



Cognitive Science 43 (2019) e12783

© 2019 The Authors. Cognitive Science published by Wiley Periodicals, Inc. on behalf of Cognitive Science Society. All rights reserved.

ISSN: 1551-6709 online

DOI: 10.1111/cogs.12783

## The Influence of Shared Visual Context on the Successful Emergence of Conventions in a Referential Communication Task

Thomas F. Müller, James Winters, Olivier Morin

*Minds and Traditions Research Group, Max Planck Institute for the Science of Human History*

Received 8 November 2018; received in revised form 27 June 2019; accepted 1 August 2019

---

### Abstract

Human communication is thoroughly context bound. We present two experiments investigating the importance of the shared context, that is, the amount of knowledge two interlocutors have in common, for the successful emergence and use of novel conventions. Using a referential communication task where black-and-white pictorial symbols are used to convey colors, pairs of participants build shared conventions peculiar to their dyad without experimenter feedback, relying purely on ostensive-inferential communication. Both experiments demonstrate that access to the visual context promotes more successful communication. Importantly, success improves cumulatively, supporting the view that pairs establish conventional ways of using the symbols to communicate. Furthermore, Experiment 2 suggests that dyads with access to the visual context successfully adapt the conventions built for one color space to another color space, unlike dyads lacking it. In linking experimental pragmatics with language evolution, the study illustrates the benefits of exploring the emergence of linguistic conventions using an ostensive-inferential model of communication.

*Keywords:* Language evolution; Shared context; Artificial language; Referential communication; Convention; Language emergence

---

---

Correspondence should be sent to Thomas F. Müller, Minds and Traditions Research Group, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, 07745 Jena, Germany. E-mail: [tmueller@shh.mpg.de](mailto:tmueller@shh.mpg.de)

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

## 1. Introduction

An outstanding puzzle for language scholars is that of its *emergence*: How does language come about from pre-linguistic or non-linguistic states (Christiansen & Kirby, 2003; Höfler, 2009; Tomasello, 2010)? Since an important function of language is communication, we can arrive at insights on language emergence by studying human communication in general (Höfler, 2009). Linguistic communication, in this view, is a special case of communication enriched by a “structured collection of conventional codes” (Scott-Phillips, 2015, p. 20). Based on this, it has been argued that the human capacity for *ostensive-inferential communication* (Sperber & Wilson, 1996) is what allows complex languages to evolve (Scott-Phillips, 2015), with our pragmatic capacity being the cognitive foundation for all of semantics, morphology, syntax, and phonology (Scott-Phillips, 2017). In ostensive-inferential communication, the sender of a message provides evidence for an intended meaning, while the receiver interprets this evidence (Sperber & Wilson, 1996). Imagine, for instance, you are listening to music over your headphones, when your friend Barbara wants to tell you something. To signal this, she makes eye contact with you, puts her hands to her ears, and mimics taking off the headphones in a slow and stylized way. Even though this specific signal might have never been used before in your shared conversational history, inference may suffice to interpret and understand the underlying message.

This interpretation of signals cannot happen in a vacuum, however; the possible meanings would be practically unlimited (Berwick, Pietroski, Yankama, & Chomsky, 2011). Consider the example outlined above again: Had you not been wearing the headphones (or had you not been aware that you were still wearing them), Barbara’s gesture would have left you quite puzzled indeed. Alternatively, suppose that after taking off the headphones, she asks you “Are you listening to this album?”, while holding up a copy of “Led Zeppelin.” Here, the demonstrative “this” could have referred to virtually any album in existence, had the meaning not been clarified by the additional visual information. As proposed in classic theories of communication (Clark, 1996; Grice, 1989; Lewis, 1969; Sperber & Wilson, 1996), interlocutors have to create and interpret messages according to their *context*, a wide range of information that includes the time and place of a message’s utterance, the interlocutors’ previous conversational history, and more.

Context is a notoriously vague notion, which has been operationalized in various ways (see e.g., the differences between Clark’s “common ground” compared to Sperber and Wilson’s “mutual cognitive environment”). In this study, we investigate the immediate *shared context* (cf. Fig. 1). Following Winters, Kirby, and Smith (2018), it is defined as the amount of relevant knowledge that the interlocutors have in common. Unlike “common ground” (Clark & Carlson, 1981), shared context does not require explicit mind reading of the “I am aware that she is aware that I am aware . . .” type. Unlike Sperber and Wilson’s mutual cognitive environment, it is restricted to information that is actually present to each interlocutor’s mind, as opposed to information that is merely accessible. Both are important dimensions of context that we choose to disregard for the purpose of

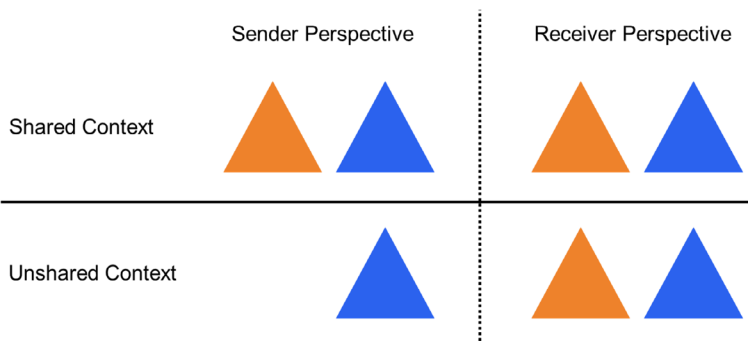


Fig. 1. Illustration of the shared context for the sender and receiver of a message in a referential communication situation. The sender in the situation is tasked with communicating the blue triangle to the receiver. The situations differ only in the availability of contextual information to the sender. With access to the shared context, the sender might refer to the triangle as “the blue triangle,” whereas without it simply “the triangle” would be sufficient from their perspective.

this study. To further simplify the issue, this paper focuses on the shared knowledge that two interlocutors have of their environment, leaving aside other aspects of shared knowledge such as shared membership in a community (cf. Clark, Schreuder, & Buttrick, 1983) or shared discourse histories (Barr & Keysar, 2002; Clark & Wilkes-Gibbs, 1986).

In line with other studies, we focus specifically on shared visual information. For two interlocutors, having a piece of visual information in common impacts communication. On the sender’s side, shared visual information allows for audience design—the tailoring of a message to its receiver’s state of knowledge (Brennan & Hanna, 2009; Galati & Brennan, 2010; Holler & Wilkin, 2009; Isaacs & Clark, 1987; Krauss & Fussell, 1991). This is a crucial aspect of the interactive alignments that characterize conversation (Garrod & Pickering, 2004). On the receiver’s side, shared information is more likely than non-shared information to be taken into account when interpreting a message (Hanna, Tanenhaus, & Trueswell, 2003).

These studies, and others like them studying the impact of shared information on communication, are based on natural language. This has two consequences. First, communicative success is usually at ceiling: Any participant can solve a simple referential communication task using words, regardless of the amount of shared information (e.g., Brennan, 2005). This makes it difficult to gauge the impact of shared information on communicative success (but see Clark & Krych, 2004; Schober & Clark, 1989; Sulik & Lupyan, 2018). Second, experiments conducted in natural language are appropriate to study the use of linguistic conventions, rather than their emergence. Conventions, whether they are linguistic or not, are solutions to repeated coordination problems (Lewis, 1969), such as referential communication (Millikan, 2005; Skyrms, 2010). They are at least partly arbitrary forms of behaviors that are sustained by the weight of precedent as opposed to any intrinsic aptness. This paper aims to investigate the impact of shared information upon the emergence of novel conventions.

### 1.1. Artificial language experiments and the emergence problem

In the past, the emergence problem of language has been addressed in the laboratory using artificial language experiments. Here, the general idea is to study interactions occurring without the presence of established communicative conventions (for studies reviewing this field, see Galantucci, 2009; Galantucci & Garrod, 2011; Galantucci, Garrod, & Roberts, 2012; Scott-Phillips & Kirby, 2010; Tamariz, 2017). Several studies have focused more closely on the form of the emerging conventions themselves, that is, the shape that the signals take, and their evolution (e.g., Galantucci, 2005; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Healey, Swoboda, Umata, & King, 2007; Scott-Phillips, Kirby, & Ritchie, 2009). For instance, Garrod et al. (2007) showed that drawings emerging *de novo* in their “Pictionary”-style experiments become simpler and more symbolic with repeated interaction, while Healey et al. (2007) focused on the extent to which drawings were abstract or iconic when participants were tasked with drawing music for each other. In comparison, this study investigates the emergence of conventions in relation to communicative success.

The contextual circumstances under which conventions arise have not been considered in the studies above. The exception to this is the “embodied communication game” by Scott-Phillips et al. (2009), who found that “the establishment of [a] default convention provides the common ground from which a signal may be created and inferred” (p. 233). In this experiment, participants had to communicate their position on a  $2 \times 2$  grid. As there was no established communication system for completing the task, participants had to use a repertoire of basic actions to signal their intention to communicate (and to then use this as a means to derive a conventional signaling system of conveying meaning). Our study differs in this respect as it focuses on the immediate shared context (as opposed to the ability of participants to leverage their shared discourse history). Additionally, Scott-Phillips et al. (2009) did not investigate the role of context directly by manipulating it experimentally. Nevertheless, the study opens up the interesting question of how conventions multiply in relation to the shared context, which is what we also try to address in this study as a secondary question. Presumably, conventions form from repeated successful usage of symbols building on the shared context (Höfler & Smith, 2009), until they become sufficiently entrenched (Langacker, 1987) and can be used as a contextual basis for novel inferences themselves (as in the case of Scott-Phillips et al., 2009). As such, we should expect the shared context to facilitate the establishing of more conventional symbols, with more conventions leading to more successful communication in turn.

In contrast to this first line of research, previous experiments have investigated contextual effects systematically, but focused on the further development of conventions after they have been established (through a training phase in the task). Several studies have demonstrated that artificial languages will optimize to the semantic dimensions relevant in context. Through simulating *iterated learning* in experiments, defined as the “process in which an individual acquires a behavior by observing a similar behavior in another individual who acquired it in the same way” (Kirby, Cornish, & Smith, 2008, p. 10681), artificial languages have been shown to develop: (a) underspecification with regard to

irrelevant dimensions in a reference space (Silvey, Kirby, & Smith, 2015), (b) overspecification when relevant dimensions are difficult to discern (Tinitis, Nölle, & Hartmann, 2017), and (c) either underspecified, holistic, or systematic linguistic structure depending on their contextual niche (Winters, Kirby, & Smith, 2015). These studies took the task's immediate perceptual context into account, but they did not manipulate the extent to which it was shared or not. Still, the general observation that specific types of context will bias artificial languages to develop a certain structure leads us to another secondary question. We want to investigate how flexibly conventions can adapt when a change in contextual environment occurs: their generalizability. As the shared context should allow interlocutors to be more successful, it might also lead to more generalizable conventions emerging from their conversation.

Winters, Kirby, and Smith (2018) specifically considered the shared immediate context in the artificial language paradigm. Using a referential communication game setup, participants were first trained in an "alien language" consisting of random syllables mapped onto a small set of referents, and then used this alien language to communicate about referents they learned as well as novel ones. Both the shared context and the generalizability of the immediate context to future contexts were manipulated. Crucially for our purpose, shared contexts fostered languages that required more contextual enrichment for interpretation than non-shared contexts. Of special interest to us are their performance results that indicate higher levels of communicative success in the shared context conditions. However, the interpretation of these results is limited, since the effect is driven by one condition that is at ceiling, while performance in the other shared context condition was as low as in the unshared conditions. Additionally, the study used a training regime to make participants learn the starting language in the experiment; a commonality it shares with all iterated learning studies mentioned above, and with another line of evidence which demonstrates the influence of shared context on the choice of referral expressions in a word-learning paradigm (Craycraft & Brown-Schmidt, 2018; Gorman, Gegg-Harrison, Marsh, & Tanenhaus, 2012; Heller, Gorman, & Tanenhaus, 2012; Wu & Keysar, 2007). In all such studies, participants are provided with pre-established mappings between the artificial language's symbols (e.g., the strings of syllables in Winters, Kirby, & Smith, 2018) and the corresponding referents. This makes it difficult to study the emergence of conventions. To resolve these issues, this study removes the training from the procedure and allows participants to freely associate and create mappings from the start.

### *1.2. Referential communication tasks and interaction*

At the core of our task is a *referential communication* paradigm. These tasks have been traditionally used in experimental pragmatics, dating back at least to Krauss and Weinheimer (1964). The basic premise is that a "sender" (alternatively, "director" or "speaker") has to communicate a target object to a "receiver" (also known as "matcher," "listener," etc.), using natural language. In our task, participants take on the role of either sender or receiver, with no role reversal (for a study experimentally investigating the effect of role reversal on conventionalization, see Moreno & Baggio, 2015). The task

consists of using black-and-white symbols to convey and identify colors (used as referents). The domain of colors has been of particular interest to studies on language evolution ever since the classic work by Berlin and Kay (1969), and it has been proven to be useful as a reference space in pragmatic experiments as early as Krauss and Weinheimer (1967). Early results using the referential communication paradigm include the fact that the descriptions become shorter with increasing conversational history (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964; Krauss & Weinheimer, 1966; Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992), but longer when referents are more similar (Krauss & Weinheimer, 1967) or when describing the referents for someone else as opposed to oneself (Fussell & Krauss, 1989). In these early studies, the intent of the research designs was to study linguistic communication in live interactions.

Later on, the focus shifted from conversational history to perceptual context, especially in the visual modality. These studies have typically been using eye-tracking in a task that involves a director instructing participants how to move objects around in a grid. Crucially, the objects are not always perceivable by both participants: Relevant items may be omitted from the director's view, or the two participants may be given access to partially different sets of items. Initial studies interpreted their findings as demonstrations of failures in matchers' usage of the context (Horton & Keysar, 1996; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Barr, Balin, & Paek, 1998), but this was contested by studies showing either an impact of the shared context on utterance comprehension, methodological problems, or both (Brown-Schmidt, 2009; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna et al., 2003; Heller, Grodner, & Tanenhaus, 2008; Nadig & Sedivy, 2002). Summarizing the debate, Brown-Schmidt (2012) states that it mainly revolves around the timing of reference resolution, which is not a central concern of our study. However, similar to this line of research, we will focus on the perceptual aspect of the shared context. The term we will use is shared visual context: The context is shared because we make it clear that interlocutors get access to the same information (or have restricted information, in our other condition), and only limited to the visual modality.

One advantage of pragmatic paradigms using natural language is that participants are often allowed to interact more freely. In particular, they are able to *repair* misunderstandings that might arise in the conversation, and will do so until they arrive at an acceptable interpretation (Clark & Schaefer, 1987). What remains poorly understood is the role that interaction might play in the early stages of emerging communication, especially in the absence of external feedback (i.e., information on the success or failure of communication, given by an outsider to the conversation, and usually provided by the experimenter, or programmed by the experimenter into the protocol's program). Most of the studies that trained participants on the initial conventions in the task have also relied on feedback provided systematically by the experimental setup (e.g., Kirby, Tamariz, Cornish, & Smith, 2015; Winters et al., 2015, 2018) or did not involve interaction between participants at all (Kirby et al., 2008; Silvey et al., 2015; Tinitis et al., 2017), whereas studies specifically concerned with the form of emerging conventions typically also privilege repair (Garrod et al., 2007; Healey et al., 2007; Scott-Phillips et al., 2009). In a similar fashion, we included basic repair mechanisms into the experiments presented here (even though we acknowledge we cannot



comprehensively cover the extent of repair strategies used in real-world communication) and refrain from providing other types of feedback to the participants.

### 1.3. *The current study*

The goal of this study is to show that the shared visual context is important for the successful emergence and use of communicative conventions. To test this idea, we conducted two referential communication experiments. Dyads of participants were tasked to accurately communicate the correct color out of an array of four colors by using novel symbols. These symbols were limited to a predetermined set of black-and-white visual signals (and combinations thereof) and some pre-established repair signals. Participants received no training on symbol meanings and no feedback by the experimental setup at any time; they had to make inferences about the most likely correct answer given the evidence, while they could use the repair mechanism to clarify or request further explanation. We manipulated the shared visual context between dyads by giving the sender access to the distractor colors in only one condition.

The main hypothesis is that, overall, the access to the visual context will influence performance in the task. Based on this, we predicted dyads in the shared visual context condition would outperform those in the unshared visual context condition (prediction 1). Furthermore, we expected pairs to make progress in performance over the time course of the experiments, as they jointly create novel conventions for communication (prediction 2). If the pairs in the shared visual context condition also subsequently profit more from building on these conventions, we should see even faster progress in that condition (prediction 3). These three predictions are tested in both experiments. In Experiment 2, we also consider the secondary hypotheses outlined in the introduction. Specifically, we ask whether the shared context would also lead to more numerous conventions, and better generalization to different contexts (predictions 4 and 5).

### 1.4. *Ethical approval and preregistration*

Both experiments received approval by the ethical committee at the FSU Jena before they were conducted. All our predictions and sample sizes were preregistered on the Open Science Framework in advance. For Experiment 1, this happened before data collection was underway; for Experiment 2, due to a technical malfunction, the registration occurred after 3 of the 48 pairs had been tested, but no changes were made to the preregistration document in the meantime. The registrations can be accessed at <https://osf.io/rb hk2/> (Experiment 1) and <https://osf.io/tn6e8/> (Experiment 2).

## 2. Experiment 1

Experiment 1 sought to establish the paradigm and provide a first test for our main hypothesis.

## 2.1. Method

### 2.1.1. Participants

In this study, 52 participants (50 of which were students) were recruited and invited to play the “Color Game” in the laboratory. Their mean age was 23 ( $SD = 3.5$ ); 35 were female and 17 male. All participants were fluent speakers of German, and all but two participants reported German as their native language. Before the main procedure, the Ishihara test for color blindness (Ishihara, 1972) was administered, since the meaning space in the experiment consisted exclusively of colors. All participants showed typical color vision in the test.

### 2.1.2. Materials

The meaning space of the experiment consisted of a continuous HSL color space (H, S, and L describing the colors in terms of Hue, Saturation, and Lightness, respectively), going full circle in  $360^\circ$  of hue. The saturation and lightness parameters were kept constant. For practical purposes, we constructed a total of 360 colors in this way (with a constant distance of  $1^\circ$  in hue, which makes neighboring colors indistinguishable). From this space, color arrays were constructed by randomly drawing a color and then choosing the three next colors with a fixed spread value of  $45^\circ$  in hue, respectively. Thus, the first color was at a distance of  $45^\circ$  from the second,  $90^\circ$  from the third, and  $135^\circ$  from the fourth color. As can be seen, the domain of color provides us with a flexible continuous meaning space that can be divided in certain ways, allowing some control over the relations between the discrete referents as well as creating similar portions of the total space participant’s experience (cf. Experiment 2).

For the signal space, participants were presented with a selection of 39 pre-constructed black-and-white symbols (see Fig. 2) to choose from. The symbols had been selected on the level of ambiguity, such that they could become associated with several different colors. For instance, the “crystal” symbol in the second to last row of Fig. 2 could be treated as a gemstone of any color imaginable. Participants were not trained on any meanings that the symbols might have and saw them for the first time just before the experiment started. This and the arbitrariness of the symbols regarding their relation to specific colors ensured that participants had to form new conventions over the course of the experiment.

### 2.1.3. Procedure

Participants were randomly paired in dyads and randomly assigned the role of sender and receiver for the entire duration of the experiment. To minimize knowledge about each other, players were seated in separate sound-proof rooms during the experiment, and one participant was scheduled to arrive to the experiment 15 min earlier than the other participant to avoid contact between them. Upon arrival, participants read the general participant information, gave informed consent regarding the experiment, and completed a short demographic questionnaire. Thereafter, the Ishihara test for color blindness (Ishihara, 1972) was administered. Just before the experimental task, participants read printed instructions that explained the rules of the game to them, and they were allowed to ask



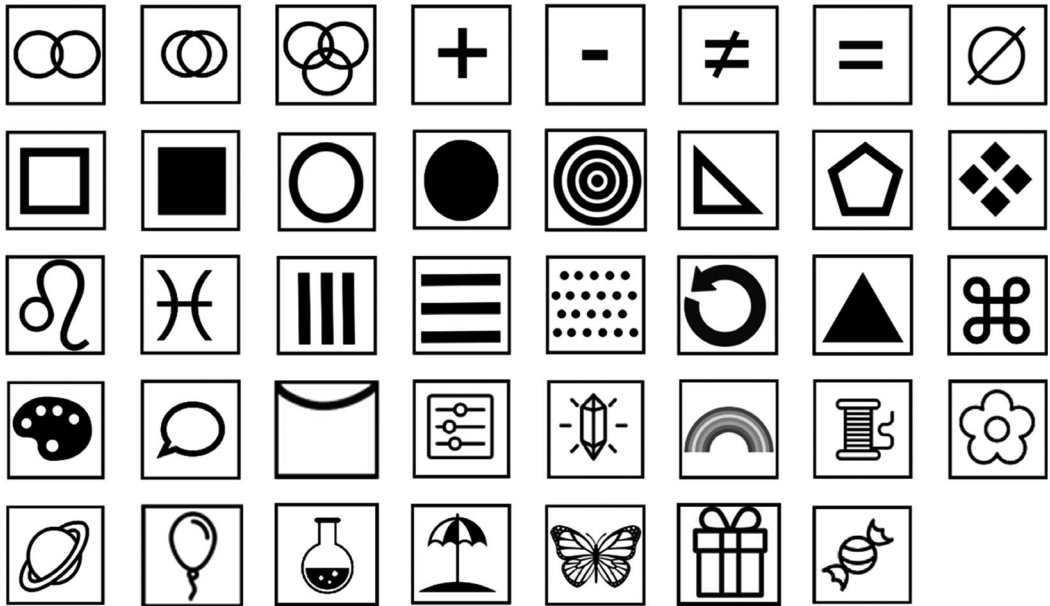


Fig. 2. List of symbols available to senders in Experiment 1. The symbols can be roughly categorized as logical symbols, abstract shapes, and symbols depicting real-world objects (from top to bottom). Crucially, all symbols were chosen to be ambiguous with regard to their association with colors.

questions for clarification. Since this study is concerned with the emergence of communication, they then proceeded immediately with the first trial of the experiment, without a training phase.

#### 2.1.4. Experimental task

On any given trial in the game, a random array of four colors was constructed in the way described above. Out of the four colors, a target color was chosen randomly and conveyed to the sender by a pointing finger next to it on the computer screen (for example screens, see Fig. 3). The sender's task always was to communicate this target color using only the symbols of the signal space; the goal for the receiver was to choose the correct color out of the array of four colors. Communication was only possible via a whiteboard application (Baiboard, created by Lightplaces Ltd.) running on two iPads that the participants were using. The iPads were connected by WLAN, allowing for live synchronization of changes made by the participants and for observation by the experimenter via a third connected iPad. Senders were free to arrange the symbols on the canvas in whatever position they wanted, and there was no limit on combinations or the number of symbols sent. The exception was that symbols were not allowed to overlap, as this would have prevented correct analysis of the messages.

Importantly, the receiver could not only passively watch the sender's message being created, but was also allowed to repair unclear messages using three simple responses that had been introduced to both players before the game started: Drawing a circle indicated

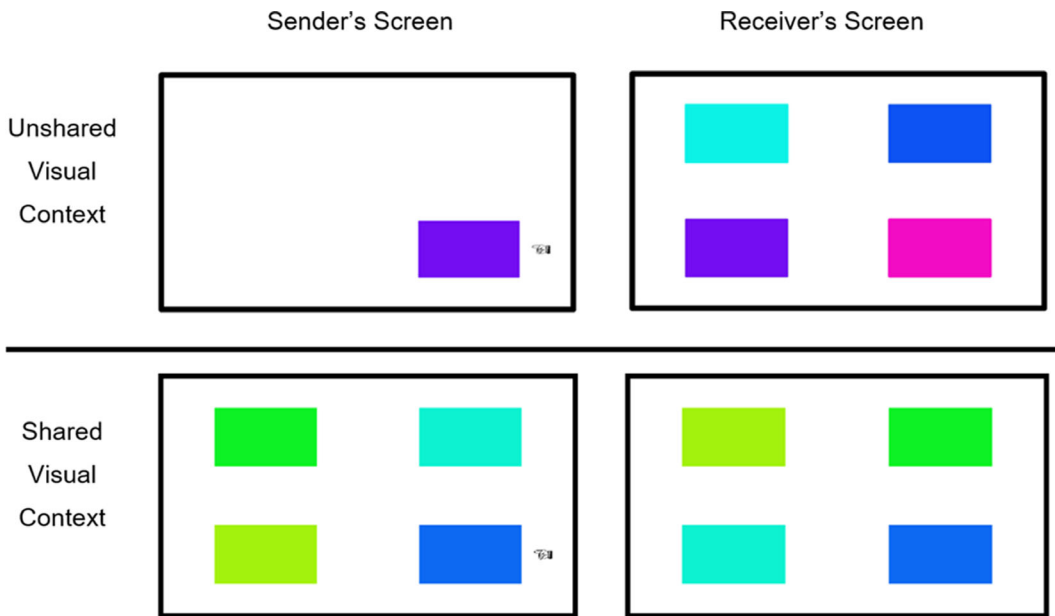


Fig. 3. Example trials in the two conditions with the corresponding screens for sender and receiver. In the top example (*unshared visual context*), the sender has to communicate the purple color, and in the bottom example (*shared visual context*), the dark blue color.

that the message had been understood; drawing an arrow was used as a prompt for clarification; and drawing a schematized hourglass meant the receiver thought that the participants were running out of time in the current trial. Likewise, the sender had two possibilities to evoke feedback from the receiver: Sending a question mark was used as a prompt for the receiver to indicate whether the message was clear, and sending an exclamation mark was an indication that the message was complete for the sender. All of these possible responses can be seen in the example trials presented in Fig. 4. In this way, we wanted to allow for interaction in the task, using rules that were the same for every pair; apart from this, there was no direct feedback that indicated correct or wrong answers, nor were senders informed about the receiver's color choice at the end of the trial. This made sure that the possibility of receivers learning their sender's code by mere memorization of correct answers, on a trial-and-error basis, was minimized. Instead, we wanted them to infer the sender's intended meaning and communicate about what they could understand and what they could not.

As displayed in Fig. 3, the experimental manipulation concerned the number of colors the sender knew about, which was either one (i.e., the sender sees only the target) or four (i.e., the sender has knowledge of the whole array). The receiver always saw all four colors. Participants were informed in the instructions about how many colors their partner in the experiment would see. The conditions varied between the 26 pairs of participants. Thus, 13 dyads experienced shared visual context (*shared condition*) and 13 dyads

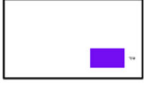



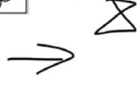
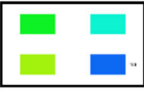



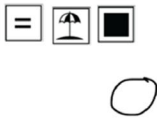
	t	t+1	t+2	t+3
				
				

Fig. 4. Two frame-by-frame examples of communication during a single trial in each condition, respectively. The examples correspond to the screens presented in Fig. 3. In the example for the unshared visual context condition (top), the receiver does not understand the “candy” symbol sent at time  $t$  and indicates so by drawing an arrow ( $t + 1$ ). The sender then tries to elaborate further by adding a “rainbow” symbol ( $t + 2$ ), but ultimately communication fails and the receiver indicates that the pair is running out of time (hourglass at  $t + 3$ ). In the example for the shared visual context condition (bottom), the sender specifies the precise meaning of “dark blue” using a combination of three different symbols ( $t$  until  $t + 2$ ). At  $t + 3$ , the receiver indicates he or she understands the intended message by drawing a circle.

unshared visual context (*unshared* condition). The only technical change in the unshared condition was that the three distractor colors were removed from the sender’s array, meaning that the target color still appeared in a random position in every trial and was marked by the finger, as in the shared condition. A consequence of this is that the sender saw a single color which was always present in the receiver’s visual context.

In total, participant pairs were presented with 64 experimental trials, divided into eight blocks (of eight trials) that were separated by a short pause, respectively. After the main experimental task, participants completed a short questionnaire in which they listed all the symbols they remembered from the main experiment (i.e., in free recall) and their corresponding meanings (suspected meanings, in case of the receiver). Finally, participants were paid 10€ plus up to 6€, depending on their success in the task, in compensation. Completion of an average experiment took between 1 and 2 hours in total.

## 2.2. Results

### 2.2.1. Does the shared visual context improve communicative success and do pairs improve over time?

Before the analysis, six trials (0.4% of the total sample) were excluded from the data for the following reasons: Three trials had reaction times below 3 seconds, probably due

to accidental button presses; and three trials were lost due to a WLAN crash. All analyses were conducted using R version 3.4.0 (R Core Team, 2017).

The mean accuracy, which is the proportion of trials with correct answers by receivers, across both conditions was  $M = 0.58$  ( $SD = 0.49$ ). There were individual differences between dyads, with the lowest scoring pair only reaching  $M = 0.31$  and the highest scoring pair reaching  $M = 0.86$  in accuracy. In the subgroups of the shared and unshared conditions, the mean outcome was higher for the dyads with shared visual context ( $M = 0.65$  and  $M = 0.51$ , respectively). Fig. 5 illustrates the mean development over time in the two conditions.

To test the predictions regarding shared visual context and communicative success (predictions 1–3), a logistic mixed effects model with the accuracy outcome was constructed. First, the two predictor variables were centered to remove collinearity between the main effects and the interaction, and thus to make parameters interpretable as the total main effects (Schielzeth, 2010). Then the model was estimated using the R package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). More precisely, accuracy was predicted by shared visual context (dummy-coded with 1 being *shared*), trial number, and their multiplicative interaction, while the maximal random effects structure was included in the model (cf. Barr, Levy, Scheepers, & Tily, 2013). This maximal structure consisted of random intercepts for pairs and random slopes for trial number.

There were significantly positive estimates for the effects of shared visual context and trial number, but not for their interaction (see Table 1). This means that participants' performance was significantly better in the shared condition (prediction 1), and significantly better the more trials they had played in the game (prediction 2); however, participants in the shared condition did not progress faster in their overall performance (no evidence for prediction 3).

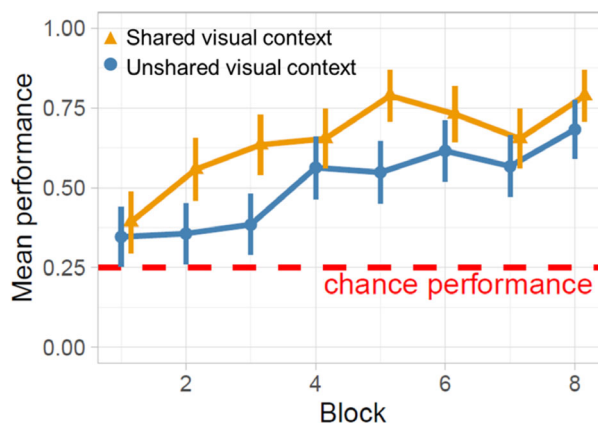


Fig. 5. Development of performance over time in Experiment 1, in blocks of eight trials. Error bars represent 95% confidence intervals. In both conditions, the mean trend is that pairs started out slightly above chance and generally improved in performance. However, the performance in the shared condition is elevated, compared to the unshared condition.

2.2.2. Exploration of the questionnaires

To get a better sense of how participants used the symbols in the task, we inspected the symbol lists they created after completing the main experiment. As can be seen exemplarily for the 10 most frequently reported symbols in Fig. 6, symbols were used by most senders as substitutes for color terms. However, there were also other reported meanings, such as symbols indicating subjective brightness, mixing of colors, correct and wrong answers, or even the fact that the target was a “basic” (i.e., primary) color. Different conventions arose in different pairs, and in fact no symbol was exclusively used for one color only.

As a proxy for conventions, we counted the number of cases in which the sender and receiver of each pair reported the same symbol and agreed on the same meaning for it. This analysis should be seen as supplementary, as there was some vagueness involved in the free descriptions provided by the participants. Moreover, the free recall meant that the players did not necessarily remember the same symbols after the game. For these

Table 1  
Estimates and *p*-values for the accuracy model in Experiment 1

Fixed Effect	$\beta$	SE	<i>p</i>
Intercept	0.39	0.12	<.002
Trial number	0.03	0.003	<.001
Shared visual context	0.69	0.25	<.005
Trial number × Shared visual context	0.002	0.007	<.733

Note. *p*-values <.05 are marked in bold.

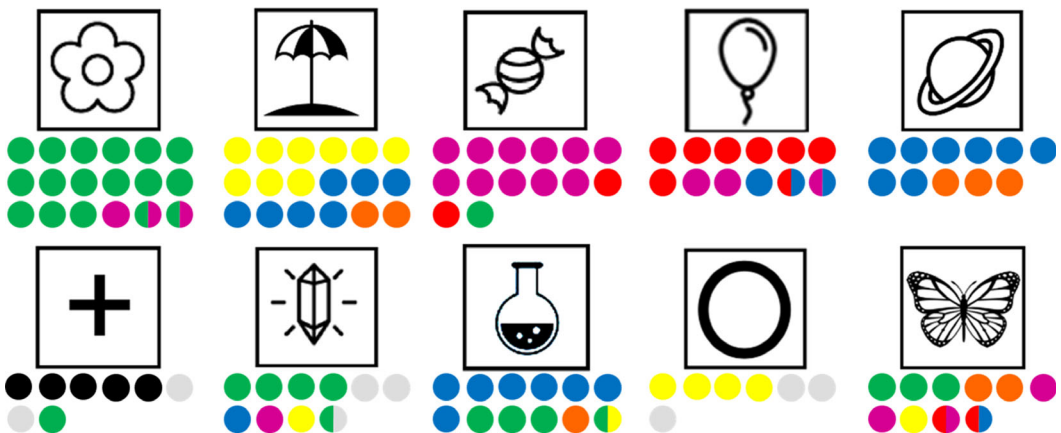


Fig. 6. Meanings of the 10 most frequently reported symbols, as recalled by the senders. Every dot below a symbol stands for one sender reporting that the symbol was used for the respective color. Double colored dots indicate that the sender reported using it for two different colors, and black or light gray dots mean dark or bright colors. In addition to the meanings presented in the figure, senders reported using the “plus” symbol on the left for mixing colors in five cases, and the “circle” symbol on the right for “basic colors” in one case.

reasons, we refrained from computing more than descriptive values for the agreements and merely wanted to get a first indication about how many conventions arose in the two conditions. On average, pairs in the unshared condition agreed on  $M = 4.08$  ( $SD = 2.28$ ) symbol meanings, and pairs in the shared condition agreed on  $M = 6.15$  ( $SD = 3.18$ ) symbol meanings.

### 2.3. Discussion

The results of Experiment 1 suggest that the shared visual context is helpful for the successful emergence of communication. Dyads in the shared condition outperformed dyads in the unshared condition. Participants managed to communicate above chance, and overall performance increased over time, indicating the formation of novel conventions. Participants mostly formed conventions for symbols by mapping them to color categories, although some were also used for the mixing of colors, lightness, or even more abstract meanings. Descriptively, these conventions also seemed to be more frequent for pairs in the shared condition.

However, this last result was merely explorative, lacking a rigorous test. We also did not investigate the generalizability of conventions in the two conditions. In addition, some methodological considerations can improve the experimental paradigm. For instance, the colors used as targets in the experimental trials (randomly chosen from a  $360^\circ$  space) might differ in their difficulty, limiting our control over this variable. The symbol space was also quite large, with some symbols clearly outperforming others and becoming very popular, while some were rarely used. We tried to address these issues in Experiment 2.

## 3. Experiment 2

Experiment 2 set out to replicate the main results of Experiment 1 with a larger sample size, improving on the paradigm in several ways, especially with regard to the meaning and signal spaces. Additionally, we aimed to test for more frequent conventions in the shared condition more rigorously by including a systematic questionnaire at the end of the experiment (prediction 4). Lastly, we predicted that conventions in the shared condition should also become easier to generalize and use in a new referential context (prediction 5), because we expected communication to be more successful. We address this by switching to a different color space after the first half of the experimental task; assuming that it is functional for successful communication to reuse symbols, symbols from the first half of the experiment should be reused more often in the shared condition.

### 3.1. Method

Where not explicitly mentioned in the subsequent paragraphs, the experimental design was the same as in Experiment 1.



### 3.1.1. Participants

This time, 96 participants (89 students) were recruited. Their mean age was 24 ( $SD = 4.0$ ); 81 were female and 15 male. Since 11 participants did not report German as their native language, we made sure (before the experiment) that all participants were fluent speakers of German and had no problems understanding the printed instructions. All participants showed typical color vision in the test for color blindness, and none had taken part in Experiment 1.

### 3.1.2. Materials

In this experiment, we improved our control over the difficulty of target referents (and their arrays) by using an artificially discretized set of colors. Similar to the first color space, physical saturation and lightness were kept constant and colors were only varied in hue. However, this time, we chose a discrete space of 32 colors by applying the CIE2000 formula (created to reflect perceptual differences; cf. Luo, Cui, & Rigg, 2001) to create a circle of perceptually equidistant colors, each in a distance of  $\Delta E = 7.8$  ( $\Delta E$  representing the distance between colors in the CIE2000 space) from their two neighbors (see Fig. 7). Additionally, this color space was split in half for each dyad to allow testing for the generalization of conventions. This resulted in two half-circles, each representing the color space for one half of the experiment, respectively. Since the location of this split in the color circle was arbitrary and might influence the results of the experiment, dyads started with different halves of the space, in total reflecting the full spectrum. This was counter-balanced between conditions (for a visualization, see Fig. 8).

The signal space in Experiment 2 consisted of a subset of the symbols used in Experiment 1; we removed those symbols that were used barely or almost constantly, leaving

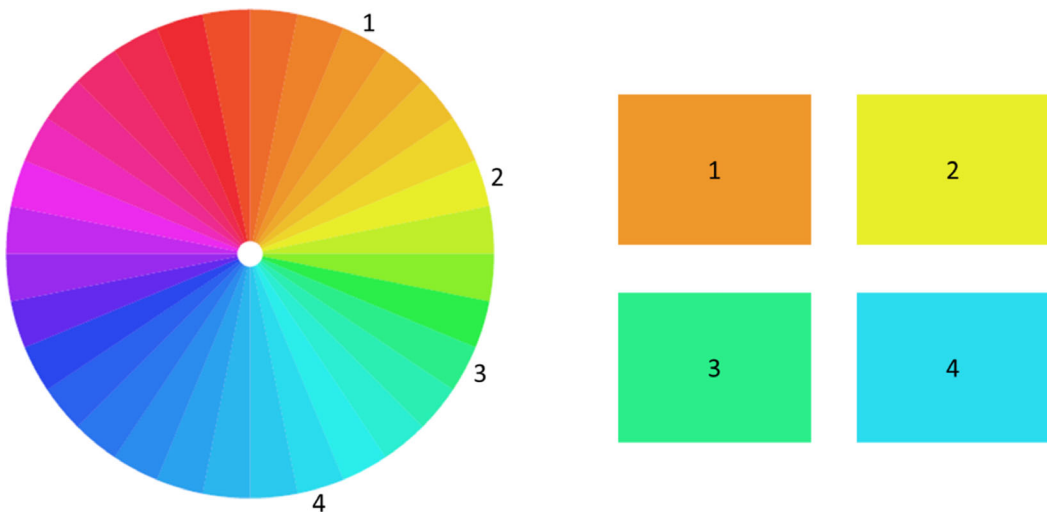


Fig. 7. Left: All 32 colors comprising the color space in Experiment 2. The space can be split in half by drawing a straight line at any border between colors. Right: A color array created from this space.

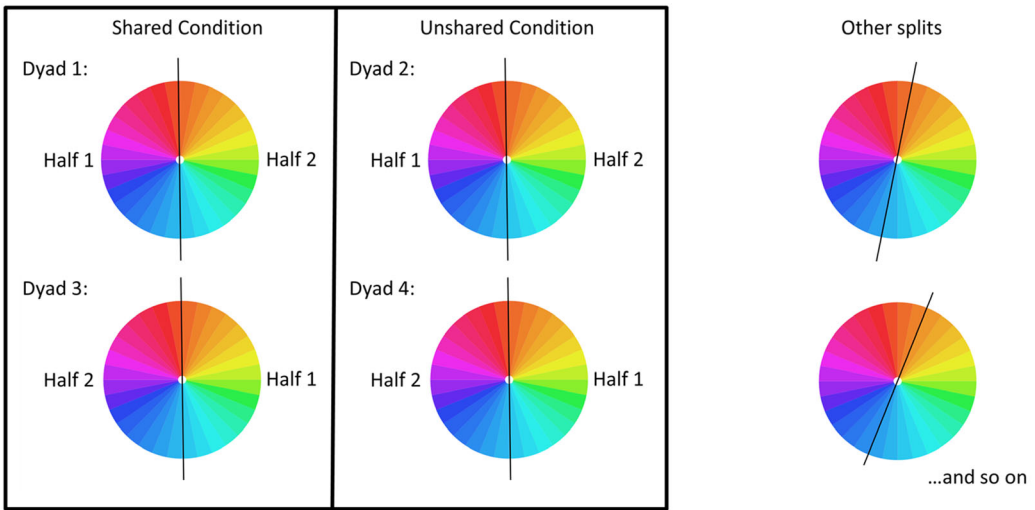


Fig. 8. Left box: Counterbalancing the color space between conditions. Each split occurred four times: once in each condition, and once in regular and reversed order. Right: Visualization of how the space can be split in different ways. Note that when the line has traversed half of the circular space ( $180^\circ$ ), we arrive at the same splitting pattern as at the example on the left ( $0^\circ$ ).

us with 23 symbols in total (see Fig. 9). The reasoning behind this was to remove the least ambiguous symbols (enabling rather similar and easy communication) and the least useful symbols (with very low usage numbers, making them less comparable).

### 3.1.3. Procedure

The procedure closely followed the design outlined for Experiment 1.<sup>1</sup> Twenty-four pairs each played in one of the shared visual context conditions. Without notice to the participants, the color arrays presented in the second half of the experiment (i.e., the second set of 32 trials, or the last 4 blocks out of 8) were switched and only drawn from the half of the color space the dyad had not encountered previously. Because the meaning space was discrete, we counterbalanced the color arrays presented and the targets chosen from them, such that each color appeared as the target twice. After completion of the experiment, an “alignment questionnaire” was handed to both participants, in which they had to tick a description for each of the 32 colors presented in the experiment.

## 3.2. Results

### 3.2.1. Can the results of Experiment 1 be replicated?

Before the analyses, 15 trials (0.5% of the total sample) were excluded from the data for the following reasons: 13 trials were lost due to crashes, and two trials had reaction times below 3 seconds (same guideline as in Experiment 1). Compared to Experiment 1, the task was slightly harder: The mean accuracy across both conditions was  $M = 0.46$

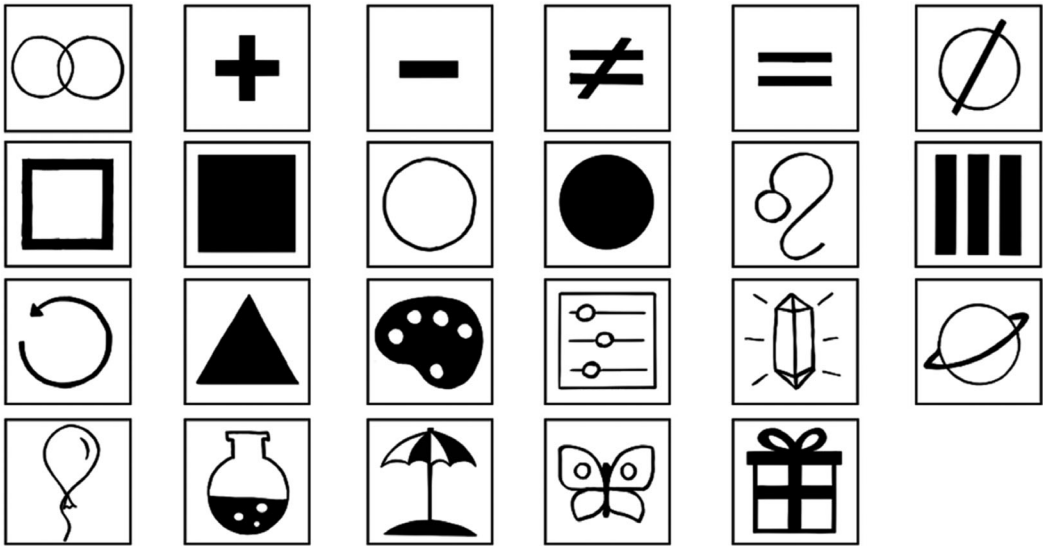


Fig. 9. List of symbols available to senders in Experiment 2.

( $SD = 0.50$ ). This time, the lowest scoring pair only reached  $M = 0.17$  in accuracy, while the highest scoring pair reached  $M = 0.80$ . In the subgroups of the different conditions, shared pairs were again better on average than unshared pairs ( $M = 0.52$  and  $M = 0.39$ ). Fig. 10 illustrates the development of performance over time.

To replicate the results of Experiment 1, a logistic mixed effects model with trial accuracy as the outcome variable was constructed, following the analytic strategy of the

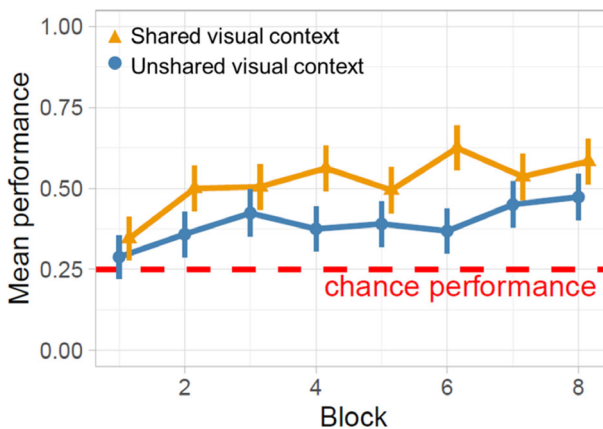


Fig. 10. Development of performance over time in Experiment 2, in blocks of eight trials. Error bars represent 95% confidence intervals. Again, pairs in both conditions generally improved in performance, but performance in the shared condition is elevated. Pairs in the shared condition showed decreased performance in block 5 (right after the change in color arrays), but recovered in the remaining blocks.

Table 2  
Estimates and  $p$ -values for the accuracy model in Experiment 2

Fixed Effect	$\beta$	$SE$	$p$
Intercept	-0.18	0.09	.059
Shared visual context	0.56	0.17	<.001
Trial number	0.01	0.003	<.001
Shared visual context $\times$ Trial number	0.003	0.007	.597

Note.  $p$ -values <.05 are marked in bold.

previous model. This time, we were able to also control for random effects of the color arrays in addition to random participant effects because of the discrete color space. There were significantly positive estimates for the effects of shared visual context and trial number, but not for their interaction, replicating the results of Experiment 1 (see Table 2).<sup>2</sup> This means that participants' performance on the accuracy outcome was significantly better in the shared condition (prediction 1), and significantly better the more trials they had played in the game (prediction 2). Again, participants in the shared condition did not progress faster in their overall performance (no evidence for prediction 3).

### 3.2.2. Does the shared visual context increase the number of conventions?

This time, the more systematic questionnaires allow us to separate the reported colors one by one instead of relying on categories chosen by the senders. This highlights the diversity of conventions in different pairs even more (cf. Fig. 11). Again, symbols were mapped to specific color hues, but also to other features such as perceived brightness levels.

We computed the normalized Levenshtein distance within dyads to measure their alignment after the experiment. This was done using the R package *stringdist* (Van der Loo, 2014) for each of the 32 colors assessed in the post-experiment questionnaire, with missing values for each color in a given pair if either of the participants had not chosen any symbol for the color. This conservative approach produced a large amount of missing values (274 cases or 17.8% of the sample). The strings compared were composed of a single (unique) letter for each symbol used for the respective color. For example distances on highly aligned and lower aligned strings, see Fig. 12. This string distance did not take the order of symbols into account, as that information was not available from our study design: During the task, symbols could be arranged freely on the whiteboard space, and thus order information was not obtained in the questionnaires. We constructed a linear mixed effects model in which the distance was predicted by the shared visual context, with random intercepts for pairs and colors. There was a positive estimate for the effect of shared visual context ( $\beta = 0.08$ ,  $SE = 0.05$ ), but it was nonsignificant ( $p = .118$ ). This means we could find no support for our prediction (4).

### 3.2.3. Are conventions developed by shared visual context pairs more generalizable?

To measure the generalization of conventions to new contexts, we look at functional symbol reuse in the second half of the experiment. Since our hypothesis was based upon the assumption that reuse was functional, we had to test this first. We computed the

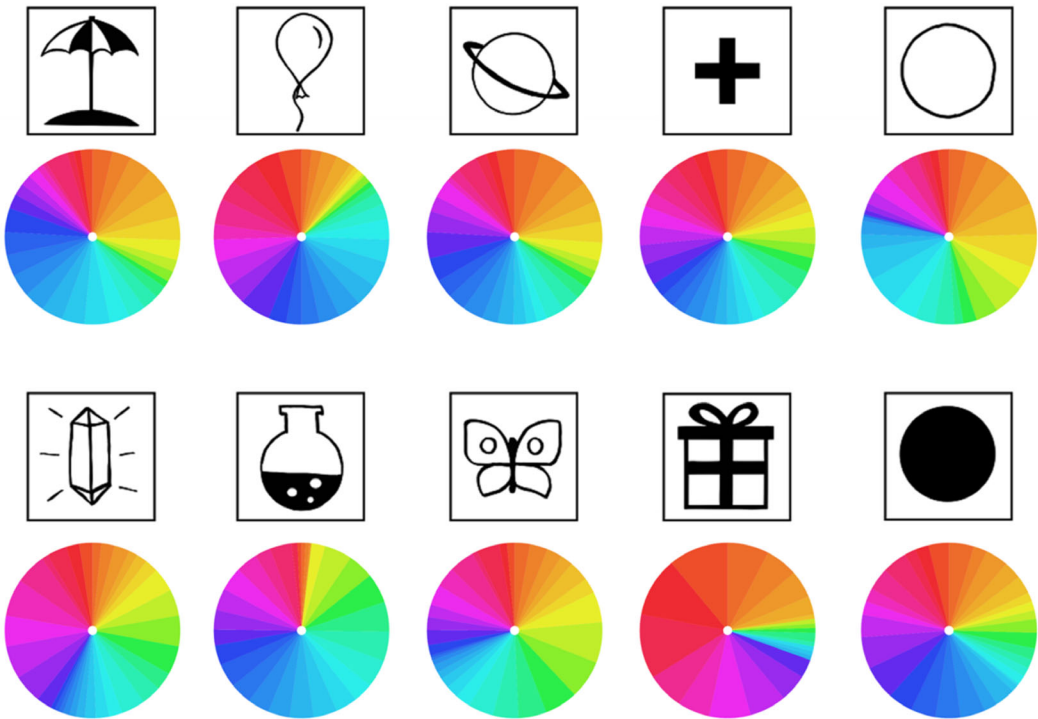


Fig. 11. Sender reports about the usage of symbols in Experiment 2 with regard to all 32 colors, for the symbols presented in Fig. 6 (+ two new ones). The size of the area occupied by every color corresponds to the number of reported uses in the questionnaires. Distributions similar to Experiment 1 can be observed, like the “planet” symbol (top middle) being used mainly for blue and orange colors, or the “flask” symbol (second from the left in the bottom) being used mainly for blue and green. The two circles (at the right in both rows) show a clear distinction for brightness: The top one is mainly used for subjectively brighter colors, and the bottom one is mainly used for subjectively darker colors.

number of symbol types reused (in the same pair) for every single trial of the experiment in half 2, that is, whether a symbol type that had been used at any time in half 1 appeared in the relevant trial. If symbol reuse was functional, this variable should be a significant predictor of accuracy in those trials. Before implementing these variables in a model, symbol reuse values were normalized by the total amount of types (reuse and novel use) appearing in the trial to account for differences in length of messages. The average proportion of symbols in the messages in half 2 that had been used in half 1 already was  $M = 0.91$  ( $SD = 0.18$ ).

We used a logistic mixed effects model in which accuracy was predicted by reuse and shared visual context, with random intercepts for pairs and target colors and a random slope for reuse (only on the intercept for pairs; for colors, a random slope was not as feasible design-wise, since color spaces varied between pairs), to test for functional reuse. There was a significantly positive estimate for shared visual context (see Table 3), replicating the accuracy result from above, and a positive but nonsignificant estimate for
















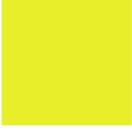




Target color	Sender description	Receiver description	Normalized Levenshtein distance
	  	  	0
	  	   	0.5
	 	 	1

Fig. 12. Example Levenshtein distances capturing sender–receiver alignment from the self-reported symbols describing specific colors in the post-experiment questionnaire. Example strings of senders and their respective receivers are in the middle columns. If a pair is perfectly aligned (top example), the normalized Levenshtein distance is 0; if strings differ completely (bottom example), it is 1.

Table 3  
Estimates and *p*-values for the fixed effects in the model for functional reuse

Fixed Effect	$\beta$	<i>SE</i>	<i>p</i>
Intercept	-0.02	0.13	.905
Shared visual context	0.69	0.24	<b>.005</b>
Reuse	0.07	0.58	.905
Shared visual context × Reuse	2.31	1.15	<b>.045</b>

Note. *p*-values <.05 are marked in bold.

reuse. Additionally, there was a significant positive interaction between shared visual context and the amount of reuse, indicating that pairs in the shared condition were performing even better when they were reusing more symbols, whereas pairs in the unshared condition were performing worse when they were reusing more symbols (for a visualization, see Fig. 13). Thus, reuse was functional for shared condition pairs, but not so for unshared condition pairs.

We then computed a new variable indicating whether any symbol used in half 1 was reused by the same pair in the second half, coded in a binary fashion ( $M = 0.64$ ,  $SD = 0.48$  for shared condition;  $M = 0.73$ ,  $SD = 0.44$  for unshared condition). This variable was predicted, in a logistic mixed effects model, by shared visual context, with



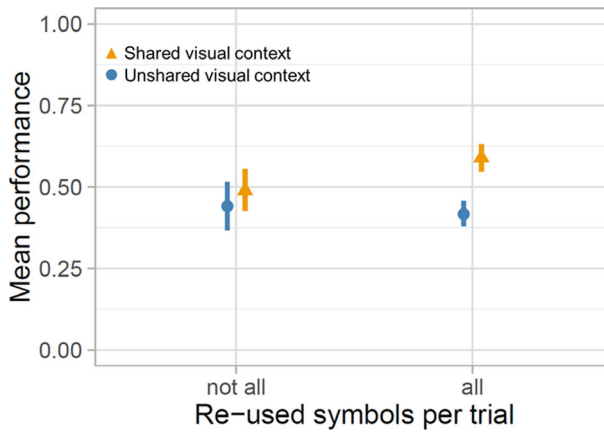


Fig. 13. Relationship between symbol reuse and accuracy in the two conditions. Because of the generally high amounts of reuse in the data, it was dichotomized into trials that consisted entirely of reused symbols and trials that saw some or no reuse of symbols, for purposes of visualization. Error bars represent 95% confidence intervals. It can be seen that shared context pairs outperform unshared context pairs whether they reuse symbols or not. Furthermore, an interaction effect is visible: Shared context pairs performed even better when they were reusing symbols, whereas unshared context pairs performed even worse when they were reusing symbols.

random intercepts for pairs and symbols. There was a significant negative effect for shared visual context ( $\beta = -0.47$ ,  $SE = 0.20$ ,  $p = .017$ ). This means that unshared condition pairs reused more symbols in half 2 than shared condition pairs (although it was not functional for them; this is contrary to prediction 5).

### 3.3. Discussion

Experiment 2 replicated the main results of Experiment 1, demonstrating for the second time the importance of the shared visual context for the successful emergence of novel communication. This could still be shown while controlling for color difficulty and the perceptual distance between colors, and with a reduced set of symbols. The task was harder overall, but dyads managed to improve in performance over time, even when generalizing to a novel reference space. We found no evidence that conventions were more frequent for pairs in the shared condition. Contrary to our expectations, pairs in the unshared condition were reusing more of their symbols in the second half; however, at the same time their reuse did not appear to be as functional as it was for pairs in the shared condition.

## 4. General discussion

In two experiments, this study demonstrates that the shared visual context between two interlocutors is useful for the emergence of communication, enabling more success when

developing a novel communication system. This is shown by our main result: the higher accuracy for dyads in the shared visual context condition, compared to the unshared condition. We demonstrated the importance of the shared context for the successful emergence of conventions even in the absence of training on symbol meanings and external feedback. Instead, participants had to rely on inference and interaction, which potentially amplified the contextual effects we wanted to investigate. Our results represent direct empirical evidence for theories emphasizing the importance of context for successful communication (e.g., Clark, 1996; Sperber & Wilson, 1996), and extend these considerations to the study of language emergence.

The *emergence* of conventions implies we saw novel conventions arising within dyads during the experiments. Participants could not achieve a high level of success without endowing vague symbols with novel meanings. While we acknowledge that providing participants with pre-established symbols means they already bring formed associations into the experiment, we argue that this cannot be the main factor behind the conventionalization. At most, it biases participants to prefer certain colors while the desired ambiguity in the selection of our symbol space remains. In other words, symbols may carry a little information from the start, but over the course of the task, they acquire much more. It is this increase in informational value that we were interested in, not possible prior associations for symbols.<sup>3</sup> The evidence we provide for this in the two experiments is twofold: First, conventions between pairs differed drastically, as seen in the results from the questionnaires in both experiments. Second, performance continually increased over time in both conditions and in both experiments, implying that participants built up a shared conversational history—and thus, conventions.

However, our alignment questionnaire in Experiment 2 failed to show a difference between the two conditions for the number of conventions arising, even though their success differed. It could be the case that the shared context does not facilitate the creation of conventions but merely boosts the successful emergence of communication. We would argue that the lack of evidence for the predicted effect is likely due to a methodological problem, however, especially since we could observe the desired pattern in the explorative results of Experiment 1. The rigorous questionnaire we employed in Experiment 2, prompting participants to tick a description for every single one of the 32 colors in the meaning space, unfortunately produced a lot of missing values. The participants felt overwhelmed by the precision required for this task, and in many cases reported being unable to remember or make an educated guess about how the symbols were used for a particular color. This was particularly true of receivers, who had to infer which meanings their sender had associated with which symbol. Because of our conservative approach in the analysis, even one missing message from either of the participants led to an exclusion of the alignment for the whole color for the pair (i.e., a given row in the data). Future studies would be well placed to investigate the importance of contextual knowledge for the formation of conventions much in the same way as we did in Experiment 1, but employ more controlled (i.e., fixed) categories instead of free recall, while not presenting participants with the entire meaning space.

In neither of the experiments did we find evidence that dyads in the shared condition progressed faster in performance than unshared visual context pairs. Following this result, we have to assume that in our specific task, shared visual context pairs were not able to capitalize more successfully on their conventions and their performance, which was already higher from the start. This would suggest that the effect of the shared visual context is fixed, elevating pairs' performance rather than multiplying it with more experience. On the flip side, it means that pairs in the unshared condition were able to improve equally well in the task, just overall below the performance of the shared condition. It would be interesting to change our experimental design to include a within-pair manipulation of the context to investigate whether dyads would immediately profit or suffer from switching to shared or unshared contexts.

Experiment 2 tested the prediction that senders in the shared condition would be more likely to reuse their symbols, with the assumption that such reuse should be functional for all dyads. We made this prediction because we believed shared information would foster the evolution of more generalizable conventions. This prediction could not be tested, because reuse turned out to be functional only for dyads in the shared condition; in the unshared condition, reuse did not help performance. Though less functional, reuse was more frequent in the unshared condition—surprisingly, in light of our initial prediction.

We do not know what made senders in the unshared condition reuse symbols more than pairs in the shared condition. Potentially, the access to the shared visual context might have tempted pairs to create conventions that rely on it to carry a good part of their intended meaning; in other words, shared visual context pairs might have made use of their opportunity for contextual enrichment (cf. Winters et al., 2018), leading to a greater need for novel symbols once the contexts changed. Alternatively, we suspect that the shared visual context could have made the transition between half 1 and half 2 more salient, encouraging senders to change their repertoire of symbols.<sup>4</sup> In one respect, however, Experiment 2 verified our expectation that the conventions evolved in the shared condition would be more generalizable: Symbol reuse resulted in better performance in the shared condition, and in that condition only. This is consistent with the general view that linguistic conventions emerge by being used in ostensive-inferential communication (Höfler, 2009), and with the specific claim that the shared visual context makes for more efficient communication, yielding more generalizable conventions.

At the center of our study was the manipulation of the shared visual context. As described in the introduction, this is merely one aspect of the general notion of context, and it ignores other types such as the historical context (e.g., Yoon, Benjamin, & Brown-Schmidt, 2016) or the basic community membership (e.g., Clark, Schreuder, & Buttrick, 1983) of participants, interesting objects of study in themselves. Interestingly, there is a case to be made for our manipulation also concerning the historical context: Dyads in the unshared condition were limited to a history of unshared contexts in addition to their immediate situation, and also switched to a new set of unshared contexts in Experiment 2. As such, we cannot separate the effects of the immediate shared context and the shared context accumulating over time. However, this is less problematic since our main interest lay in the evaluation of the shared effect, which entails both of these confounded aspects.

Our manipulation to the shared context was achieved by removing all distractor colors from the arrays of senders in the task, so that they only knew what the target in the current trial was. It is important to note that different operationalizations of the shared context would have been possible: For instance, another option would have been to present entirely different contexts to sender and receiver (with the same target color), but keep the amount of colors the same. We decided not to do this because (a) it is difficult to keep the differences between and within conditions constant with this design and (b) participants would probably have to be deceived about them not seeing the same colors in this case. In contrast, we settled for an open and informed quantitative manipulation of the shared context, such that only the amount of colors varied. As such, we expected and found quantitative differences in the performance of dyads as well; nevertheless, the question whether this result generalizes to other operationalizations of the shared context would need to be addressed empirically by future studies.

Another open question concerns the cognitive representations underlying the more successful communication in the shared condition. Do interlocutors take their partner's knowledge into account to communicate accurately? Some theories suggest that they should (Clark, 1996; Lewis, 1969). However, as outlined in the introduction, this point has been challenged by a line of research studying reference resolution with the eye-tracking method (e.g., Horton & Keysar, 1996; Keysar et al., 2000). As such, the results of our experiments are also in line with a more parsimonious explanation (Keysar, 1997): Senders could simply be better at the task because there is more knowledge available to them. It is important to note that this is still in agreement with Sperber and Wilson's relevance theory (Sperber & Wilson, 2002). Here, shared representations are not always necessary for communication, but the individual representation of contextual information for both interlocutors is often sufficient. In our case, the senders' messages could simply be built on their contextual information, and likewise, the receivers' inferences could be drawn from the message combined with what they see on their screen. Minimally, then, we have shown the benefits of the shared context for the successful emergence of communication, but we are not making any claims as to how this context is used by the interlocutors exactly.

We grounded our experimental design and the general research question about contextual influences on the emergence of language in an ostensive-inferential model of communication. This led to a number of design choices, most notably the absence of training for any symbol meanings and the reliance on repair mechanisms combined with a lack of external feedback. By doing this, we aimed to come closer toward the emergence problem of communication. Participants were encouraged to create novel conventions through interaction. Although we acknowledge that there might still be biases from interference with their natural language (a general problem for artificial language experiments), we think it is necessary to eliminate as many alternative mechanisms for the formation of novel conventions as possible in an experimental setting. All in all, we think the current study provides a firm basis for how future studies can utilize the ostensive-inferential framework to investigate the emergence of language.

## 5. Conclusion

In this paper, we set out to investigate the influence of the shared visual context on the successful emergence of communication. To this end, we combined pragmatic concepts with the methods of experimental semiotics. We constructed two artificial language experiments and found that participants performed better in a referential task when they had access to the visual context. This has implications for the emergence of language, and it is in accordance with an ostensive-inferential model of communication: To successfully create and interpret a novel convention, interlocutors build on the contextual information. In the second experiment, we also found that participants sharing the visual context adapted their conventions more successfully to new contexts than those lacking the context. At the same time, unshared visual context pairs reused more of their conventions, the reasons for which remain unclear. On the methodological side, our experiments demonstrate how an ostensive–inferential framework can be used to inform choices in the designs used by artificial language experiments, emphasizing inferential processes and interaction.

## Acknowledgments

We thank Helene Kreysa and Dana Schneider for their role in supervising Thomas Müller's master thesis, partly reflected in Experiment 1. My special thanks to Lisa Jeschke for her help with organizing and conducting Experiment 2.

## Data and code availability

All data and R code are available on the Open Science Framework: <https://osf.io/ts4ka/files/>.

## Notes

1. There were two different experimenters this time, who followed the same parallel procedure when conducting the study.
2. Adding an effect for the two experimenters that had conducted the study did not reveal any differences between them.
3. For a supplementary analysis regarding the effect that biased associations might have on successful or unsuccessful communication, see the appendix.
4. An additional suggestion brought forward during review is that reuse in shared context pairs might have focused on symbol *combinations* rather than single symbols. We address this idea with a supplementary analysis in the appendix.

## References

- Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, 46(2), 391–418. <https://doi.org/10.1006/jmla.2001.2815>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: California University Press.
- Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, 35(7), 1207–1242. <https://doi.org/10.1111/j.1551-6709.2011.01189.x>
- Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell, & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 95–129). Cambridge, MA: MIT Press.
- Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, 1(2), 274–291. <https://doi.org/10.1111/j.1756-8765.2009.01019.x>
- Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, 61(2), 171–190. <https://doi.org/10.1016/j.jml.2009.04.003>
- Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27(1), 62–89. <https://doi.org/10.1080/01690965.2010.543363>
- Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, 107(3), 1122–1134. <https://doi.org/10.1016/j.cognition.2007.11.005>
- Christiansen, M. H., & Kirby, S. (2003). *Language evolution*. Oxford, UK: Oxford University Press.
- Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.
- Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Hillsdale, NJ: Erlbaum.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81. <https://doi.org/10.1016/j.jml.2003.08.004>
- Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1), 19–41. <https://doi.org/10.1080/01690968708406350>
- Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2), 245–258. [https://doi.org/10.1016/S0022-5371\(83\)90189-5](https://doi.org/10.1016/S0022-5371(83)90189-5)
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39. [https://doi.org/10.1016/0010-0277\(86\)90010-7](https://doi.org/10.1016/0010-0277(86)90010-7)
- Craycraft, N. N., & Brown-Schmidt, S. (2018). Compensating for an inattentive audience. *Cognitive Science*, 42(5), 1504–1528. <https://doi.org/10.1111/cogs.12614>
- Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25(3), 203–219. [https://doi.org/10.1016/0022-1031\(89\)90019-X](https://doi.org/10.1016/0022-1031(89)90019-X)
- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5), 737–767. [https://doi.org/10.1207/s15516709cog0000\\_34](https://doi.org/10.1207/s15516709cog0000_34)
- Galantucci, B. (2009). Experimental semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, 1(2), 393–410. <https://doi.org/10.1111/j.1756-8765.2009.01027.x>
- Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience*, 5, 1–11. <https://doi.org/10.3389/fnhum.2011.00011>



- Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, 6(8), 477–493. <https://doi.org/10.1002/lnc3.351>
- Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62(1), 35–51. <https://doi.org/10.1016/j.jml.2009.09.002>
- Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6), 961–987. <https://doi.org/10.1080/03640210701703659>
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1), 8–11. <https://doi.org/10.1016/j.tics.2003.10.016>
- Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2012). What's learned together stays together: Speakers' choice of referring expression reflects shared experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 843. <https://doi.org/10.1037/a0029467>
- Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, 49(1), 43–61. [https://doi.org/10.1016/S0749-596X\(03\)00022-6](https://doi.org/10.1016/S0749-596X(03)00022-6)
- Healey, P. G., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, 31(2), 285–309. <https://doi.org/10.1080/15326900701221363>
- Heller, D., Gorman, K. S., & Tanenhaus, M. K. (2012). To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2), 290–305. <https://doi.org/10.1111/j.1756-8765.2012.01182.x>
- Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, 108(3), 831–836. <https://doi.org/10.1016/j.cognition.2008.04.008>
- Höfler, S. (2009). Modelling the role of pragmatic plasticity in the evolution of linguistic communication. Doctoral dissertation, University of Edinburgh, Edinburgh.
- Höfler, S., & Smith, A. D. M. (2009). The pre-linguistic basis of grammaticalisation: A unified approach to metaphor and reanalysis. *Studies in Language*, 33(4), 886–909. <https://doi.org/10.1075/sl.33.4.03hoe>
- Holler, J., & Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24(2), 267–289. <https://doi.org/10.1080/01690960802095545>
- Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, 59(1), 91–117. [https://doi.org/10.1016/0010-0277\(96\)81418-1](https://doi.org/10.1016/0010-0277(96)81418-1)
- Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1), 26–37. <https://doi.org/10.1037/0096-3445.116.1.26>
- Ishihara, S. (1972). *The series of plates designed as a test for colour-blindness*. Tokyo, Japan: Kanehara & Co., Ltd.
- Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, 24(2–3), 253–270. <https://doi.org/10.1080/01638539709545015>
- Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1), 32–38. <https://doi.org/10.1111/1467-9280.00211>
- Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, 39(1), 1–20. <https://doi.org/10.1006/jmla.1998.2563>
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686. <https://doi.org/10.1073/pnas.0707835105>
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102. <https://doi.org/10.1016/j.cognition.2015.03.016>

- Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, 9(1), 2–24. <https://doi.org/10.1521/soco.1991.9.1.2>
- Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1–12), 113–114. <https://doi.org/10.3758/BF03342817>
- Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3), 343–346. <https://doi.org/10.1037/h0023705>
- Krauss, R. M., & Weinheimer, S. (1967). Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, 6(3), 359–363. [https://doi.org/10.1016/S0022-5371\(67\)80125-7](https://doi.org/10.1016/S0022-5371(67)80125-7)
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford, CA: Stanford University Press.
- Lewis, D. (1969). *Convention*. Cambridge: Harvard University Press.
- Luo, M. R., Cui, G., & Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5), 340–350. <https://doi.org/10.1002/col.1049>
- Millikan, R. G. (2005). *Language: A biological model*. New York: Oxford University Press.
- Moreno, M., & Baggio, G. (2015). Role asymmetry and code transmission in signaling games: An experimental and computational investigation. *Cognitive Science*, 39(5), 918–943. <https://doi.org/10.1111/cogs.12191>
- Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4), 329–336. <https://doi.org/10.1111/j.0956-7976.2002.00460.x>
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>. Accessed June 27, 2019.
- Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1(2), 103–113. <https://doi.org/10.1111/j.2041-210X.2010.00012.x>
- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21(2), 211–232. [https://doi.org/10.1016/0010-0285\(89\)90008-X](https://doi.org/10.1016/0010-0285(89)90008-X)
- Scott-Phillips, T. C. (2015). *Speaking our minds: why human communication is different, and how language evolved to make it special*. Houndmills, Basingstoke, Hampshire; New York, NY: Palgrave Macmillan.
- Scott-Phillips, T. C. (2017). Pragmatics and the aims of language evolution. *Psychonomic Bulletin & Review*, 24(1), 186–189. <https://doi.org/10.3758/s13423-016-1061-2>
- Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9), 411–417. <https://doi.org/10.1016/j.tics.2010.06.006>
- Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2), 226–233. <https://doi.org/10.1016/j.cognition.2009.08.009>
- Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1), 212–226. <https://doi.org/10.1111/cogs.12150>
- Skyrms, B. (2010). *Signals: Evolution, learning, and information*. New York: Oxford University Press.
- Sperber, D., & Wilson, D. (1996). *Relevance: Communication and cognition* (2nd ed.). Oxford, UK; Cambridge, MA: Blackwell Publishers.
- Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1–2), 3–23. <https://doi.org/10.1111/1468-0017.00186>
- Sulik, J., & Lupyan, G. (2018). Perspective taking in a novel signaling task: Effects of world knowledge and contextual constraint. <https://doi.org/10.31234/osf.io/ftz94>
- Tamariz, M. (2017). Experimental studies on the cultural evolution of language. *Annual Review of Linguistics*, 3(1), 389–407. <https://doi.org/10.1146/annurev-linguistics-011516-033807>
- Tinits, P., Nölle, J., & Hartmann, S. (2017). Usage context influences the evolution of overspecification in iterated learning. *Journal of Language Evolution*, 2(2), 148–159. <https://doi.org/10.1093/jole/lzx011>
- Tomasello, M. (2010). *Origins of human communication*. Cambridge, MA: MIT Press.

- Van der Loo, M. P. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111–122.
- Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2), 183–194. [https://doi.org/10.1016/0749-596X\(92\)90010-U](https://doi.org/10.1016/0749-596X(92)90010-U)
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7(03), 415–449. <https://doi.org/10.1017/langcog.2014.35>
- Winters, J., Kirby, S., & Smith, K. (2018). Contextual predictability shapes signal autonomy. *Cognition*, 176, 15–30. <https://doi.org/10.1016/j.cognition.2018.03.002>
- Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, 31(1), 169–181. <https://doi.org/10.1080/03640210709336989>
- Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, 154, 102–117. <https://doi.org/10.1016/j.cognition.2016.05.011>

## Appendix

Here, we report two supplementary and exploratory analyses suggested during review. First, we tried to address the question of whether a reason for shared condition pairs reusing fewer symbols in the second half of Experiment 2 might be that they develop conventions for *combinations* of symbols rather than the single symbols themselves. We investigated this by treating symbol use in the second experiment on the level of entire messages, that is, unique symbol combinations instead of single symbols, ignoring reduplications of the same symbol. Descriptively, the diversity of unique combinations used in the second half of the experiment does not differ between the two conditions, as suggested by both the relative proportion of combinations that are duplicates of previous messages (45% vs. 47%) and the conditional entropy of combinations given participant pairs (3.68 bits vs. 3.64 bits). We also repeated our analysis on symbol reuse on the level of unique combinations (as opposed to individual symbols), predicting the reuse of every combination used in half 1 of the experiment (as a binary variable) by condition while adding a random intercept for participant pairs. This model revealed no effect of condition ( $\beta = -0.39$ ,  $SE = 0.34$ ,  $p = .25$ ), leading us to tentatively conclude that pairs in both conditions reused unique combinations of symbols at a similar rate, based on this post hoc analysis.

Second, we ran an exploratory analysis based on Experiment 2 to find out whether players are more successful when using symbols that usually show strong associations with a given color range (cf. Fig. 11). Biased associations were assessed by computing the conditional entropy on the frequencies of colors given symbols: Here, for each symbol, higher values mean more diversity in symbol associations (i.e., a less biased distribution of colors). Interestingly, most symbols end up with very high values of entropy, calculated this way ( $>.9$  on the normalized variable, which takes values between 0 and 1). This can be seen as further evidence that symbol associations were not straightforward for participants and not limited in reference to a selective part of the color space. What we find in the model is that we can replicate the known effects for condition and trial number, but we do not see a significant effect for the mean entropy per trial, even though the parameter points into the expected direction (i.e., higher accuracy for trials that inhibit symbols with more biased associations).