus round, non-indices comprise 80% of sign introductions compared to 77% in ambiguous regular session maps and 47% in base teaching maps. Moreover, the difference in average weighted score between base and bonus 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 3.1). That is, having developed non-indices closes the performance gap between base and bonus teaching maps.

We have examined the effect of number of signs and number of non-indices on generalization. In doing so, we take an alternative approach to prior works, which instead frame generalization as a function of compositionality [4]. In our setting, communication protocols are in a nascent stage where pairs do not yet develop enough signs, and we believe analyzing generalization in terms of compositionality would better suit a later stage of development. For example, over a long time, we would expect a much larger vocabulary size to exert a

downward pressure on the feasible number of tokens, favoring compositionality. Then, we will be able to analyze the generalization capability of a reduced, compositional sign set to novel tasks.

Chapter 4

Conclusion and Future Work

The advent of symbolic communication is thought to be a crucial first step in the evolution of human language. While a number of studies have investigated the emergence of symbolic communication in humans, ours is the first to do so over a task space expanded to first-order-logic and over a continuous, embodied communication channel. Our primary results show that a sign's category influences the difficulty of establishing the sign; that indices systematically saturate before icons and symbolic signs; and that icons and symbolic signs arise when players need to communicate displaced information in ambiguous scenarios. These phenomena likely arise from the nature of mappings from form to meaning in each sign category.

Our high-level findings mirror theories that human cognition co-adapted with changes in ecological niche [34, 25, 5]. This is reflected in several overarching trends. First, a sign's category and morphology correspond heavily with its meaning and grounding environment. This implies that the grounding environment has a strong effect on when certain signs enter the lexicon. Different environments then beget different signs, and **sign production substantially changes in response to environmental shifts**. The particular shift to a setting that required communicating displaced information motivated the development of icons and symbolic signs, mirroring the theory espoused in (Bickerton, 2009) [5].

4.1 Future Work

The direction we currently prioritize is to gather enough human data in order to perform statistical tests. Further extensions to our experiments include determining when symbolic signs saturate in our setting (section 3.4.1); furnishing a longer curriculum of distributional shifts, especially expanding vocabulary so that icons may be re-introduced as symbols (section 3.4.1); and making inconvenient settings actually infeasible to teach indexically as a pressure for sign introduction. With more data, further analyses include substantiating our quantitative analyses with statistical inferences about the population, using e.g. multiple regression or causal inference; predicting perceptibility of signs using information-theoretic metrics (section 3.3); and determining whether a machine learning model can infer the meanings of player movements by watching game playback.

Finally, we may better understand how robots communicate non-verbally by training pairs of artificial agents to play our communication game. We are actively investigating training models in a multi-agent setting to determine which architectures, feature representations, and reward functions can enable agents to communicate in the context of our game. We were able to perform a preliminary experiment training a single agent to converge on the task *Touch any object* and in a fixed map. Once transitioned to the multi-agent setting, we can compare human and robot performance in symbol production and segmentation, asking questions such as: to what extent can robots perceive gestures as communicative? Does robot communication similarly evolve from indexical, to iconic, to symbolic communication? Can robots produce behavior that humans interpret as meaningful and symbolic?

If robots are able to spontaneously develop symbols, it will be possible for robots and humans to interact on human terms, that is, by communicating symbolically. Perhaps by first understanding human symbolic communication, we are one step closer to building robots that are able to communicate meaningfully with humans.

Appendix A

Tables

Pair Performance Data

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

Table A.1: Breakdown across pairs of average weighted score, average raw score, percent correct student guesses, and mutually understood signs across the entire game (regular and bonus sessions). A student guess is *correct* if its denotation is the same as that of the task, given the universe of predicates in Table 2.1. 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 Sign Introduction

	Base	Inconvenient	Ambig. Predicate	Ambig. Quantifier	Bonus
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. % Symbolic	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 # Icon	8	5	3	15	11
Total # Symbolic	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 (note that sampling a base, inconvenient, ambiguous predicate, or ambiguous quantifier map is equally likely). In the first section, the average number of signs introduced per task is shown with the standard error. 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 bonus setting. In the second section, for each pair and type of map we compute the proportion of introduced signs that are indices, icons, and symbolic signs, then report the averages and standard error across all pairs. In the third section, the total count of indices, icons, and symbolic signs over all pairs is shown.

Uses until Perception and Understanding for Perceived and Understood Signs

	Nonsymbolic icon	Nonsymbolic index	Symbolic
# Uses until Perceived	0.26 (0.05)	-0.28 (0.02)	0.03 (0.03)
# Uses until Understood	0.34 (0.06)	-0.41 (0.03)	0.05 (0.08)

Table A.3: For all perceived signs (row 1), we report the number of uses until perceived in z-value. For all understood signs (row 2), we report the number of uses until understood. The standard error is reported in parentheses. Evidence of perception/understanding is taken from the student's reflection form.

Forms per Meaning

Meaning	Total # of Forms	# Duplicate Forms	# Pairs
ONE	6	3	3
TWO	2	1	1
ALL	3	2	1
NO	2	1	1
AVOID	2	1	1
COLOR	2	1	1
RED	2	1	1
PHANTOM	2	1	1
EMPHASIS	4	2	2

Table A.4: For all meanings for which multiple forms were developed, we report the number of forms developed across all pairs, sorted roughly by lexical category. The first 7 signs are all symbolic and the following 2 signs are indices for all iterations of the sign. Pairs 3, 5, 8, 9, 11, and 13 developed multiple forms for one meaning. For all 13 duplicate forms (forms that are not the original for a meaning), 3 are introduced when the previous form is understood, 4 are introduced when the previous form is only perceived, and 5 are introduced when the previous form is neither understood nor perceived.

Rubric for scoring student guess

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

Table A.5: 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 2.1. 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.

Appendix B

Figures

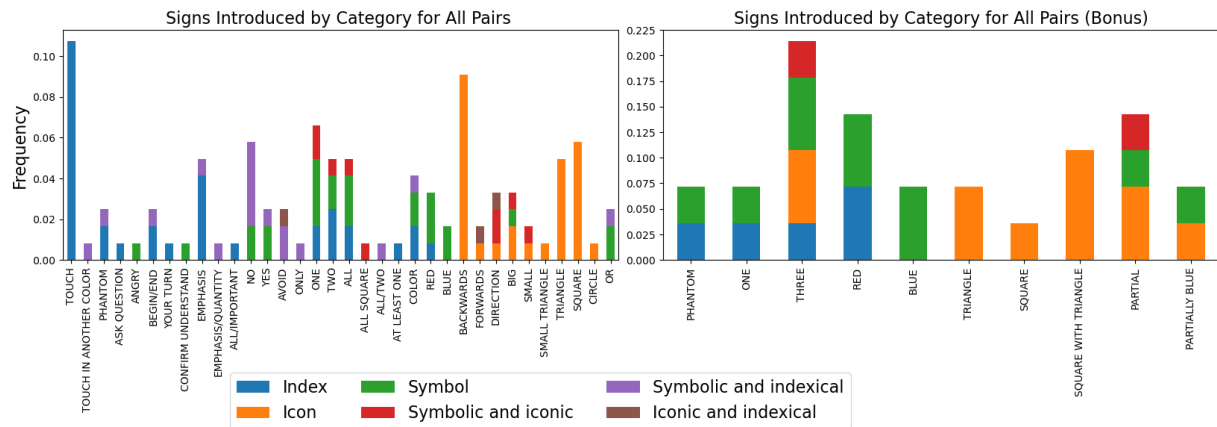


Figure B-1: The frequency of signs introduced by pairs in the regular session (left) and bonus session (right). There were no signs introduced for the category *symbolic*, *iconic*, and *indexical*.

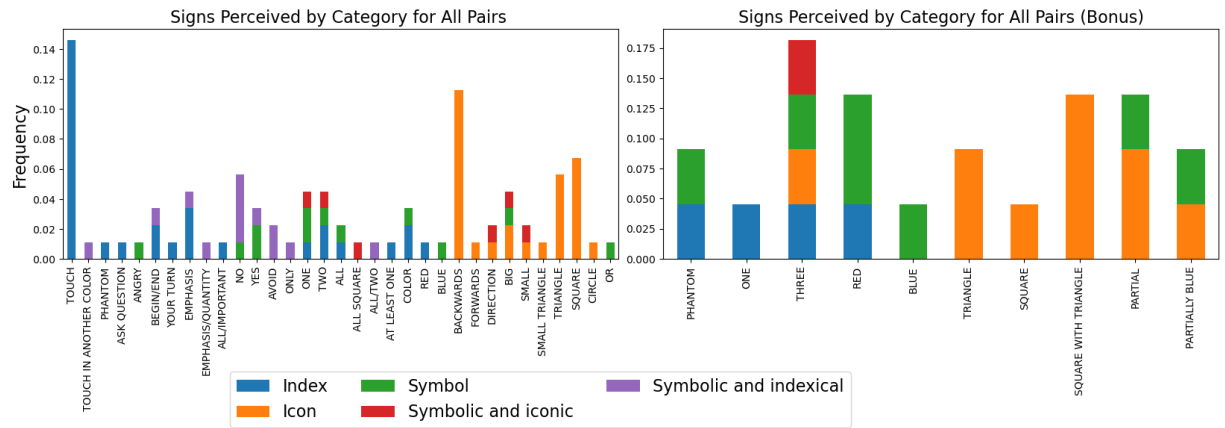


Figure B-2: The frequency of signs perceived by pairs in the regular session (left) and bonus session (right). There were no signs perceived for the category *iconic and indexical*, nor *symbolic, iconic, and indexical*.

Total # of sign uses until perceived and understood

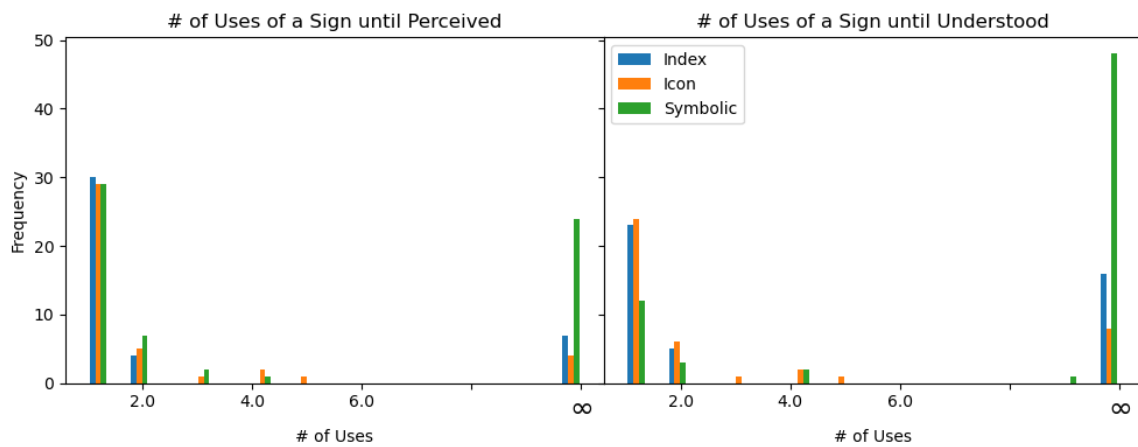


Figure B-3: Distributions of the total # of uses of a sign until perceived and understood, split by sign category across all pairs. Symbolic signs are by far the hardest to notice and understand.

Cumulative sign introductions over task index

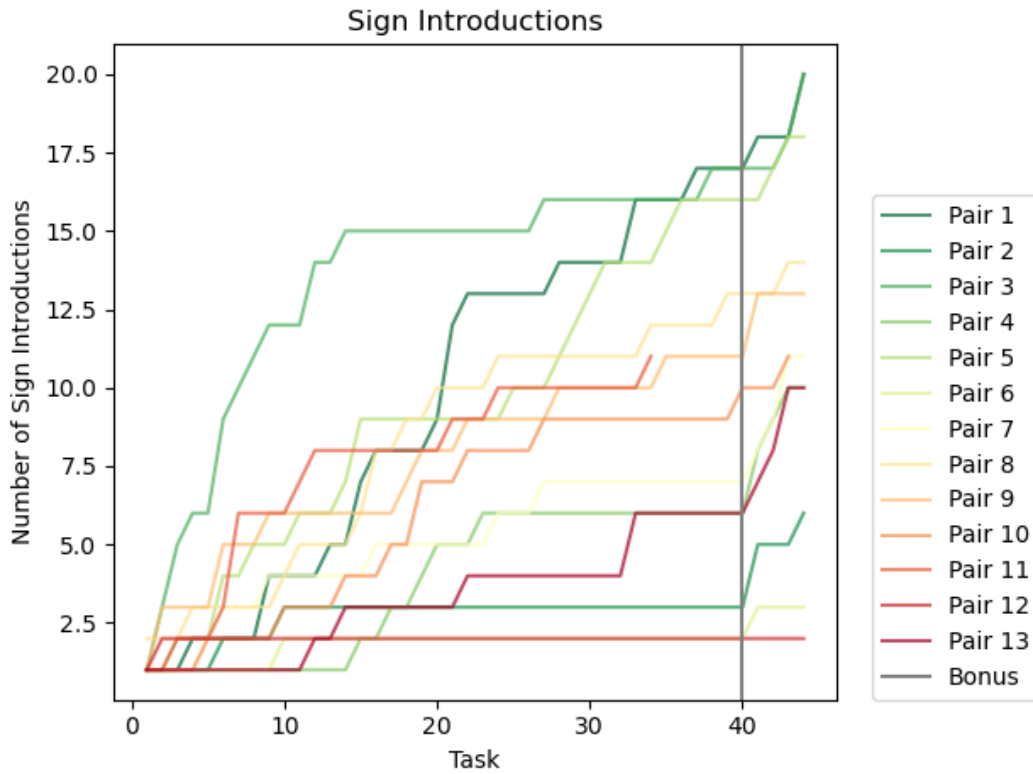


Figure B-4: Comparative view of cumulative sign introductions over tasks for Pairs 1-13, where the pairs are sorted in decreasing order of weighted score. High scoring pairs (green) are closer to the top and low scoring pairs (red) are closer to the bottom (see table 3.2). We observe a separation of trajectories into three groups by the end of the regular session ($x = 40$), roughly corresponding to color: number of signs introduced $y > 15$, between $6 < y < 13$, and $y < 3$.

Evolution of player lexicon over task index

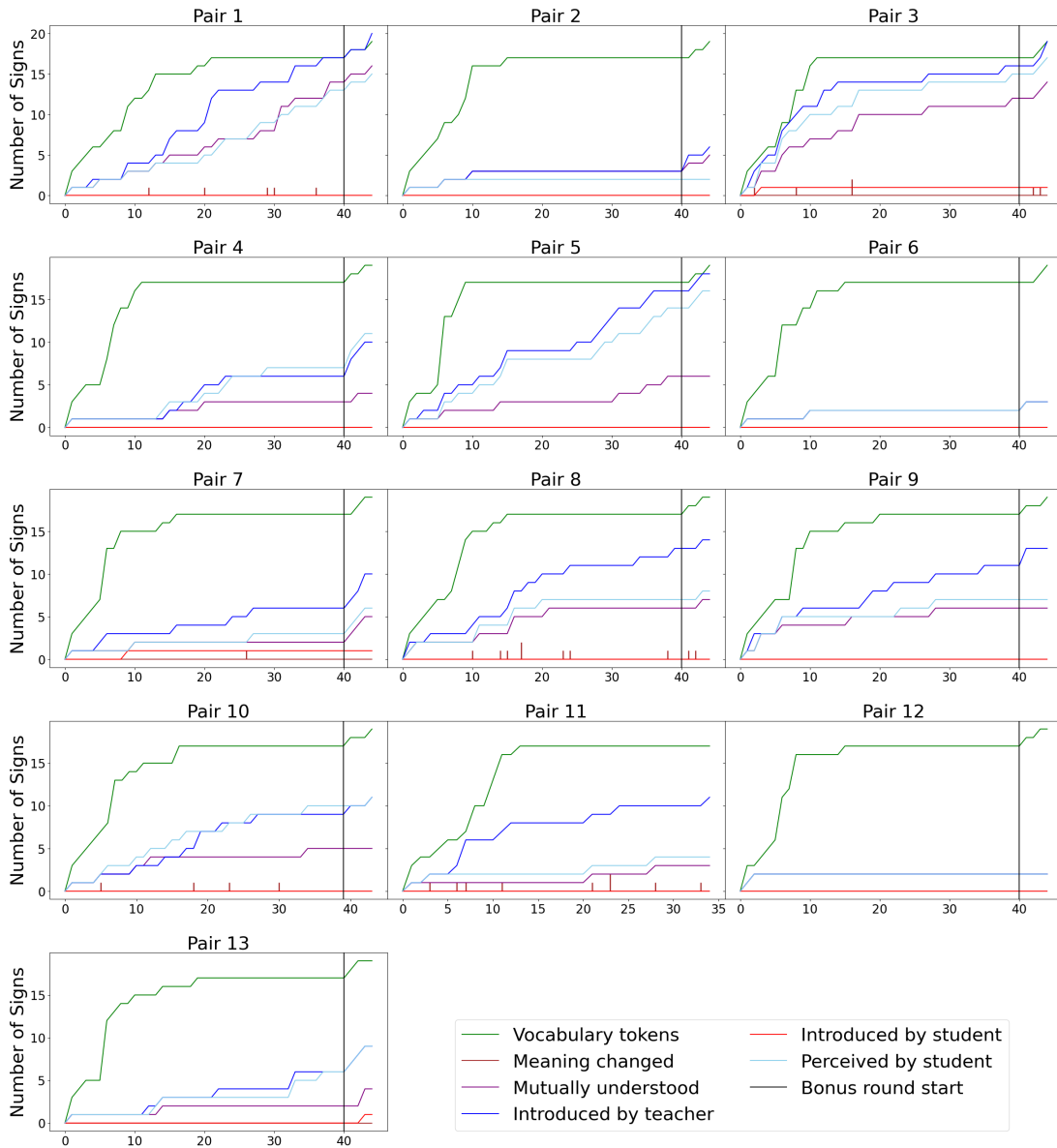


Figure B-5: Evolution of player lexicon over task index for Pairs 1-13.

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 *vocabulary tokens* in a task description are those in Table 2.1; 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 $t' \leq t$; A sign’s meaning is *changed* at task t if in either player’s reflection form for t , the player updates the sign’s description. We consider two registered actions sufficiently similar to be the same sign if their natural language descriptions have the same meaning and the drawings are similar. All plots except for *meaning changed* are cumulative.

Sign introduction over task index by sign category

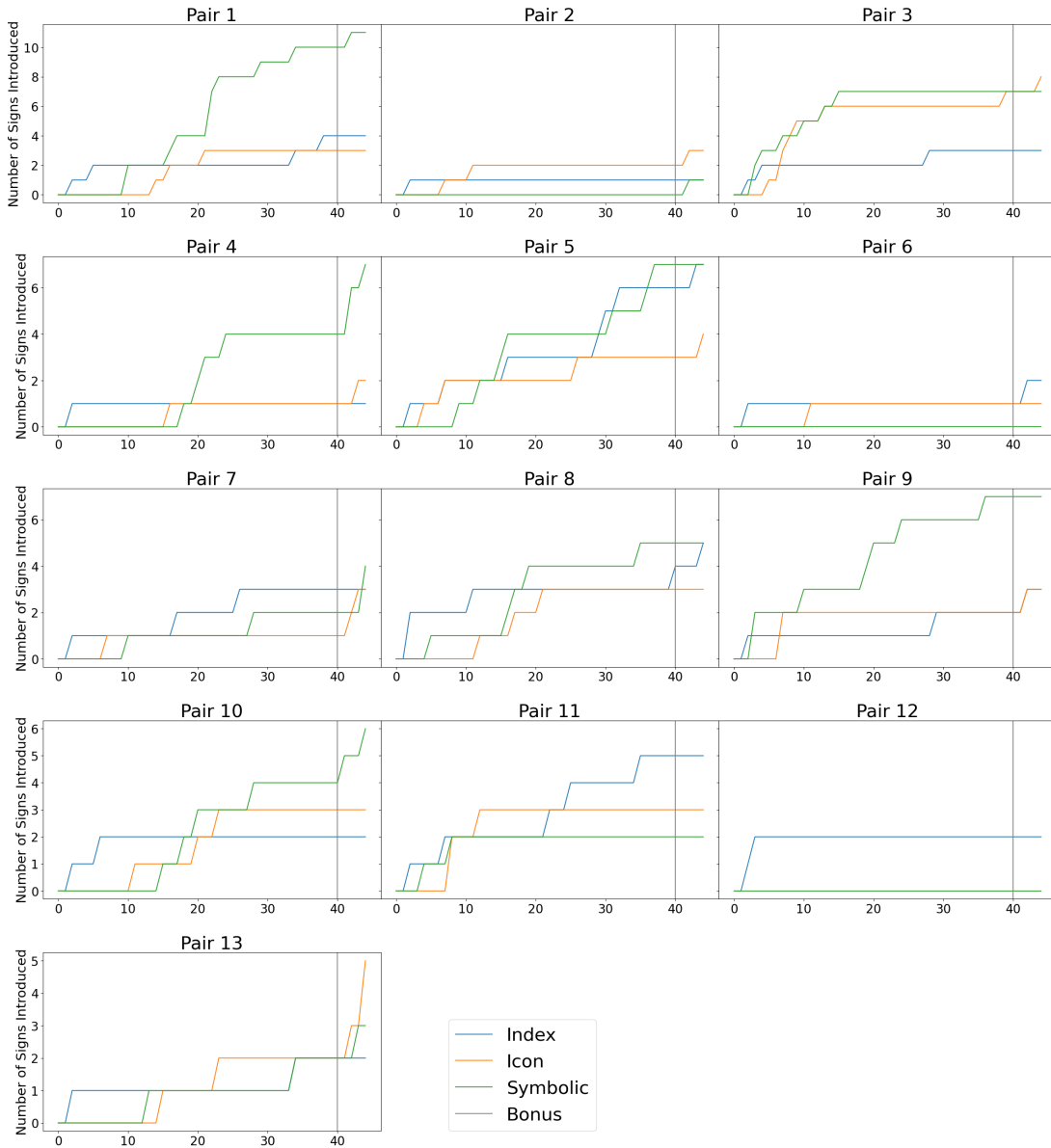


Figure B-6: Sign introductions over task index for Pairs 1-13, split by sign category. In each pair, the order of saturation is indices, then icons, then symbolic signs with the exception of Pairs 7 and 11.

Shared lexicon over task index by sign category

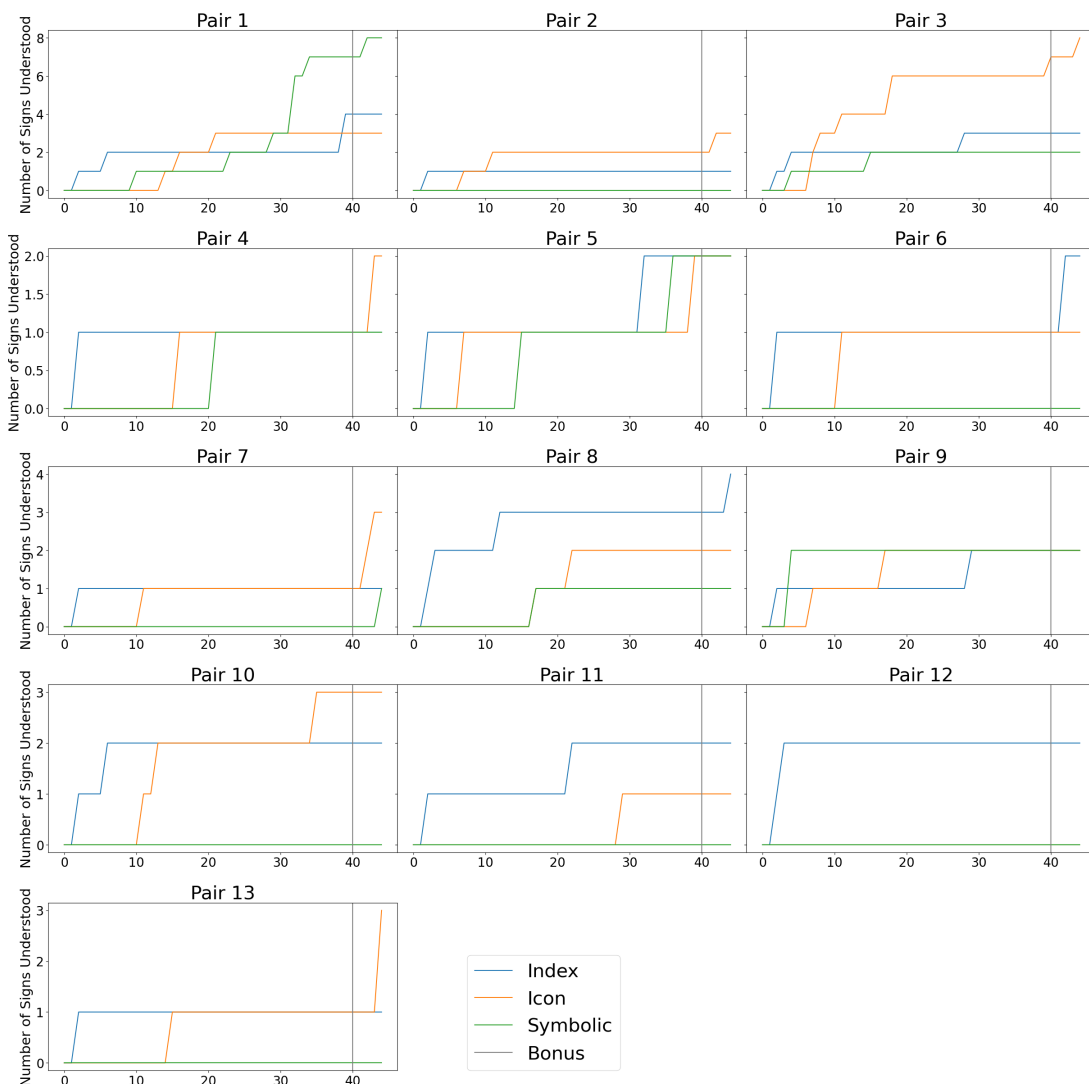
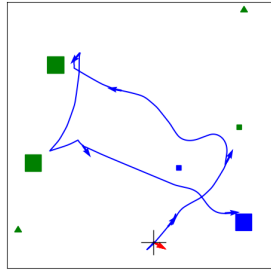


Figure B-7: Mutually understood signs over task index for Pairs 1-13, split by sign category. In each pair, the order of saturation is indices, then icons, then symbolic signs with the exception of Pairs 3, 8, and 9. In all pairs, indices saturate before non-indices.

Teaching Map

Teacher Response

Student Response



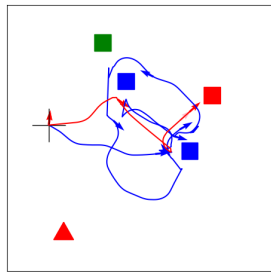
Iconic

/Indexical *Touch all objects that are big.*



I used it to mean don't touch this type of object because i turned away from it

*Unnoticed by student

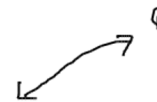


Symbolic

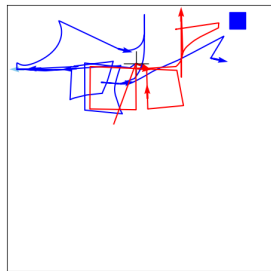
/Indexical *Touch all objects that are blue.*



I believe that the student used this action to ask "should I touch this?"



This was asking about the circled object, and if it's part of the set.



Symbolic

/Iconic *Touch exactly one object that is square.*



I ran in to the wall once to indicate that she needed to touch exactly one square.



Number of times the car is hit against the wall = number of shapes I should hit (I think this is what he means)

Figure B-8: Several teaching map traces and corresponding signs registered by Pairs 3 (symbolic/indexical), 7 (iconic/indexical), and 1 (symbolic/iconic). Each row is (left to right) a teaching map, a sign registered by the teacher immediately thereafter, and the corresponding sign registered by the student.

Appendix C

Probabilistic FOL 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 , Pred' (x)	1		→ green	0.33
NegPred	→ Pred	0.5	Size	→ big	0.5
	→ not Pred'	0.25		→ small	0.5
	→ NegPred or NegPred	0.25	Shape	→ square	0.5
Pred'	→ Size'	0.6		→ triangle	0.5
	→ Pred' or Pred'	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			

Figure C-1: Probabilistic grammar for task generation. As the grammar is recursive, we impose a negative bias on sample depth so tasks are simpler. We do so by first multiplying the highest daughter node probability by a multiplicative factor of $\alpha = 1.1$ after each sample, then re-balancing the remaining daughter probabilities proportionally. Furthermore, at runtime, we cull predicates 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.*

Bibliography

- [1] David Armstrong, Stokoe W., and S. Wilcox. Signs of the origin of syntax. *Current Anthropology*, 35(4), 1994.
- [2] Albert Atkin. Peirce's Theory of Signs. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2013 edition, 2013.
- [3] Dukes P. Caccamise D. Banich, M. T. *Generalization of knowledge: Multidisciplinary perspectives*. Psychology Press, 2010.
- [4] M. Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 2020.
- [5] Derek Bickerton et al. *Adam's tongue: how humans made language, how language made humans*. Macmillan, 2009.
- [6] Rudolf P. Botha and Chris Knight. *The cradle of language*. Oxford University Press, 2009.
- [7] Perreault C. and Mathew S. Dating the origin of language using phonemic diveristy. *PLoS One*, 07 2012.
- [8] Cristiane Cäsar, Klaus Zuberbuehler, Robert John Young, and Richard William Byrne. Titi monkey call sequences vary with predator location and type. *Biology Letters*, 9(5), October 2013.
- [9] N. Chomsky. *Language and mind (3rd ed.)*. Cambridge University Press., 2006.
- [10] Noam Chomsky. Linguistics and cognitive science: Problems and mysteries. In Aka Kasher, editor, *The Chomskyan Turn*, pages 26–53. Blackwell, 1991.
- [11] Morten H. Christiansen and Simon Kirby. Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300–307, 2003.
- [12] M.C. Corballis. *The lopsided ape: Evolution of the generative mind*. Oxford University Press, 1991.

- [13] Jan De Ruiter, Matthijs Noordzij, Sarah Newman-Norlund, Peter Hagoort, Stephen Levinson, and Ivan Toni. Exploring the cognitive infrastructure of communication. *Interaction Studies*, v.11, 51-77 (2010), 11, 03 2010.
- [14] Ferdinand de Saussure. *Course in General Linguistics*. McGraw-Hill, 1959.
- [15] Terrence William. Deacon. *The symbolic species: the co-evolution of language and the brain*. International Society for Science and Religion, 2007.
- [16] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.
- [17] Bruno Galantucci. An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5):737–767, 2005.
- [18] Bruno Galantucci and Simon Garrod. Experimental semiotics: a review. *Frontiers in Human Neuroscience*, 5(11), 2011.
- [19] Mitchell Green. Speech Acts. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.
- [20] H.P. Grice. *Logic and conversation*, 1975.
- [21] Paul Grouchy, Gabriele M. T. D’Eleuterio, Morten H. Christiansen, and Hod Lipson. On the evolutionary origin of symbolic communication. *Scientific Reports*, 6(1), 2016.
- [22] The “Five Graces Group”, Clay Beckner, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. Language is a complex adaptive system: Position paper. *Language Learning*, 59(s1):1–26, 2009.
- [23] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 2149–2159. Curran Associates, Inc., 2017.
- [24] C.D. Hockett. The origin of speech. *Scientific American*, 203(3):88–96, 1960.
- [25] Kevin N Laland and Marcus W Feldman. *Niche construction: The neglected process in evolution*. Princeton University Press, 2003.
- [26] Antoine Arnauld Claude Lancelot. *General and Rational Grammar: The Port-Royal Grammar, translated by Jacques Rieux and Bernard E. Rollin*. Mouton, 1975.
- [27] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era, 2020.

- [28] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *CoRR*, abs/1612.07182, 2016.
- [29] D. Lewis. *Convention. A Philosophical Study*. Harvard University Press, 1969.
- [30] Arbib MA, K Liebal, and S Pika. Primate vocalization, gesture, and the evolution of human language. *Current Anthropology*, 49, 12 2008.
- [31] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. 03 2017.
- [32] Andreas Nieder. Prefrontal cortex and the evolution of symbolic reference. *Current Opinion in Neurobiology*, 19(1):99–108, 2009.
- [33] Charles Sanders Peirce. On a new list of categories". volume 7. American Academy of Arts and Sciences, 1868.
- [34] Steven Pinker. The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):8993–8999, 2010.
- [35] Michael Rescorla. The Language of Thought Hypothesis. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition, 2019.
- [36] Stjernfelt F. Deacon T. Schilhab, T. *The Symbolic Species Evolved*. Springer, 2012.
- [37] Thomas Scott-Phillips, Simon Kirby, and Graeme Ritchie. Signalling signalhood and the emergence of communication. *Cognition*, 113:226–233, 11 2009.
- [38] Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition, 2020.