Cultural evolution of precise and consensual semantic conventions in a multiplayer gaming app

Olivier Morin*1,2, Thomas F. Müller2,3, Tiffany Morisseau4,5, James Winters2,6

* Corresponding author: morin@shh.mpg.de

1. Institut Jean Nicod, Département d’études cognitives, ENS, EHESS, CNRS, PSL University, UMR 8129.
2. Minds and Traditions Research Group, Max Planck Institute for the Science of Human History, Jena, Germany
3. Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany.
4. Université de Paris, LAPEA, Boulogne-Billancourt, France
5. LAPEA, Université Gustave Eiffel, IFSTTAR, Versailles, France
6. UM6P School of Collective Intelligence, 43150 Ben Guerir, Morocco.

ORCID identifiers:

Olivier Morin https://orcid.org/0000-0002-6216-1307
Thomas Müller https://orcid.org/0000-0002-8663-1035
James Winters https://orcid.org/0000-0003-2982-2991
Tiffany Morisseau https://orcid.org/0000-0002-5239-6049
Cultural evolution of precise and consensual semantic conventions in a multiplayer gaming app

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 studied the evolution of semantic conventions, in the form of variable mappings between symbols and their referents, in a multiplayer gaming app built around a referential communication game, where a sender had to indicate a color to a receiver, using only black and white symbols. We expected that the players’ mappings between symbols and colours would grow more specific over time, precision being measured as sense entropy, 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 careers. 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.

Background

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 “thing” is less precise, thus less informative, than the word “teaspoon”. A wide variability in the number of potential meanings per word (from “teaspoon” to “thing”) is be observed in many human languages [1–3], and in non-human signals [4].

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 [1,3,5], partly explaining why informativeness is so unevenly distributed among words (this reasoning also applies to non-linguistic symbols, such as coin designs [6]). The second research tradition considers diachronic changes in word meanings: historical trends that lead to semantic narrowing, or the opposite, semantic broadening [7]. Relatively little work has asked what consequences the functional constraints studied by the first tradition may have on the evolutionary trends that concerns the second line of work. 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. Words that are less costly to produce and process may tend to be re-used more easily, thus acquiring new meanings or extending the scope of their original meaning [3]. This conjecture has the advantage of accounting for the three-way relationship between word lengths, frequencies, and number of meanings identified by Zipf [1]. Pragmatic inferences based on context are
another plausible cause for the narrowing or broadening of word meanings [8,9]. 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 usually conveyed by that phrase [9]. 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 [10–14]. 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 [15], or in experiments where anonymous interlocutors are made to communicate with minimal cues [11,12], 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 [16]. One obstacle on this path is the lack of a unified and consensual definition or metric for the informativeness of symbols. Linguists and philosophers traditionally distinguish two ways that a word may be uninformative or ambiguous: polysemy and vagueness [17–19]. 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”, “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 [19,20]. 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 towards a quantitative metric for ambiguity in the word sense disambiguation literature [21], where “sense entropy” measures, in information-theoretic terms, the dispersion of word uses between distinct meanings (see also [3]).

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 [22] points out, some words are “tolerable friends”, carrying distinct (though related) meanings to different speakers. The English word “peruse” standardly means “to read or examine, typically with great care”, but the 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 [23], while an entirely new meaning, “browsing the internet”, is also gaining ground. Unlike false friends, Enfield’s “tolerable friends” aren’t due to cross-linguistic misunderstandings: they may co-exist 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.

Even though ambiguity may be advantageous for communication (when precision is costly or context provides sufficient information [3,11,12,24,25]), a convention linking a symbol with the same precise meaning for all users is, all else being equal, of clear benefit for information transmission [26]. 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 [3,17,24,25]. How do such unambiguous semantic conventions evolve? Along with formal models (e.g. [27]), experiments in artificial language evolution and experimental semiotics [28,29] provide tools to answer this question. Experimental set-ups 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 [12].

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 [30,31]. The key distinction between transformative and selectionist evolutionary processes concerns the role of individual insight and creativity. Both accounts agree that evolutionary change feeds upon variation, but disagree upon the nature of variation: constrained by cognition in one case, random or unconstrained in the other. 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 non-existent [30]. In the case of communicative conventions, where agents’ primary goal is to align their behaviour 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 [32–34]. There is also evidence that, in artificial language evolution experiments, selectionist processes matter much more than transformative ones [32,35]. In a selectionist scenario, precise semantic conventions would evolve because they are more likely to be imitated than others, thus becoming more widespread.

This selectionist account suggests a solution to the problem of consensus. In Enfield’s hypothesis [22], 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, and the more numerous such occasions, the easier it is to converge upon similar meanings. In a selectionist view, precise conventions gain more adepts, making them more likely to be frequently encountered. If Enfield’s conjecture is right, being frequent could in turn make these conventions more consensual, reducing variance between different individuals’ conception of a word’s meaning.
We tested the hypothesis according to which conventions mapping a symbol with a narrow range of referents are favoured in a setting where participants must communicate with unfamiliar interlocutors, using unfamiliar symbols, with little contextual information to rely on. We expected cultural evolution to favour precise conventions, and we expected this to happen in a selectionist fashion. In addition, we expected to see more agreement between interlocutors over the range of meanings of the most frequently used symbols.

Methods

We studied the evolution of communicative conventions over 11 months in an online multiplayer gaming app, the “Color Game” app [36,37]. 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 Methods file.

The app allowed players from more than 100 nationalities to participate in a series of referential communication games [38], where one player (the Sender) had to communicate a colour (the target, indicated by a dot) to a Receiver presented with an array that included the target and three other colours. Participants could only communicate through black and white symbols. Points were earned 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 ten 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 symbols. Through the app, players could freely choose to associate with willing co-players, and could switch between the roles of Sender and Receiver if they found other players willing to play the opposite role, an interaction structure closer to real communication dynamics than a typical experiment. To ensure statistical robustness, the player pool was divided into two closed “halves”: a player could only play with co-players from her half. A system of points and a user-friendly design ensured that players were endogenously motivated.

Each of the game’s 32 colours was drawn from the CIE2000 colour space [39], chosen because it provides a metric for distance between colour hues (“Delta E”) built to reflect perceptual distance, as opposed to merely physical quantities. The colours were equal in luminance and saturation, with a constant perceptual distance between any colour and its two neighbours of Delta E = 7.8 (Fig. 3). Thirty-two colour arrays were formed from this set of 32 colours by picking every fourth colour along the dimension of hue, until a four-colours array was formed, using each of the 32 colours as starting point. In this way, all colours occurred in exactly four arrays. The array present on any given trial was randomly picked, all arrays being equiprobable. Which of the array’s four colours served as the target was likewise randomised. In addition to the target colour (which she always saw), the Sender could see some or all of the colours visible in the Receiver's array. This quantity varied from one (only the target) to four (the full array). The number and nature of the colours shown to Sender were also randomised.
Senders had to communicate using black and white symbols bearing no straightforward
association with any single hue of colour. These symbols had been experimentally tested to
make sure that they would be neither too easy (evoking too narrow a range of colours), nor
too difficult (allowing no colour 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 colours [40]. To maximize the variability in symbol use, as well as provide the
game with a reward structure, the players who start 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.

Predictions

We preregistered this study in three waves. The first preregistration 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 pre-test our predictions on a subsample of the data
(second registration). The Color Game did not incentivise 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 pre-test 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). 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,
including data exclusions and model specifications, were registered at this stage. Our
predictions were the following:

Prediction 1. As they get more experienced, players increasingly prefer to use more
informative mappings between symbols and colours. (Informativeness is defined below, as
low sense entropy.) This prediction reflects the view that, when communicating with
anonymous coplayers in a situation of scarce contextual information, players will gradually
evolve highly informative conventions. It also reflects the view that this evolution will be
underpinned by a selectionist process: the rise of informative symbols, we assumed, would be
driven by more frequent use for symbols that mapped to a narrower range of colours. These
mappings would then be imitated and spread selectively.

Prediction 2. When a symbol has been frequently been used by a pair of players, they are
more likely to use it to indicate the same colours. This prediction is meant to test Enfield’s

---

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

**Sense entropy for a continuous and circular meaning space**

To measure the degree to which the conventions linking a symbol to a range of colours are precise or vague, we use a measurement tool that is in some respects similar to “sense entropy” in the word disambiguation literature [3,21]. To calculate the sense entropy corresponding to the various meanings of a word, one counts how many times the word was used to mean sense 1, sense 2, etc. This yields a probability distribution over which a standard entropy calculation is performed. A higher sense entropy means that the target colours associated with a symbol by a player are unpredictable. Thus, a high-entropy symbol is less (not more) informative than a low-entropy one.

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 colours 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 in a continuous and circular space. (The colour space is assumed to be continuous because colours in the CIE2000 space are designed to be perceptually equidistant.) Like standard sense entropy, this measure uses Shannon’s entropy [41]. It was, however, adapted by one of us (JW) to fit the structure of the colour space (Fig. 1). This alternative measure suits our purposes better than existing measures of entropy or angular dispersion on a circle [42–46]. Our measure extends upon standard measures of conditional entropy whilst preserving both the granularity of the colour space and its discrete circular ordering, as per equation (2):

\[
(1) \ H_w(C|S_i) = \frac{H(C|S_i)}{\max_{c \in C} H(C)} \times \frac{\max_{S_i} H(W|S_i)}{\max_{c \in C} H(C)}
\]

- \( H_w(C|S_i) \) is the “category-entropy” for a set of colours C (the 32 colours of our colour space) for 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 equation (2):

\[
(2) \ H(C|S_i) = \sum_{c \in C} P(c) \log P(c)
\]

where \( c \in C \) is one colour of the set of colours 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 colour from C.

- \( \max_{c \in C} H(C) \) is the maximum entropy that can obtain over the set of all possible colours. In other words, it is the entropy of the probability distribution that we get by assuming all the game’s 32 colours to be equiprobable.
\[ \max_{c \in W} H(W|S_i) \] corresponds to the maximum entropy of the colour category structure \((W)\) for symbol \(S_i\). A symbol’s “category structure” refers to the set of contiguous colours that form the shortest path linking all the colours that a symbol is associated with (see Fig. 1).

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).

Fig. 1. 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 colour 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 colours 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 colours on the colour wheel, giving the same values for distributions 1 & 2, or to 3 & 4. Our modified measure (bottom row) also considers the width of the range of colours that a symbol is associated with, the “category structure” or “\(w\)”, indicated by the circular black line (see Equation 1). This results in different entropy values for distributions 1 to 4.

Results

For both predictions, we started from the "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. 75 635 trials were removed for this reason (271 971 trials remaining). Only for measuring one specific variable (relevant for prediction 1) were those trials kept.

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

We extracted data on all the Senders who used at least one identical symbol on at least 20 trials, after reaching the level that unlocks the full keyboard. There were 209 such Senders, for a total of 2,154 data points (each data point being 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"), log-transformed to avoid convergence failures and other scaling-related issues. 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 —[47,48]) 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, thus this second version was not considered further. 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 information) 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 residuals plots did not change this effect. The second prediction 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 residuals 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 colours.
One consequence of this unexpected pattern is that one of our original expectations could not be properly tested. Prediction 1 was premised upon the view that players would gradually come to use symbols in a more informative 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). More experienced players use their symbols in a more informative way, to refer to a narrower range of colours (Figure 1).

![Figure 1: Average sense entropy of the player's symbols against player experience (log-transformed)](image)

**Fig. 2.** 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).

As a further follow-up, we considered how sense entropy evolved through the course of a player's career, from the time when they unlock the full keyboard to 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 points 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. We then built a linear mixed effects model to predict SENSEENTROPY in the use of a given 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). More experienced players have more informative symbols because experience made their symbols more informative (Fig. 3). We replicated this result after excluding six outlier...
players who played many more trials as Senders than all other players (and thus have a much longer history of playing).
**Fig. 3. 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 player used 20 times or more during their first and their last 1000 trials. The distribution of target colours associated with each symbol by players 1 to 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.

**Prediction 2: Frequent symbols are consensual.**

For this prediction, we also discarded the trials for which there was no Receiver. These are trials that were played by a Sender, who saw a target colour and sent a series of symbols to help some Receiver pick the colour, 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' behaviour exclusively.) Removing them caused 8,694 trials to be excluded (resulting 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 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 [49]) between the colour distributions associated with the symbol by each of the two player (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 informative fashion by both players ( informativeness being measured using categorical entropy, as for prediction 1) are more likely to be used by both players to signal the same colours.

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 informatively, disagreement can be very high but it can also be quite low. Conversely, when players use a symbol in a very uninformative fashion, corresponding to a high degree of
dispersion of the associated colours, the distance between player 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 consensual than that of little used symbol. The effect of AVERAGEENTROPY remained strongly positive in this final model.

Discussion

Our participants, faced with a referential communicational 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 colours grow more precise in the course of a player’s career, and more experienced players associate symbols with a smaller range of colours as a result. There does not seem to be any equivalent for such a general narrowing of semantic conventions in the history of natural languages. Lexical ambiguity changes through time for particular words, through narrowing or broadening [7], but there is no evidence of entire vocabularies gradually becoming more specific [17]. 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 ([3,10–12], but see [50]). Our results support this view. Contextual information in the app was kept to a minimum: Senders’ symbols were the only cue that Receivers could exploit to pick the target colour from the three distractors. This put a strong pressure over the symbols to be maximally informative, and they evolved accordingly. The way this evolution took place was surprising, given the strong emphasis the literature gives to selectionist as distinct from transformative evolutionary processes [32,33]. We expected participants to copy one another’s use of the symbols fairly closely, since communication can only gain from coordination. Precise mappings between colours 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', '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 (intra-class correlation for symbols: 0.24; for players: 0.32; computed with irrNA [51]). 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
informative fashion. This process of semantic narrowing [7] took place despite the fact that high-entropy mappings were more widespread than low-entropy ones.

The strong positive correlation that we observe, between the sense entropy of the mapping associated with a symbol and that symbol’s frequency, makes retrospective sense. Numerous previous findings indicate a positive link between a word’s frequency and the number of meanings it accepts — a correlation known as “Zipf’s Law of Meaning” [1] (also known as the “principle of economic versatility” [8]). This relation holds when controlling for confounds such as word length [3]. Our findings extend and deepen Zipf’s Law of Meaning in two ways. We show how it applies to a non-verbal symbolic language, using a measurement of ambiguity that goes beyond the standard polysemy measurement (i.e. counting meaning) and integrates sense entropy as well as vagueness, taking advantage of the continuous structure of the set of possible meanings in our task.

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 minimise their production effort and the comprehension effort of their interlocutors. Building upon this, Piantadosi [3] proposes that frequent words, being easier to use and process (since their frequency makes them easier to memorise), 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 (frequency distributions approximate a power law).

A second type of explanation focuses on the organisation 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 [52]. 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 seems 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 memorise and process by 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 informativeness 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 colours 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. [17]: 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 words, 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.

Conclusion

Why are communicative signals less than optimally informative? 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 colours. On the other hand, even highly experienced participants kept using some symbols in quite uninformative 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 [30,31,53].


Supplementary Materials: We append a Supplementary Methods document (specific to this paper), and a general presentation of the Color Game project and dataset (shared with all Color Game projects).

Author contributions. OM, JW and TFM conceived the original study and wrote the preregistration. JW conceived the special entropy measurement. OM, JW and TFM organized data collection through the Color Game app. OM, TM and JW carried out the statistical analysis. OM drafted the manuscript, with critical revisions by the other authors. All authors gave final approval for publication.

Ethics. This study has been approved by the ethical committee attached to the president of the Max Planck Society (advice n° 2017_05). Special attention was paid to data privacy in designing and running the study.

Acknowledgements. We thank all the persons involved in the Color Game project, listed here (https://osf.io/nsxu4/), and all the players. This work has received funding from the "Frontiers in Cognition" EUR grant, ANR-17-EURE-0017 EUR.
References


Supplementary methods for the paper

“Cultural evolution of precise and consensual semantic conventions in a multiplayer gaming app”

An open-ended iterated referential communication game. The Color Game app enabled its players to take part in open-ended series of referential communication games with other players of their choice. At its core, the game is a standard referential communication task (Figure S1). One player (the Sender) is presented with a "target colour", indicated by a dot. This colour corresponds to one of the colours in an array of four that is seen by the other player (the Receiver). The sender's goal is to communicate with the Receiver to help them pick the target colour. A successful pick earns points for the Receiver. Senders must communicate using black and white symbols that bear no straightforward association with any single hue of colour. Unusually for an artificial language evolution experiment, our app does not provide players with trial-by-trial feedback on the success or failure of communication. A block of 10 trials must be played by both Sender and Receiver for points to be earned. After every block, the Receiver is told how many of the preceding ten trials they got right, but not which ones. If both players were playing synchronously (see below) both players get that information, otherwise the Sender merely knows that someone played with their messages, and how many points they earned as a result. In laboratory versions of this task, participants performed the task repeatedly in pairs, without changing partners. In contrast, the app’s multiplayer functionalities and the large number of available players allowed us to let players choose their partners and interactions freely.

![Sender and Receiver screens](image)
**Fig. S1.** An example of a trial in the Color Game, showing Sender’s screen (left) and Receiver’s screen (right).

**Synchronous vs. asynchronous play.** The app left players free to choose their interlocutors and the format of their interaction. It allowed for both “synchronous” (i.e., live, or in real time) and “asynchronous” play. Senders playing asynchronously simply typed black and white symbols corresponding to the target colour in a window seen by both Sender and Receiver. The symbols were then sent to Receivers, along with the corresponding colour array, but with no indication about the target. The Receiver then had to figure out the target from the symbols sent by the Sender. These *asynchronous puzzles* remained available inside the app indefinitely, but each puzzle disappeared as soon as one Receiver had played it. *Synchronous play*, on the other hand, required Sender and Receiver to contact each other and stay connected for as long as they played; this enabled them to communicate in real time, and exchange repair signals consisting of the signs ”?!” and ”!”. (The players were not told what these punctuation marks should mean in the context of the game, and we expected them to be used idiosyncratically; in fact, they served little.) In both cases, once a receiver completed a puzzle (a set of 10 trials), this puzzle ceased to be available for other Senders. Completion means that the Receiver had made a guess for each of the 10 target colours corresponding to the 10 messages sent by the Sender. A bug made it possible for some receivers to complete some senders’ asynchronous puzzles more than once, but these trials were discarded (see Data exclusions). Some trials were never solved, because receivers did not pick them up, or because they formed parts of puzzles that were incomplete (less than 10 trials). These uncompleted trials are not taken into account when we consider the players’ performances in finding the target, but we do take them into account for prediction 2, which is only concerned with Senders’ behavior.

**Partner choice in the Color Game.** The app’s players were free to choose their partners from a vast pool of players (**Fig. S2**). To play with another contact, a player either invited them for synchronous play and waited for the invitation to be accepted, or sent an asynchronous message, which could be broadcast to the whole group or sent to a specific individual contact who could open it at a later time. To ensure that the app always contained a sufficient number of high-quality asynchronous puzzles, Receivers paid Senders for their asynchronous puzzles, which they “bought” with the points they had previously earned by playing. Points were typically earned by successfully completing puzzles, the rewards for playing well being higher than the typical puzzle’s price. The Sender received these points regardless of the Receiver’s performance, incentivizing participants to play as Senders. Extra points could be earned by playing an especially difficult, time-limited "speed mode", available to players after a certain level. This points system was meant to increase player’s enjoyment, and also to discourage the creation of low-quality puzzles or the absorption of high-quality puzzles by negligent players.
player anonymity. Strict player anonymity was ensured by a system of pseudonyms and "cover names". The app recorded no personal identifiers (such as names or nicknames), but gave each player a public pseudonym, displayed on the "top players" score board. However, that pseudonym could not be used to identify a player as a contact to play with. Instead, contacts were only ever known to one another by means of "cover names" that varied from contact to contact and from player to player. Player X knew her contact through a list of randomly generated names that reliably identified contact A, contact B, contact C, and so on. Player Y could also identify the same contacts A, B, C, but by completely different names, so that A's name for X did not correspond in any way to A's name for Y, even though both players could reliably identify and contact A. This ensured that players could not circumvent the app to contact one another in real life, on social networks, etc.

Division of the player pool. Players were divided between two big pools of players, with each new player being assigned randomly to one of these to “halves”. All interactions between players were restricted to their half, for as long as the app was on. 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.

Languages. When released in April 2018, the app was available in five languages: English, Chinese, French, German, and Spanish. The Color Game makes sparse use of natural language, so translations chiefly concerned the tutorial and short mentions found on the home screen (Fig. S2) and on the screens informing players about their performance. In version 1.10 of the app, released on September 15th, 2018, we added languages Portuguese (Brazil), Russian, and Japanese. However, for most of these languages we did not manage to get the
number of players required by the study’s preregistered exclusion thresholds. Only speakers who chose to play the app in English, French, German, or Spanish were retained (see below, Data exclusions).

Tutorial: 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 re-used in the normal course of the game (figure S3, bottom row), and a colour array randomly picked from the 32 possible arrays. The player was presented with a dummy symbol and a four-colours array, and asked to point which colour 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 colour. What symbol or colour the players 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 (cf. Fig. S2).

Ethics, data privacy & open data. Participants agreed to have their data collected in anonymous format and for research purposes in a consent form approved at the start of the game. The app and the six preregistered studies were approved by the Max Planck Society headquarters’ ethical committee (advice n° 2017_05). The app’s source code is open and the anonymous data will be made publicly available after a period of embargo. In addition to recording all the players’ moves as Sender or Receiver, the app also records notes taken by the players who wish to make use of “the notebook”, a feature of the app that allowed players to write down the meaning of a particular symbol (as they figure it). (See the full statement on data privacy here: http://osf.io/guwyr/.)

App launch and termination. Once it was ready, and after a lengthy period of piloting and debugging (all beta players involved in the piloting and debugging, friends and collaborators of the authors and developers, were barred from taking part in subsequent stages), the App was deployed in two steps. A ‘technical’ launch took place on April 26th, 2018: private invitations were sent to selected guests — journalists, influencers, participants to gaming conventions. 80 players were thus brought in. They allowed us to fine-tune the app’s points system, and seeded the app with puzzles that new players would be able to solve. On May the 15th 2018 (midnight), the app was accessible on the Android and iOS app stores, the embargo on the press release was lifted, and the wider public started downloading the game. This marked the end of the technical launch period. As planned at preregistration, we discarded all the trials that were completed before May the 15th (see below, Data exclusions). The app was originally scheduled to be live until May 15th, 2019, but as traffic slowed down it was decided to shut it on April the 10th, 2019. At that point there were still several players a day playing on the App.

Symbols. Senders had to communicate using black and white symbols bearing no straightforward association with any single hue of colour (Fig. S3). These symbols had been experimentally tested to make sure that they would be neither too easy (evoking too narrow a range of colours), nor too difficult (allowing no colour 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 colours [40]. To maximize the variability in
symbol use, as well as provide the game with a reward structure, the players who start 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.

![Symbols](image)

Fig. S3. The 35 symbols used in the game (first four rows). Bottom row, in grey: the five
symbols used for the tutorial and for the videos advertising the game (these symbols are for
tutorials only).

**Colour space and colour arrays.** Each of the game’s 32 colours was drawn from the CIE2000
colour space [39], chosen because it provides a metric for distance between colour hues
(“Delta E”) built to reflect perceptual distance, as opposed to merely physical quantities. The
colours were equal in luminance (L = 55) and saturation (S = 85), with a constant perceptual
distance between any colour and its two neighbors of Delta E = 7.8 (Fig. 3). Thirty-two colour
arrays were formed from this set of 32 colours by picking every fourth colour along the
dimension of hue, until a four-colours array was formed, using each of the 32 colours as
starting point (Figs. S4 & S5). This way, all colours occurred in exactly four arrays. The array
present on any given trial was randomly picked, all arrays being equiprobable. Which of the
array’s four colours served as the target was likewise randomized. In addition to the target
colour (which she always saw), the Sender could see some or all of the colours visible in
Receiver's array. This quantity varied from one (only the target) to four (the full array). The number and nature of the colours shown to Sender were also randomized.

Fig. S4. The game’s colour space. Each colour is given its associated Hex code (as used by the app).
Fig S5. How colour arrays were built. Top row: The composition of two colour arrays, one marked by white dots, the other by black dots, is shown relative to the colour space. Bottom row: Six contiguous colour arrays (out of 32), including the white-dot and black-dot ones.