Iterated learning and communication jointly explain efficient color naming systems

Emil Carlsson (caremil@chalmers.se)
Devdatt Dubhashi (dubhashi@chalmers.se)
Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics, UC Berkeley, Berkeley, CA, USA

Abstract

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 iterated learning alone, and communication alone, do not yield the same outcome as clearly.

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

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 (e.g. 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 & Regier, 2012), container names (Y. Xu et al., 2016), names for seasons (Kemp et al., 2019), and numeral systems (Y. Xu et al., 2020). Zaslavsky et al. (2018) gave this proposal a firm theoretical foundation by grounding it in an independent information-theoretic principle of efficiency, the Information Bottleneck (IB) principle (Tishby et al., 1999); they also showed that color naming systems across languages are efficient in the IB sense, that optimally IB-efficient systems resemble those found in human languages, and that 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 (IL; e.g. Scott-Phillips & Kirby, 2010). In IL, 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. J. 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 explicitly address efficiency, Carstensen et al. (2015) drew that link explicitly: they reanalzyed the data of J. Xu et al. (2013) and showed that the color naming systems produced by IL 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) argued, on the basis of a Bayesian model of IL and experiments with human participants, that learning actually contributes simplicity rather than informativeness. Overall, there is support for the idea that IL can lead to efficient semantic systems, with continuing debate over how and why. There are also recent proposals that non-iterated learning – e.g. in the context of a dyad of communicating agents (e.g. Kågebäck et al., 2020; Chaabouni et al., 2021; Tucker et al., 2022), or in a single agent without communication (e.g. Steinert-Threlkeld & Szymanik, 2020; Gyevnar et al., 2022) – can explain efficient color naming systems. These recent contributions build on an important line of earlier work using agent-based simulations cast as evolutionary models, without explicitly addressing efficiency (e.g. Steels & Belpaeme, 2005; Belpaeme & Bleys, 2005; Dowman, 2007; Jameson & 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, p. 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

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

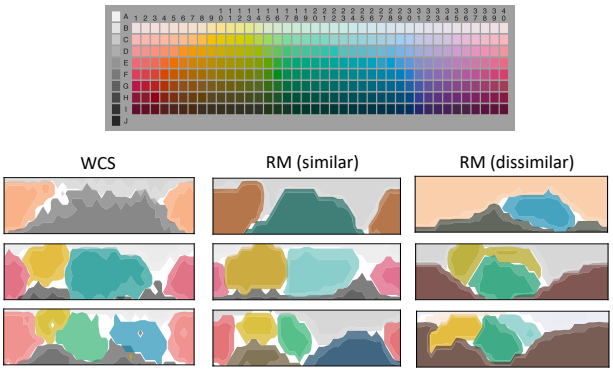

Figure 1: Top: Color naming stimulus grid. Bottom: 9 color naming systems displayed relative to this grid. The left column contains color naming systems from 3 languages in the WCS (from top to bottom: Bete, Colorado, Dyimini). Colored regions indicate category extensions, and the color code used for 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 systems of complexity comparable to that of the WCS system in the same row. The middle column shows random systems that are similar to the WCS system in the same row. The right column shows random 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 random system.

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 by other means. We explore that question here.

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 existence of 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 it 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 IL, instantiated in communicating neural networks, 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, and communication 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.

Not all efficient systems are human-like

We considered a class of artificial color naming systems related to one considered by Zaslavsky et al. (2022). In the class we consider, each named category w is modeled as a spherical Gaussian-shaped kernel with mean (prototype) x_w in 3-dimensional CIELAB color space, such that the distribution over words w given a color chip c 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 $K = 3\ldots10$ categories each, by first sampling η randomly from a uniform distribution over the interval $[0.001, 0.005]$ 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 at the top of Figure 1. In analyzing these systems, we drew on the following three quantities from the IB framework as presented by Zaslavsky et al. (2018): the complexity of a category system, gNID (a measure of dissimilarity between two category systems), and ε (a measure of inefficiency, or deviation from the theoretical limit of efficiency). We noted that the range of complexity (in the IB sense) for systems in the World Color Survey (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 systems as RM, for random model.

The lower panels of Figure 1 show that some of these RM systems are similar to, and others quite dissimilar to, natural systems in the WCS. 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, meaning that it is dissimilar to all WCS systems. Thus, there exist RM systems that are quite dissimilar to naturally occurring systems. To quantify this pattern, we separated the RM 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 (RM dissimilar) columns of Figure 1, which is 0.29. We then grouped all RM systems with gNID to the closest WCS system below this threshold into one group, RM_s (for similar to WCS), and the other RM systems into another group, RM_d (for dissimilar to WCS). 38% of the RM systems fell in RM_d and they spanned the complexity range $[0.86, 2.26]$. Thus, a substantial proportion of the RM systems are at least as dissimilar to WCS systems as are those in the right column of Figure 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 RM systems. It can be seen that all RM systems are highly efficient in the IB sense – i.e. 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 RM systems tend to exhibit greater efficiency (lower inefficiency ε) than do the WCS systems in the same complexity range ($P \ll .001$), and (2) that the RM_d systems, which are dissimilar to WCS systems, are also more efficient than WCS systems ($P \ll .001$, one-

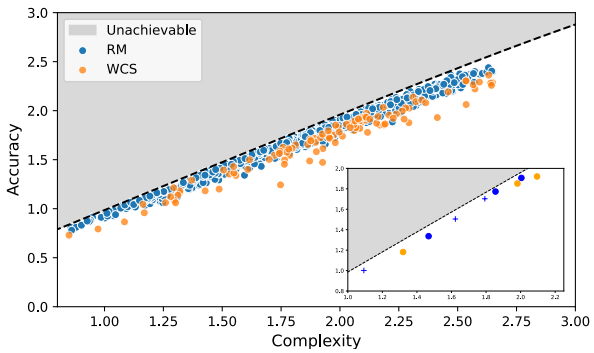

Figure 2: Efficiency of color naming, following Zaslavsky et al., 2018. The dashed line is the IB theoretical limit of efficiency for color naming, indicating the greatest possible accuracy for each level of complexity. The color naming systems of the WCS are shown in orange, replicating the findings of Zaslavsky et al., 2018. Our RM systems are shown in blue. It can be seen that the RM systems are often closer to the IB curve than the WCS systems are. The inset shows the 9 color systems of Figure 1, with the dissimilar random systems shown as +.

sided), and slightly to marginally more efficient than RM_s systems ($P = .019$ one-sided; Bonferroni corrections do not change the qualitative outcome). 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 sought an evolutionary process that would yield both efficiency in the IB sense, and similarity to human systems, without specifying IB optimization as a part of that process (cf. Tucker et al., 2022).

Iterated learning and communication

As noted above, iterated learning (IL; e.g. 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 IL (e.g. Levinson, 2012; Carstensen et al., 2015; Carr et al., 2020). However other work we have reviewed instead addresses cultural evolution through communication within a single generation (e.g. Kågebäck et al., 2020; Chaabouni et al., 2021; Tucker et al., 2022). We wished to explore the roles of both IL and communication, and so we adopted an approach that 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). In the NIL algorithm, artificial agents are implemented as neural networks that communicate with each other within a generation, and 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.[1]

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 Figure 1. A reward is given to the listener depending on how perceptually similar the selected chip is to the original color. (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). We represent both the speaker and listener as neural networks with one hidden layer consisting of 25 units with a sigmoidal activation function. Individual colors are represented in 3-dimensional CIELAB space when supplied as input to the speaker, and category names as one-hot encoded vectors. 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, 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 & 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 1000 gradient steps. We stop the NIL algorithm once the maximum difference in IB complexity and accuracy over the ten latest generations is smaller than 0.1 bit,

[1] 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.

Algorithm 1 Neural Iterated Learning

Initialize D_1 uniformly at random
for $t = 1\ldots$ **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

i.e. when the last ten generations are all within a small region of the IB plane. Algorithm 1 presents a schematic overview of the NIL algorithm, and Ren et al. (2020) present a detailed description. The hyperparameters were tuned empirically by studying which parameters yielded the highest reward in a small set of experiments. We found very little difference between different sizes of the network.

For each vocabulary size $K = 3\ldots10$ 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 Figure 3 (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 RM systems we considered above (median inefficiency $\varepsilon = 0.09$), which in turn are more efficient than the human systems of the WCS (see above). Thus, iterated learning plus communication as formalized in the NIL algorithm leads to semantic systems that are efficient in the IB sense. This is not entirely surprising: the reward during the signaling game favors informativeness (higher reward for similar colors, following Kågebäck et al., 2020), and it has been argued that learning favors simplicity (e.g. 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 $K = 100$.

J. Xu et al. (2013) showed that chains of iterated human learners 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 RM system. Figure 4 shows that IL+C systems tend to be similar to WCS systems to a greater extent than RM 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, RM.

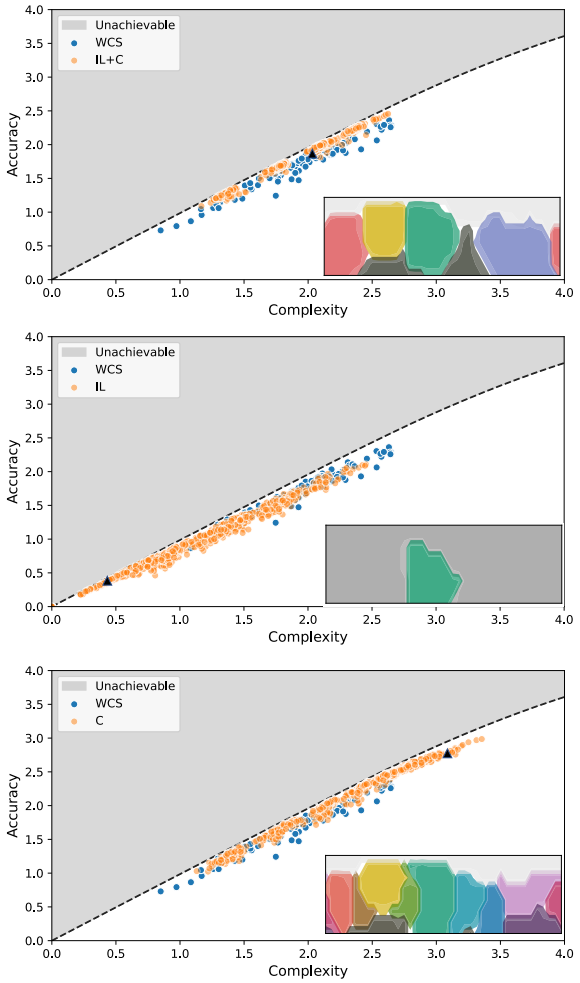

Figure 3: Efficiency of the (top) IL+C, (middle) IL, and (c) C evolved color naming systems, in each case compared with the natural systems of the WCS. The black triangle indicates the end state of one run, shown in the inset color map.

We also asked whether NIL would transform efficient systems that were dissimilar to those of the WCS (namely those of RM_d) into comparably efficient systems that were more similar to the WCS. To test this, we initialized the NIL algorithm with a system sampled from RM_d, ran the NIL algorithm, and compared the initial system to the one that resulted from NIL. Figure 5 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 6 shows the same general pattern but aggregated over all RM_d

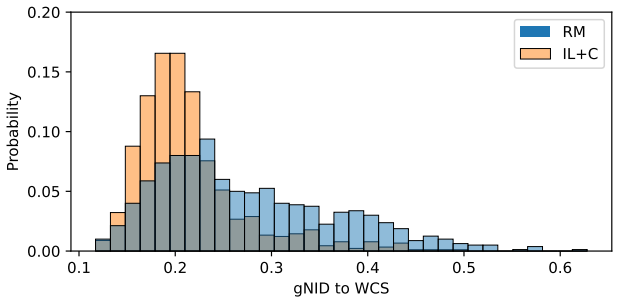

Figure 4: Distribution of dissimilarity to WCS systems (minimum gNID to any WCS system), shown for IL+C and RM systems. The RM systems include both $\mathrm{RM_s}$ and $\mathrm{RM_d}$.

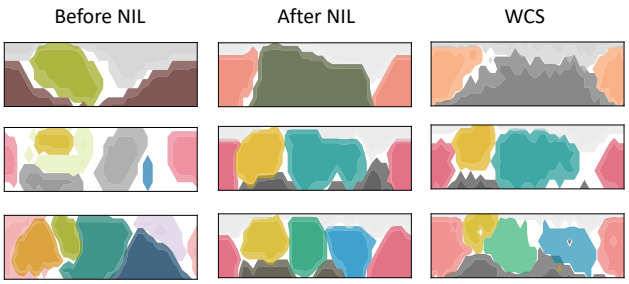

Figure 5: NIL transforms efficient color naming systems to become more similar to the WCS. In each row, the left column shows an $\mathrm{RM_d}$ system that was used to initialize NIL, the middle column shows the result of running NIL from that initialization state, and the right column shows a WCS system (from top to bottom: Bete, Colorado, Dyimini) that is similar to the NIL result.

systems. For each NIL chain initialized with an $\mathrm{RM_d}$ system, we measured the dissimilarity (gNID) of that initialized system to the most similar WCS system, and the gNID of the end result of NIL to its most similar WCS system. It can be seen that NIL transforms $\mathrm{RM_d}$ 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, $n = 302$, $T = 1113$, $P \ll .001$). The median inefficiency of $\mathrm{RM_d}$ 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 $\mathrm{RM_d}$ systems slightly more efficient (one-sided (paired) Wilcoxon signed-rank test, $n = 302$, $T = 7716$, $P \ll .001$). 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 3 terms converges to a system that is similar to a 3-term WCS system (see the top row of Figure 5), because 3-term systems are the one

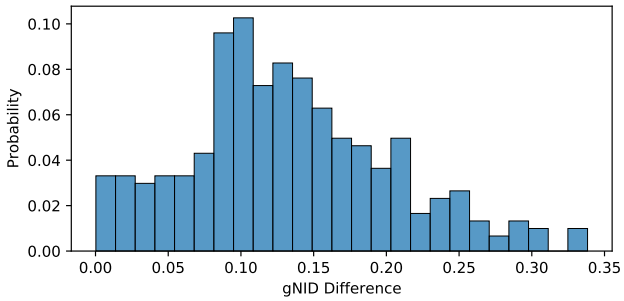

Figure 6: NIL transforms efficient $\mathrm{RM_d}$ color naming systems to become more similar to the WCS. The difference score is dissimilarity to WCS (minimum gNID to any WCS system) before NIL, minus the same quantity after NIL. A higher value indicates that NIL has moved the systems closer to the WCS. There are no values below 0, meaning that NIL never caused a system to become less similar to the WCS.

case in which IB optimal systems qualitatively diverge from human data (Zaslavsky et al., 2018, p. 7941). Thus, this is a case in which NIL appears to provide a better qualitative fit to the data than IB does (see also Regier et al., 2007, 2015).

Other possible evolutionary processes

So far, we have seen evidence that 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, or communication alone, would yield comparable results.

To find out, we ran two variants of the NIL 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). All other aspects of the algorithm were unchanged. 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).

Comparison of the three panels of Figure 3 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 human systems of the WCS. In contrast, the IL systems (middle 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 do-

main. This skew toward simplicity is compatible with Carr et al.'s (2020) claim 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 provides some bias toward informativeness. Finally, the C systems (bottom 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. We found that IL+C systems are slightly more efficient (mean $\varepsilon = 0.07 \pm 0.02$) than IL (mean $\varepsilon = 0.15 \pm 0.08$) or C (mean $\varepsilon = 0.11 \pm 0.04$) systems, where the $\pm$ indicates plus or minus one standard deviation. IL+C systems were also closer to the most similar WCS system (mean gNID = 0.21 ± 0.05) than were IL (mean gNID = 0.57 ± 0.17) or C (mean gNID = 0.27 ± 0.07) 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, as these ideas are formalized in the NIL algorithm.

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, and communication 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 — which 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. This in turn highlights the importance of identifying cultural evolutionary processes that avoid these local near-optima and instead converge toward systems we find in human languages.

We have argued that iterated learning plus communication, as cast in the NIL algorithm, is such a process, and that it provides a better account than either iterated learning alone, or communication alone. This idea, and our findings supporting it, may help to resolve a question 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 (see also Kirby et al., 2015 for a similar argument concerning linguistic form). Our finding that both forces are needed to account for the data aligns with Carr et al.'s (2020) claim. However our finding that learning alone also converges to efficient systems – although to overly simple ones – helps to make sense of Carstensen et al.'s (2015) findings.

It is natural to think of NIL, or any such process of cultural evolution, as a means by which the abstract computational goal of optimal efficiency might be approximated — and for the most part, that seems an accurate and useful way to frame the matter. 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, there is an important exception to this pattern. As noted above, in the case of 3-term systems, the IB optimal system qualitatively differs from the color naming patterns found in the WCS (Zaslavsky et al., 2018, p. 7941), whereas IL+C systems qualitatively match them (see e.g. the top row of Figure 5, middle and right panels). 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 possibility anticipated by Kemp and Regier (2012, p. 1054).

A major question left open by our findings is exactly why we obtain the results we do. NIL is just one possible evolutionary process, and we have seen that that 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 (Carr et al., 2020; Kirby et al., 2015) — 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 (e.g. Tucker et al., 2022). A related direction for future research concerns the fact that the evolutionary process we have explored 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. Finally, we have focused here on the 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.

Acknowledgments

We thank Noga Zaslavsky and 3 anonymous reviewers for helpful comments on an earlier version of this paper. Any errors are our own. Author contributions: EC, DD, and TR designed the research; EC performed the research; EC analyzed the data; and EC, DD, and TR wrote the paper. EC was funded by Chalmers AI Research (CHAIR) and the Sweden-America Foundation (SweAm). Computing resources used for the experiments were provided by the Swedish National Infrastructure for Computing (SNIC).

References

Baronchelli, A., Gong, T., Puglisi, A., & 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., & Bleys, J. (2005). Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior*, *13*(4), 293-310.

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

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.*

Chaabouni, R., Kharitonov, E., Dupoux, E., & 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., & Regier, T. (2005). The World Color Survey database: History and use. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (p. 223-241). Amsterdam: Elsevier.

Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, *31*(1), 99-132.

Gyevnar, B., Dagan, G., Haley, C., Guo, S., & Mollica, F. (2022). Communicative efficiency or iconic learning: Do acquisition and communicative pressures interact to shape colour-naming systems? *Entropy*, *24*(11). doi: 10.3390/e24111542

Jameson, K. A., & Komarova, N. (2009, 07). 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., & Regier, T. (2019). Season naming and the local environment. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.*

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049-54.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*.

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., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87-102.

Kågebäck, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020). A reinforcement-learning approach to efficient communication. *PLoS ONE*, *15*(7), 1–26.

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

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, *118*(49).

Regier, T., Kay, P., & 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.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The Handbook of Language Emergence*(January 2015), 237–263.

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & 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* (p. 27-48). New York: Lawrence Erlbaum Associates.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411-417.

Smith, K., Kirby, S., & Brighton, H. (2003, 02). Iterated learning: A framework for the emergence of language. *Artificial life*, *9*, 371-86.

Steels, L., & 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., & Szymanik, J. (2020). Ease of learning explains semantic universals. *Cognition*, *195*, 104076.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* (p. 368–377).

Tucker, M., Levy, R. P., Shah, J., & 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.*

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

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term univer-

sals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Xu, Y., Regier, T., & 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., & Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*.

Zaslavsky, N., Kemp, C., Regier, T., & 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., & 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., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual conference of the Cognitive Science Society.*