
# Cultural evolution via iterated learning and communication explains efficient color naming systems

*Journal of Language Evolution*, 2024, **XX(XX)**, 1–18
https://doi.org/10.1093/jole/lzae010
Advance access publication 22 November 2024
**Research article**

# Cultural evolution via iterated learning and communication explains efficient color naming systems

**Emil Carlsson[1,*], Devdatt Dubhashi[1] and Terry Regier[2]**

[1]Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Göteborg, Sweden
[2]Department of Linguistics, UC Berkeley, Berkeley, CA, 94720-2284 United States

*Corresponding author: Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Göteborg, Sweden.
E-mail: caremil@chalmers.se

**Associate Editor:** Dan Dediu

It has been argued that semantic systems reflect pressure for efficiency, and a current debate concerns the cultural evolutionary process that produces this pattern. We consider efficiency as instantiated in the Information Bottleneck (IB) principle, and a model of cultural evolution that combines iterated learning and communication. We show that this model, instantiated in neural networks, converges to color naming systems that are efficient in the IB sense and similar to human color naming systems. We also show that some other proposals such as iterated learning alone, communication alone, or the greater learnability of convex categories, do not yield the same outcome as clearly. We conclude that the combination of iterated learning and communication provides a plausible means by which human semantic systems become efficient.

**Keywords:** cultural evolution; iterated learning; efficient communication; semantic categories; color naming

## 1. Introduction

Semantic categories vary across languages, and it has been proposed that this variation can be explained by functional pressure for efficiency. On this view, systems of categories are under pressure to be both simple and informative (Rosch, 1978), and different languages arrive at different ways of solving this problem, yielding wide yet constrained cross-language variation. There is evidence for this view from semantic domains such as kinship (Kemp and Regier, 2012), container names (Xu et al. 2016), names for seasons (Kemp et al. 2019), indefinite pronouns (Denić et al. 2022), modals (Imel and Steinert-Threlkeld, 2022), numeral systems (Xu et al. 2020, and relatedly Denić and Szymanik, 2024). Zaslavsky et al. (2018) gave this proposal an independent theoretical foundation by grounding it in an information-theoretic principle of efficiency, the Information Bottleneck (IB) principle (Tishby et al. 1999); they also showed that (1) color naming systems across languages are efficient in the IB sense, (2) optimally IB-efficient systems resemble those found in human languages, and (3) the IB principle accounts for important aspects of the data that had eluded earlier explanations. Subsequent work has shown that container naming (Zaslavsky et al. 2019), grammatical categories of number, tense, and evidentiality (Mollica et al. 2021), and person systems (Zaslavsky et al. 2021) are also efficient in the IB sense.

In a commentary on this line of research, Levinson (2012) asked how semantic systems evolve to become efficient and suggested that an important role may be played by iterated learning (Scott-Phillips and Kirby, 2010). In iterated learning, a cultural convention is learned by one generation of agents, who then provide training data from which the next generation learns, and so on. The convention changes as it passes through

generations, yielding a cultural evolutionary process. The idea that such a process could eventually lead to efficient semantic systems has since been explored and broadly supported. Xu et al. (2013) showed that chains of human learners who were originally given a randomly generated color category system eventually produced systems that were similar to those of the World Color Survey (WCS; Cook et al. 2005), a large dataset of color naming systems from 110 unwritten languages. Although this study did not directly address efficiency, Carstensen et al. (2015) drew that link explicitly: they reanalzyed the data of Xu et al. (2013) and showed that the color naming systems produced by iterated learning not only became more similar to those of human languages—they also became more informative; the same paper also presented analogous findings for semantic systems of spatial relations. In response, Carr et al. (2020), building on earlier work by Kirby et al. (2015) and others, argued that iterated learning primarily contributes simplicity rather than informativeness—but that a bias for simplicity can nonetheless sometimes result in an increase in informativeness. Overall, there is support for the idea that iterated learning can lead to efficient semantic systems, with continuing debate over how and why. There are also recent proposals that non-iterated learning—for example, in the context of a dyad of communicating agents (Kågebäck et al. 2020; Chaabouni et al. 2021; Tucker et al. 2022), or in a single agent without communication (Steinert-Threlkeld and Szymanik, 2020; Gyevnar et al. 2022)—can explain efficient color naming systems. In particular, Steinert-Threlkeld and Szymanik (2020) argued that "[e]ase of learning explains semantic universals" (see also Gentner and Bowerman, 2009). To support this claim, Steinert-Threlkeld and Szymanik (2020) first noted that earlier proposals (Gärdenfors 2000; Jäger, 2010) had argued for the importance of convexity in conceptual space as an important constraint on human semantic categories; they then demonstrated the greater learnability, in a neural network, of convex as opposed to non-convex color categories. These recent contributions, and the present one, build on an important line of earlier work using agent-based simulations cast as evolutionary models, without explicitly addressing efficiency (Steels and Belpaeme, 2005; Belpaeme and Bleys, 2005; Dowman, 2007; Jameson and Komarova, 2009; Baronchelli et al. 2010).

Several of these prior studies have engaged efficiency in the IB sense, and two are of particular relevance to our own work. Chaabouni et al. (2021) showed that a dyad of neural network agents, trained to discriminate colors via communication, eventually arrived at color naming systems that were highly efficient in the IB sense. However, these systems did not always resemble those of human languages: their categories "depart to some extent from those typically defined by human color naming" (Chaabouni et al. 2021: 11 of SI). Tucker et al. (2022) explored a similar color communication game, and found that their neural agents gravitated to color naming systems that are both essentially optimally efficient in the IB sense, and similar to human color naming systems from the WCS. They achieved this by optimizing an objective function that is based on the IB objective. To our knowledge, earlier work leaves open whether both high IB efficiency and similarity to human languages can be achieved through processes and principles that are independent of IB. We explore that question here. We also wish to establish here whether such independent principles may address the one case in which IB-optimal color naming systems deviate to some extent from empirical observation: the case of three-term systems (Zaslavsky et al. 2018,: 7941). Overall, we wished to ascertain whether a natural model of cultural evolution might account both for the many cases in which IB matches the data, and for the one case in which it deviates from the data to some extent.

A natural candidate model of cultural evolution was advanced by Kirby et al. (2015), and the ideas we pursue here build on that general model. Specifically, Kirby et al. (2015) proposed a model of cultural evolution that interleaves two kinds of learning touched on above: (1) learning that occurs during transmission of a linguistic system from one generation to the next, and (2) learning that occurs during communication among agents within a single generation. That formulation allowed them to isolate the effect of each of the two kinds of learning, and to examine their combination. They were interested in particular in what evolutionary forces could give rise to compositional structure of the sort found in human language. In computational simulations and experiments with human participants, they found that transmission from one generation to the next exerted pressure for simplicity, that within-generation communication exerted pressure for informativeness—and that only the two forces operating together gave rise to compositional structure. Here, we apply the same general cultural evolutionary model to a different question, that of color naming systems in human languages.

In what follows, we first demonstrate that there exist many possible color naming systems that are highly efficient in the IB sense, but do not closely resemble human systems. The fact that there exist such efficient-yet-not-human-like systems is not surprising given that IB is a non-convex optimization problem (Tishby et al. 1999; Zaslavsky et al. 2018), but appreciating the

prevalence of such systems may be helpful in understanding how Chaabouni et al. (2021) achieved high IB efficiency with systems that deviate from human ones. We then show that the general cultural evolutionary model of Kirby et al. (2015), instantiated in neural networks (Ren et al. 2020), gravitates toward efficiency and, within the class of efficient systems, gravitates more toward human color naming systems than toward others. Finally, we show that iterated learning alone, communication alone, and convexity alone, do not yield that outcome as clearly. We conclude that iterated learning and communication jointly provide a plausible explanation of how human color naming systems become efficient.

## 2. Not all efficient systems are human like

We considered a natural class of artificial color naming systems (Abbott et al. 2016; Zaslavsky et al. 2022). In this class, each named category $w$ is modeled as a spherical Gaussian-shaped kernel with mean (prototype) $x_w$ in three-dimensional CIELAB color space (Fig. 1, top right panel), such that the distribution over words $w$ given a color chip $c$ at location $x_c$ in CIELAB space is:

$$S(w|c) \propto e^{-\eta||x_c - x_w||_2^2} \quad (1)$$

where $\eta > 0$ is a parameter controlling the precision of the Gaussian kernel. We then generated artificial color category systems with $K = 3, \ldots, 10$ categories each, by first sampling $\eta$ randomly from a uniform distribution over the interval $[0.001, 0.005]$ for each system, using the same $\eta$ for all categories in a given system, and then sampling the prototype $x_w$ of each category $w$ randomly, without replacement, from a uniform distribution over the cells of the color naming grid shown in the top left panel of Fig. 1. This figure shows the same set of colors as in the top right panel, but now in a 2D array. In analyzing these systems, we draw on four quantities from the IB framework as presented by Zaslavsky et al. (2018) and reviewed below in the Appendix: the complexity of a category system, the accuracy of a category system, $\epsilon$ (a measure of the inefficiency of a category system, or its deviation from the theoretical limit of efficiency), and gNID (a measure of dissimilarity between two category systems). We noted that the range of complexity (in the IB sense) for systems in the WCS was $[0.84, 2.65]$, and also noted that our random model sometimes generated systems outside this range; we only considered artificial systems with complexity within this range, and generated 100 such systems for each $K$; we refer to these randomly generated systems as Gaussian systems.

The lower panels of Fig. 1 compare natural color naming systems to artificial Gaussian systems. The leftmost column shows three attested color naming systems from the WCS, from top to bottom: Bété (iso: bev, Côte d'Ivoire), Colorado / Tsafiki (iso: cof, Ecuador), and Dyimini (iso: dyi, Côte d'Ivoire). The middle column shows randomly generated Gaussian systems that are similar to the WCS system in the same row, and the rightmost column shows Gaussian systems that are dissimilar to the WCS system in the same row but of about the same complexity. In each row, the rightmost system, which is dissimilar to the WCS system in that row, is nonetheless more similar to that WCS system than to any other WCS system; this means it is dissimilar to all WCS systems. Thus, there exist Gaussian systems that are quite similar to naturally occurring systems, and other Gaussian systems that are quite dissimilar to naturally occurring systems. To quantify this pattern, we separated the Gaussian systems into two groups, based on whether their gNID to the closest WCS system exceeded a threshold. We set this threshold to the smallest gNID between systems in the left (WCS) and right (Gaussian dissimilar) columns of Fig. 1, which is 0.29. We then grouped all Gaussian systems with gNID to the closest WCS system below this threshold into one group, Gaussian[S] (for similar to WCS), and the other Gaussian systems into another group, Gaussian[D] (for dissimilar to WCS). We found that 38% of the Gaussian systems fell in Gaussian[D] and they spanned the complexity range $[0.86, 2.26]$. Thus, a substantial proportion of the randomly generated Gaussian systems are at least as dissimilar to WCS systems as are those in the right column of Fig. 1.

Figure 2 shows the results of an IB efficiency analysis of the WCS systems (replicating Zaslavsky et al. 2018, and assuming their least-informative prior), and also of our Gaussian systems. It can be seen that all Gaussian systems are highly efficient in the IB sense—that is, they are close to the IB curve that defines the theoretical limit of efficiency in this domain. Mann–Whitney $U$ tests revealed (1) that the Gaussian systems tend to exhibit greater efficiency (lower inefficiency $\epsilon$) than do the WCS systems in the same complexity range ($P \ll .001$), and (2) that the Gaussian[D] systems, which are dissimilar to WCS systems, are also more efficient than WCS systems ($P \ll .001$, one sided), and slightly to marginally more efficient than Gaussian[S] systems ($P = .019$ one sided; Bonferroni corrections do not change the qualitative outcome) (Throughout the paper we use one-sided tests when comparing different sets of color naming systems to the Gaussian systems. The reason for this is that we are interested in knowing whether various systems generated by an evolutionary process exceed a

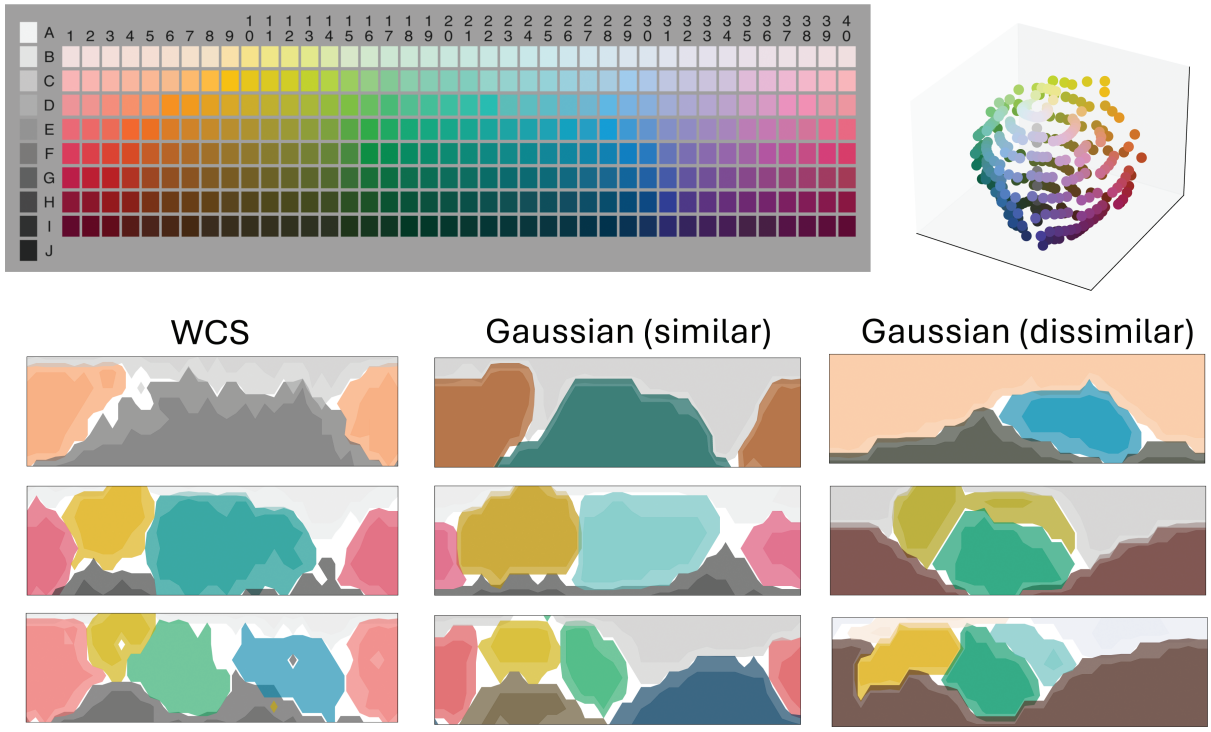


**Figure 1.** Top: Color naming stimulus grid (left), and stimuli plotted in CIELAB space (right). Bottom: nine color naming systems displayed relative to the grid. The left column contains color naming systems from three languages in the World Color Survey (WCS). Colored regions indicate category extensions, and the color code used for each category is the mean of that category in CIELAB color space. The named color categories are distributions, and for each category we highlight the level sets between 0.75–1.0 (unfaded area) and 0.3–0.75 (faded area). The middle and right columns contain randomly generated Gaussian systems of complexity comparable to that of the WCS system in the same row. The middle column shows random Gaussian systems that are similar to the WCS system in the same row. The right column shows random Gaussian systems that are dissimilar to the WCS system in the same row; at the same time, there is no other WCS system that is more similar to this Gaussian system.

random baseline when it comes to either efficiency or similarity to human systems.). These findings suggest that there is a substantial number of color naming systems that are dissimilar to those of human languages, yet more efficient than them. This in turn may help to make sense of Chaabouni et al.'s (2021) finding that their evolutionary process yielded systems that were highly efficient but not particularly similar to human ones: our analysis illustrates that there are many such systems. Given this, we wished to determine whether a natural evolutionary process would yield both efficiency in the IB sense, and similarity to human systems.

## 3. Iterated learning and communication

As noted above, iterated learning (Kirby, 2001; Smith et al. 2003) is a cultural evolutionary process in which a cultural convention is learned first by one generation of agents, who then pass that convention on to another generation, and so on—and the convention changes during inter-generational transmission. Some of the work we have reviewed above addresses iterated learning (Levinson, 2012; Carstensen et al. 2015). However, other work we have reviewed instead addresses cultural evolution through communication within a single generation (Kågebäck et al. 2020; Chaabouni et al. 2021; Tucker et al. 2022). We wished to explore the roles of both iterated learning and communication, and so we adopted the general approach of Kirby et al. (2015), which involves both in a way that allows the role of each to be highlighted. Specifically, we adopted the recently proposed *neural iterated learning* (NIL) algorithm Ren et al. 2020, which can be seen as a neural network implementation of the approach of Kirby et al. (2015). In the

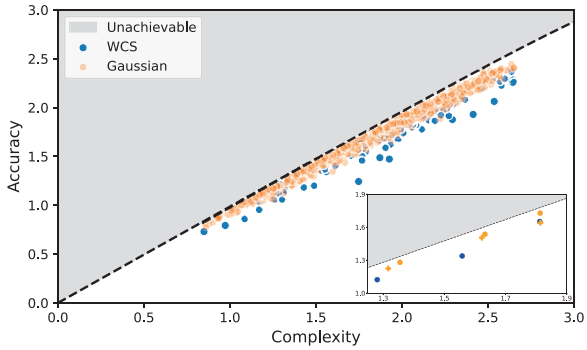


**Figure 2.** Efficiency of color naming, following Zaslavsky et al. (2018). The dashed line is the IB theoretical limit of efficiency for color naming, indicating the greatest possible accuracy for each level of complexity. The color naming systems of the WCS are shown in blue, replicating the findings of Zaslavsky et al. (2018). Our randomly generated Gaussian systems are shown in orange. The Gaussian systems are often closer to the IB curve than the WCS systems are. The inset shows the nine color systems of Fig. 1, with the dissimilar Gaussian systems shown as +.

NIL algorithm, illustrated in overview in Fig. 3, artificial agents are implemented as neural networks that communicate with each other within a generation, and transmit information across generations. Cultural convention (in our case, a color naming system) evolves both from within-generation communication and from inter-generational transmission, as the convention is iteratively passed down through generations of artificial agents, with each new generation learning from the previous one (NIL, or neural iterated learning, is therefore not an entirely informative name for this process, as it does not explicitly label the important element of within-generation communication.).

In the NIL algorithm, each generation $t$ (for time step) consists of two artificial agents, a speaker $S_t$ and a listener $L_t$. The NIL algorithm operates in three phases. (1) In the first phase, the *learning phase*, both agents are exposed to the naming convention of the previous generation. This is done by first training the speaker $S_t$, using cross-entropy loss, on color-name pairs generated by the speaker of the previous generation. The listener $L_t$ is then trained via reinforcement learning in a few rounds of a signaling game while keeping $S_t$ fixed: that is, the speaker learns from the previous generation, and the listener then learns from the speaker. We had the agents play the signaling game used by Kågebäck et al. (2020), in which the speaker is given a color chip $c$, sampled from a prior distribution over color chips, and produces a category name describing that color. The listener then attempts to identify the speaker's intended color based on the name produced, by selecting a color chip $\hat{c}$ from among those of the naming grid shown in Fig. 1. A reward is given to the listener depending on how perceptually similar the selected chip is to the original color, following equation 2. (2) In the second phase, the *interaction phase*, the agents play the same signaling game but this time both agents receive a joint reward and update their parameters during communicative interactions. (3) In the third phase, the *transmission phase*, color-name pairs are generated by sampling colors from the prior distribution and obtaining names for them from the speaker $S_t$. These color-name pairs are then passed on to the next generation of agents. In all three phases, color chips are sampled according to the least-informative prior of Zaslavsky et al. (2018). Algorithm 1 presents a schematic overview of the NIL algorithm, and Ren et al. (2020) present a detailed description. Both the NIL algorithm and the setting explored by Kågebäck et al. (2020) build on important earlier work exploring the emergence of communication in neural network models (Foerster et al. 2016; Havrylov and Titov, 2017; Lazaridou et al. 2017; Mordatch and Abbeel, 2018).

In our experiments, we represent both the speaker and listener as neural networks with one hidden layer consisting of 25 units with a sigmoidal activation function followed by a softmax output layer. Individual colors are represented in three-dimensional CIELAB space when supplied as input to the speaker, and category names as one-hot encoded vectors. The speaker's network parameterizes a conditional distribution over categories given a color. To produce an utterance during communication, the speaker samples a category from this distribution and conveys it to the listener. The input to the listener is the category uttered by the speaker, represented as a one-hot encoded vector. The output of the listener's network is a probability distribution over the stimulus set, and the listener produces a guess by sampling from this distribution. For the reinforcement learning parts of NIL we use the classical algorithm REINFORCE (Williams, 1992). For the transmission phase we sample 300 color-name pairs with replacement, out of the 330 chips in the entire stimulus set; this ensures that the new generation will have seen examples from most of color space but it is impossible for them to have seen all color-name pairs. To optimize the neural networks, we use the optimizer Adam (Kingma and Ba, 2015), both in the learning and interaction phase, with learning rate 0.005 and batch size 50. For each phase in the NIL algorithm, we take 1,000 gradient steps. We stop the NIL algorithm either after 250 generations or once the maximum difference in IB complexity and accuracy over the ten latest generations is smaller than 0.1 bit, that is, when the last

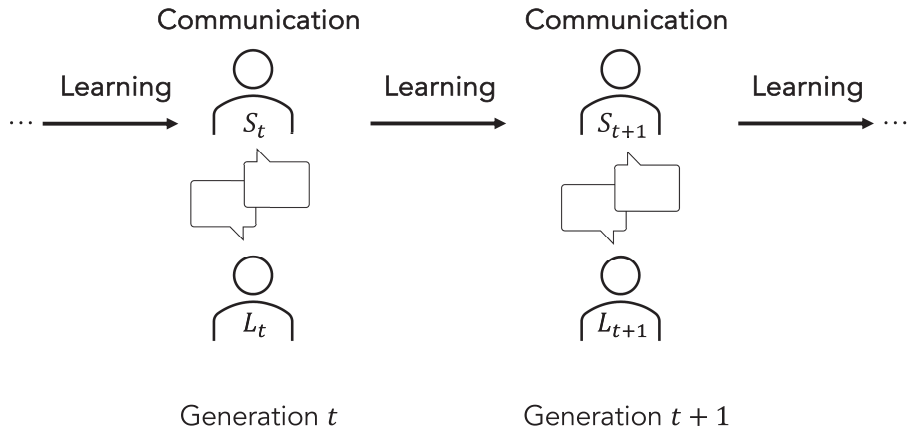


**Figure 3.** Illustration of the neural iterated learning (NIL) algorithm (Ren et al. 2020). The algorithm alternates between communication within a generation, and learning that is iterated across generations. The speaker ($S$) in each generation learns from the speaker in the previous generation, and communicates with the listener ($L$) in their own generation.

**Algorithm 1** Neural Iterated Learning

Initialize dataset $D_1$ uniformly at random
**for** $t = 1...$ **do**
  ***Learning Phase***
  Randomly initialize $S_t$ and $L_t$.
  Train $S_t$ on $D_t$ using stochastic gradient descent and cross-entropy loss.
  Play signaling game between $S_t$ and $L_t$ and update parameters of only $L_t$ using the rewards.
  ***Interaction Phase***
  Play signaling game between $S_t$ and $L_t$ and update parameters of **both** agents using the rewards.
  ***Transmission Phase***
  Create transmission dataset $D_{t+1}$ consisting of color-name pairs, $(c, w)$ by sampling colors from the prior $p(c)$ and providing them as input to $S_t$.
**end for**

ten generations are all within a small region of the IB plane. Note that NIL is not guaranteed to converge in the IB plane and might oscillate back and forth. This is because the transmission dataset is finite and randomly sampled, so the next generation might only be able to approximately reconstruct the naming system of the previous generation.

The reward function: The reward function of Kågebäck et al. 2020, which we use here, takes the form:

$$r(c, \hat{c}) = e^{-\gamma ||x_c - x_{\hat{c}}||_2^2} \quad (2)$$

where $c$ is the chip sampled by the speaker, $\hat{c}$ is the chip chosen by the listener as their interpretation of the chip intended by the speaker, $x_c$ is the location in CIELAB space of chip $c$, and $\gamma$ is a parameter that controls how precise the listener's choice $\hat{c}$ has to be. As $\gamma \to \infty$ the above reduces to a binary reward function, that is, the listener has to perfectly reconstruct the color to get any reward. On the other hand, if $\gamma = 0$ the reward function is vacuous in the sense that any possible reconstruction yields a reward of 1. We use $\gamma = 0.001$ which was originally used by Kågebäck et al. 2020 and motivated by the analysis by Regier et al. (2007).

## 4. Analyses and results

### 4.1 Iterated learning and communication operating together

For each vocabulary size $K = 3, \ldots, 10$ and $K = 100$, we ran 100 independent instances of the NIL algorithm. For each instance, we considered the color naming system of the last speaker to be the result of that instance—we call these systems IL+C, as they are the result of iterated learning plus communication, and we evaluated the IL+C systems in the IB framework. As can be seen in Fig. 4 (top panel), the IL+C systems are highly efficient in the IB sense: they lie near the theoretical efficiency limit (median inefficiency $\varepsilon = 0.07$), and they are no less efficient than the random Gaussian systems we considered above (median inefficiency $\varepsilon = 0.09$), which in turn are more efficient than the human systems of the WCS (see Section 2). Thus, iterated learning plus communication as formalized in the NIL algorithm leads to semantic systems that are efficient in the IB sense. This is consistent with existing proposals: the reward during the signaling game favors informativeness (higher reward for similar colors, following Kågebäck et al.

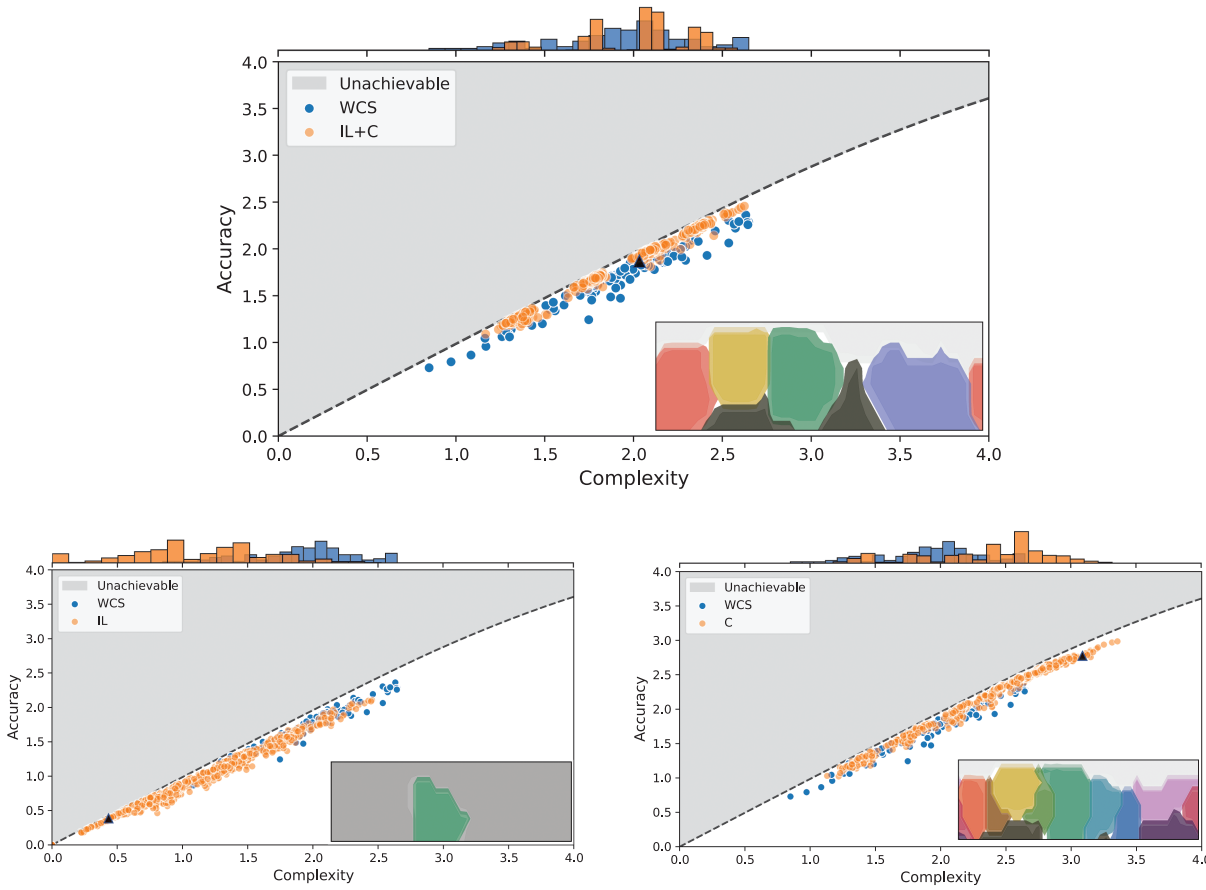


**Figure 4.** Efficiency of the (top) IL+C, (bottom left) IL, and (bottom right) C evolved color naming systems (orange dots), in each case compared with the natural systems of the WCS (blue dots). The black triangle indicates the end state of one run, shown in the inset color map. The histograms above each figure indicate the proportion of systems at the corresponding complexity level.

2020), and it has been argued that iterated learning favors simplicity (Kirby et al. 2015; Carr et al. 2020). Interestingly, all the resulting systems lie within the complexity range of the WCS systems even though NIL could theoretically produce much more complex systems, especially when initialized with $K = 100$.

Xu et al. (2013) examined how color naming systems evolved through chains of iterated human learners without within-generation communication, but with the number of categories constrained. They found that these lab-evolved systems tended to gravitate toward color naming systems that were similar to those of the WCS, and we wished to know whether the same was true of computational agents in the NIL framework. For each IL+C system, we determined the dissimilarity (gNID) between that system and the most similar (lowest gNID) WCS system. We also determined the analogous quantity (dissimilarity to the most similar WCS system) for each random Gaussian system. Figure 5 shows that IL+C systems tend to be similar to WCS systems to a greater extent than Gaussian systems do, and this was confirmed by a one-sided Mann–Whitney $U$ test ($P \ll .001$). Thus, the NIL process tends to gravitate toward human (WCS) systems to a greater extent than a random but efficient baseline, the Gaussian systems (We found that 14% of the IL+C experiments ran for the maximum number of generations without converging in the IB plane. Excluding these systems from the analysis and only considering the IL+C runs that did converge does not change the qualitative outcome of the analysis above.).

We also asked whether NIL would transform efficient systems that were dissimilar to those of the WCS (namely those of Gaussian[D]) into comparably efficient systems that were more similar to the WCS. To test this, we initialized the NIL algorithm with a

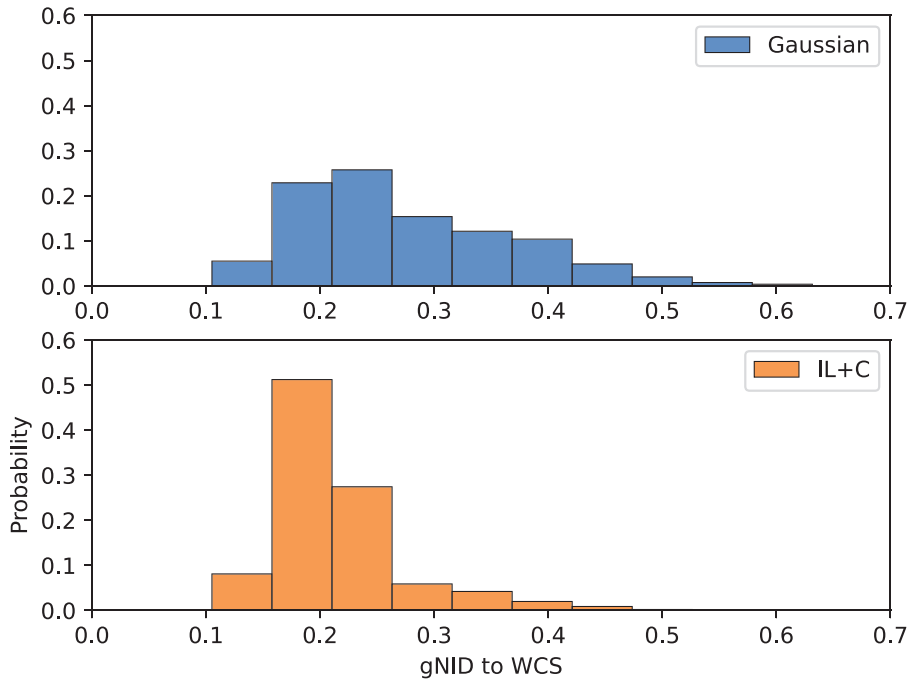


**Figure 5.** Distribution of dissimilarity to WCS systems (minimum gNID to any WCS system) shown for IL+C and Gaussian systems. The Gaussian systems include both Gaussian[S] and Gaussian[D]. Evolved IL+C systems tend to be more similar to attested WCS systems than are random but highly efficient Gaussian systems.

Gaussian[D] system, ran the NIL algorithm, and compared the initial system to the one that resulted from NIL. Figure 6 illustrates the beginning and end points of this process for a small set of systems, and shows that NIL transforms systems that are efficient but unlike the WCS into systems that are similar to particular WCS systems. Figure 7 shows that the same general pattern also holds over Gaussian[D] systems taken as a whole. For each Gaussian[D] system, we created an NIL chain, and initialized the chain with that Gaussian[D] system. For each such NIL chain, we measured the dissimilarity (gNID) of its initial Gaussian[D] system to the most similar WCS system, and the gNID of the end result of NIL to its most similar WCS system. We found that NIL tends to transform Gaussian[D] systems into systems that are more similar to the human systems of the WCS. The mean gNID to WCS was 0.38 before NIL and 0.25 after, and the reduction in dissimilarity to WCS after applying NIL was significant (one-sided (paired) Wilcoxon signed-rank test, $n = 302$, $T = 1,113$, $P \ll .001$). The median inefficiency of Gaussian[D] systems is $\varepsilon = 0.09$ and the median inefficiency of the results of NIL is slightly lower at $\varepsilon = 0.07$, meaning that NIL made the already-efficient Gaussian[D] systems slightly more efficient (one-sided (paired) Wilcoxon signed-rank test, $n = 302$, $T = 7,716$, $P \ll .001$). Thus, NIL moves already-efficient systems closer to the attested systems of the WCS, while maintaining and even slightly improving efficiency. Finally, it is noteworthy that NIL with three terms converges to a system that is similar to a three-term WCS system (see the top row of Fig. 6), because three-term systems are the one case in which IB optimal systems qualitatively diverge from human data (Zaslavsky et al. 2018: 7941; see also Fig. 8 and accompanying text). Thus, this is a case in which NIL appears to provide a better qualitative fit to the data than IB does.

## 4.2 Iterated learning alone and communication alone

So far, we have seen evidence that the Kirby et al. (2015) model of cultural evolution, as implemented in the NIL algorithm, may provide a plausible model of the cultural evolutionary process by which human color naming systems become efficient. We have referred to the result of the full NIL algorithm as IL+C systems, because these systems result from both iterated learning (IL) and communication (C). This raises the question whether iterating learning alone,

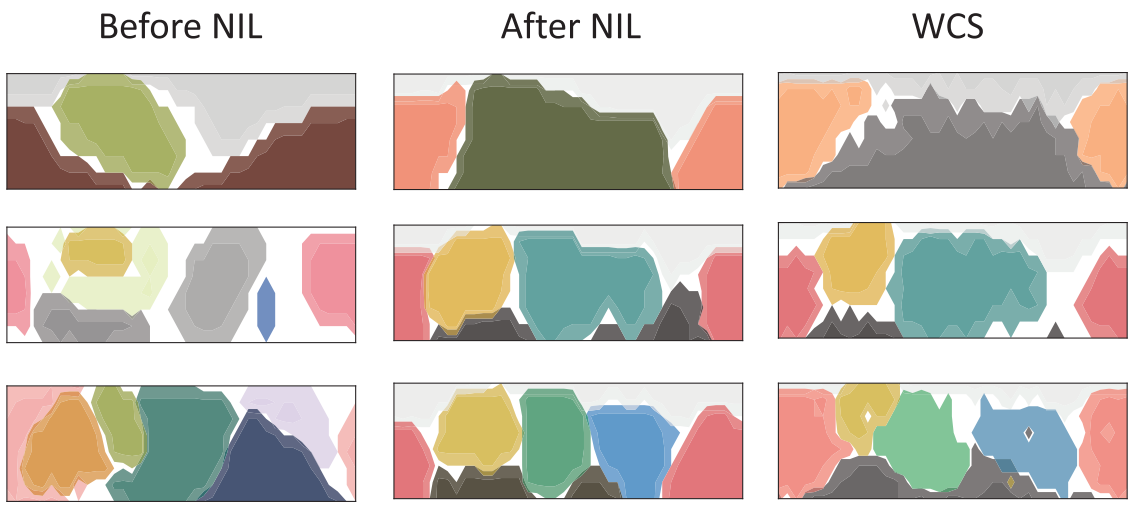


**Figure 6.** NIL transforms efficient color naming systems to become more similar to the WCS. In each row, the left column shows a Gaussian[D] system that was used to initialize NIL, the middle column shows the result of running NIL from that initialization state, and the right column shows a WCS system (from top to bottom: Bété, Colorado, Dyimini) that is similar to the NIL result.

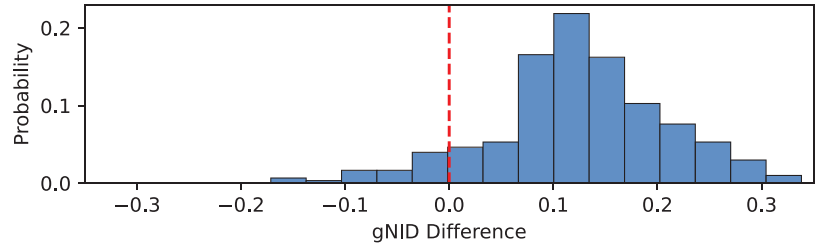


**Figure 7.** NIL tends to transform efficient Gaussian[D] color naming systems to become more similar to the WCS. The difference score is dissimilarity to WCS (minimum gNID to any WCS system) before NIL, minus the same quantity after NIL. Values above zero (marked by the dashed vertical red line) indicate that NIL has brought a system closer to the systems of the WCS. There is a clear trend toward positive values, indicating that NIL tends to transform already-efficient systems into systems that are more human like.

or communication alone, would yield comparable results.

To find out, following Kirby et al. (2015), we ran two variants of this culural evolutionary algorithm. One variant included only iterated learning but no communication (i.e., lines 6–8 of Algorithm 1 were omitted). The other variant included communication but no iterated learning (i.e., there was only one pass through the main loop, which stopped at line 9); this is exactly the experiment that was performed by Kågebäck et al. (2020). We refer to the results of the iterated-learning-only algorithm as IL (for iterated learning), and the results of the communication-only algorithm as C (for communication). For the C experiments, we trained each dyad of agents for at most 250,000 batches but stopped the training once the agents satisfied the stopping criterion used for IL+C. Note that Kågebäck et al. (2020) only trained each dyad for 50,000 steps without any early stopping criterion. We found that 99.6% of the C experiments converged before reaching the maximum number of batches. All the IL experiments converged in the IB plane before reaching 250 generations.

Comparison of the three panels of Fig. 4 reveals that there are qualitative differences in the profiles of the systems produced by the 3 variants of the NIL algorithm (IL+C, IL, and C). We have already seen that IL+C systems (top panel) are both efficient and similar to human systems; we also note that they lie within roughly the same complexity range as the

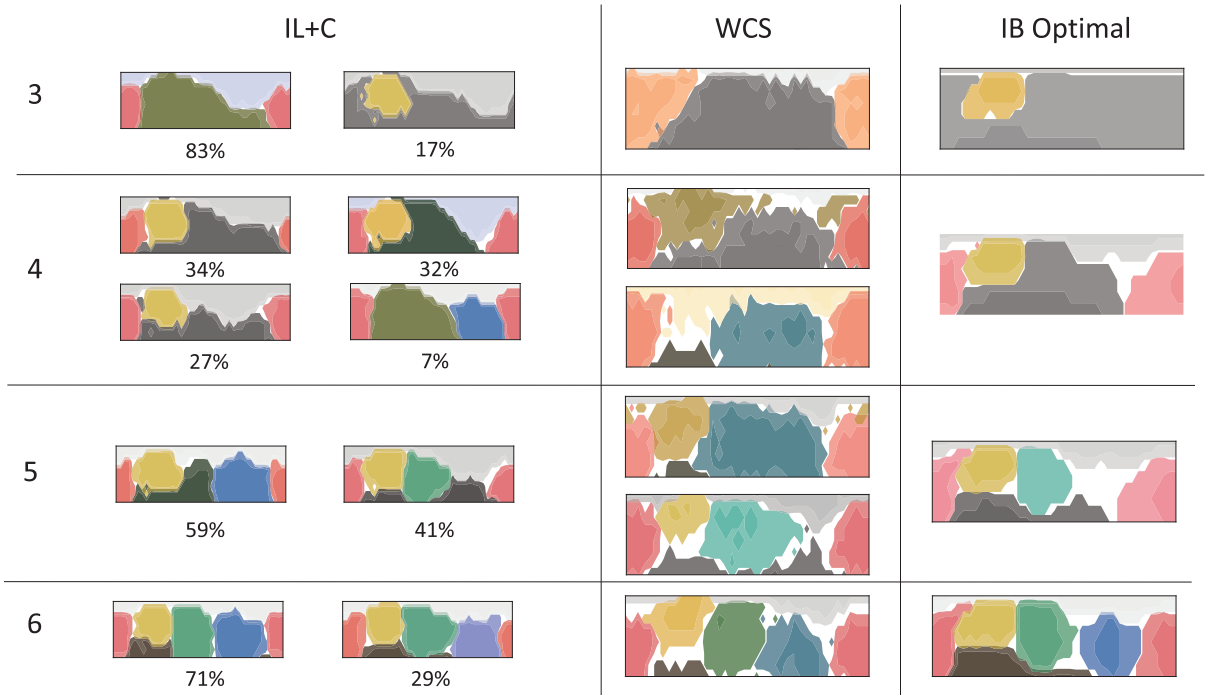


**Figure 8.** Representative IL+C systems (left column), WCS systems (middle column) and IB optimal systems (right column), with 3, 4, 5, and 6 color terms (rows). The % under each IL+C system indicates the percentage of IL+C systems in the corresponding cluster. The WCS systems are, from top to bottom: Nafaanra (iso: nfr, Ghana), Culina (iso: cul, Peru, Brazil), Waorani (iso: auc, Ecuador), Jicaque (iso: jic, Honduras), Berik (iso: bkl, Indonesia), and Kalam (iso: kmh, Papua New Guinea).

human systems of the WCS. In contrast, the IL systems (bottom left panel) skew toward lower complexity than is seen in human systems, and in fact about 6% of the IL systems lie at the degenerate point $(0, 0)$ in the IB plane, at which there is a single category covering the entire color domain. This skew toward simplicity is compatible with the view (Kirby et al. 2015; Carr et al. 2020) that iterated learning provides a bias toward simplicity. At the same time, the IL systems are not only simple but also quite efficient (i.e., informative for their level of complexity), which is in turn compatible with Carstensen et al.'s (2015) claim that iterated learning can produce informativeness, and with Carr et al.'s (2020) proposal that a process that primarily drives toward simplicity can sometimes also result in greater informativeness. Finally, the C systems (bottom right panel) show the opposite pattern: a bias toward higher informativeness, at the price of higher complexity, extending well above the complexity range observed in the human systems of the WCS.

Taken together, these results suggest that iterated learning alone over-emphasizes simplicity, communication alone over-emphasizes informativeness, and iterated learning with communication provides a balance between the two that aligns reasonably well with what is observed in human color naming systems. Overall, these results suggest that iterated learning plus communication is a more plausible model of the cultural evolutionary process that leads to efficient human color naming systems than is either iterated learning alone, or communication alone. These findings echo those of Kirby et al. (2015), who found that compositional structure evolved in a communicative system only under the combination of iterated learning and within-generation communication, and not under either process taken alone.

### 4.3 The distribution of systems produced by IL+C

To further explore the distribution of systems produced by IL+C we grouped all IL+C systems from the main experiment based on the number of color terms, $K$, in the systems. For each number of color terms, we clustered the systems using spectral clustering (von Luxburg, 2007) with gNID as the

dissimilarity measure. To find the appropriate number of clusters for each number of color terms, we performed spectral clustering with $C = 2, 3, 4$ clusters and reported the clustering with the highest silhouette score (Rousseeuw, 1987) which is standard in clustering. Since spectral clustering does not return cluster centers, we take the system that minimizes the average pairwise gNID to all other systems in the cluster as a representative sample of that cluster. The resulting systems, for $K = 3,\ldots,6$, are presented in Fig. 8 along with some WCS systems and the optimal IB systems. The number under each representative IL+C system indicates the percentage of systems contained in the corresponding cluster.

Interestingly, we see that the IL+C systems with three color terms appear in two clusters: a larger cluster that corresponds reasonably well to 3-term systems observed in the WCS, and a smaller cluster that is similar to the unattested IB optimal system. This suggests that there are two different optima that IL+C converges to: one human-like and the other corresponding to the IB optimal solution. The fact that the cluster corresponding to the IB solution is much smaller suggests that IL+C has a bias toward systems that are more similar to the WCS systems. These results are compatible with the idea that the attested 3-term systems represent a local optimum that is easier to reach through a process of cultural evolution than is the IB optimal solution. Related ideas have also been proposed in connection with kin terminologies, Epling et al. (1973), Kemp and Regier (2012: 1054).

For the four term systems, we observe that 93% of the IL+C systems end up in clusters that correspond fairly well with the optimal IB system and one of the WCS systems shown in Fig. 8. The last 7% of the systems end up in a cluster that does not map clearly onto the WCS data. For both $K = 5$ and $K = 6$, we observe that at least one of the IL+C clusters seems to correspond fairly well with systems in the WCS and with IB optimal systems.

### 4.4 Learnability and convexity

As mentioned above in our review of relevant literature, an influential idea holds that human categories form convex regions in a given conceptual space (Gärdenfors, 2000). In the case of color, a natural space for testing this claim is CIELAB space (Fig. 1, top right panel), and Jäger (2010) has shown that the natural color categories found in the WCS are convex sets in CIELAB space—supporting the convexity claim of Gärdenfors (2000) in the domain of color. More recently, Steinert-Threlkeld and Szymanik (2020) have extended this line of thought by arguing that convex color categories are easier to learn than are non-convex ones, and that this greater learnability helps to explain why human color categories tend to be convex.

We sought to situate this argument relative to the one we have been advancing here. Intuitively, it seems plausible that the artificial Gaussian systems we have considered above should also be convex, because they are based on spherical Gaussian-shaped kernels—but as we have seen, many of these Gaussian systems are quite dissimilar to the human systems of the WCS. This suggests that convexity may be a necessary but not sufficient criterion for characterizing human-like semantic categories, a suggestion with which proponents of the convexity argument are comfortable (P. Gärdenfors, G. Jäger, personal communication; see also Gärdenfors (2024)). To probe this possibility further, we assessed the convexity, the (non-iterated) learnability, and the efficiency of the WCS systems, the randomly generated Gaussian systems, and an additional set of baseline systems that draw category distinctions based only on hue. These hue-based systems were designed to be convex but not similar to human systems. Specifically, for vocabulary sizes $K = 3, \ldots, 10$ we divided the Munsell chart into equally sized categories by grouping together color chips based on their hue only; in case equally sized categories were not possible we created $K - 1$ equally sized categories and added the remaining color chips to the last category. Example hue-based systems are shown in Fig. 9: these are deterministic systems in which hue column fully determines the category to which a given chip belongs.

To assess the convexity of a color naming system, we adopted the measure of Steinert-Threlkeld and Szymanik (2020). They took the degree of convexity of a single category, named by a word $w$, to be:

$$\mathrm{dcc}(w) := \frac{|w|}{|\mathrm{ConvHull}(w)|}$$

where $|\cdot|$ is the size of a set, that is, the number of color chips in that set, and $\mathrm{ConvHull}(w)$ is the convex hull, in CIELAB space, of those chips in category $w$. Thus, $\mathrm{dcc}(w)$ gives us the proportion of those chips in the convex hull of category $w$ that are also in the category $w$ itself. For a perfectly convex category, this proportion will be 1. Steinert- Threlkeld and Szymanik (2020) then defined the degree of convexity of an entire system $S$ of categories to be the average, weighted by category size, of $\mathrm{dcc}(w)$ across categories $w$ in $S$:

$$\mathrm{dc}(S) := \frac{\sum_{w \in S} |w| \cdot \mathrm{dcc}(w)}{\sum_{w \in S} |w|}$$

**Figure 9.** Hue-based artificial systems, with 3 (left) and 10 (right) categories.

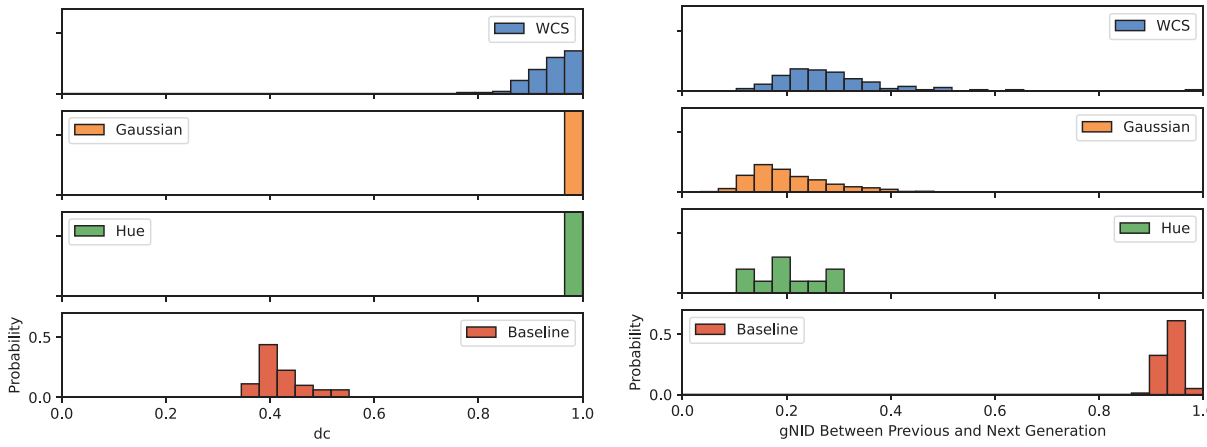


**Figure 10. Left panel: Convexity.** Convexity for different types of category systems. The natural systems of the WCS, artificial Gaussian systems, and artificial hue-based systems, are all highly convex when compared with a baseline of randomly generated systems in which each color chip is assigned to a category selected uniformly at random (labeled "Baseline"). We generated such baseline random systems with $k = 3, \ldots, 10$ color categories and for each $k$ we drew 10 random systems. **Right panel: Learnability.** Ease of learning is assessed by how well a learner generalizes, and generalization is measured by gNID between a learned system and the system from which training data was drawn. Artificial Gaussian and hue-based systems show generalization that is no worse than that of natural WCS systems.

A dc($S$) value of 1 corresponds to a system of perfectly convex color categories (This method assumes deterministic rather than probabilistic category membership. When applying this method to probabilistic systems, we first converted the probabilistic system to a deterministic one by assigning each chip to the modal category for that chip; we then applied this convexity measure to the resulting deterministic system.).

To assess the (non-iterated) learnability of a color naming system, we took a system to be easily learned to the extent that a neural network learner *generalizes* the system well—that is, to the extent that the learned system matches the one from which training data was sampled. We assessed this by considering only the learning phase of the NIL algorithm, and considering only the speaker's learning (specifically lines 3–5 of Algorithm 1), leaving all parameters unchanged. We then measured the gNID between the learned system and the system from which training data was drawn. During training, the agent sees only part of the entire system, so this gNID is a measure of how well the agent generalizes from the data it receives. To mitigate possible effects caused by sampling the training dataset, we performed each experiment over 10 independent runs and averaged.

We assessed the convexity, the learnability, and the IB efficiency of the (natural) WCS, (artificial) Gaussian, and (artificial) hue-based systems. Convexity results are shown in Fig. 10 (left panel), and learnability results are shown in Fig. 10 (right panel). All three types of system are highly convex, with the artificial Gaussian and hue-based systems being slightly more convex than the natural WCS systems—perhaps because the natural systems include noise. Moreover, in line with the expectation that convex systems will be learnable, all three types of system show good generalization, with no advantage for the natural WCS systems over the artificial Gaussian and hue-based systems. These results confirm that convex systems tend to be highly learnable, and also highlight that something beyond convexity and (non-iterated) learnability must play a role in differentiating human systems from artificial semantic systems that do not resemble them. Finally, Fig. 11 shows that artificial hue-based systems are not especially efficient—in contrast with artificial Gaussian systems and natural WCS systems. We take these results

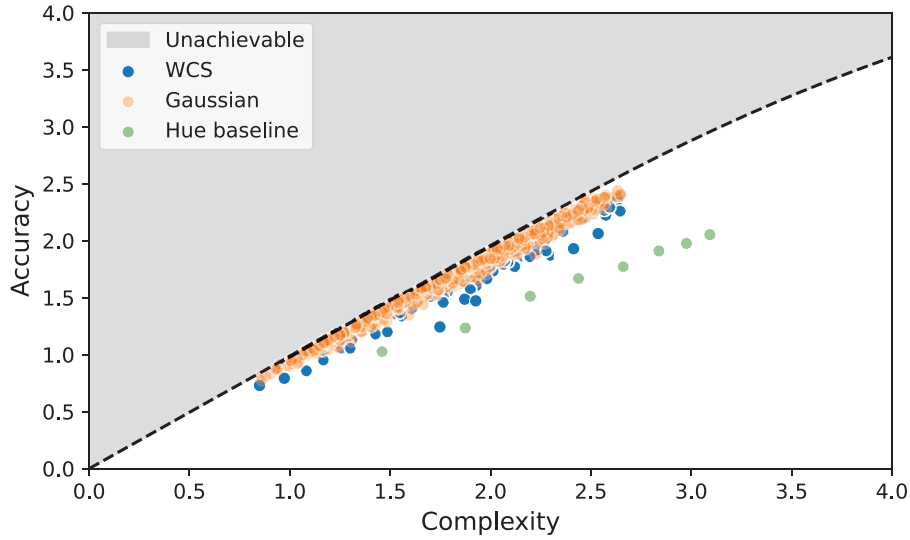


**Figure 11.** Some convex and learnable category systems are not efficient. Efficiency of the artificial hue-based systems (green dots), compared with that of the artificial Gaussian (orange dots) and natural WCS (blue dots) systems.

to suggest that convexity and learnability provide a partial answer to the question of what characterizes human semantic categories—and that a fuller answer may be provided by iterated learning and communication operating together, as a model of cultural evolution that leads toward efficient and human-like systems of semantic categories.

## 5. Discussion

We have shown (1) that there exists a reasonably sized class of color naming systems that are highly efficient in the IB sense but dissimilar from human systems; (2) that iterated learning plus communication, as captured in the NIL algorithm, leads to color naming systems that are both efficient in the IB sense and similar to human systems; and (3) that iterated learning alone, communication alone, and convexity alone, do not yield that result as clearly. These findings help to answer some questions, and also open up others.

As we have noted, the existence of highly efficient systems that do not align with human ones is not in itself surprising. IB is a non-convex optimization problem (Tishby et al. 1999; Zaslavsky et al. 2018), so multiple optima and near-optima are to be expected. However we feel that our identification of such systems may nonetheless be helpful, because it highlights just how many such systems exist, and just how dissimilar from human systems they sometimes are. In particular, this helps to make sense of Chaabouni et al.'s (2021) finding that simulations of cultural evolution can lead to color naming systems that exhibit high IB efficiency but deviate to some extent from human systems—something that we also sometimes find, as seen above in Fig. 8. This in turn highlights the importance of identifying cultural evolutionary processes that tend to avoid these outcomes and instead converge toward systems we find in human languages.

We have argued that iterated learning plus communication, as proposed by Kirby et al. (2015) and implemented in the NIL algorithm (Ren et al. 2020), is such a process, and that it provides a better account of cross-language color naming data than either iterated learning alone, or communication alone. Our findings supporting this idea thus generalize Kirby et al.'s (2015) argument, which concerned compositionality in language, to a different aspect of language. Our findings also confirm a proposed resolution to a tension in the literature. As we have noted, Carstensen et al. (2015) argued that iterated learning alone can lead to informative semantic systems, whereas Carr et al. (2020) argued that iterated learning provides a bias for simplicity, and communication provides a bias for informativeness. However, Carr et al. (2020) also found that a bias for simplicity—such as that provided through iterated learning—can produce systems that are informative as well as simple, and they suggested that this helps to resolve the tension. Specifically, they suggested

that an increase in informativeness through iterated learning, as observed by Carstensen et al. (2015), can result from a process (iterated learning) the primary outcome of which is a drive toward simplicity. Our finding that both forces are needed to account for the data aligns with Carr et al.'s (2020) central position. In addition, our finding that iterated learning alone also converges to efficient and thus informative systems—although often to overly simple ones—qualitatively replicates the findings of Carstensen et al. (2015), in a way that confirms Carr et al.'s (2020) proposed resolution of the tension: iterated learning does lead to simplicity, as suggested, but it also leads to informativeness to some extent.

It is natural to think of cultural evolution as a mean by which the abstract computational goal of optimal efficiency might be attained or approximated. The optimally efficient color naming systems on the IB curve closely resemble those in human languages (Zaslavsky et al. 2018), and the IL+C systems are likewise highly efficient and similar to those in human languages. However, as noted above, there is an exception to this pattern. In the case of three-term systems, the IB optimal system qualitatively differs from the color naming patterns found in the WCS (Zaslavsky et al. 2018: 7941), whereas IL+C systems often qualitatively match attested 3-term systems (recall the top rows of Figs 6 and 8). Thus, in this one case, it appears that human languages do not attain the optimal solution or something similar to it, and instead attain a somewhat different near-optimal solution that is apparently more easily reached by a process of cultural evolution.

A major question left open by our findings is exactly why we obtain the results we do. The general model of Kirby et al. (2015), as implemented in the NIL algorithm, is just one possible cultural evolutionary process, and we have seen that the process accounts for existing data reasonably well. It makes sense intuitively that NIL strikes a balance between the simplicity bias of iterated learning and the informativeness bias of communication—but what is still missing is a finer-grained sense for exactly which features of this detailed process are critical, vs. replaceable by others, and what the broader class of such processes is that would account well for the data (Tucker et al. 2022). A related direction for future research concerns the fact that the evolutionary process we have explored here is somewhat abstract and idealized, in that agents communicate with little context or pragmatic inference. Actual linguistic communication is highly context-dependent, and supported by rich pragmatic inference—it seems important to understand whether our results would still hold in a more realistic and richer environment for learning and interaction. Our agents also have designated roles: an agent acts either as a speaker or as a listener, and a direction for future research is to extend our setting to a more realistic model in which agents can alternate between the two roles. In addition, in our idealized setup a given agent interacts with only one other agent, whereas in human social interaction, communication within a generation happens in social networks such that an agent interacts with many other agents throughout their lifetime. An interesting direction for future research would be to explore what biases are introduced by certain population structures and whether varying the population structure can account for the variance observed in human color naming data.

Another important issue concerns the situating of this evolutionary account relative to the classic account of Berlin and Kay (1969). Our work here inherits, from the work of Zaslavsky et al. (2018) on which we build, an important connection to that earlier classic account: a trace along the IB curve reveals a sequence of color naming systems that gradually increase in complexity and that recapitulate the Berlin and Kay (1969) hierarchy, while also capturing aspects of competing accounts (MacLaury, 1997; Levinson, 2000). However, the mapping of that connection to fine-grained empirical data concerning language change over historical time has only recently begun (Zaslavsky et al. 2022), and a connection to the evolutionary model we explore here has not to our knowledge been attempted. Finally, we have focused on the semantic domain of color, but the ideas we have pursued are not specific to color, so another open question is the extent to which our results generalize to other semantic domains.

## Acknowledgements

We thank the two anonymous reviewers and the editor for their valuable comments on this paper. An earlier version of this paper appeared in the Proceedings of the 45th Annual Meeting of the Cognitive Science Society (2023). We thank Noga Zaslavsky and three anonymous reviewers for helpful comments on that earlier version of the paper. We also thank Gerhard Jäger and Peter Gärdenfors for helpful discussion of category convexity.

## Funding

EC was funded by Chalmers AI Research (CHAIR) and the Sweden-America Foundation (SweAm). Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. We

thank the library of Chalmers University of Technology for covering publication costs.

## Author contributions

E.C., D.D., and T.R. designed the research; E.C. performed the research; E.C. analyzed the data; and E.C., D.D., and T.R. wrote the paper.

## Data availability

The WCS data can be found in the WCS Data Archive https://www1.icsi.berkeley.edu/wcs/data.html. We used the original code of Zaslavsky et al. (2018), available at https://github.com/nogazs/ib-color-naming, to compute the IB objective, the inefficiency of the languages, and the gNID between languages. Our code is available here: https://github.com/e-carlsson/iterated-learning-color-naming

## References


Abbott, J. T., Griffiths, T. L., and Regier, T. (2016). *Focal Colors Across Languages Are Representative Members of Color Categories. Proceedings of the National Academy of Sciences*, 113, 11178–11183.

Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). Modeling the Emergence of Universality in Color Naming Patterns. *Proceedings of the National Academy of Sciences*, 107(6), 2403–2407.

Belpaeme, T., and Bleys, J. (2005). Explaining Universal Color Categories Through a Constrained Acquisition Process. *Adaptive Behavior*, 13(4), 293-310.

Berlin, B., and Kay, P. (1969). *Basic color term. their universality and evolution*. University of California Press; Berkeley; Los Angeles. (2010)

Carr, J. W., Smith, K., Culbertson, J., and Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, 202, 104289.

Carstensen, A., Xu, J., Smith, C., and Regier, T. (2015). Language evolution in the lab tends toward informative communication. In D. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 303–308). Austin, TX: Cognitive Science Society.

Chaabouni, R., Kharitonov, E., Dupoux, E., and Baroni, M. (2021). Communicating Artificial Neural Networks Develop efficient Color-Naming Systems. *Proceedings of the National Academy of Sciences*, 118, e2016569118.

Cook, R. S., Kay, P., and Regier, T. (2005). The World Color Survey Database: History and Use. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science*, pp. 223-241. Elsevier: Amsterdam.

Denić, M., Steinert-Threlkeld, S., and Szymanik, J. (2022). Indefinite Pronouns Optimize the Simplicity/ Informative -ness Trade-off. *Cognitive Science*, 46(5), e13142.

Denić, M., and Szymanik, J. (2024). Recursive Numeral Systems Optimize the Trade-Off Between Lexicon Size and Average Morphosyntactic Complexity. *Cognitive Science*, 48(3), e13424.

Dowman, M. (2007). Explaining Color Term Typology With an Evolutionary Model. *Cognitive Science*, 31(1), 99-132.

Epling, P., Kirk, J., and Boyd, J. P. (1973). Genetic Relations of Polynesian Sibling Terminologies. *American Anthropologist*, 75(5), 1596–1625.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. (2016). Learning to Communicate With Deep Multiagent Reinforcement Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2145–2153). Curran Associates Inc: Red Hook, NY, USA.

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press: Cambridge, MA.

Gärdenfors, P. (2024). Natural Concepts and the Economics of Cognition and Communication. *Philosophia*, 1–18. https://doi.org/10.1007/s11406-024-00734-4

Gentner, D., and Bowerman, M. (2009). Why Some Spatial Semantic Categories Are Harder to Learn Than Others: The Typological Prevalence Hypothesis. In *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, pp. 465–480. Psychology Press: Hove, UK.

Gyevnar, B., Dagan, G., Haley, C., Guo, S., and Mollica, F. (2022). Communicative Efficiency or Iconic Learning: Do Acquisition and Communicative Pressures Interact to Shape Colour-Naming Systems? *Entropy*, 24(11), 1542.

Havrylov, S., and Titov, I. (2017). Emergence of Language With Multi-Agent Games: Learning to Communicate With Sequences of Symbols. In I. Guyon et al., (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2146–2156). Curran Associates.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.

Imel, N., and Steinert-Threlkeld, S. (2022). Modal Semantic Universals Optimize the Simplicity/Informativeness Tradeoff. In *Proceedings of SALT 32 (Semantics and Linguistic Theory)*, pp. 227-248.

Jäger, G. (2010). Natural Color Categories Are Convex Sets. In M. Aloni, H. Bastiaanse, T. de Jager, & K. Schulz (Eds.), *Logic, Language and Meaning*, pp. 11–20. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jameson, K. A., and Komarova, N. (2009). Evolutionary Models of Color Categorization I: Population Categorization Systems Based on Normal and Dichromat Observers. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 26, 1414–23.

Kemp, C., Gaby, A., and Regier, T. (2019). Season Naming and the Local Environment. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Kemp, C., and Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, 336, 1049-54.

Kingma, D. P., and Ba, J. (2015). Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations*. San Diego: ICLR.

Kirby, S. (2001). Spontaneous Evolution of Linguistic Structure–An Iterated Learning Model of the Emergence of Regularity and Irregularity. *IEEE Transactions on Evolutionary Computation*, 5, 102–110.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and Communication in the Cultural Evolution of Linguistic Structure. *Cognition*, 141, 87–102.

Kågebäck, M., Carlsson, E., Dubhashi, D., and Sayeed, A. (2020). A Reinforcement-Learning Approach to Efficient Communication. *PLoS ONE*, 15(7), 1–26.

Lazaridou, A., Peysakhovich, A., and Baroni, M. (2017). Multi-Agent Cooperation and the Emergence of (Natural) Language. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. pp. 1–11. Toulon, France.

Levinson, S. C. (2000). Y´el`ı Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1), 3-55.

Levinson, S. C. (2012). Kinship and human thought. *Science*, 336(6084), 988-989.

MacLaury, R. E. (1997). Color and cognition in Mesoamerica: *Constructing categories as vantages*. University of Texas Press.

McKinney, W., et al. (2010). Data Structures for Statistical Computing in Python. In *Scipy* (Vol. 445, pp. 51–56).

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., and Kemp, C. (2021). The Forms and Meanings of Grammatical Markers Support Efficient Communication. *Proceedings of the National Academy of Sciences*, 118(49).

Mordatch, I., and Abbeel, P. (2018). Emergence of Grounded Compositional Language in Multi-Agent Populations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32) New Orleans, Louisiana, US: The Association for the Advancement of Artificial Intelligence.

Paszke, A., Gross, S., Massa, F., Lerer, A., et al. (2019). Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Regier, T., Kay, P., and Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4), 1436–1441.

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., and Kirby, S. (2020). Compositional Languages Emerge in a Neural Iterated Learning Model. In *International Conference on Learning Representations*.

Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyd (eds.), *Cognition and categorization*, pp. 27-48. Lawrence Erlbaum Associates: New York.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Scott-Phillips, T. C., and Kirby, S. (2010). Language Evolution in the Laboratory. *Trends in Cognitive Sciences*, 14(9), 411-417.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated Learning: A Framework for the Emergence of Language. *Artificial Life*, 9, 371-86.

Steels, L., and Belpaeme, T. (2005). Coordinating Perceptually Grounded Categories Through Language: A Case Study for Colour. *Behavioral and Brain Sciences*, 28(4), 469–488.

Steinert-Threlkeld, S., and Szymanik, J. (2020). Ease of Learning Explains Semantic Universals. *Cognition*, 195, 104076. Amsterdam, Netherlands: Elsevier.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The Information Bottleneck Method. In *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, pp. 368–377.

Tucker, M., Levy, R. P., Shah, J., and Zaslavsky, N. (2022). Trading off Utility, informativeness, and Complexity in Emergent Communication. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (eds.), *Advances in Neural Information Processing Systems*. Vol. 35, pp. 22214–22228. Curran Associates, Inc.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. von Luxburg, U. (2007). A Tutorial on Spectral Clustering. Statistics and Computing, 17(4), 395–416.

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), 3021.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3), 229–256.

Xu, J., Dowman, M., and Griffiths, T. L. (2013). Cultural Transmission Results in Convergence Towards Colour Term Universals. *Proceedings of the Royal Society B: Biological Sciences*, 280(1758), 20123073.

Xu, Y., Liu, E., and Regier, T. (2020). Numeral Systems Across Languages Support Efficient Communication: From Approximate Numerosity to Recursion. *Open Mind*, 4, 57–70.

Xu, Y., Regier, T., and Malt, B. C. (2016). Historical Semantic Chaining and Efficient Communication: The Case of Container Names. *Cognitive Science*, 40, 2081-2094.

Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., and Regier, T. (2022). The Evolution of Color Naming Reflects Pressure for Efficiency: Evidence from the Recent Past. *Journal of Language Evolution* 7(2), 184–199.

Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient Compression in Color Naming and Its Evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(31), 7937–7942.

Zaslavsky, N., Maldonado, M., and Culbertson, J. (2021). Let's Talk (efficiently) About Us: Person Systems Achieve Near-Optimal Compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Zaslavsky, N., Regier, T., Tishby, N., and Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 41. Montreal, Canada.

# Appendix

## The framework of Zaslavsky et al. (2018)

Zaslavsky et al. (2018) cast the notion of efficiency in terms of an independent information-theoretic principle, the Information Bottleneck (IB) principle (Tishby et al. 1999). In the framework of Zaslavsky et al. (2018), a semantic system is considered efficient to the extent that it achieves an optimal tradeoff between the complexity of a system, and the accuracy of communication that that system supports. These notions are grounded in the communicative scenario illustrated in Fig. A.1, in which a speaker attempts to communicate with a listener about referents in a given domain universe $U$, in our case the domain of color. Here, the speaker considers a specific target color $t \in U$ and holds it in mind in the form of a mental representation $m_t$, which is a probability distribution over color space (CIELAB; recall Fig. 1), centered at $t$. To communicate that mental representation, the speaker utters a word $w$, drawn from a language-specific probabilistic encoder $q(w|m_t)$ that maps from meanings $m_t$ to words $w$; this encoder $q(w|m_t)$ is the semantic system by which the speaker and listener communicate. The listener then produces, on the basis of the uttered word $w$, a mental representation $\hat{m}_w$ that is the listener's reconstruction of the speaker's original representation $m_t$. Casting this simple communicative scenario in terms of the IB principle results in formal definitions of four quantities that are central to the IB formalization of efficiency, and on which we rely in our work: *complexity, accuracy, ε*, and *gNID*.

The complexity of a semantic system $q$ is given by $I_q(M_t; W)$, that is, the mutual information between the speaker's mental representation $m_t$ and the word $w$ used to express it. The greater the complexity of the system, the more information the word $w$ carries about the speaker's mental representation $m_t$. The accuracy of a semantic system is given by $I_q(W; U)$, which can be shown to capture the similarity of the speaker's and listener's mental representations (see Zaslavsky et al. 2018: 7939). The core idea of efficiency in this framework is to obtain the greatest accuracy possible for a given level of complexity—that is, to communicate as precisely as possible for a given amount of information sent. An optimally efficient semantic system $q$ is thus one that minimizes the IB objective function:

$$\mathcal{F}_\beta[q] = I_q(M_t; W) - \beta I_q(W; U)$$

where $\beta \geq 0$ is a tradeoff parameter that controls the relative weight given to complexity and accuracy. Those systems $q^*$ that minimize this objective function for different values of $\beta$ yield the IB theoretical limit of efficiency; that is, these are the systems with the greatest possible accuracy for each level of complexity. Zaslavsky et al. (2018) showed that human color naming systems achieve near-optimal efficiency in the IB sense, and that fully IB-optimal systems often closely correspond to color naming systems in human languages.

In our analyses, we also make use of two other quantities from the framework of Zaslavsky et al. (2018). First, $\varepsilon_q$ measures the inefficiency of a semantic system, or its deviation from optimal efficiency, as described on p. 7939 of their article:

$$\varepsilon_q = \frac{1}{\beta}\left(\mathcal{F}_\beta[q] - \mathcal{F}^*_\beta\right)$$

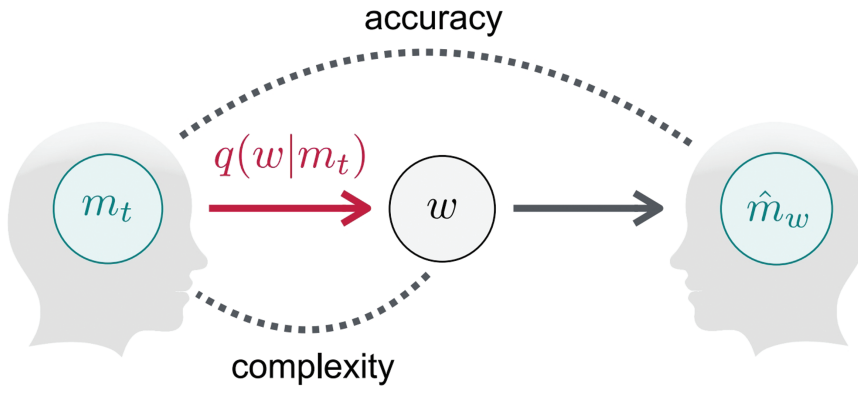


**Figure A.1.** The framework of Zaslavsky et al. (2018). A speaker communicates a specific referent to a listener by producing a word. The IB principle provides formal specifications of various quantities associated with this communicative act; see text for details. The figure is from Zaslavsky et al. (2021).

Here $\mathcal{F}_\beta^*$ is the optimal value of the IB objective for a given value of $\beta$, and $\beta$ is chosen to minimize the difference $\mathcal{F}_\beta[q] - \mathcal{F}_\beta^*$ for a given semantic system $q$. Finally, we follow Zaslavsky et al. (2018) in using their gNID measure to measure the dissimilarity between two semantic systems, as described on p. 7942 of their article. This measure assumes that a single meaning $m$ is assigned a name by each of two semantic systems $q_1$ and $q_2$: $W_1 \sim q_1(w_1|m)$ and $W_2 \sim q_2(w_2|m)$. Then the dissimilarity between $q_1$ and $q_2$ is given by

$$\mathrm{gNID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\left\{I(W_1; W_1'), I(W_2; W_2')\right\}}.$$

Here, $W_i'$ corresponds to another independent speaker using the system $q_i$.