Cultural Evolution of Precise and Agreed-Upon Semantic Conventions in a Multiplayer Gaming App

Olivier Morin, Thomas Müller, Tiffany Morisseau, James Winters

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13113

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13113

Cultural Evolution of Precise and Agreed-Upon Semantic Conventions in a Multiplayer Gaming App

Olivier Morin,[a,b] Thomas F. Müller,[c] Tiffany Morisseau,[d] James Winters[e]

[a]*Minds and Traditions Research Group, Max Planck Institute for the Science of Human History*
[b]*Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University*
[c]*Center for Humans and Machines, Max Planck Institute for Human Development*
[d]*Université de Paris and Université Gustave Eiffel, LaPEA*
[e]*School of Collective Intelligence, UM6P*

Received 5 February 2021; received in revised form 30 November 2021; accepted 21 January 2022

Abstract

The amount of information conveyed by linguistic conventions depends on their precision, yet the codes that humans and other animals use to communicate are quite ambiguous: they may map several vague meanings to the same symbol. How does semantic precision evolve, and what are the constraints that limit it? We address this question using a multiplayer gaming app, where individuals communicate with one another in a scaled-up referential game. Here, the goal is for a sender to use black and white symbols to communicate colors. We expected that the players' mappings between symbols and colors would grow more specific over time, through a selection process whereby precise mappings are preferentially copied. We found that players become increasingly more precise in their use of symbols over the course of their interactions. This trend did not, however, result from selective copying of precise mappings. We explore the implications of this result for the study of lexical ambiguity, Zipf's Law of Meaning, and disagreements over semantic conventions.

Keywords: Language evolution; Sense entropy; Lexical semantics; Zipf's Law of Meaning; Experimental semiotics

Correspondence should be sent to Olivier Morin, Minds and Traditions Research Group, Max Planck Institute for the Science of Human History, 10, Kahlaische strasse, 07745 Jena, Germany. E-mail: morin@shh.mpg.de

1. Introduction

The conventions we use to communicate carry information to the extent that they are specific. Some words or symbols do this better than others. The word "line," having several meanings, is less precise, thus less informative, than the word "teaspoon." A wide variability in the number of potential meanings per word (from "teaspoon" to "line") is observed in many human languages (Calude & Pagel, 2011; Piantadosi, Tily, & Gibson, 2012; Zipf, 1949), and in nonhuman signals (Ferrer-i-Cancho & Lusseau, 2009).

This study asks how precise, informative, and agreed-upon semantic conventions evolve––that is, how people, when they communicate, manage to map signs to meanings that are restricted enough for their purposes, while still agreeing with one another about the meanings of these signs. We explore quantitatively two key aspects of referential communication: the information carried by symbols when they are used to refer to things, and agreement between interlocutors on those symbols' meanings. We expected that, in a repeated referential communication game played with scant feedback and little contextual information, participants would produce increasingly precise ways to use symbols to refer to their targets. We tested two hypotheses regarding this process. First, we expected that precision in symbol use would grow because certain symbols, which lend themselves more to precise referential use, would become increasingly popular—a selection-like process, where people copy the use of informative symbols from each other. Second, we predicted that the most frequently used symbols would also be those whose meaning is most likely to be agreed-upon.

Two main lines of research have investigated informativeness in the conventions that bind meanings to signals. One considers the costs and benefits of informativeness: more precise words tend to be longer and more costly to produce and process (Piantadosi et al., 2011, 2012; Zipf, 1949), partly explaining why informativeness is so unevenly distributed in the lexicon. The second research tradition considers diachronic changes in word meanings: historical trends that lead to semantic narrowing, or its opposite, semantic broadening (Traugott & Dasher, 2009). Relatively little work has asked how the costs and benefits of lexical precision may constrain language evolution.

Why do some words become more precise in meaning, while others follow the opposite trend? Most possible answers revolve around two key variables: cost and context.

Cost: Words that are less costly to produce and process may be reused more easily, thus acquiring new meanings or extending the scope of their original meaning (Piantadosi et al., 2012). This conjecture has the advantage of accounting for the three-way relationship between word lengths, frequencies, and number of meanings identified by Zipf (1949): longer words tend to be less frequent, and frequent words tend to have more meanings.

Context: Pragmatic inferences are another plausible cause for the narrowing or broadening of word meanings (Levinson, 2000; Wilson, 2003). For instance, the use of the phrase "a temperature" in a medical context has produced the narrowed-down meaning of "high body temperature," because that is the most contextually relevant information that this phrase usually conveys (Wilson, 2003). This hypothesis does not specify under what circumstances, or for which words, broadening tendencies should prevail over narrowing tendencies. But there is evidence (both historical and experimental) that highly informative codes tend to evolve

in situations where contextual information is limited (Nölle, Staib, Fusaroli, & Tylén, 2018; Trudgill, 2011; Winters & Morin, 2019; Winters, Kirby, & Smith, 2015; Wray & Grace, 2007). When two interlocutors share an important background of common memories and communicative habits, imprecise messages can suffice to activate precise representations. On the contrary, when shared contextual information is rarefied, for instance, in written communication (Morin, Kelly, & Winters, 2020), or in experiments where anonymous interlocutors are made to communicate with minimal cues (as is the case in many language evolution experiments), a message cannot carry all the information required unless it is encoded with precision.

Understanding how the communicative conventions that attach meanings to symbols or to words become more precise is a question of growing importance in a world where communication is increasingly digital, written, and decontextualized (McCulloch, 2019). One obstacle on this path is the lack of a unified and agreed-upon definition or metric for the precision of symbols. Linguists and philosophers traditionally distinguish two ways that a word may be uninformative or ambiguous: polysemy and vagueness (Geeraerts, 1993; Tuggy, 1993; Wasow, Perfors, & Beaver, 2005). While polysemy refers to the number of distinct meanings that a word may have, vagueness points to the fact that a single meaning may be partly underdetermined, with borderline cases that fit a given meaning in a dubious fashion. "Red" has several meanings—compare "red hair," "red light," and "red eyes"—each of which is somewhat vague (admitting of borderline cases). This twofold characterization of semantic ambiguity suffers from two problems for our purpose. It is difficult to quantify: counting and individuating meanings is a hazardous task, and few measures of vagueness have been put forward. In addition, the distinction between polysemy and vagueness is itself vague (Geeraerts, 1993, 2010). How do we tell whether two instances of word use count as nuances of the same meaning or as two different meanings—is a red-hot iron red in the same sense as a red light? Progress has been made toward a quantitative metric for ambiguity in the word sense disambiguation literature (Edmonds, 2009), where "sense entropy" measures, in information-theoretic terms, the dispersion of word uses between distinct meanings (see also Piantadosi et al., 2012). This study builds on this research.

Precision in using a word is not guaranteed to make that word informative. One needs to make sure that others understand the same word in the same way. An understudied aspect of ambiguity (and its converse, informativeness) is agreement or consensus, which can be defined as a word's capacity to convey a precise meaning from a speaker to an addressee if both of them understand it in the same way. As Enfield (2010) points out, some words are "tolerable friends," carrying distinct but related meanings to different speakers. The English word "peruse" standardly means "to read or examine, typically with great care," but the opposite meaning, "to glance over; to skim," has been gaining in prevalence in the last decades, to the point where a majority of native English speakers now find the sentence "I only had a moment to peruse the manual quickly" acceptable (The American Heritage Dictionary of the English Language, 2011). Unlike false friends, Enfield's "tolerable friends" are not due to cross-linguistic misunderstandings: they may coexist inside one language. Even though "tolerable friends" can support partial communication up to a certain point, they also occasion misunderstandings. Here again, we lack precise metrics to quantify this phenomenon. Yet,

if we want to understand how useful semantic conventions can emerge, we need to understand not just the precision of symbol meanings, but also the degree of overlap between the meanings that users attach to symbols.

Ambiguity may be advantageous for communication, when precision is costly or when context provides sufficient information (Brochhagen, 2020; Piantadosi et al., 2012; Santana, 2014; Winters & Morin, 2019; Winters, Kirby, & Smith, 2018). Ambiguity can also be strategically employed, as is the case when a speaker intentionally aims to be general or vague. Nevertheless, a convention linking a symbol with the same precise meaning for all users is, all else being equal, of clear benefit for information transmission (Gibson et al., 2019). The key function of the conventions linking symbols with meanings is the transmission of information. A symbol that conveys a wide range of meanings is, out of context, less informative than one conveying a narrower signification (Brochhagen, 2020; Piantadosi et al., 2012; Santana, 2014; Wasow et al., 2005). How do such unambiguous semantic conventions evolve? Along with formal models (e.g., O'Connor, 2014; Skyrms, 2010), experiments in artificial language evolution and experimental semiotics (Galantucci, 2005; Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Murthy, Hawkins, & Griffiths, 2021; Raviv, Meyer, & Lev-Ari, 2019; Scott-Phillips & Kirby, 2010) provide tools to answer this question. Experimental setups allow us to simulate the evolution of entirely novel communicative conventions, in controlled situations where the informative power of codes can be measured at every step of their evolution, paralleling techniques developed in the study of natural language lexicon (Regier, Kemp, & Kay, 2015; Zaslavsky, Kemp, Regier, & Tishby, 2018).

Two main types of cultural-evolutionary processes can be distinguished: those where most of evolution is driven by individuals creating novel cultural items and transforming the cultural items that they pass on to others ("transformative" processes); and those where change is due to selective copying, some items being more likely to spread than others because they get imitated to a greater extent—"selectionist" processes (Claidière, Scott-Phillips, & Sperber, 2014; Nettle, 2020). Both accounts agree that evolutionary change feeds upon variation, but disagree upon the nature of variation—to what extent it is constrained by human cognition. In transformative processes, evolution is driven by agents' inventions (introducing novel variants) or directed transformations (changing the variants that they transmit). In selectionist processes, such changes are either random or nonexistent (Nettle, 2020). In the case of communicative conventions, where agents' primary goal is to align their behavior on those of others, it is not unreasonable to assume that agents blindly copy arbitrary conventions, and that successful conventions are imitated to a greater extent than unsuccessful ones. Accordingly, in language evolution research, it has been claimed that all language evolution is ultimately based on replication and selection (Croft, 2019; Tamariz, 2019; Tamariz, Ellison, Barr, & Fay, 2014). There is also evidence that, in artificial language evolution experiments, selectionist processes matter much more than transformative ones (Tamariz et al., 2014, 2017). In a selectionist scenario, precise semantic conventions would evolve because they are more likely to be retained than others, thus becoming more widespread.

This selection-based account suggests a solution to the problem of consensus. In Enfield's hypothesis (Enfield, 2010), it is easier for speakers to associate the same meaning to the same word when the word in question is frequent. Each occasion to hear the word is an

occasion to learn about its meaning. Such learning (the hypothesis goes) is made easier when learners have access to a diverse sample of instances of word use. Typically (although not necessarily), a word that is often used is more likely to be heard in a variety of contexts, making convergence upon similar meanings more likely. This follows from a more general linguistic argument: more frequently used words can be learnt multiple times, with multiplied opportunities to correct irregularities—especially when a word is heard from different people (Bybee, 2010; Diessel, 2007; Lev-Ari, 2017). In a selectionist view, precise conventions gain more users, making them more likely to be frequently encountered. If Enfield's conjecture is right, being frequent could in turn make these conventions more agreed-upon, reducing variance between different individuals' conception of a word's meaning. This hypothesis is exceedingly hard to test on natural language data, not least because of the challenges linked to measuring overlap between mental representations.

We tested the hypothesis according to which conventions mapping a symbol with a narrow range of referents are favored in a setting where participants must communicate with unfamiliar interlocutors, using unfamiliar symbols, with little contextual information to rely on aside from an array of four colors. We expected cultural evolution to favor precise conventions, and we expected this to happen in a selectionist fashion. The rise of precise symbols, we assumed, would be driven by more frequent use for symbols that mapped to a narrower range of colors (Tamariz et al., 2014). These mappings would then be imitated and spread selectively. Two predictions followed from this. First, the symbols that are used with precision should be popular, that is, used more frequently. Second, this effect should be more pronounced with more experienced players. Precision was defined on information-theoretic grounds, as the entropy of the frequency distribution of the referents a symbol was used to designate.

In addition, we expected to see more agreement between interlocutors over the range of meanings of the most frequently used symbols. This prediction is meant to test Enfield's conjecture that the amount of overlap between individual representations of the meaning of a symbol is influenced by the frequency at which this symbol is used (Enfield, 2010). A symbol's meaning, in a referential communication task, can be understood as its extension: as the range of referents that the symbol is used for. Likewise, agreement between two interlocutors on the meaning of a symbol can be measured as the distance between the range of referents that each interlocutor associates with the symbol.

2. Methods

We studied the evolution of communicative conventions over 11 months in an online multiplayer gaming app, the "Color Game" app (Morin et al., 2018; Morin et al., 2020). This study was part of a larger registration that involved six projects making distinct predictions on different aspects of the Color Game App data, five of which were carried to completion. (More detail can be found in the Supplementary Materials.)

The app allowed players from more than 100 nationalities to participate in a series of referential communication games (Lewis, 1969), where one player (the Sender) had to communicate a color (the target, indicated by a dot) to a Receiver presented with an array that included

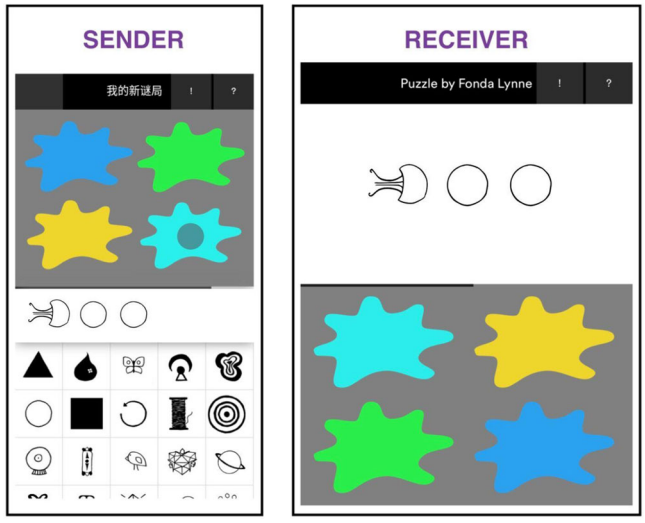

Fig 1. An example of a trial in the Color Game, showing Sender's screen (left) and Receiver's screen (right). (Adapted from Morin et al., 2020).

the target and three other colors. Participants could only communicate through black and white symbols (Fig. 1). Points were earned by both players when a Receiver made a successful pick. The app did not provide players with trial-by-trial feedback on the success or failure of communication. A block of 10 trials had to be played by both Sender and Receiver for points to be earned. After every block, the Receiver was told how many of the preceding 10 trials they got right, but not which ones. This made it more difficult for Receivers to learn by mere association what meaning Senders associated with which symbols, making the task more challenging and, arguably, more naturalistic (in real-life communication, misunderstandings are seldom resolved by outside interventions).

Through the app, players could freely choose to associate with willing coplayers, and could switch between the roles of Sender and Receiver if they found other players willing to play the opposite role. To ensure statistical robustness, the player pool was divided into two closed "halves": a player could only play with coplayers from her half. (Which "half" a player came from did not make a difference to any of the analyses reported here: the corresponding variable did not make the models more informative.) A system of points and a user-friendly design ensured that players were endogenously motivated.

Each of the game's 32 colors was drawn from the CIE2000 color space (Luo, Cui, & Rigg, 2001), chosen because it provides a metric for distance between color hues ("Delta E") built to reflect perceptual distance, as opposed to merely physical metrics. The colors were equal in luminance and saturation, with a constant perceptual distance between any color and its two neighbors of Delta E = 7.8. Thirty-two different color arrays of four colors each were formed from this set of 32 colors, by picking every fourth color along the dimension of hue, until a four-colors array was formed, using each of the 32 colors as starting point. In this

way, each color was present in exactly four arrays. The array present on any given trial was randomly picked, all arrays being equiprobable. Which of the array's four colors served as the target for communication was likewise randomized. In addition to the target color (which she always saw), the Sender could see some or all of the colors visible in the Receiver's array. The Receiver always saw all four colors in the array. This quantity varied from one (only the target) to four (the full array). The number and nature of the colors shown to Sender were also randomized.

Senders had to communicate using a keyboard of black and white symbols (35 different symbols overall). The messages composed on this keyboard could consist of one or several symbols. Use of several different symbols in one message was possible and common; so were repetitions. These symbols had been experimentally tested to make sure that they would be neither too easy (evoking too narrow a range of colors), nor too difficult (allowing no color associations whatsoever). Laboratory experiments show that the symbols are as ambiguous as desired, since different pairs of participants can use them to solve the communication task above chance, but distinct pairs will associate the same symbol with different colors (Müller, Winters, & Morin, 2019). To maximize variability in symbol use, as well as provide the game with a reward structure, players at the start of the game were only provided with a random sample of 10 symbols (out of 35), earning the right to use additional symbols progressively as they earned points and ascended to new levels. For this study, we removed from the dataset all the trials where the Sender played with an incomplete keyboard, because she had not (or not yet) reached the level that unlocks all symbols.

Every new player, on their first opening of the app (but not later) was greeted with a short tutorial explaining the basics of the game. The tutorial simulated a referential communication game, using dummy symbols that were never reused in the normal course of the game (Fig. 2, bottom row), and a color array randomly picked from the 32 possible arrays. The player was presented with a dummy symbol and a four-colors array, and asked to point which color it might refer to. Then, the player was asked to play as Sender and use one of the dummy symbols to refer to a target color. What symbol or color the payers used at this stage did not matter: they were told they had completed that step of the tutorial. After this, the players were given a guided tour of the home screen, with pop-up messages in their chosen language.

2.1. Preregistration

We preregistered this study in four waves. The first preregistration[1] made a number of assumptions about the data that would be collected through the app, which turned out to be unwarranted when we attempted to pretest our predictions on a subsample of the data (second registration[2]). The Color Game did not incentivize or mandate participants to play. One consequence was that, contrary to our assumptions, activity was far from evenly spread in time. This and other unexpected findings led us to abandon the pretest and any type of test of our original predictions, and to propose a new set of predictions, better suited to the structure of the data (third registration[3]). These predictions were made and registered after the greater part of the data was collected, but before the relevant measurements (sense entropy, earth-mover distances, symbol frequency of use, etc.) were taken. All details of the analyses,

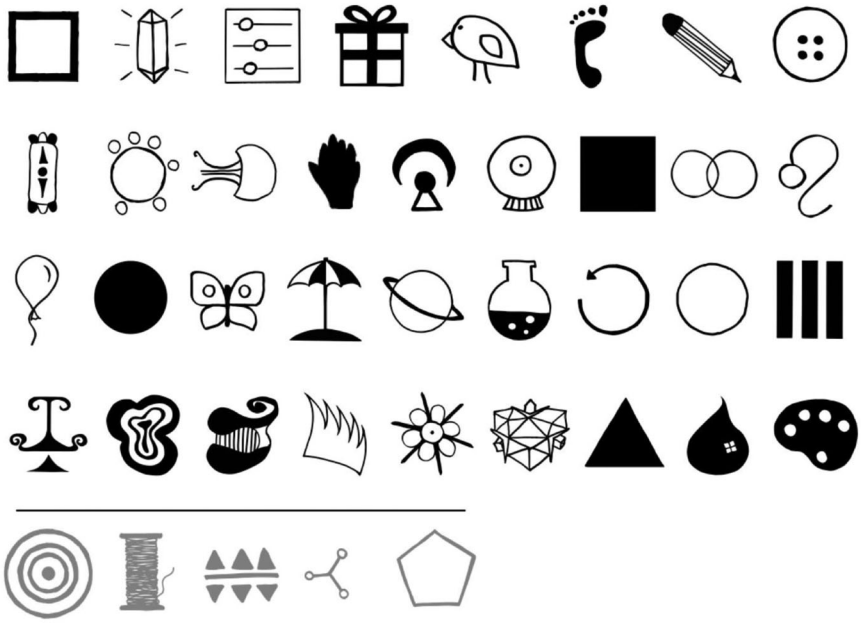

Fig 2. The 35 symbols used in the game (first four rows). Bottom row, in gray: the five symbols used for the tutorial and for the videos advertising the game (these symbols are for tutorials only).

including data exclusions and model specifications, were registered at this stage. Later, a fourth registration[4] was written and recorded in response to reviewers' comments, bearing essentially on the robustness of our results to arbitrary methodological choices.

To measure the degree to which the conventions linking a symbol to a range of colors are precise or vague, we use a measurement tool that is in some respects similar to "sense entropy" in the word disambiguation literature (Edmonds, 2009; Piantadosi et al., 2012). To calculate the sense entropy corresponding to the various meanings of a word, one counts how many times the word has been used to mean sense 1, sense 2, and so on. This yields a probability distribution over which a standard entropy calculation is performed. In the current study, a higher sense entropy means that the target colors associated with a symbol by a player are less predictable. In the same way, we calculated the sense entropy of a symbol, as used by a player over a set of trials, by considering the entropy of the target colors over those trials.

Our measure differs from standard sense entropy in that it takes advantage of the fact that meanings, in the Color Game, were located at evenly spaced intervals on a circular space. (The color space is assumed to be continuous because colors in the CIE2000 space are designed to be perceptually equidistant.) Like standard sense entropy, this measure uses Shannon's entropy (Shannon, 1948). It was, however, adapted by one of us (JW) to fit the structure of the color space (Fig. 3). This alternative measure suits our purposes better than existing measures of entropy or angular dispersion on a circle (Brunsdon & Corcoran, 2006; Gudmundsson & Mohajeri, 2013; Takahashi et al., 2015; Tastle & Wierman, 2007; Timme & Lapish, 2018).

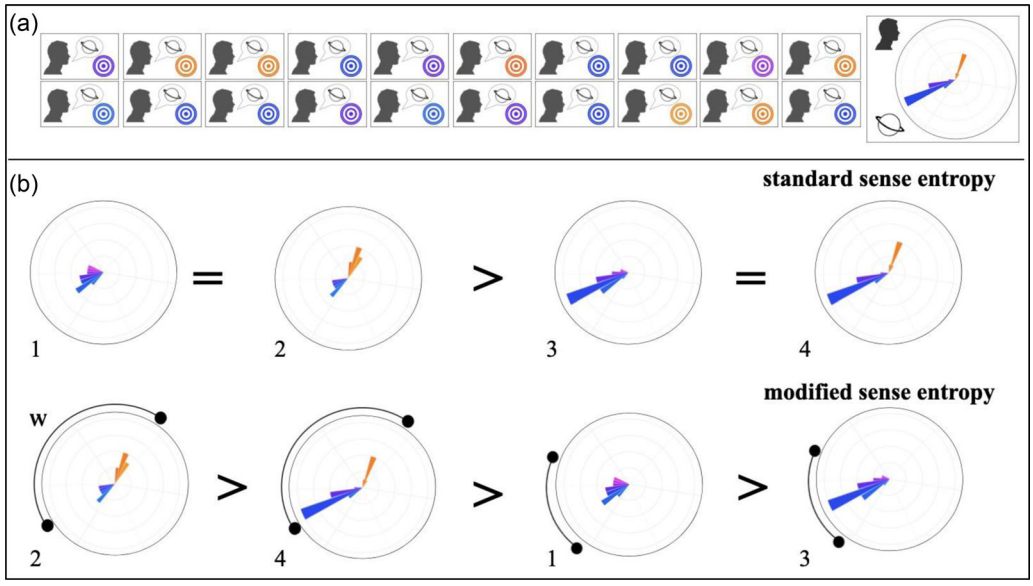

Fig 3. Sense entropy for Color Game symbols. Top panel (a) To calculate sense entropy for a participant's use of a symbol, we first consider what the target color was on the last 20 times the participant used the symbol (alone, or in combination with other symbols). The result (far right box) is a distribution of the colors associated with the symbol by that user. Bottom panel (b) We then compute the entropy of that distribution. A standard entropy measure (top row) would take no account of the location of colors on the color wheel, giving the same values for distributions 1 and 2, or to 3 and 4. Our modified measure (bottom row) also considers the width of the range of colors that a symbol is associated with, the "category structure" (w), indicated by the black arc (see Eq. 1). This results in different entropy values for distributions 1–4.

Our measure extends upon standard measures of conditional entropy while preserving both the granularity of the color space and its discrete circular ordering, as per Eq. (1):

$$H_w(C|S_i) = \frac{H(C|S_i)}{\max_{c\in C} H(C)} \times \frac{\max_{c\in W,S_i} H(W|S_i)}{\max_{c\in C} H(C)} \tag{1}$$

- $H_w(C|S_i)$ is the "category-entropy" for a set of colors C (the 32 colors of our color space) and a symbol S_i (one of the game's 35 symbols).
- $H(C|S_i)$ is the conditional entropy of C given S_i. It is calculated with Eq. (2):

$$H(C|S_i) = \sum\nolimits_{c\in C} P(c)\, log\, P(c), \tag{2}$$

where $c \in C$ is one color of the set of colors that S_i is used in connection with, and $P(c)$ is the frequency with which S_i is used in connection with c, as opposed to another color from C.

- $\max_{c \in C} H(C)$ is the maximum entropy that can obtain over the set of all possible colors. In other words, it is the entropy of the probability distribution that we get by assuming all the game's 32 colors to be equiprobable. Its value is constant ($\log^2(32) = 5$ bits). Division by $\max_{c \in C} H(C)$ allows us to normalize our entropy values between 0 and 1.
- $\max_{c \in W, S_i} H(W|S_i)$ corresponds to the maximum entropy of the color category structure (W) for symbol S_i. A symbol's "category structure" refers to the set of contiguous colors that form the shortest path linking all the colors that a symbol is associated with (see Fig. 3b, the semi-circle named W).
- Like any entropy measure, this one is vulnerable to confounds due to sample sizes, because small samples tend to be noisier than large ones, all else being equal. To avoid this, all our sense entropy measurements were performed over sets of 20 trials (the last trials for a given player, symbol, and, as applicable, time period), a threshold that was preregistered. We later replicated all our analyses with a different threshold (see below).

3. Results

For both predictions, we started from the Color Game's "Canonical dataset" (347,606 trials). This dataset is a cleaned-up version of the raw data outputted by the app, that all six of our projects use as a starting point (see "Open data and code"). For this study, we removed from the Canonical dataset all the trials where the Sender played with an incomplete keyboard, because they had not (or not yet) reached the level that unlocks all symbols. In total, 75,635 trials were removed for this reason (271,971 trials remaining). Only for measuring one specific variable (relevant for prediction 0) were those trials kept. We later made sure that this exclusion did not affect our results.

3.1. Prediction 0: High-precision symbols are more popular, especially with experienced players

Our original prediction, that there should be cultural selection for high-precision symbols, implied that symbols used with high precision would be the most frequently used symbols, especially among experienced players. This prediction was refuted.

We extracted data on all the Senders who used the same identical symbol (on its own or accompanied by other symbols) on at least 20 trials, after reaching the level that unlocks the full keyboard. There were 209 such Senders, for a total of 2154 data points (each data point representing one player's use of one particular symbol). All 35 symbols were represented. To calculate a player's experience, we considered how many trials the Sender played overall (PLAYERXP). This count includes trials that were played before the player unlocked the full keyboard, and thus it was computed over the complete set of trials ($n = 347{,}606$), unlike other variables. Each symbol's frequency of use (SYMBOLFREQUENCY) for each given player was computed as the ratio of the total number of trials where the player played as Sender and used the relevant symbol (at least once), over the total number of trials that this player played as

Sender. The amount of information carried by each symbol, as used by a given Sender over her last 20 trials (SENSEENTROPY), was computed as indicated above.

We built a linear mixed effects model (lme4 package for R—Bates, Mächler, Bolker, & Walker, 2015; R Core Team 2018) to predict SYMBOLFREQUENCY, using SENSEENTROPY and PLAYERXP. Both SYMBOLFREQUENCY and PLAYERXP were log-transformed, to satisfy the assumptions of a linear mixed effects model and avoid convergence issues. This model included a random effect for individual players and another for individual symbols. Lastly, we included an interaction term, SENSEENTROPY * PLAYERXP. A second version of the model was run with two additional controls, the "half" that the player belonged to, and the number of symbols that a player uses on average on one trial. None of these additional controls made for a more informative model (using a threshold of $\Delta_{AIC} > 2$), thus this second version was not considered further.

We also tested (as a follow-up) whether the number of trials after which a symbol was added to the Sender's keyboard makes a difference to SYMBOLFREQUENCY. Since symbols are not made available all at once to every Sender (some are present from the get-go, others are progressively unlocked), it was possible that Senders would have acquired habits with the very first symbols they had access to. But including this "Symbol Age" variable did not make our model more informative; it was dropped from subsequent analyses.

Here and in all other models described in this report, we attempted to add random slopes (one by one), in addition to random effects, to model the interactions between our random effects and our fixed effects, but each of these attempts resulted in a model that either failed to converge or produced singular fits, and did not prove more informative than a simpler version.

The prediction that SENSEENTROPY would be negatively correlated with SYMBOLFREQUENCY was refuted. High entropy (i.e., lower precision) in the use of a symbol is *positively* and clearly correlated with that symbol's frequency of use (Beta weight for SENSEENTROPY: + 3.4, 95%, $SE = 0.73$, $t = 4.7$, CI: + 2.01 to + 4.88). Removing a set of outliers revealed by the residual plots did not change this effect. The second prediction, that a positive correlation between SENSEENTROPY and SYMBOLFREQUENCY would be stronger in more experienced players, could, therefore, not be tested. We did find a negative interaction, whereby the effect of SENSEENTROPY over SYMBOLFREQUENCY was modulated by PLAYERXP (i.e., that effect was weaker for more experienced players), but this interaction term became weaker and changed direction when removing a set of outliers revealed by the residual plots. Thus, we found no selection affecting high-information symbols. Instead, Senders tend to map the symbols that they use more frequently to a broader range of colors.

Is the positive relationship that we found between SENSEENTROPY and SYMBOLFREQUENCY a robust result? We ran a series of supplementary analyses to find out, in the manner of a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). In a multiverse analysis, a statistical test is replicated as many times as needed to explore the main degrees of freedom that can be exploited. Degrees of freedom are arbitrary decisions made by the analyst that can make a difference to the outcome. We identified the following degrees of freedom:

-The size of the threshold for entropy calculation (10 trials or 20);
-Whether or not to include messages consisting of more than one symbol;
-Whether or not to log-transform SYMBOLFREQUENCY;
-Whether or not to eliminate the outlier data points that appear when SYMBOLFREQUENCY is log-transformed.

Crossing all these degrees of freedom yields a set of 12 analyses. All analyses fully confirmed our original result, with one exception. In all analyses, the effect of SENSEENTROPY over SYMBOLFREQUENCY is positive at the lowest boundary of the 95% confidence interval. In all analyses but one, adding SENSEENTROPY to the model makes it more informative (all $\Delta_{AIC} > 2$, but $\Delta_{AIC} = 1.8$ for the exception). The effect of SENSEENTROPY is clearly much weaker if we only consider one-symbol messages, which dramatically reduces our statistical power, dividing the number of data points by two or three (Table S1).

3.2. Prediction 1: More experienced players are more likely to use symbols precisely

Thus, we found the opposite of our original prediction to be true—and this is a robust finding. One consequence of this unexpected pattern of results is that one of our original expectations could not be properly tested. Prediction 0 was premised upon the view that players would gradually come to use symbols in a more specific fashion. Two post-hoc analyses confirmed that this was the case. We first considered whether PLAYERXP was a good predictor of SENSEENTROPY. A linear mixed effects model was built to predict SENSEENTROPY with PLAYERXP, with random effects added for individual players and symbols. It showed a clear negative effect of PLAYERXP (Beta weight: –0.022, $SE = 0.004$, $t = –4.57$, 95% CI: –0.03 to –0.01). As before, adding a control for the number of trials after which Sender unlocked the symbol ("Symbol Age") did not change this result, and did not make the model more informative. More experienced players use their symbols in a more precise way, to refer to a narrower range of colors (Fig. 4).

Here again, we submitted this finding to a multiverse analysis, to make sure that it is robust, considering all the degrees of freedom manipulated in the previous multiverse analysis, and adding one more manipulation (including, instead of excluding, the trials that players played before they unlocked the full keyboard). The results fully verified our prediction in 27 out of 28 cases (see Supplementary Materials).

The previous analysis focused on the Senders' last 20 (or 10) trials, and their level of experience at this last stage of their career. It could not test diachronically the claim that Senders become more precise with experience. To test this, we considered how sense entropy evolved over the course of a player's gaming history, from the time when they unlock the full keyboard, until their last trial. We divided each player's trials into bins of 1000 trials each. Bins were numbered consecutively from earliest to latest: these bin numbers make up the BIN variable, which indicates what stage in a player's "career" the bin was taken from. Inside each one of each player's bins, we considered all the symbols that had been used 20 times or more. We calculated SENSEENTROPY for the last 20 trials involving a given symbol in a given BIN for a given Sender. This yielded a dataset of 4871 trials for 209 Senders. We then built a linear mixed effects model to predict SENSEENTROPY in the use of a given

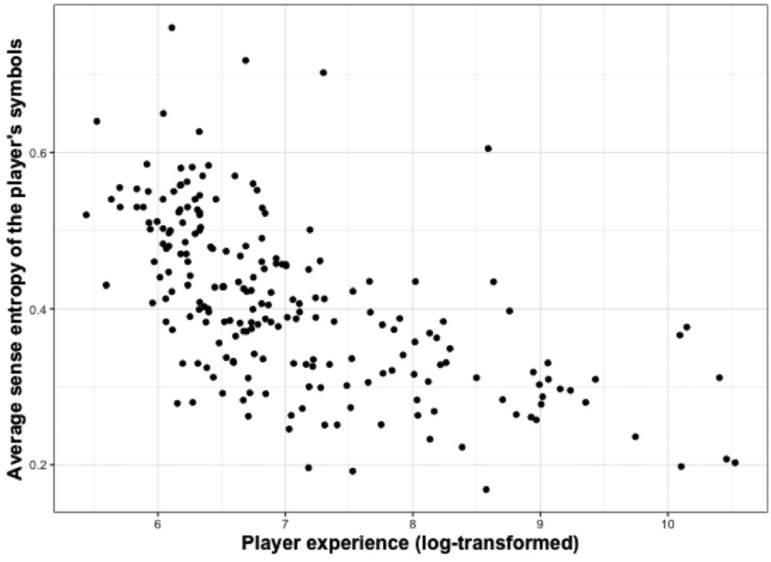

Fig 4. Symbols have lower sense entropy when used by more experienced players. Each dot shows, for one Color Game player ($n = 209$), the average entropy of the symbols they used (considering only symbols used 20 times or more), plotted against the player's experience (total number of trials played in the game, log-transformed).

symbol by a given player, depending on the player's progress (as indicated by BIN). This model included random effects for individual players and individual symbols (adding random slopes caused convergence failures). It gave a clearly negative estimate for the effect of BIN (log-transformed) over SENSEENTROPY (Beta weight: –0.02, $SE = 0.002$, $t = –9.4$, 95% CI: –0.03 to –0.02) (Fig. 5). More experienced players use symbols more specifically because experience made them use symbols with more precision (Fig. 6).

A multiverse analysis shows that this result is robust to variations alongside the degrees of freedom we identified previously, as well as two additional variables proper to this analysis: whether or we remove six outlier players (the most productive players, responsible for all the later trials), and whether or not we remove the players who produced only one bin's worth of data. The analysis fully verified our prediction in 62 out of 64 cases (see Supplementary Materials).

3.3. Alternative ways to measure entropy

The way we measure the amount of information carried by a symbol is but one possible way of assessing it. It has (at least) one limitation. It ignores an important piece of information visible by the Receiver: the array of four colors that the target is included in. This information can change the symbol's precision, since it narrows down the range of possible guesses considerably. We tested two alternative measures of symbol precision that take the array into account.

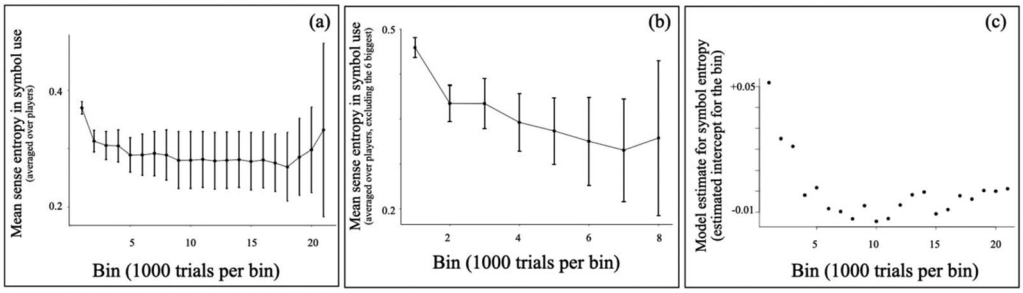

Fig 5. Changes in symbol entropy in the course of the players' career. In (a), the average entropy of symbols across players is plotted against bins, one bin being a set of trials that a player played with the same symbol inside the same window of 1000 consecutive trials ("bin"). To get each data point, we averaged over players the sense entropy that each player displays in using symbols, inside the relevant bin. Bins are arranged chronologically from 1 to 21. Only players who played a sufficient number of trials appear in the relevant bin: a player who played 5000 trials will only be counted in bins 1–5. The vast majority of players played less than 9000 trials overall, which explains the increasing error in the right portion of the graph. Panel (b) shows the same data but removes the six players who played more than 9000 trials. Error bars, here as in panel (a), stand for 95% confidence intervals. These graphs represent raw data and do not account for variation due to specific players or symbols. Panel (c) shows how a linear mixed effects model estimates the change in SENSE ENTROPY from bin to bin. A linear mixed effects model was run predicting sense entropy with three random intercepts, one for players, one for symbols, and one for bin. The figure shows the value of the random intercept for each bin.

Our first alternative measure, in-context sense entropy (INCONTEXTENTROPY), considers, like SENSEENTROPY, a set of trials where a given Sender used a given symbol, but unlike SENSEENTROPY, it is only computed for the trials played with one given array of four colors. Since the combination of same Sender, same symbol, and same array is not so frequently encountered, we lower our minimum number of trials to measure to 10 consecutive trials (instead of 20). We consider all the Senders that satisfy the condition of having played a minimum of 10 trials with the same array, having sent the same symbol (on its own, or accompanied by others) on each of the 10 trials. For each of these eligible Senders, we considered the latest 10 trials in which they sent the same symbol for the same given array. We repeated this for all eligible symbols and arrays. We counted how many times each of the array's colors was the target in those past 10 trials. We then computed the entropy of this distribution. One of the motivations for using this alternative measure was to address concerns that our SENSEENTROPY measure was not standard enough, so this measure uses a simple and standard entropy measurement that neglects the distances between colors—Eq. (3):

$$H(C) = -\sum\nolimits_{c \in C} P(c) \log P(c), \tag{3}$$

where $c \in C$ is one of the array's four colors and $P(c)$ is the proportion of trials where color c was the target color, among the past 10 trials where Sender used the symbol we are interested in.

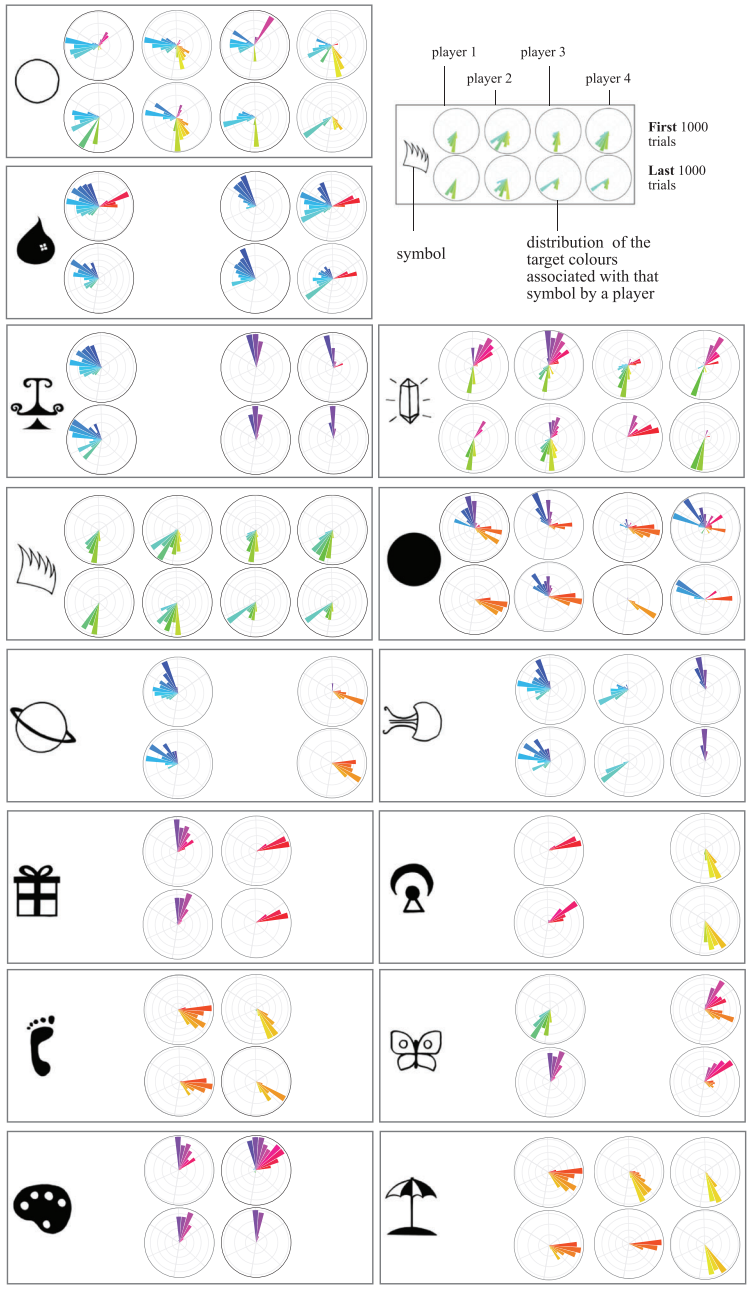

Fig 6. Evolution of symbol use for four big Color Game players. This figure considers, for the four biggest players of Half B (one of the two groups of players that could interact with each other), all the symbols that several players used 20 times or more during their first and their last 1000 trials—20 trials being our threshold for measuring symbol meanings. (When a player used a symbol less than 20 times, the data are not shown.) The distribution of target colors associated with each symbol by players 1–4 (left to right) is given for the first 1000 trials (top row) and last 1000 trials. For all players except player 2, sense entropy for most symbols decreases between the first and last 1000 trials.

We attempted to replicate our two main effects using this alternative measure: Do more experienced players use symbols with more precision? Does symbol specificity increase in the course of a player's career?

Do more experienced players use symbols with more precision? Yes. We replicated our original analysis, replacing SENSEENTROPY with INCONTEXTENTROPY. For this analysis, we had 10,522 data points from 98 Senders. A linear mixed effects model was built to predict INCONTEXTENTROPY with random effects added for individual Senders, symbols, and arrays, using as controls (as previously), the average length of the Sender's messages, and the frequency of the symbol for that Sender. Adding PLAYERXP to this model made it more informative ($\Delta_{AIC} = 12$) and yielded a model with a negative estimate for the effect of PLAYERXP, up to the highest boundary of the 95% confidence interval. Log-transforming PLAYERXP to get better-behaving residuals does not change this result, and neither does discretizing INCONTEXTENTROPY (converting all non-0 values to 1, which is justified given the large number of 0 values).

Does symbol precision increase in the course of a player's career? Yes. We replicated our previous analysis, replacing SENSEENTROPY with INCONTEXTENTROPY. We divided each player's trials into bins of 1000 trials each, numbered consecutively from earliest to latest. Inside each one of each player's bins, we considered all the symbols that had been used 10 times or more with the same array. We calculated INCONTEXTENTROPY for the last 10 trials involving a given symbol in a given bin, for a given Sender, with a given array. This yielded a dataset of 2880 trials from 99 Senders. We then built a linear mixed effects model to predict INCONTEXTENTROPY, with random effects for individual Senders, individual symbols, and individual arrays. Adding the player's progress (as indicated by log-transformed BIN) to this model made it more informative ($\Delta_{AIC} = 30$) and yielded a model with a negative estimate for the effect of PLAYERXP, up to the highest boundary of the 95% confidence interval. Not log-transforming BIN did not change this basic result. As previously, we retested the model having removed the six biggest players, as well as removing players who play only one bin, and the result remains robust.

Our second alternative measure of symbol entropy also considers the information carried by a symbol in context, that is, given the color array it is associated with; but this one takes the Receiver's perspective instead of the Sender's. It starts by asking what the probability distribution for the target's location is, for a given Receiver, a given array, and a given symbol. It then confronts this distribution with the correct location of the target. We used this alternative measure, "IN-CONTEXT PRECISION," to test whether symbols became more precise from the Receivers' perspective, with time.

We split our dataset into Receiver-specific subsets. Each subset gathered all the trials for which a specific player played as Receiver. Each of these subsets was then broken into bins of 1000 trials each (BIN variable), the trials being considered in chronological order. We considered the last 10 trials where the Receiver played with the array of interest, having gotten a message from Sender that included the symbol of interest. (All messages were considered, not just one-symbol messages.) This entailed excluding Receivers who did not meet the criterion of having played 10 trials as Receiver on the same array, having been exposed to the same symbol. The Receiver's estimated probability distribution for the target's location, given

a symbol and an array, was computed directly from their choices of colors on the last 10 trials played with this array and symbol. We then computed the divergence between this distribution and the target location (as Kullback–Leibler divergence—KLD). This measure tells us how much information is gained by replacing the Receiver's distribution of color picks, on the last 10 trials where they saw the symbol of interest, with the real target location. If the Receiver's expectations allow them to pick the target with certainty, then, KLD = 0. A higher KLD corresponds to a greater distance between the Receiver's belief and the target's true location, meaning that the symbol is less informative (in context). The KLD divergence between the correct location of the target, and Receiver's distribution of past choices for this array and this symbol, gives us the IN-CONTEXT PRECISION measure.

A Receiver's past experience with a symbol is a relatively good guide to its future meaning. Overall, the divergence between the Receiver's past 10 choices on seeing a given symbol paired with a given array, and the true location of the target on their next choice with the same symbol and array, is relatively small (0.16 on average). But contrary to our expectation, we found no clear sign that IN-CONTEXT PRECISION decreases in the course of a Receiver's career. We built a linear mixed effects model predicting IN-CONTEXT PRECISION with random effects for individual Senders, individual symbols, and individual arrays. Adding the BIN variable to this model did not make it more informative ($\Delta_{AIC} < 0$), although the estimated effect of BIN was negative as predicted. This finding seems retrospectively obvious, in light of two things. First, whether symbols are used with precision depends on Senders, not Receivers, and an experienced Receiver is not necessarily more likely to interact experienced Senders. Second, Receivers interacted with different Senders who used symbols in different ways, and pooling together a Receiver's past trials regardless of who they interacted with may not make much sense.

3.4. Prediction 2: Frequent symbols are agreed-upon

As a way to test Enfield's conjecture that the amount of overlap between individual representations of the meaning of a symbol is influenced by the frequency at which this symbol is used, we measured the overlap between the range of referents (colors) that each player in a pair associated with the symbol of interest.

For this prediction, we discarded the trials for which there was no Receiver. These are trials that were played by a Sender who saw a target color and sent a series of symbols to help some Receiver pick the color, but whose message was never picked up to be solved by an actual Receiver. (These trials were not excluded when we tested Prediction 1, which focuses on Senders' behavior exclusively.) Removing them caused 8,694 trials to be excluded (remaining $n = 263{,}277$ trials). We considered all the pairs of players in which both players had used the same symbols at least 20 times each. This was necessary in order to get sufficient data on the use of symbols by each Sender. This exclusion criterion only allowed us to get data concerning 156 pairs (713 data points in total, each individual data point representing the use of one symbol by one pair of players).

The amount of disagreement between players over the meaning of a symbol was quantified as the distance (using Earth Mover Distance, computed with emdist in R [Urbanek & Rubner,

2012]) between the color distributions associated with the symbol by each of the two players (DISAGREEMENT). To avoid biases due to sample size, distances were measured using only, for each player, the last 20 trials on which the symbol was used. Distances were log-transformed (to meet the assumptions of a linear mixed effects model). The frequency at which a symbol was used was given by the ratio of the number of the pair's trials where the symbol was used, over the total number of trials (FREQUENCY). A series of linear mixed effects models were built to predict DISAGREEMENT. The first "null" model nested each data point by pairs (the identity of the two players playing together) and by symbols (which picture the pair used: Butterfly, Drop, etc.). We then tested a series of controls: the total number of trials played by the pair, the total number of the pair's trials involving the symbol, and lastly, the sense entropy of the symbol as used by each Sender (AVERAGEENTROPY). Only this last variable made the model more informative ($\Delta_{AIC} = 29$; for the other two variables $\Delta_{AIC} < 2$). Symbols that are used, on average, in a more specific fashion by both players (specificity, or precision, being measured using our modified sense entropy measure, as in Fig. 3) are more likely to be used by both players to signal the same colors.

The model estimates the effect of AVERAGEENTROPY upon DISAGREEMENT to be positive (Beta weight: + 1.9, $SE = 0.34$, $t = 5.72$, 95% CI: +1.26 to +2.65). Visual inspection of the residuals shows a considerable degree of heteroscedasticity. When players use a symbol very precisely, disagreement can be very high but it can also be quite low. Conversely, when players use a symbol in a very imprecise fashion, corresponding to a high degree of dispersion of the associated colors, the distance between players can neither be very high nor very low.

The test model added FREQUENCY to this last model. It did not prove more informative than the last one ($\Delta_{AIC} < 2$). The effect of FREQUENCY was in the right direction, but weak (Beta weight: –0.62, $SE = 0.54$, $t = –1.14$, 95% CI: –1.69; +0.45). Thus, the meaning of frequently used symbols was not clearly more agreed-upon than that of little used symbols. The effect of AVERAGEENTROPY remained strongly positive in this final model.

4. Discussion

Our participants, faced with a referential communication task where they had to interact with strangers without the backing of rich contextual cues, spontaneously evolved increasingly precise semantic conventions. The associations between symbols and colors grew more precise in the course of a player's trajectory: more experienced players matched symbols with a smaller range of colors. The symbols available to players of the Color Game differed from natural language words in a number of respects; most importantly, they carried clear associations with color referents even before the game started—although none were precise enough to be associated with a single one of the game's 32 color hues. Some spoken words are similarly iconic, but iconicity is arguably more pronounced in this particular task. If anything, this aspect of the task would have worked against the effects we document. If we assume that Color Game symbols already had highly precise meanings for players before the game started, there would be no reason to expect symbol use to increase in precision with time.

There does not seem to be any equivalent for such a general narrowing of semantic conventions in the history of spoken languages. Lexical ambiguity changes with time for particular words, through narrowing or broadening (Traugott & Dasher, 2009; Urban, 2015), with many well-attested instances of words (like "excellent" or "fantastic") whose meanings become less precise or attenuated through repeated use (Deo, 2015). But there is no evidence of entire vocabularies gradually becoming more specific (Wasow et al., 2005), with the possible exception of sign languages emerging *de novo,* like Nicaraguan Sign Language or Al Sayyid Bedouin Sign Language (Sandler & Meir, 2005; Senghas, Kita, & Ozyürek, 2004), where signs became more conventionalized over time. Why then do conventions gain in precision in the Color Game app? The best established explanations for semantic ambiguities in natural language rest upon the abundance of contextual information in normal conversations, making a certain amount of ambiguity tolerable, even efficient (Piantadosi et al., 2012; Winters & Morin, 2019; Winters et al., 2015, 2018) (but see O'Connor, 2015). Our results support this view. Contextual information in the app was kept low: Senders' symbols were the only cue that Receivers could exploit to pick the target color from the three distractors. This put pressure over the symbols to be maximally specific, and they evolved accordingly.

The way this evolution took place was surprising, given the strong emphasis the literature places on selectionist as distinct from transformative evolutionary processes (Tamariz, 2019; Tamariz et al., 2014). We expected participants to copy one another's way of using the symbols fairly closely, since communication can only gain from close coordination. Precise mappings between colors and symbols being more informative, such mappings should be reproduced more than others, resulting in low-entropy, high-information conventions becoming more frequent at the expense of low-information ones. The data depart from these predictions in at least two ways. First, precision in symbol use is not faithfully copied. A given symbol (e.g., "Butterfly" or "Grass") can be used with great precision by one player but not by the next player. The share of variance in SENSEENTROPY that is accounted for by symbols is less than the share accounted for by players (intraclass correlation for symbols: 0.24; for players: 0.32; computed with irrNA—Brueckl & Heuer, 2018). Second, low-information mappings were more frequently used (and thus, seen) than high-information ones: there was a strong positive correlation between sense entropy and frequency. If participants had copied the mappings they encountered most frequently, the result would have been a decrease in the overall precision of the conventions used in the Color Game. In spite of this, Senders gradually came to use their symbols in a more specific fashion. This process of semantic narrowing (Traugott & Dasher, 2009) took place despite the fact that high-entropy mappings were more widespread than low-entropy ones.

Frequently used symbols tend to be used for a greater variety of colors—to be less precise. This finding makes retrospective sense. In hindsight, the "tolerable friends" hypothesis neglected the fact that frequently used words are not simply words we can easily learn the meaning of, being exposed to a more varied sample of word uses. Frequent words, at least as far as the lexicon is concerned, are also likely to have a broader meaning, or a wider variety of meanings. Numerous previous results indicate a positive link between a word's frequency and the number of dictionary meanings it is associated with—a correlation known as "Zipf's Law of Meaning" (Zipf, 1949) (also known as the "principle of economic versatility"—Levinson,

2000). This relation holds when controlling for confounds such as word length (Piantadosi et al., 2012). Our findings extend and deepen Zipf's Law of Meaning in two ways. First, we show how it applies to a nonverbal symbolic language. The symbols of the Color Game, rather remote in many ways from the words of natural languages (among other things, they lack a morphology and do not clearly obey syntactic rules shared between players), do follow the Zipfian principle, in that more frequently used symbols refer to a broader range of colors. Second, we use a measurement of ambiguity that goes beyond the standard polysemy measurement. Standard tests of Zipf's Law of Meaning have to rely on counting the number of meanings associated with a word. Counting meanings neglects two important dimensions of word meaning: the fact that some meanings are more precise than others, and the fact that meanings may be more or less close to one another. With our measure of sense entropy, which integrates polysemy and vagueness into one quantitative metric thanks to the use of a controlled space of referents, we can explore Zipf's Law of Meaning in a properly quantitative fashion.

Two main types of causal interpretations have been put forward to explain Zipf's Law of Meaning. The first type starts from Zipf's original intuition: speakers tend to minimize their production effort and the comprehension effort of their interlocutors. Building upon this, Piantadosi et al. (2012) propose that frequent words, being easier to use and process (since their frequency makes them easier to memorize), are more likely to be used in novel contexts, thus acquiring new meanings. This explanation has the merit to link Zipf's Law of Meaning with Zipf's other "laws," the law of abbreviation (frequent words are shorter) and the law of frequency (word frequency distributions approximate a power law).

A second type of explanation focuses on the organization of the concepts encoded by words. If one classifies a set of objects using a hierarchical structure of categories, while avoiding synonyms, a Zipfian relation between frequency and broadness of meaning obtains (Manin, 2008). Our study was not meant to adjudicate between these two views of Zipf's Law of Meaning, but its results are nonetheless indicative. In the standard Zipfian perspective, semantic ambiguity is useful because speakers can rely on rich contextual information, and because precise signals are costly to produce. Neither of these conditions seem clearly to obtain in this study. It could be argued that frequently used conventions are easier to produce and process because their frequency makes them easier to memorize, and to process, for both Senders and Receivers. Against this, however, we found no clear indication that players showed more agreement over frequently used symbols. In our view, a more likely reason for ambiguous semantics is the fact that we allowed Senders to combine several different symbols to form messages. Such combinatorial use allows agents to encode high quantities of information through symbols whose individual meaning is ambiguous. Further analyses of the open Color Game dataset could shed light on the players' various strategies for combining symbols (asking, e.g., whether they obey compositional principles).

While we could neither confirm nor refute Enfield's conjecture that agreement between agents is greater for frequently used conventions, we found an interesting interplay between specificity and agreement. Very high levels of disagreement and agreement obtain only for low-entropy mappings, but become increasingly unlikely as the conventions linking symbols and colors become less precise. In other words, disagreement on vague conventions cannot

be very high, because the two interlocutors find some common ground in ambiguity. This is reminiscent of a point made by Wasow et al. (2005): lexical ambiguity can be functional when different people or communities associate different meanings with a given word. By entertaining multiple or vague meanings for a given word, an agent may be able to communicate with other agents who hold a variety of precise but discrepant views about the word's meaning. Imprecision fosters a modicum of consensus.

5. Conclusion

The origins and functions of linguistic ambiguity have puzzled researchers since the dawn of logic. Some likely sources of ambiguity are the availability of contextual information in the environment, the cost of producing and processing precise messages, and the possibility of combining ambiguous signals together to create a more precise message. Our findings both illustrate and qualify the impact of the first two sources of ambiguity. In a referential communication game where contextual information was minimal, and production costs were equally small for all symbols, participants gradually devised ever more precise conventions mapping symbols to colors. On the other hand, even highly experienced participants kept using some symbols in quite imprecise ways. The evolutionary process leading to lower ambiguity was not driven by the selective copying of precise conventions. Individual players gradually made their own mappings more precise without copying the mappings they most frequently saw others using. This illustrates how cultural evolution can produce sophisticated and efficient outcomes without selection being at play (Acerbi, Charbonneau, Miton, & Scott-Phillips, 2021; Claidière et al., 2014).

Acknowledgments

We would like to thank all the people who contributed to the Color Game, either in its creation or as a participant. A summary of acknowledgments regarding these contributions can be found here: https://osf.io/nsxu4/. We also wish to thank the reviewers and editor at Cognitive Science for their helpful comments.

Open access funding enabled and organized by Projekt DEAL.

WOA Institution: Max-Planck-Gesellschaft Blended DEAL: Projekt DEAL

Funding

This research was funded by the Max Planck Society. OM acknowledges the support of the "Frontiers in Cognition" EUR grant, ANR-17-EURE-0017 EUR.

Conflicts of interest

The authors have no conflicts to disclose.

Notes

1 https://osf.io/547bp/
2 https://osf.io/28nxb/
3 https://osf.io/s7y2v/
4 https://osf.io/enhsc/

References

Acerbi, A., Charbonneau, M., Miton, H., & Scott-Phillips, T. (2021). Culture without copying or selection. *Evolutionary Human Sciences*, *3*, E50. https://doi.org/10.1017/ehs.2021.47

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

Brochhagen, T. (2020). Signalling under uncertainty: Interpretative alignment without a common prior. *British Journal for the Philosophy of Science*, *71*(2), 471–496. https://doi.org/10.1093/bjps/axx058

Brueckl, M. & Heuer, F. (2018). irrNA: Coefficients of interrater reliability - Generalized for randomly incomplete datasets (0.1.4) [Computer software]. Retrieved from https://CRAN.R-project.org/package=irrNA

Brunsdon, C., & Corcoran, J. (2006). Using circular statistics to analyse time patterns in crime incidence. *Computers, Environment and Urban Systems*, *30*(3), 300–319. https://doi.org/10.1016/j.compenvurbsys.2005.11.001

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Calude, A. S., & Pagel, M. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1101–1107. https://doi.org/10.1098/rstb.2010.0315

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How Darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B*, *369*(1642), 20130368. https://doi.org/10.1098/rstb.2013.0368

Croft, W. (2019). All social behavior is replication: Comment on 'Replication and emergence in cultural transmission' by Monica Tamariz. *Physics of Life Reviews*, *30*, 72–73. https://doi.org/10.1016/j.plrev.2019.08.018

Deo, A. (2015). Formal semantics/pragmatics and language change. In C. Bowern (Ed.), *The Routledge handbook of historical linguistics* (pp. 393–409). Routledge.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*(2), 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002

Edmonds, P. (2009). Disambiguation. In K. Allan (Ed.), *Concise encyclopedia of semantics* (pp. 223–239). Elsevier.

Enfield, N. J. (2010). Tolerable friends. *Proceedings of the 33rd Annual Meeting of the Berkeley Linguistics Society*.

Ferrer-i-Cancho, R., & Lusseau, D. (2009). Efficient coding in dolphin surface behavioral patterns. *Complexity*, *14*(5), 23–25. https://doi.org/10.1002/cplx.20266

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

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

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, *4*(3), 223–272. https://doi.org/10.1515/cogl.1993.4.3.223

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gudmundsson, A., & Mohajeri, N. (2013). Entropy and order in urban street networks. *Scientific Reports*, *3*, 3324. https://doi.org/10.1038/srep03324

Lev-Ari, S. (2017). Talking to fewer people leads to having more malleable linguistic representations. *PLoS One*, *12*(8), e0183593. https://doi.org/10.1371/journal.pone.0183593

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, D. (1969). *Convention: A philosophical study*. Wiley-Blackwell.

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

Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*(7), 1075–1098. https://doi.org/10.1080/03640210802020003

McCulloch, G. (2019). *Because Internet: Understanding the new rules of language*. Penguin Publishing Group.

Morin, O., Kelly, P., & Winters, J. (2020). Writing, graphic codes, and asynchronous communication. *Topics in Cognitive Science*, *12*, 727–743. https://doi.org/10.1111/tops.12386

Morin, O., Winters, J., Müller, T. F. & Morisseau, T. (2020). An overview of the "Color Game" App project. Retrieved from https://osf.io/preprints/socarxiv/cjaxw/

Morin, O., Winters, J., Müller, T. F., Morisseau, T., Etter, C., & Greenhill, S. J. (2018). What smartphone apps may contribute to language evolution research. *Journal of Language Evolution*, *3*(2), 91–93. https://doi.org/10.1093/jole/lzy005

Müller, T. F., Winters, J., & Morin, O. (2019). The influence of shared visual context on the successful emergence of conventions in a referential communication task. *Cognitive Science*, *43*(9), e12783. https://doi.org/10.1111/cogs.12783

Murthy, S. K., Hawkins, R. D. & Griffiths, T. L. (2021). Shades of confusion: Lexical uncertainty modulates ad hoc coordination in an interactive communication task. arXiv preprint arXiv:2105.06546.

Nettle, D. (2020). Selection, adaptation, inheritance and design in human culture: The view from the Price equation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1797), 20190358. https://doi.org/10.1098/rstb.2019.0358

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93–104. https://doi.org/10.1016/j.cognition.2018.08.014

O'Connor, C. (2014). The evolution of vagueness. *Erkenntnis*, *79*(4), 707–727. https://doi.org/10.1007/s10670-013-9463-2

O'Connor, C. (2015). Ambiguity is kinda good sometimes. *Philosophy of Science*, *82*(1), 110–121. https://doi.org/10.1086/679180

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

R Core Team. (2018). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262. https://doi.org/10.1098/rspb.2019.1262

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Brady (Eds.), *The handbook of language emergence* (pp. 237–263). John Wiley & Sons, Ltd.

Sandler, W., & Meir, I. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences*, *102*(7), 2661–2665.

Santana, C. (2014). Ambiguity in cooperative signaling. *Philosophy of Science*, *81*(3), 398–422. https://doi.org/10.1086/676652

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

Senghas, A., Kita, S., & Ozyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*(5691), 1779–1782. https://doi.org/10.1126/science.1100199

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Takahashi, K., Kim, S., Coleman, T. P., Brown, K. A., Suminski, A. J., Best, M. D., & Hatsopoulos, N. G. (2015). Large-scale spatiotemporal spike patterning consistent with wave propagation in motor cortex. *Nature Communications*, *6*, 7169. https://doi.org/10.1038/ncomms8169

Tamariz, M. (2019). Action replication ultimately supports all cultural transmission. *Physics of Life Reviews*, https://doi.org/10.1016/j.plrev.2019.10.009

Tamariz, M., Ellison, T. M., Barr, D. J., & Fay, N. (2014). Cultural selection drives the evolution of human communication systems. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1788), 20140488. https://doi.org/10.1098/rspb.2014.0488

Tamariz, M., Roberts, S. G., Martínez, J. I., & Santiago, J. (2017). The interactive origin of iconicity. *Cognitive Science*, *42*(1), 334–349. https://doi.org/10.1111/cogs.12497

Tastle, W. J., & Wierman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, *45*(3), 531–545. https://doi.org/10.1016/j.ijar.2006.06.024

The American Heritage Dictionary of the English Language. (2011). Houghton Mifflin Harcourt.

Timme, N. M., & Lapish, C. (2018). A tutorial for information theory in neuroscience. *eNeuro*, *5*(3), 1–40. https://doi.org/10.1523/ENEURO.0052-18.2018

Traugott, E., & Dasher, R. (2009). *Regularity in semantic change*. Cambridge University Press.

Trudgill, P. (2011). *Sociolinguistic typology*. Oxford University Press.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, *4*(3), 273–290. https://doi.org/10.1515/cogl.1993.4.3.273

Urban, M. (2015). Lexical semantic change and semantic reconstruction. In C. Bowern & B. Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 374–392). Routledge.

Urbanek, S., & Rubner, Y. (2012). *emdist: Earth Mover's Distance* (0.3-1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=emdist

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*. CSLI Publications.

Wilson, D. (2003). Relevance and lexical pragmatics. *Rivista Di Linguistica*, *15*(2), 273–291.

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

Winters , J., & Morin, O. (2019). From context to code: Information transfer constrains the emergence of graphic codes. *Cognitive Science*, *43*, e12722. https://doi.org/10.1111/cogs.12722

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of sociocultural influences on linguistic form. *Lingua*, *117*, 543–578.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942. https://doi.org/10.1073/pnas.1800521115

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Materials: We append a Supplementary Materials document, available here: https://osf.io/qtdgx/ as well as a general presentation of the Color Game project and dataset (shared with all Color Game projects), available here: https://osf.io/preprints/socarxiv/cjaxw/.

Fig. S1. The app's home screen (left) with legend (right). The colorful logos that identify each contact are randomly generated from a set of black and white pictures and a set of colors.

Fig. S2. The game's color space. Each color is given its associated Hex code (as used by the app).

Fig. S3. How color arrays were built. Top row: The composition of two color arrays, one marked by white dots, the other by black dots, is shown relative to the color space. Bottom row: Six contiguous color arrays (out of 32), including the white-dot and black-dot ones.

Table S1. Multiverse analysis for the effect of SENSEENTROPY on SYMBOLFREQUENCY. The first four columns show the methodological decisions that were taken for a given set of analyses. The last five columns show the results of the relevant analyses. We did not test for the removal of outliers when SYMBOLFREQUENCY was not log-transformed, because our outliers were not defined for this case. In blue: original preregistered analysis. In green: best behaved model. In red: the one analysis which fails to confirm our prediction.

Table S2. Multiverse analysis for the effect of PLAYERXP over SENSEENTROPY. The first five columns show the methodological decisions that were taken for a given set of analyses. The last five columns show the results of the relevant analyses. In blue: original preregistered analysis. In green: best behaved model. In red: the one analysis which fails to confirm our prediction.

Table S3. (This page and the previous one.) Multiverse analysis for the effect of BIN over SENSEENTROPY.

The first six columns show the methodological decisions that were taken for a given set of analyses. The last five columns show the results of the relevant analyses. In blue: original preregistered analysis. In red: the two analyses which fail to confirm our prediction.