Information-Theoretic Principles in the Evolution of Semantic Systems

Thesis submitted for the degree
"Doctor of Philosophy"
by

Noga Zaslavsky

Submitted to the Senate of the Hebrew University of Jerusalem
August 2019

This work was carried out under the supervision of

Prof. Naftali Tishby

Acknowledgements

First and foremost, I am deeply grateful to my PhD advisor and academic parent, Naftali Tishby. My journey with Tali started when I was a shy undergraduate student. Ever since, Tali has been patiently guiding me through the latent beauty of information theory and statistics, and has given me the unique opportunity to join his fearless quest for the deepest principles of intelligence. With time, I learned to borrow a seed of his courage and vantage, and grow that into my own research. This thesis is a fruit of that process.

This thesis would not have taken its current form without Terry Regier, who in many ways has been like another advisor to me. One day in the midst of my graduate studies, I showed up at Terry's office with some thoughts for a project. That meeting later unfolded into a very productive international collaboration, spanning much of this thesis. I am deeply grateful to Terry for giving me the opportunity to work with him, and for hosting me for what turned out to be a long-term visit at UC Berkeley. I have grown as a researcher much thanks to his profound insight and perceptive guidance.

I have been fortunate to have Charles Kemp as a close collaborator. Remarkably, our productive collaboration started more than two years before we actually met in person. Charles' razor sharp scientific mind has never allowed me to overlook a loose point in an argument, and every discussion with him has intellectually enriched and strengthened my work.

I owe the third chapter of this thesis to Karee Garvin, who traveled to West Africa a couple of times to do fieldwork, collecting the extraordinary Nafaanra color naming data that lie at the heart of that chapter. It was a pleasure to collaborate with Karee, and I learned from her a lot about the fascinating ways linguists view the world. I am also grateful to Delwin Lindsey and Angela Brown for sharing with me their English color naming data, which accompanied me in multiple chapters and have added great value to this thesis.

My advisory committee members, Amir Globerson and Yonatan Loewenstein, as well as Yoram Burak, provided extremely helpful feedback along the way, and each one of them has been uniquely supportive and encouraging. I am very grateful to them for that.

During the course of my graduate studies, I have also had the privilege to work with several outstanding people to whom I owe special thanks. Idan Segev and Siwei Wang, with whom I collaborated at an early stage of my studies, have broadened my view of the role of information theory in neuroscience. Idan has also sparked my interest in Neural Computation long before I joined the Hebrew University. Inbal Arnon, with whom I started to collaborate toward the

end of my graduate studies, has provided valuable feedback on my work and added her unique perspective on language development. In between, I served as a TA in Shai Shalev-Shwartz's introduction to machine learning course. Working with Shai has deepened my understanding in machine learning. I also learned a lot from Shai about teaching.

I thank Nori Jacoby for very helpful advice at a couple of critical crossroads, and for stimulating discussions on information theory and cognitive science. Daniel Reichman has also been a great source for discussions and advice, and I am particularly grateful to Daniel for referring me to Terry's work and making the personal introduction. I am also very thankful to several other friends and colleagues who have been particularly supportive and with whom I have had many inspiring discussions: Yoav Wald, Hadar Levi-Aharoni, Roy Fox, Pedro Ortega, Michal Moshkovitz, Nadav Amir, Shlomi Agmon, Ravid Shwartz-Ziv, Ori Lavi-Rotbain, Oren Peles, Shai Berman, Josh Abbott, and Geoff Bacon. I also thank Tom Griffiths for welcoming me to his lab meetings when they were at UC Berkeley, and Collin Baker for welcoming me to ICSI.

I thank Bruno Olshausen and the other organizers of the Brain and Computation program at the Simons Institute for the Theory of Computing for hosting me in this program. It was a truly thought provoking research experience. I also thank Noam Slonim for being my sponsor for an IBM PhD Fellowship, and Yonatan Bilu for mentoring me during an exciting summer internship at IBM Research.

Finally, I thank my dear family for their love and support. My mother, Orit, has provided countless invaluable advice along the way. I can now prove mathematically that her blind belief in me carries no information, however its value may be unbounded.

And last but not least, huge thanks to the one person on earth with whom I share every journey, my partner for life, Roy Fox. His deep insight and endless support have safely guided me around every obstacle along the way; his love and companionship have made this endeavor possible and worthwhile.

Abstract

Across the world, languages enable their speakers to communicate effectively using relatively small lexicons compared to the complexity of the environment. How do word meanings facilitate this ability across languages? The forces that govern how languages assign meanings to words, i.e., human semantic systems, have been debated for decades. Recently, it has been suggested that languages are adapted for efficient communication. However, a major question has been left largely unaddressed: how does pressure for efficiency relate to the evolution of semantic systems? This thesis addresses this question by identifying fundamental information-theoretic principles that may underlie semantic systems and their evolution. The main results and contributions of this thesis are structured in three parts, as detailed below.

Part I presents our information-theoretic approach to semantic systems and demonstrates its predictive power and empirical advantages. We argue that languages efficiently encode meanings into words by optimizing the Information Bottleneck (IB) tradeoff between the complexity and accuracy of the lexicon. We begin by testing this hypothesis in the domain of color naming, and show that color naming across languages is near-optimally efficient in the IB sense. Furthermore, this finding suggests (1) a theoretical explanation for why empirically observed patterns of inconsistent naming and stochastic categories, which introduce ambiguity, are efficient for communication; and (2) that languages may evolve under pressure for efficient coding through an annealing-like process that synthesizes continuous and discrete aspects of previous accounts of color category evolution. This process generates quantitative predictions for how color naming systems may change over time. These predictions are directly supported by an analysis of recent data documenting changes over time in the color naming system of a single language. In addition, we show that this general approach also applies to two qualitatively different semantic domains: names for household containers, and for animal categories. Taken together, these findings suggest that pressure for efficient coding under limited resources, as defined by IB, may shape semantic systems across languages and across domains.

Part II presents an information-theoretic approach for characterizing communicative need. Communicative need is a central component in many efficiency-based approaches to language, including the IB approach mentioned above. It is formulated as a prior distribution over elements in the environment that reflects the frequency in which they are referred to during communication. There is evidence that this component may have substantial influence on semantic systems, however it has not been clear how to characterize and estimate it. We address

this problem by invoking two general information-theoretic principles: the capacity-achieving principle, and the maximum-entropy principle. As before, we test this approach in the domain of color naming. First, an analysis based on the capacity-achieving principle suggests that color naming may be shaped by communicative need in interaction with color perception, as opposed to traditional accounts that focused mainly on perception and recent accounts that focused mainly on need. Second, by invoking the maximum-entropy principle with word-frequency constraints, we show that linguistic usage may be the most relevant factor for characterizing the communicative need of colors, as opposed to the statistics of colors in the visual environment. This approach is domain-general, and so it may also be used to characterize communicative need in other semantic domains.

Part III touches on the foundations of the IB framework by extending the mathematical understanding of the structure and evolution of efficient IB representations. This contribution is important given the growing evidence for the applicability of the IB principle not only to language, but also to deep learning, neuroscience, and cognition. Here, we consider specifically the case of discrete, or symbolic, representations, as in our application of IB to the evolution of human semantic systems. We characterize the structural changes in the IB representations as they evolve via a deterministic annealing process; derive an algorithm for finding critical points; and numerically explore the types of bifurcations and related phenomena that occur in IB. These phenomena and the theoretical justification for this approach apply to efficient symbolic representations in both humans and machines. Therefore, we believe that this approach could potentially guide the development of artificial intelligence systems with human-like semantics.

In conclusion, this thesis presents a mathematical approach to semantic systems that is comprehensively grounded in information theory and is supported empirically. Pressure for efficient coding arises as a major force that may shape semantic systems across languages, suggesting that the same principles that govern low-level neural representations may also govern high-level semantic representations.

Contents

Chapter 1

Introduction

This thesis aims to identify general computational principles that underlie the structure and evolution of semantic systems. To this end, we seek independently motivated optimization principles that generate quantitative predictions, and then test these predictions on cross-linguistic data. Our theoretical motivation is grounded in information theory, which was initially introduced by Shannon in 1948 as "A Mathematical Theory of Communication" (Shannon, 1948) and renamed in the following year to "The Mathematical Theory of Communication" (Shannon and Weaver, 1949). This seemingly minor change reflects the appreciation of Shannon's axiomatic and platform-independent approach, which situates it as *the* theory of communication.

However, since the early days of information theory, the immediate appeal of invoking it as a theory of human language — the most remarkable communication system — has been widely contested (Miller, 2003; Luce, 2003; Bentz, 2018). Particularly noteworthy in our context is the claim that information theory only pertains to the engineering aspects of communication, and not to the semantic aspects of communication (Shannon and Weaver, 1949). This picture has started to change over the years (e.g., Pereira et al., 1993; Pereira, 2000; Plotkin and Nowak, 2000; Ferrer i Cancho and Solé, 2003), and more recently there has been a surge of evidence that information-theoretic approaches may explain a wide range of linguistic phenomena (for review: Gibson et al., 2019). In particular, it has been argued that semantic systems are adapted for efficient communication (Kemp et al., 2018), and this notion of efficiency has been formulated partially in information-theoretic terms. However, a major question has been left largely unaddressed: how does pressure for efficiency relate to the evolution of semantic systems?

In this thesis, we address this question by presenting a mathematical approach to semantic systems which is comprehensively grounded in information theory and is supported empirically. Central to this approach is the Information Bottleneck principle (Tishby et al., 1999) that arises as the link between semantics and the branch of information theory called Rate–Distortion theory (Shannon, 1948, 1959), which addresses the problem of data compression. Our theoretical account is novel, to our knowledge, and shows that fundamental information-theoretic principles may explain the different ways human languages encode meanings into words. This account has several important implications for the evolution of human semantic

systems, as well as potential applications for the development of human-like semantic systems in machines. Preparatory to discussing these results, we first review the information-theoretic principles on which this thesis builds. We then review previous studies that applied related ideas to semantic systems, as well as other relevant applications of information theory to language. We conclude this chapter with an outline and overview of the main contribution of this thesis.

1.1 Information theory

We begin with an overview of the main results in information theory that are particularly relevant for this thesis. Many details are left out, including proofs. For a comprehensive discussion of this topic see Cover and Thomas (2006). In this section, and generally throughout this thesis, we use upper-case letters to denote random variables (e.g., X), calligraphic letters to denote their support (e.g., $\mathcal{X}$), and lower-case letters to denote a specific realization (e.g., x). For simplicity, all random variables are assumed to be discrete, unless stated otherwise. In addition, we use the notation $p(x)$ to denote either the probability mass distribution of X, or the probability of a realization $x \in \mathcal{X}$ according to this distribution. This abuse of notation is standard, and can be disambiguated from context. For brevity, we occasionally denote distributions by lower-case letters (e.g., p) when the intention is clear.

1.1.1 Informational measures

The formal notion of information follows from several basic measures. These measures are model-independent in the sense that they depend only on the probability distribution of some random variables without making any assumptions about the shape of these distributions. The first informational measure we define is entropy, which measures the uncertainty induced by a given distribution.

Definition 1. *The entropy of a random variable $X \sim p(x)$ is defined as*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \,.$$

Shannon derived this definition axiomatically by defining three intuitive properties that any measure of uncertainty should satisfy (continuity, additivity, and monotonicity), and proving that entropy is the only[1] function that satisfies these properties. Intuitively, entropy reflects the minimum description length of X, because it is (roughly) the minimal number of bits or binary questions (e.g., "is $X \in \mathcal{A}$?") that are needed in order to determine the exact value of X. Entropy is maximal when $p(x)$ is uniform, in which case $H(X) = \log |\mathcal{X}|$. It is minimal

[1] This holds up to the choice of the base of the logarithm that determines the units. Here we assume that the log is in base 2, namely the units are bits.

when $p(x)$ is deterministic, i.e. when there is only one possible value with $p(x) = 1$, in which case there is no uncertainty and $H(X) = 0$.

The derivation of entropy as a measure of uncertainty forms the basis of the *maximum entropy principle* (MaxEnt: Jaynes, 1982). MaxEnt states that if $p(x)$ is unknown, then the most justified estimator $\tilde{p}(x)$ is the one that requires minimal assumptions, or equivalently, leaves maximal uncertainty about X. In other words, $\tilde{p}(x)$ is the distribution that maximizes $H(X)$. One particularly interesting property of entropy in this context, is that it is a concave functional of $p(x)$, which implies that the MaxEnt distribution is unique. In Chapter 6 we invoke this principle by introducing a MaxEnt-type of principle for characterizing and estimating the communicative need of elements in the environment.

The next informational measure we define is the Kullback-Leibler (KL) divergence, also known as relative entropy, which is a measure of the divergence between two distributions.

Definition 2. *The KL divergence between two distributions $p(x)$ and $q(x)$ is defined as*

$$D[p \parallel q] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} .$$

Note that this definition is based on the convention that $0 \log \frac{0}{0} = 0$.

It is possible to show that $D[p \parallel q] \geq 0$, and that equality holds if and only if $p = q$. The KL divergence is not a metric because it is not symmetric and does not obey the triangle inequality. However, it is still useful to think about it as the natural "distance" between distributions for several reasons. First, notice that $D[p \parallel q]$ is the expected log–likelihood ratio between $p(x)$ and $q(x)$, and thus it controls the discriminability between these two distributions, when p is the true underlying distribution of X. Second, the KL divergence reflects the difference between the minimal description length of $X \sim p(x)$ and the description length if $q(x)$ is mistakenly used instead of $p(x)$. To see this, notice that $D[p \parallel q] = \sum_x p(x) \log \frac{1}{q(x)} - H(X)$, and recall that $H(X)$ is the minimal description length. Third, Pinsker's inequality implies that the KL divergences upper bounds the L_1 distance between p and q. Therefore, it also bounds the L_2 distance, and in general any L_ρ distance for $\rho \geq 1$. Fourth, it was shown that the KL divergence is unique in the sense that it is the only divergence measure for probabilities that satisfies several desired properties (Harremoës and Tishby, 2007).

The last informational measure we introduce here is mutual information, which is the key measure for characterizing the theoretical limits of communication.

Definition 3. *Let $(X, Y) \sim p(x, y)$, and let $p(x)$ and $p(y)$ be their marginal distributions respectively. The mutual information between X and Y is defined as*

$$I(X;Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} .$$

Notice that mutual information is closely related to the other informational measures we

previous defined. One way of thinking about mutual information is in terms of the KL divergence between $p(x,y)$, and the hypothetical joint distribution had X and Y been independent of each other, i.e. $q(x,y) = p(x)p(y)$. To see this, notice that $I(X;Y) = D[p(x,y) \parallel p(x)p(y)]$ by definition. It thus follows that $I(X;Y) \geq 0$ and $I(X;Y) = 0$ if and only if X and Y are independent. In addition, it is possible to show that $I(X;Y) = H(X) - H(Y|X)$, and from symmetry it also holds that $I(X;Y) = H(Y) - H(X|Y)$. Therefore, another interpretation is that the mutual information captures the reduction of uncertainty about one variable as a result of knowing the other variable. In addition, since $H(X|Y), H(Y|X) \geq 0$, it holds that $I(X;Y) \leq \min\{H(X), H(Y)\}$. If $H(X|Y) = 0$, i.e. X is completely known given Y, then $I(X;Y) = H(X)$. As a special case, it holds that $I(X;X) = H(X)$. In this sense, mutual information is more general than entropy, and entropy is sometimes referred to as self-information.

The fundamental theorems in information theory involve the maximization or minimization of mutual information. In particular, this thesis invokes several optimization principles of this type. The following theorem implies that maximizing $I(X;Y)$ with respect to $p(x)$ is a concave optimization problem, and minimizing it with respect to $p(y|x)$ is a convex optimization problem.

Theorem 1 (Cover and Thomas (2006), Theorem 2.7.4)**.** *Let* $(X,Y) \sim p(x,y) = p(x)p(y|x)$. *The mutual information* $I(X;Y)$ *is a concave function of* $p(x)$ *for fixed* $p(y|x)$ *and a convex function of* $p(y|x)$ *for fixed* $p(x)$.

Another important property of mutual information is the Data Processing Inequality (DPI), which implies that the information Y contains about X cannot increase by processing Y. Before stating this formally, we need to define what it means to process Y. We say that (X,Y,Z) form a Markov chain if their joint distribution can be decomposed as $p(x,y,z) = p(x,y)p(z|y)$. That is, Z is a function of Y and given Y it is independent of X. Therefore, we consider Z to be a processed version of Y.

Theorem 2 (DPI)**.** *If* (X,Y,Z) *form a Markov chain, then* $I(X;Y) \geq I(Z;X)$.

It is easy to show that $p(x,y)p(z|y) = p(y,z)p(x|y)$, and therefore if (X,Y,Z) form a Markov chain so do (Z,Y,X). We denote these Markov chains by $X - Y - Z$. Notice that this implies that if $X - Y - Z$, then $I(X;Y) \geq I(Z;X)$ and $I(Y;Z) \geq I(Z;X)$.

1.1.2 The fundamental problem of communication

The fundamental problem of communication is how to transmit information from a sender to a receiver over an imperfect communication channel, while ensuring sufficiently small error and minimal channel uses. Shannon's formulation of this problem is based on an abstract communication model (Figure 1) which in its simplest form is composed of the following components: (1) An information source that generates a message $M \sim p(m)$. (2) A sender

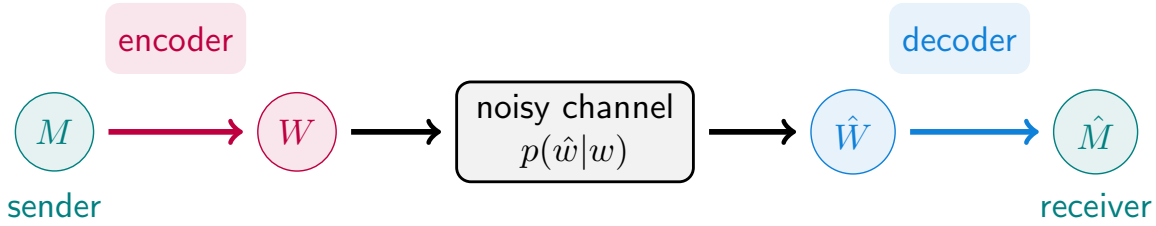

Figure 1. Shannon's basic communication model (Shannon, 1948).

that takes a message and encodes it into a transmittable form. (3) A communication channel[2] that is defined by an input alphabet $\mathcal{W}$, an output alphabet $\hat{\mathcal{W}}$, and the transition probability distribution $p(\hat{w}|w)$. Note that this abstraction captures the noise characteristics of the channel regardless of its physical medium. (4) A receiver that decodes the output of the channel by reconstructing from it the transmitted messages. Since the channel is noisy and limited it may inject errors, and so this reconstruction may be inaccurate.

For any given source distribution $p(m)$ and channel $p(\hat{w}|w)$, the objective is to find a code, i.e. an encoder and decoder pair, that attains sufficiently small reconstruction error while minimizing the number of channel symbols that needs to be transmitted. One of Shannon's key insights was that this problem can be broken into two parts: source coding, or data compression, that removes redundancies so that only the minimal representation of the source would need to be transmitted; and channel coding, that adds redundancy in a constrained manner in order to correct errors due to the channel's noise. The source coding problem also applies to cases in which some distortion is permitted between the sent and received message. For example, cellular phones reduce the amount of transmitted data by compressing speech signals with some degree of distortion that is insignificant for understanding what is being said. This type of compression is called lossy compression, and is characterized by Rate–Distortion (RD) theory (Shannon, 1959).

Shannon's coding theorems characterize the theoretical limits of communication, which are attained when messages are encoded in large sequences rather than individually (block coding). The key result is that despite the channel's noise, it is possible to communicate with arbitrarily small error if $H(M)$ does not exceed[3] some quantity C called the channel capacity, which we will define later. In the case of lossy compression, $H(M)$ is replaced with $I(M;\hat{M}) = H(M) - H(M|\hat{M})$. Shannon's separation theorem (see Cover and Thomas, 2006, section 7.13) implies that combining the source coding solution under an assumption of a noiseless channel, with the channel coding solution under the assumption of a uniform source, can be done (asymptotically) without loss in performance compared to joint source-channel coding.

The following two sections elaborate on channel capacity and RD, as they both play an important role in this thesis. The principle of achieving the channel capacity is central to our

[2]Strictly speaking, we refer here to a discrete memoryless channel.

[3]Following Cover and Thomas (2006), we assume that a sequence of n source messages is mapped to a sequence of $k = n$ channel symbols. In the more general case, the requirement is $H(M) \leq kC/n$.

characterization of the notion of communicative need, as discussed in Part II of this thesis. In Parts I and III we invoke the Information Bottleneck (IB) principle (Tishby et al., 1999), which can be cast in terms of RD (see Section 1.2), in order to characterize human semantic systems and symbolic representations more generally. However, we invoke the two principles separately, that is, in each case we consider a different instance of the general communication model. While these two communication models are coupled in some sense, the channel we consider for characterizing communicative need is different than the implicit noiseless channel in the IB source coding problem. Therefore, in our following discussion on channel capacity we use X and Y to denote the input and output of the channel, and leave the notation we used thus far for our discussion on RD and IB.

1.1.3 Channel capacity

The capacity of a channel is the maximal bits per symbol (on average) that can be transmitted over the channel. In other words, the maximum number of distinct messages that can be transmitted accurately with n channel uses is roughly 2^{nC}, where C is the channel capacity. Perhaps the most celebrated result in information theory is the characterization of the channel capacity in terms of the mutual information between the input and output of the channel.

Definition 4. *The capacity of a channel $p(y|x)$ is defined as*

$$C = \max_{p(x)} \; I(X;Y)\,,$$

where the maximum is taken over all prior distributions $p(x)$ for the channel's input.

Recall that maximizing $I(X;Y)$ with respect to $p(x)$ for fixed $p(y|x)$ is a concave optimization problem (Theorem 1). Strictly speaking, the above definition refers to the "information channel capacity." Shannon's channel coding theorem shows that the information channel capacity is equal to the operational definition mentioned above, i.e. the maximal achievable rate of bits per channel use. The channel capacity is an important characteristic of the channel, independent of the specific channel codes that are used in practice. Here we focus on this characterization of the channel, and in particular on the prior that achieves the capacity. We refer to this prior as the capacity-achieving prior (CAP).

In general, there is no closed form solution for the channel capacity and the CAP. However, in the cases considered in this thesis, they can be evaluated numerically via the Blahut–Arimoto (BA) algorithm (Blahut, 1972; Arimoto, 1972). The main idea behind this algorithm is to introduce an auxiliary variable $p(x|y)$ and to differentiate $I(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x|y)}{p(x)}$

with respect to $p(x)$ and $p(x|y)$. This gives the following coupled equations:

$$\begin{cases} p(x) &= \frac{1}{Z}\exp\left(-\sum_y p(y|x)\log p(x|y)\right) \\ p(x|y) &= \frac{p(x)p(y|x)}{\sum_{x'}(y|x')p(x')} \end{cases}, \tag{1.1}$$

where Z is a normalization factor. Since the problem is concave, satisfying these equations self-consistently is a necessary and sufficient condition for optimality. The BA algorithm finds the optimal solution by iteratively updating these two equations until convergence.[4]

In this thesis we invoke the capacity-achieving principle with respect to a naming channel, that is, a channel that takes as input an object in the environment and outputs a word that is associated with it. This is a semantic channel implemented by speakers of a language, rather than a low-level physical channel. The CAP in this case induces a prior distribution over objects in the environment. This prior distribution is a property of the given channel, and so it may reveal communicative aspects that are implicit in the naming systems of different languages. Another motivation for considering capacity achieving priors in our context is the relation of such priors to MaxEnt priors and to the notion of least informative priors in Bayesian inference (Bernardo, 2005). These connections are laid out in chapter 2 (Supporting Information, Section 2).

1.1.4 Rate–Distortion theory

The dual problem to transmitting information over a noisy channel is data compression. Shannon's Rate–Distortion (RD) theory addresses this problem and characterizes the optimal source coding schemes in the case of lossy compression, i.e. when some distortion between the sent and received messages, $d(m, \hat{m})$, may be tolerated. Lossless compression is obtained as a special case when no distortion is allowed. However, in many cases the source contains more information than what is needed for successful communication, as in the speech example mentioned earlier. In other cases, even a noticeable amount of error may be beneficial, if the cost of such errors is lower than the cost or effort involved in transmitting all the data. Here we are interested in such cases.

Loosely speaking, the RD problem is to find the minimal number of bits that are required for representing M, while not exceeding an allowed level of distortion; or equivalently, to find the minimal expected distortion that can be achieved with a given budget of bits that are used for the representation.

To formulate this more precisely, consider an encoder f that takes a sequence of n messages and represents it using nR bits; i.e., $f : \mathcal{M}^n \to \{1, \ldots, 2^{nR}\}$. Assume that this representation is transmitted over a noiseless channel, and is then decoded by the receiver using a decoder

[4] It can be shown that each iteration increases the mutual information, and we have already seen that this function is bounded. Therefore, asymptotical convergence is guaranteed.

$g : \{1, \ldots, 2^{nR}\} \to \hat{\mathcal{M}}^n$. R is called the rate of the code, and it is the number of bits per message. Low values of R imply high compression rates, however this may result in high expected distortion, defined by

$$D_{f,g} = \sum_{m^n \in \mathcal{M}^n} p(m^n) \frac{1}{n} \sum_{i=1}^{n} d\left(m_i, g(f(m^n))_i\right) . \tag{1.2}$$

We can simplify this expression by rewriting it as

$$D_{f,g} = \sum_{m,\hat{m}} p(m) P_{f,g}(\hat{m}|m) d(m, \hat{m}) , \tag{1.3}$$

where $P_{f,g}(\hat{m}|m)$ is the conditional probability distribution induced by the code. To see this, first note that the probability that the i-th reconstructed messages is $\hat{m}$ given that the i-th sent message was m is obtained by summing over all possible sequences as follows

$$p\Big(g(f(M^n))_i = \hat{m} | M_i = m\Big) = \frac{\sum_{m^n \in C^i_{m,\hat{m}}} p(m^n)}{\sum_{m^n \in C^i_m} p(m^n)} . \tag{1.4}$$

where $C^i_m = \{m^n \in \mathcal{M}^n : m_i = m\}$ and $C^i_{m,\hat{m}} = \{m^n \in C^i_m : g(f(m^n))_i = \hat{m}\}$. Taking the expectation over i with respect to $p(i) = \frac{1}{n}$ gives

$$P_{f,g}(\hat{m}|m) = \sum_i p(i) p\Big(g(f(M^n))_i = \hat{m} | M_i = m\Big) . \tag{1.5}$$

We refer to this distribution as the *probabilistic signature* of the code.

RD theory characterizes the achievable[5] region of (R, D) pairs, as well as the probabilistic signature of optimal codes at the limit of large n. The rate–distortion function $R(D)$ is defined by the infimum over all rates that are achievable with distortion at most D. Remarkably, Shannon showed that $R(D)$ is equal to the information rate–distortion function which is defined by the following constrained optimization problem:[6]

$$\begin{aligned} R^{(I)}(D) \quad = \quad & \min_{p(\hat{m}|m)} \; I(M; \hat{M}) \\ & \text{such that } \sum_m p(m) p(\hat{m}|m) d(m, \hat{m}) \leq D \end{aligned} \tag{1.6}$$

We therefore consider $R^{(I)}(D)$ as the rate, or representational complexity, because it roughly corresponds to the expected number of bit that are needed to encode M using the representation $\hat{M}$. Note that rates below $R^{(I)}(D)$ are unachievable with distortion at most D.

Shannon's formulation and proofs are based on coding in large blocks, however from a

[5] The term "achievable" is used here loosely, and we rely on its intuitive interpretation. See Cover and Thomas (2006) for a formal definition of achievability.

[6] For proof and more details see Cover and Thomas (2006), Theorem 10.2.1.

cognitive perspective this assumption is unlikely (Luce, 2003). In some cases it is possible to reach the theoretical limit with finite blocks or even with single-message coding (Gastpar et al., 2003), but it is also possible that the "cognitively achievable" region is smaller than the region defined by the RD curve. While we do not know what the actual cognitively achievable region is, we can still compare the probabilistic signature of the optimal codes with the probabilistic signature of human-generated coding schemes (in our case, naming systems). If the two signatures are similar, then this suggests that the human-generated coding schemes may lie near the theoretical limit of communication that was derived by Shannon.

Shannon's proofs are not constructive, and finding the optimal codes may be computationally hard. However, finding the probabilistic signature of the optimal codes is tractable in our case. Recall that the mutual information is convex in $p(\hat{m}|m)$ and note that the constraint on the distortion is linear in $p(\hat{m}|m)$. Therefore equation equation (1.6) is a convex optimization problem. Solving this constrained optimization problem amounts to minimizing the Lagrangian

$$\mathcal{L}[p(\hat{m}|m);\beta] = I(M;\hat{M}) + \beta \underset{\substack{m\sim p(m)\\ \hat{m}\sim p(\hat{m}|m)}}{\mathbb{E}} [d(M,\hat{M})] \tag{1.7}$$

where β is the Lagrange multiplier that corresponds to the constraint on the distortion. The minimum of this Lagrangian can be found via the same BA algorithm we discussed in Section 1.1.3, however in this case $p(\hat{m}|m)$, rather than $p(m)$, has an exponential form:

$$\begin{cases} p(\hat{m}|m) &= \dfrac{p(\hat{m})}{Z_\beta(m)} \exp\left(-\beta d(m,\hat{m})\right) \\ p(\hat{m}) &= \sum_m p(m)p(\hat{m}|m) \end{cases}, \tag{1.8}$$

where $Z_\beta(m)$ is a normalization factor.

While RD theory is a powerful framework, it does not provide a means to determine the distortion measure and instead assumes that the distortion is specified by the designer of the communication system. However, is it not always clear how to choose the right distortion measure. In the following section we discuss an influential framework that addresses this question, and is the main framework on which this thesis builds.

1.2 The Information Bottleneck principle

The Information Bottleneck (IB) principle was introduced by Tishby et al. (1999) as a method for finding a concise representation of an input variable which is maximally informative about some target variable. IB can be cast in terms of RD theory, with a unique distortion measure that arises naturally from the statistics of the input and target variables. In addition, it has been broadly applied across multiple disciplines, including neuroscience (Buesing and Maass, 2010; Palmer et al., 2015; Wang et al., 2017), signal processing (Hecht and Tishby, 2005), functional

harmony (Jacoby et al., 2015), language (Slonim and Tishby, 2000), deep learning (Tishby and Zaslavsky, 2015; Alemi et al., 2017; Shwartz-Ziv and Tishby, 2017), and machine learning more generally (Chechik et al., 2005; Shamir et al., 2010).

In this thesis, we build on the IB framework in order to account for semantic systems and their evolution. This application of IB is novel to our knowledge, and it is based on the RD, or lossy compression, interpretation of IB. We first present the standard IB formulation, and then show how this formulation corresponds to the communication model in Figure 1 and to the RD problem discussed in Section 1.1.4.

Let X be an input random variable, Y a target variable, and $p(x,y)$ their joint distribution. A representation T is a stochastic function of X defined by a mapping $p(t|x)$. In other words, $Y - X - T$ form a Markov chain. Note that the Data Processing Inequality (DPI, see Section 1.1.1) implies that $I(X;Y) \geq I(T;Y)$ regardless of how $p(t|x)$ is chosen. We refer to $I(T;Y)$ as the *relevant information* in T about Y, or the *accuracy* of the representation. If T is a copy of X, then its accuracy is maximal. However, if T is a compressed version of X that losses relevant bits about Y, then these bits cannot be restored and the inequality in the DPI is strict. The *representational complexity* of T is roughly the expected number of bits that it keeps about X. In IB, as in RD, this complexity term is measured by the informational rate $I(X;T)$.

According to the IB principle, optimal representations satisfy a tradeoff between compressing X, i.e. minimizing the representational complexity, and maximizing the relevant information about Y. Formally, the IB optimization problem is to minimize the following objective function:

$$\mathcal{F}_\beta[p(t|x)] = I(X;T) - \beta I(T;Y)\,, \tag{1.9}$$

where β is the tradeoff parameter. Equivalently, this can be formulated as a constrained optimization problem, similar to equation (1.6), where β is the Lagrange multiplier that corresponds to a constraint on the amount of required relevant information.

It can be shown (see e.g. Chapter 7, Lemma 1) that:

$$I(T;Y) = I(X;Y) - \mathop{\mathbb{E}}_{\substack{x\sim p(x)\\ t\sim p(t|x)}} \left[D[p(y|x) \parallel p(y|t)]\right]. \tag{1.10}$$

Therefore, the expected KL divergence between $p(y|x)$ and $p(y|t)$ defines the accuracy loss, and minimizing $\mathcal{F}_\beta$ is equivalent to minimizing

$$\mathcal{L}_{\mathrm{IB}}[p(t|x)] = I(X;T) + \beta \mathop{\mathbb{E}}_{\substack{x\sim p(x)\\ t\sim p(t|x)}} \left[D[p(y|x) \parallel p(y|t)]\right]. \tag{1.11}$$

Notice that equation (1.11) has the same form as equation (1.7), where $D[p(y|x) \parallel p(y|t)]$ emerges as the distortion measure.

The RD problem that corresponds to the IB framework is obtained when the messages them-

selves are defined by distributions (Harremoës and Tishby, 2007). In this case, the KL divergence is indeed the natural distortion measure between M and $\hat{M}$ (see Section 1.1.1 for detailed justification). To see exactly this relation between RD and IB, let $\mathcal{M}$ be the set of probability distributions $m(y)$ over $\mathcal{Y}$ that is induced by $p(y|x)$, i.e., $\mathcal{M} = \{m(y) : \exists x, m(y) = p(y|x)\}$. Let $\phi : \mathcal{X} \to \mathcal{M}$ be the mapping from x to its corresponding message, i.e. $\phi(x) = p(y|x)$. Similarly, we map each t to the corresponding reconstruction. Formally, let $\hat{\mathcal{M}}$ be the simplex of probability distributions over $\mathcal{Y}$ and let $\mu : \mathcal{T} \to \hat{\mathcal{M}}$ such that $\mu(t) = p(y|t)$. In other words, X corresponds to the sent messages, and T corresponds to its reconstructed representation.

In Chapter 2 we apply this formulation to naming systems, that is, mappings from objects in the environment to words. We assume that the speaker (sender) mentally represents each object by a distribution M over some relevant features, and wishes to communicate this mental representation to the listener (receiver). This is done by encoding the speaker's mental representation into a codeword W and transmitting the codeword to the listener. In this setting T corresponds to W, and the listener's interpretation of the word corresponds to $\hat{M}$. This setting suggests a slightly different relation between IB and RD compared to the relation that was shown by (Harremoës and Tishby, 2007), however we have shown (see Chapter 2, Supporting Information) that the differences between these two formulations are not substantial. Based on these theoretical foundations, we can derive the optimal naming systems, i.e., systems in which W is an optimal IB representation of M, and compare these systems with actual naming systems across languages.

Finding optimal IB representations can be done via the IB method, which is a variant of the BA algorithm. Denote by $p_\beta(t|x)$ an optimal IB representation, which is necessarily a stationary point of equation (1.9) (or equivalently, of equation (1.11)) for a given value of β. Tishby et al. (1999) have shown that a necessary condition for optimality is that p_β satisfies the following self consistent equations:

$$\begin{cases} p_\beta(t|x) &= \dfrac{p_\beta(t)}{Z_\beta(x)} \exp\left(-\beta D[p(y|x) \parallel p_\beta(y|t)]\right) \\ p_\beta(t) &= \displaystyle\sum_{x \in \mathcal{X}} p(x) p_\beta(t|x) \\ p_\beta(y|t) &= \displaystyle\sum_{x \in \mathcal{X}} p_\beta(x|t) p(y|x)\,. \end{cases} \tag{1.12}$$

The IB method starts with some initial condition $p_0(t|x)$, and iteratively updates these equations until convergence. Notice that the first two equation are similar to the BA algorithm for RD (1.8), however here we have a third update equation for $p_\beta(y|t)$. In addition, the distortion measure in this case, i.e. the KL divergence in the exponent of the first update equation, depends on the solution through $p_\beta(y|t)$. In other words, the distortion measure in IB is not constant, as in RD, but rather depends on the properties of the problem (Gilad-Bachrach et al., 2003). This powerful feature however comes at a price, namely that the IB optimization problem is non-convex. Therefore, only local convergence is guaranteed in the IB method. This

issue can be mitigated using a method called deterministic annealing (Rose et al., 1990; Rose, 1998, and see also Chapter 7 in this thesis).

Denote by $I_\beta(X;T)$ and $I_\beta(T;Y)$ the complexity and accuracy of the optimal IB representation defined by p_β. The IB theoretical limit is defined by these Pareto-optimal tradeoffs $\Big(I_\beta(X;T), I_\beta(T;Y)\Big)$ as a function of β. This theoretical limit is discussed in detail in Chapters 2 and 7. In the following section we complement that by discussing a geometric interpretation of IB that provides useful intuition for interpreting our main results in Parts I and III.

1.2.1 Geometric interpretation

One of the first applications of IB has been in the context of distributional clustering (Pereira et al., 1993; Slonim and Tishby, 2001). IB can be seen as soft clustering of points in the simplex of probability distributions over $\mathcal{Y}$, denoted by $\triangle(\mathcal{Y})$. The KL divergence is the natural "distance" measure in this clustering problem, for the reasons discussed in Section 1.1.1. To see this clustering interpretation of IB, notice that the update equations of the IB method (1.12) can be rewritten as follows:

$$p(t|x) \;\propto\; p(t)\exp\left(-\beta d(\phi(x),\mu(t))\right) \tag{1.13}$$

$$p(t) \;=\; \sum_x p(x)p(t|x) \tag{1.14}$$

$$\mu(t) \;=\; \sum_x p(x|t)\phi(x)\,, \tag{1.15}$$

where $d(\phi(x),\mu(t)) = D[\phi(x) \,\|\, \mu(t)]$. These three equations correspond to soft clustering of the points $\phi(x)$ in $\triangle(\mathcal{Y})$. In this interpretation, x behaves as the index of the point $\phi(x)$ and

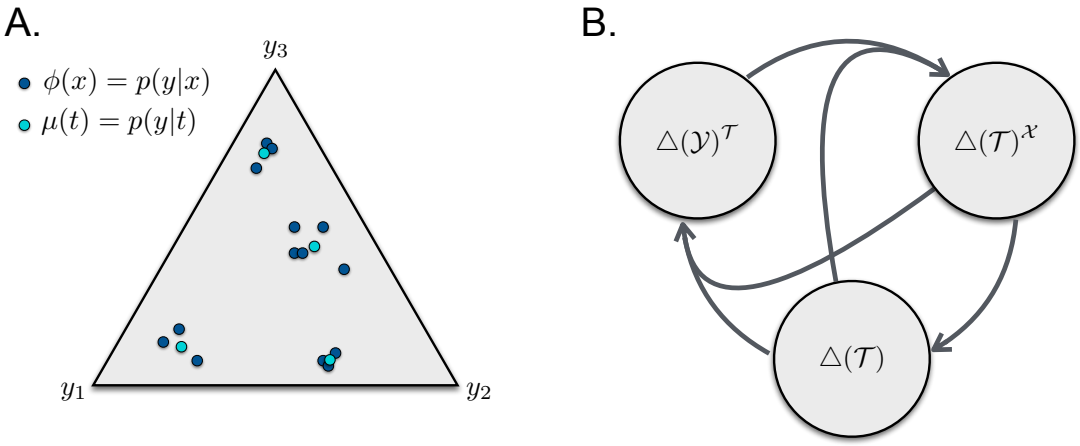

Figure 2. Geometric interpretation of IB. A. Illustration of distributional clustering for $|\mathcal{Y}| = 3$. The gray triangle represent $\triangle(\mathcal{Y})$. Each x is mapped to the point $\phi(x) \in \triangle(\mathcal{Y})$, and each t is mapped to a point $\mu(t) \in \triangle(\mathcal{Y})$. The IB method clusters the points $\phi(x)$, by iteratively updating the cluster assignment probabilities $p(t|x)$, the cluster weights $p(t)$, and the cluster centroids $\mu(t)$. B. Illustration of the IB method as alternating minimization. Each arrow corresponds to a projection onto a set of distributions.

t behaves as an index of a cluster. Equation (1.13) gives the cluster assignment probabilities, equation (1.14) gives the cluster weights, and equation (1.15) gives the cluster centroids that lie in $\triangle(\mathcal{Y})$. This geometric interpretation is illustrated in Figure 2A.

Another closely related geometric interpretation of the IB method is based on the fact that it is an instance of the alternating minimization algorithmic scheme (Csiszár and Shields, 2004). Given an initial clustering assignment, the optimal weights (1.14) are obtained by minimizing $\mathcal{F}_\beta$ with respect to $p(t)$ for fixed $p(t|x)$ and fixed centroids. This is a projection onto the simplex of probability distributions over $\mathcal{T}$, i.e. $\triangle(\mathcal{T})$. Given the clustering assignment and weights, the optimal centroids (1.15) are obtained by minimizing $\mathcal{F}_\beta$ only with respect to $p(y|t) = \mu(t)$, which is a projection onto $\triangle(\mathcal{Y})^{\mathcal{T}} = \triangle(\mathcal{Y}) \times \cdots \times \triangle(\mathcal{Y})$. Finally, the update of the clustering assignment (1.13) is obtained by minimizing $\mathcal{F}_\beta$ with respect to $p(t|x)$ for fixed weights and centroids. This alternating minimization process is illustrated in Figure 2B.

1.2.2 Relation to distributional semantics

A very influential idea in computational semantics is that the meaning of a word is determined by the context in which it appears (Wittgenstein, 1953), or "by the company that it keeps" (Firth, 1957). This intuition was formalized by Harris (1954) in his distributional hypothesis, which suggests that the semantic similarity between words is determined by their distributional similarity, i.e. the similarity between the distributions they induce over their context. The idea of distributional clustering of words with respect to a KL "distance" (Pereira et al., 1993) naturally formalizes this notion of distributional semantics. Therefore, IB arises as the natural method for distributional clustering of words. This approach has been successfully applied in the context of natural language processing (e.g. Slonim and Tishby, 2000, 2001), and in particular, it has been shown to reveal semantic hierarchies from corpus statistics (Pereira et al., 1993; Slonim, 2002). In this thesis we build on a similar idea, namely that meanings can be represented by distributions and that semantic similarities can be measured in terms KL divergence. However, a key difference between our approach and previous applications of IB in this context is that here we will consider *grounded* meaning representations. That is, the meaning of a word is determined by the distribution it induces over a set of perceptual or conceptual features which are grounded in the environment, rather than by the distribution it induces over other words.

1.3 Semantic variation, language evolution, and information

Languages assign meanings to words in many different ways. For example, English has separate terms for "wood" and "tree," whereas Hebrew has only one term for this pair. Such cross-linguistic variation in word meanings appears widely across the lexicon, and has been studied for several decades in semantic domains such as color (Berlin and Kay, 1969; Kay and McDaniel, 1978), folk biology (Berlin, 1992; Brown, 1984), and kinship (Murdock, 1949).

Intriguingly, this wide semantic variation appears to be constrained, and general tendencies in word meanings have been identified across languages (von Finter and Matthewson, 2008). This type of constrained semantic variation has often been held to reflect different linguistic stages along a common evolutionary trajectory (Berlin and Kay, 1969; Brown, 1976, 1984).

One of the most renowned examples in this context is Berlin and Kay's (1969) implicational hierarchy for color naming. Berlin and Kay observed that languages vary widely in their color terms, but at the same time, languages with similar number of color terms tend to divide up the color spectrum in a similar way. That is, languages with two color terms would have terms for black/dark and white/light; languages with three color terms tend to add a term for red; languages with four terms tend to add a term for either green or yellow, and so on. Berlin and Kay conjectured that color terms may evolve via this evolutionary sequence. This influential proposal also inspired similar approaches in other semantic domains (e.g., Brown, 1976, 1984).

These observations suggest that there may be universal principles that underlie systems of word meanings across languages, i.e., human semantic systems. Recently, it has been proposed that such a principle may be the need to communicate efficiently (Kemp and Regier, 2012; Regier et al., 2015). In this view, languages are pressured to optimize a tradeoff between cognitive effort (or complexity) and communicative accuracy. This intuitive idea can be traced back to Zipf's least effort principle (Zipf, 1949), and to Rosch's view of semantic categories (Rosch, 1999). The formulation of this idea by Regier et al. (2015) for semantic systems is partially based on information-theoretic terms, and has gained empirical support in several semantic domains, including kinship, color, and numeral systems (see Kemp et al., 2018, for review). This approach is closely related to the IB framework, however a key difference is the specification of the complexity measure. While Regier et al. (2015) define complexity in a domain-specific manner, in the IB framework the complexity measure is domain-general and grounded in Rate–Distortion theory (Shannon, 1948). This difference yields qualitatively different predictions, especially regarding the probabilistic nature of semantic categories. In Chapter 2 (Supporting Information, Section 4) we show the mathematical relation between these two frameworks in the case of color naming, and discuss the differences in their predictions.

More broadly, related information-theoretic approaches have been applied to a wide range of linguistic phenomena, such as the evolution of word forms (Plotkin and Nowak, 2000), scaling and criticality (Ferrer i Cancho and Solé, 2003), sentence processing (Levy and Jaeger, 2007; Jaeger, 2010), lightness terms (Baddeley and Attewell, 2009), word lengths (Piantadosi et al., 2011), word order (Gibson et al., 2013), symbol grounding (Corominas-Murtra et al., 2014), and compositionality (Kirby et al., 2015). Taken together, this growing body of work reflects the importance of information theory for studying language. However, most of these approaches are based on the problem of data transmission over a noisy channel, rather than on the problem of lossy data compression which is the basis for the approach laid out in this thesis. Furthermore, the relation between a drive for efficient communication and the evolution of semantic systems has been left largely unexplored. This thesis aims to advance toward closing

this gap, while comprehensively grounding the notion of semantic efficiency in fundamental information-theoretic principles.

1.3.1 The case of color naming

A large part of this thesis focuses on the special case of color naming. While our theoretical framework is not specific to color and has been applied to other semantic domains (Chapter 4), color provides a particularly useful and important test case for three main reasons. First, this domain has been central to the study of linguistic diversity and its relation to language evolution, as well as of the interaction between language and other cognitive functions. Second, this is a unique case in which fine-grained naming data exists for over 100 languages of non-industrialized societies (Kay et al., 2009), in addition to western languages such as English (Lindsey and Brown, 2014). Third, several information-theoretic approaches has previously been proposed for this domain (Lindsey et al., 2015; Regier et al., 2015; Gibson et al., 2017, and see also Zhaoping, 2007), as well as various computational models for the emergence of color categories (e.g., Steels and Belpaeme, 2005; Dowman, 2007; Loreto et al., 2012). This rich literature calls for a general theoretical account of color naming and its evolution.

1.4 Overview and main contributions

This thesis presents a mathematical approach to semantic systems and their evolution, which is comprehensively grounded in information theory and is supported empirically. The main results and contributions of this thesis are structured in three parts, as detailed below.

Part I: Efficient compression in the lexicon. The first and main contribution of this thesis is a principled information-theoretic account of the structure and evolution of human semantic systems. Chapter 2 presents this general approach and begins by testing it in the domain of color naming. Specifically, we argue that languages efficiently encode meanings into words by optimizing the Information Bottleneck (IB) tradeoff between the complexity and accuracy of the lexicon. Using a rigorous quantitative evaluation, we show that color naming across languages is near-optimally efficient, as predicted by the IB principle. Furthermore, this finding suggests (1) a theoretical explanation for why empirically observed patterns of inconsistent naming and stochastic categories, which introduce ambiguity, are efficient for communication; and (2) that languages may evolve under pressure for efficient coding through an annealing-like process that synthesizes continuous and discrete aspects of previous accounts of color category evolution. This process generates quantitative predictions for how color naming systems may change over time. In Chapter 3, we directly test these predictions in one language, Nafaanra, by analyzing current color naming data for this language. We compare these data with similar data for this language collected 40 years ago, documenting recent language change and showing

that this change has occurred in a manner consistent with our predictions. In Chapter 4, we show that our approach generalizes to two qualitatively different domains: names for household containers, and for animal categories. Taken together, these findings suggest that pressure for efficient coding under limited resources, as defined by IB, may shape semantic systems across languages and across semantic domains.

Part II: Information-theoretic approach to communicative need. The second contribution of this thesis is an information-theoretic approach for characterizing communicative need. Communicative need is a central component in many efficiency-based approaches to language, including those reviewed in Section 1.3 and the IB approach presented in Part I. This component is formulated as a prior distribution over elements in the environment that reflects the frequency in which they are referred to during communication. There is evidence that this component may have substantial influence on semantic systems, however it has not been clear how to characterize and estimate it. We address this problem by invoking two general information-theoretic principles: the capacity-achieving principle, and the maximum-entropy (MaxEnt) principle. As before, we test this approach in the domain of color naming. In Chapter 5 we analyze communicative need through the lens of the capacity-achieving principle. This analysis suggests that color naming may be shaped by communicative need in interaction with color perception, as opposed to traditional accounts that focused mainly on perception and recent accounts that focused mainly on need. In Chapter 6, we present a systematic evaluation of several factors that may reflect the communicative need of colors in the environment: capacity constraints, linguistic usage, and the statistics of colors in the visual environment. By invoking the MaxEnt principle with word-frequency constraints, we show that linguistic usage may be the most relevant factor for characterizing the communicative need of colors. This MaxEnt approach is domain-general, and so it may also apply to communicative need in other semantic domains.

Part III: Evolution of compressed representations. The third contribution of this thesis touches on the theoretical foundations of Part I, by extending the mathematical understanding of the structure and evolution of optimal IB representations. This contribution is important also in a broader context, given the growing evidence for the applicability of the IB principle not only to language, but also to deep learning, neuroscience, and cognition. In Chapter 7, we study the case of discrete, or symbolic, IB representations, which corresponds to our application of IB to semantic systems. We characterize the structural changes in the IB representations as they evolve via a deterministic annealing process; derive an algorithm for finding critical points; and explore numerically the types of bifurcations and related phenomena that occur in IB. These phenomena and the theoretical justification for this approach apply to efficient symbolic representations in both humans and machines. Therefore, we believe that this approach could potentially guide the development of artificial intelligence systems with human-like semantics.

Part I

Efficient Compression in the Lexicon

Chapter 2

Efficient Compression in Color Naming and its Evolution

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937– 7942. DOI: 10.1073/pnas.1800521115.

Efficient compression in color naming and its evolution

Noga Zaslavsky[a,b,1], Charles Kemp[c,2], Terry Regier[b,d], and Naftali Tishby[a,e]

[a]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem 9190401, Israel; [b]Department of Linguistics, University of California, Berkeley, CA 94720; [c]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213; [d]Cognitive Science Program, University of California, Berkeley, CA 94720; and [e]The Benin School of Computer Science and Engineering, The Hebrew University, Jerusalem 9190401, Israel

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved June 18, 2018 (received for review January 11, 2018)

We derive a principled information-theoretic account of cross-language semantic variation. Specifically, we argue that languages efficiently compress ideas into words by optimizing the information bottleneck (IB) trade-off between the complexity and accuracy of the lexicon. We test this proposal in the domain of color naming and show that (*i*) color-naming systems across languages achieve near-optimal compression; (*ii*) small changes in a single trade-off parameter account to a large extent for observed cross-language variation; (*iii*) efficient IB color-naming systems exhibit soft rather than hard category boundaries and often leave large regions of color space inconsistently named, both of which phenomena are found empirically; and (*iv*) these IB systems evolve through a sequence of structural phase transitions, in a single process that captures key ideas associated with different accounts of color category evolution. These results suggest that a drive for information-theoretic efficiency may shape color-naming systems across languages. This principle is not specific to color, and so it may also apply to cross-language variation in other semantic domains.

information theory | semantic typology | color naming | categories | language evolution

Languages package ideas into words in different ways. For example, English has separate terms for "hand" and "arm," "wood" and "tree," and "air" and "wind," but other languages have single terms for each pair. At the same time, there are universal tendencies in word meanings, such that similar or identical meanings often appear in unrelated languages. A major question is how to account for such semantic universals and variation of the lexicon in a principled and unified way.

One approach to this question proposes that word meanings may reflect adaptation to pressure for efficient communication—that is, communication that is precise yet requires only minimal cognitive resources. On this view, cross-language variation in semantic categories may reflect different solutions to this problem, while semantic commonalities across unrelated languages may reflect independent routes to the same highly efficient solution. This proposal, focused on linguistic meaning, echoes the invocation of efficient communication to also explain other aspects of language (e.g., refs. 1–4).

Color is a semantic domain that has been approached in this spirit. Recent work has relied on the notion of the "informativeness" of word meaning, has often cast that notion in terms borrowed from information theory, and has accounted for several aspects of color naming across languages on that basis (5–10). Of particular relevance to our present focus, Regier, Kemp, and Kay (ref. 8, henceforth RKK) found that theoretically efficient categorical partitions of color space broadly matched major patterns of color naming seen across languages—suggesting that pressure for efficiency may indeed help to explain why languages categorize color as they do.

However, a fundamental issue has been left largely unaddressed: how a drive for efficiency may relate to accounts of color category evolution. Berlin and Kay (11) proposed an evolutionary sequence by which new terms refine existing partitions of color space in a discrete order: first dark vs. light, then red, then green and yellow, then blue, followed by other basic color categories. RKK's efficient theoretical color-naming systems correspond roughly to the early stages of the Berlin and Kay sequence, but they leave the transitions between stages unexamined and are based on the false (9, 12, 13) simplifying assumption that color-naming systems are hard partitions of color space. In actuality, color categories are a canonical instance of soft categories with graded membership, and it has been argued (12, 13) that such categories may emerge gradually in parts of color space that were previously inconsistently named. Such soft category boundaries introduce uncertainty and therefore might be expected to impede efficient communication (9). Thus, it remains an open question whether a hypothesized drive for efficiency can explain not just discrete stages of color category evolution, but also how systems evolve continuously from one stage to the next, and why inconsistent naming patterns are sometimes observed.

Here, we argue that a drive for information-theoretic efficiency provides a unified formal explanation of these phenomena. Specifically, we argue that languages efficiently compress ideas into words by optimizing the trade-off between the complexity and accuracy of the lexicon according to the information bottleneck (IB) principle (14), an independently motivated formal principle with broad scope (15–17), which is closely related (ref. 18 and *SI Appendix*, section 1.3) to rate distortion theory (19). We support this claim by showing that cross-language variation in color naming can be explained in IB terms. Our findings suggest that languages may evolve through a trajectory of efficient solutions in a single process that synthesizes, in formal terms, key ideas from Berlin and Kay's (11) theory and from more continuous accounts (12, 13) of color category evolution. We also show that soft categories and inconsistent naming can be information-theoretically efficient.

Significance

Semantic typology documents and explains how languages vary in their structuring of meaning. Information theory provides a formal model of communication that includes a precise definition of efficient compression. We show that color-naming systems across languages achieve near-optimal compression and that this principle explains much of the variation across languages. These findings suggest a possible process for color category evolution that synthesizes continuous and discrete aspects of previous accounts. The generality of this principle suggests that it may also apply to other semantic domains.

Author contributions: N.Z., C.K., T.R., and N.T. designed research; N.Z. performed research; N.Z. and N.T. contributed new reagents/analytic tools; N.Z. analyzed data; and N.Z. and T.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

[1]To whom correspondence should be addressed. Email: noga.zaslavsky@mail.huji.ac.il.

[2]Present address: School of Psychological Sciences, The University of Melbourne, Parkville, Victoria 3010, Australia.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1800521115/-/DCSupplemental.

Published online July 18, 2018.

Our work focuses on data compression, in contrast with work that views language in information-theoretic terms but focuses instead on channel capacity (2–4, 7, 20), including work on language evolution (21). Our work also further (e.g., refs. 7 and 22) links information theory to the study of meaning, a connection that has been contested since Shannon's (23) foundational work. IB has previously been used to find semantically meaningful clusters of words (ref. 15; see also ref. 22), but has not previously been used to account for word meanings as we do here.

Communication Model

To define our hypothesis precisely, we first formulate a basic communication scenario involving a speaker and a listener. This formulation is based on Shannon's classical communication model (23), but specifically concerns messages that are represented as distributions over the environment (Fig. 1). We represent the environment, or universe, as a set of objects $\mathcal{U}$. The state of the environment can be any object $u \in \mathcal{U}$, and we let U be a random variable that represents a possible state. We define a meaning to be a distribution $m(u)$ over $\mathcal{U}$ and assume the existence of a cognitive source that generates intended meanings for the speaker. This source is defined by a distribution $p(m)$ over a set of meanings, $\mathcal{M}$, that the speaker can represent. Each meaning reflects a subjective belief about the state of the environment. If the speaker's intention is $m \in \mathcal{M}$, this indicates that she wishes to communicate her belief that $U \sim m(u)$. We consider a color communication model in which $\mathcal{U}$ is restricted to colors and each $m \in \mathcal{M}$ is a distribution over colors.

The speaker communicates m by producing a word w, taken from a shared lexicon of size K. The speaker selects words according to a naming policy $q(w|m)$. This distribution is a stochastic encoder that compresses meanings into words. Because we focus on the uncertainty involved in compressing meanings into words, rather than the uncertainty involved in transmission, we assume an idealized noiseless channel that conveys its input unaltered as its output. This channel may have a limited capacity, which imposes a constraint on the available lexicon size. In this case, the listener receives w and interprets it as meaning $\hat{m}$ based on her interpretation policy $q(\hat{m}|w)$, which is a decoder. We focus on the efficiency of the encoder and therefore assume an optimal Bayesian listener with respect to the speaker (see *SI Appendix*, section 1.2 for derivation), who interprets every word w deterministically as meaning

$$\hat{m}_w(u) = \sum_{m \in \mathcal{M}} q(m|w)m(u), \qquad [1]$$

where $q(m|w)$ is obtained by applying Bayes' rule with respect to $q(w|m)$ and $p(m)$.

In this model, different color-naming systems correspond to different encoders, and our goal is to test the hypothesis that encoders corresponding to color-naming systems found in the world's languages are information-theoretically efficient. We next describe the elements of this model in further detail.

Encoders. Our primary data source for empirically estimating encoders was the World Color Survey (WCS), which contains color-naming data from 110 languages of nonindustrialized societies (24). Native speakers of each language provided names for the 330 color chips shown in Fig. 2, *Upper*. We also analyzed color-naming data from English, collected relative to the same stimulus array (25). We assumed that each color chip c is associated with a unique meaning m_c and therefore estimated an encoder $q_l(w|m_c)$ for each language l from the empirical distribution of word w given chip c (see data rows in Fig. 4 for examples). Each such encoder corresponds to a representative speaker for language l, obtained by averaging naming responses over speakers.

Meaning Space. In our formulation, colors are mentally represented as distributions. Following previous work (6, 8), we ground these distributions in an established model of human color perception by representing colors in 3D CIELAB space (Fig. 2, *Lower*) in which Euclidean distance between nearby colors is correlated with perceptual difference. We define the meaning associated with chip c to be an isotropic Gaussian centered at c, namely $m_c(u) \propto \exp\left(-\frac{1}{2\sigma^2}\|u-c\|^2\right)$. m_c reflects the speaker's subjective belief over colors that is invoked by chip c, and the scale of these Gaussians reflects her level of perceptual uncertainty. We take $\sigma^2 = 64$, which corresponds to a distance over which two colors can be comfortably distinguished (*SI Appendix*, section 6.3).

Cognitive Source. The cognitive source $p(m)$ specifies how often different meanings m must be communicated by a speaker. In principle, different cultures may have different communicative needs (8); we leave such language-specific analysis for future work and instead consider a universal source for all languages. Previous studies have used the uniform distribution for this purpose (8, 10); however, it seems unlikely that all colors are in fact equally frequent in natural communication. We therefore consider an alternative approach, while retaining the uniform distribution as a baseline. Specifically, we focus on a source that is derived from the notion of least informative (LI) priors (*Materials and Methods*), a data-driven approach that requires minimal assumptions. This approach also accounts for the data better than another approach based on image statistics (*SI Appendix*, section 7.2).

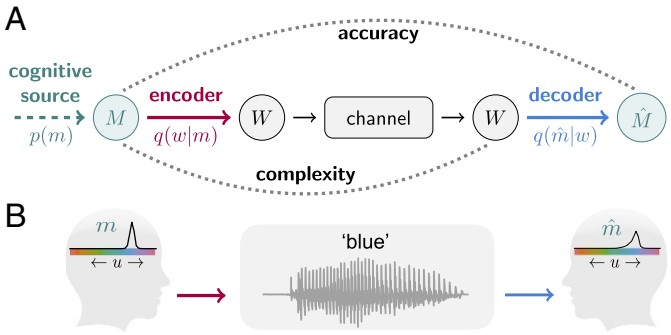

Fig. 1. (*A*) Shannon's (23) communication model. In our instantiation of this model, the source message M and its reconstruction $\hat{M}$ are distributions over objects in the universe $\mathcal{U}$. We refer to these messages as meanings. M is compressed into a code, or word, W. We assume that W is transmitted over an idealized noiseless channel and that the reconstruction $\hat{M}$ of the source message is based on W. The accuracy of communication is determined by comparing M and $\hat{M}$, and the complexity of the lexicon is determined by the mapping from M to W. (*B*) Color communication example, where $\mathcal{U}$ is a set of colors, shown for simplicity along a single dimension. A specific meaning m is drawn from $p(m)$. The speaker communicates m by uttering the word "blue," and the listener interprets blue as meaning $\hat{m}$.

Bounds on Semantic Efficiency

From an information-theoretic perspective, an optimal encoder minimizes complexity by compressing the intended message M as much as possible, while maximizing the accuracy of its interpretation $\hat{M}$ (Fig. 1*A*). In general, this principle is formalized by rate distortion theory (RDT) (19). In the special case in which messages are distributions, the IB principle (14) provides a natural formalization. In IB, as in RDT (*SI Appendix*, section 1.3), the complexity of a lexicon is measured by the number of bits of information that are required for representing the intended meaning. In our formulation the speaker represents her intended

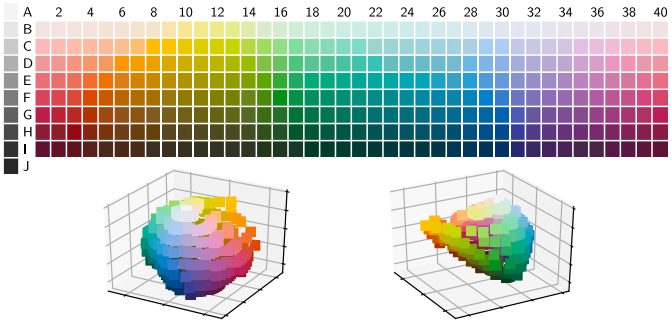

Fig. 2. (*Upper*) The WCS stimulus palette. Columns correspond to equally spaced Munsell hues. Rows correspond to equally spaced lightness values. Each stimulus is at the maximum available saturation for that hue/lightness combination. (*Lower*) These colors are irregularly distributed in 3D CIELAB color space.

meaning M by W, using an encoder $q(w|m)$, and thus the complexity is given by the information rate

$$I_q(M;W)=\sum_{m,w}p(m)q(w|m)\log\frac{q(w|m)}{q(w)}, \qquad [2]$$

where $q(w)=\sum_{m\in\mathcal{M}}p(m)q(w|m)$. Minimal complexity, i.e., $I_q(M;W)=0$, can be achieved if the speaker uses a single word to describe all her intended meanings. However, in this case the listener will not have any information about the speaker's intended meaning. To enable useful communication, W must contain some information about M; i.e., the complexity $I_q(M;W)$ must be greater than zero.

The accuracy of a lexicon is inversely related to the cost of a misinterpreted or distorted meaning. While RDT allows an arbitrary distortion measure, IB considers specifically the Kullback–Leibler (KL) divergence,

$$D\left[m\|\hat{m}\right]=\sum_{u\in\mathcal{U}}m(u)\log\frac{m(u)}{\hat{m}(u)}, \qquad [3]$$

which is a natural distortion measure between distributions. [For a general justification of the KL divergence see ref. 26, and in the context of IB see ref. 18.] Note that this quantity is 0 if and only if the listener's interpretation is accurate; namely, $\hat{m}\equiv m$. The distortion between the speaker and the ideal listener is the expected KL divergence,

$$\mathbb{E}_q\left[D[M\|\hat{M}]\right]=\sum_{m,w}p(m)q(w|m)D\left[m\|\hat{m}_w\right]. \qquad [4]$$

In this case, the accuracy of the lexicon is directly related to Shannon's mutual information,

$$\mathbb{E}_q\left[D[M\|\hat{M}]\right]=I(M;U)-I_q(W;U). \qquad [5]$$

Since $I(M;U)$ is independent of $q(w|m)$, minimizing distortion is equivalent to maximizing the informativeness, or accuracy, of the lexicon, quantified by $I_q(W;U)$. This means that mutual information appears in our setting as a natural measure both for complexity and for semantic informativeness.

If the speaker and the listener are unwilling to tolerate any information loss, the speaker must assign a unique word to each meaning, which requires maximal complexity. However, between the two extremes of minimal complexity and maximal accuracy, an optimal trade-off between these two competing needs can be obtained by minimizing the IB objective function,

$$\mathcal{F}_\beta[q(w|m)]=I_q(M;W)-\beta I_q(W;U), \qquad [6]$$

where $\beta\geq 1$ is the trade-off parameter. Every language l, defined by an encoder $q_l(w|m)$, attains a certain level of complexity and a certain level of accuracy. These two quantities can be plotted against each other. Fig. 3 shows this information plane for the present color communication model. The maximal accuracy that a language l can achieve, given its complexity, is bounded from above. Similarly, the minimal complexity that l can achieve given its accuracy is bounded from below. These bounds are given by the complexity and accuracy of the set of hypothetical IB languages that attain the minimum of Eq. **6** for different values of β. The IB curve is the theoretical limit defined by these optimal languages, and all trade-offs above this curve are unachievable.

Predictions

Near-Optimal Trade-offs. Our hypothesis is that languages evolve under pressure for efficient compression, as defined by IB, which implies that they are pressured to minimize $\mathcal{F}_\beta$ for some value of β. If our hypothesis is true, then for each language l there should be at least one value, β_l, for which that language is close to the optimal $\mathcal{F}^*_{\beta_l}$. If we are able to find a good candidate β_l for every language, this would support our hypothesis, because such an outcome would be unlikely given systems that evolved independently of $\mathcal{F}_\beta$. A natural choice for fitting β_l is the value of β that minimizes $\Delta\mathcal{F}_\beta=\mathcal{F}_\beta[q_l]-\mathcal{F}^*_\beta$. We measure the efficiency loss, or deviation from optimality, of language l by $\varepsilon_l=\frac{1}{\beta_l}\Delta\mathcal{F}_{\beta_l}$.

Structure of Semantic Categories. Previous work (e.g., ref. 8) has sometimes summarized color-naming responses across multiple speakers of the same language by recording the modal naming response for each chip, resulting in a hard categorical partition of the stimulus array, called a mode map (e.g., Fig. 4*A*). However, IB predicts that if some information loss is allowed, i.e., $\beta<\infty$, then an efficient encoder would induce soft rather than hard categories. This follows from the structure of the IB optima (14), given by

$$q_\beta(w|m)\propto q_\beta(w)\exp(-\beta D[m\|\hat{m}_w]), \qquad [7]$$

which is satisfied self-consistently with Eq. **1** and with the marginal $q_\beta(w)$. We therefore evaluate how well our model accounts for mode maps, but more importantly we also evaluate how well it accounts for the full color-naming distribution across

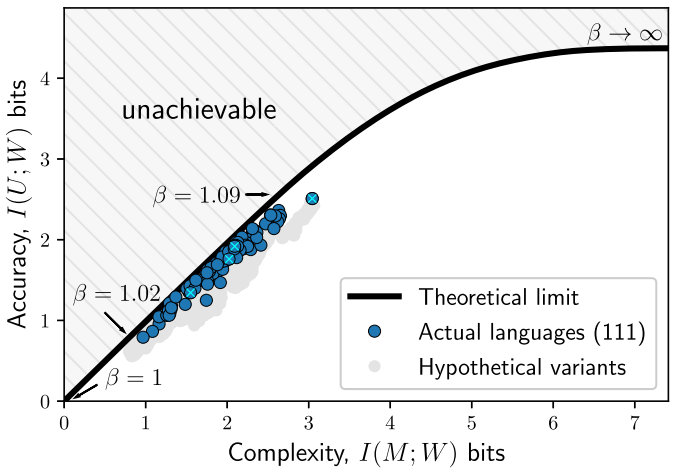

Fig. 3. Color-naming systems across languages (blue circles) achieve near-optimal compression. The theoretical limit is defined by the IB curve (black). A total of 93% of the languages achieve better trade-offs than any of their hypothetical variants (gray circles). Small light-blue Xs mark the languages in Fig. 4, which are ordered by complexity.

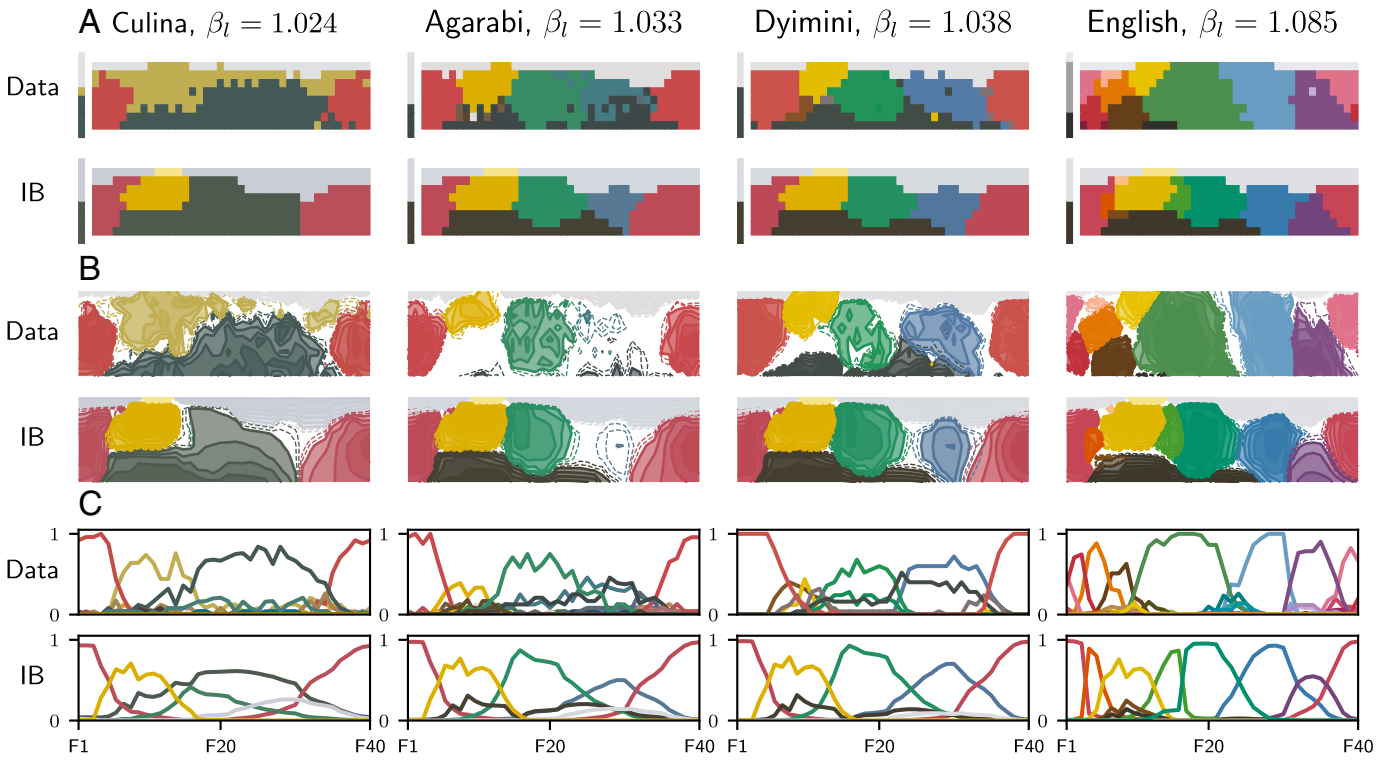

Fig. 4. Similarity between color-naming distributions of languages (data rows) and the corresponding optimal encoders at β_l (IB rows). Each color category is represented by the centroid color of the category. (*A*) Mode maps. Each chip is colored according to its modal category. (*B*) Contours of the naming distribution. Solid lines correspond to level sets between 0.5 and 0.9; dashed lines correspond to level sets of 0.4 and 0.45. (*C*) Naming probabilities along the hue dimension of row F in the WCS palette.

speakers of a given language. If languages achieve near-optimal trade-offs, and their category structure is similar to that of the corresponding IB encoders, this would provide converging support for our hypothesis. We evaluate the dissimilarity between the mode maps of q_l and q_{β_l} by the normalized information distance (NID) (27) and the dissimilarity between their full probabilistic structures by a generalization of NID to soft partitions (gNID) (*Materials and Methods*).

Results

We consider the color communication model with the IB objective of efficient compression (IB model) and, as a baseline for comparison, with RKK's efficiency objective (RKK+ model, see *SI Appendix*, section 4). We consider each model with the LI source and again with the uniform source. Because the LI source is estimated from the naming data, it is necessary to control for overfitting. Therefore, we performed fivefold cross-validation over the languages used for estimating the LI source. Table 1 shows that IB with the LI source provides the best account of the data. Similar results are obtained when estimating the LI source from all folds, and therefore the results with this source (*SI Appendix*, Fig. S1) are used for the figures. Table 1 and Fig. 3 show that all languages are near-optimally efficient with β_l that is only slightly greater than 1; this means that for color naming, maximizing accuracy is only slightly more important than minimizing complexity. These trade-offs correspond to the steepest part of the IB curve, in which every additional bit in complexity contributes the most to the accuracy of communication. In this sense, naturally occurring color-naming systems lie along the most active area of the curve, before the point of diminishing returns.

IB achieves 74% improvement in ε_l and 61% improvement in gNID compared to RKK+ with the LI source; however, the difference in NID is not substantial. Similar behavior appears with the uniform source. This result makes sense: The RKK+ bounds correspond to deterministic limits of suboptimal IB curves in which the lexicon size is restricted (*SI Appendix*, section 4.6). Because RKK's objective predicts deterministic color-naming systems, it can account for mode maps but not for full color-naming distributions.

Although Table 1 and Fig. 3 suggest that color-naming systems in the world's languages are near-optimally efficient, a possible objection is that perhaps most reasonable naming systems are near optimal according to IB, such that there is nothing privileged about the actual naming systems we consider. To rule out the possibility that IB is too permissive, we follow ref. 6 and construct for each language a control set of 39 hypothetical variants of that language's color-naming system, by rotating that naming system in the hue dimension across the columns of the WCS palette (*SI Appendix*, section 8). A total of 93% of the languages achieve better trade-offs than any of their hypothetical variants, and the remaining 7% achieve better trade-offs than most of their variants (Fig. 3).

The quantitative results in Table 1 are supported by visual comparison of the naming data with IB-optimal systems. Fig. 4 shows that IB accounts to a large extent for the structure of

Table 1. Quantitative evaluation via fivefold cross-validation

Source	Model	ε_I	gNID	NID	β_I
LI	IB	**0.18 (±0.07)**	**0.18 (±0.10)**	**0.31 (±0.07)**	1.03 (±0.01)
	RKK+	0.70 (±0.23)	0.47 (±0.10)	0.32 (±0.10)	
U	IB	0.24 (±0.09)	0.39 (±0.12)	0.56 (±0.07)	1.06 (±0.01)
	RKK+	0.95 (±0.22)	0.65 (±0.08)	0.50 (±0.10)	

Shown are averages over left-out languages ±1 SD for the LI and uniform (U) source distributions. Lower values of ε_I, gNID, and NID are better. Best scores are in boldface.

color naming in four languages with increasing complexity. Similar results for all languages are presented in *SI Appendix*, section 10. The category colors in Fig. 4 correspond to the color centroids of each category, and it can be seen that the data centroids are similar to the corresponding IB centroids. In addition, the IB encoders exhibit soft category boundaries and sometimes leave parts of color space without a clearly dominant name, as is seen empirically (9, 13). Note that the qualitatively different solutions along the IB rows are caused solely by small changes in β. This single parameter controls the complexity and accuracy of the IB solutions.

Tracking the IB centroids along the IB curve (Fig. 5) reveals a hierarchy of color categories. These categories evolve through an annealing process (28), by gradually increasing β (*SI Appendix*, Movie S1). During this process, the IB systems undergo a sequence of structural phase transitions, in which the number of distinguishable color categories increases—corresponding to transitions between discrete stages in Berlin and Kay's (11) proposal. Near these critical points, however, one often finds inconsistent, low-consensus naming—consistent with more continuous views of color category evolution (9, 12, 13). It is in this sense that the IB principle provides a single explanation for aspects of the data that have traditionally been associated with these different positions.

By assigning β_l to each language we essentially map it to a point on this trajectory of efficient solutions. Consider for example the languages shown in Figs. 4 and 5 (see *SI Appendix*, Movie S2 for more examples). Culina is mapped to a point right after a phase transition in which a green category emerges. This new green category does not appear in the mode maps of Fig. 4*A*, *Left* (data and IB), because it is dominated by other color categories, but it can be detected in Fig. 4*C*. Such dominated categories could easily be overlooked or dismissed as noise in the data, but IB predicts that they should exist in some cases. In particular, dominated categories tend to appear near criticality, as a new category gains positive probability mass. The color-naming systems of Agarabi and Dyimini are similar to each other and are mapped to two nearby points after the next phase transition, in which a blue category emerges. These two languages each have six major color categories; however, IB assigns higher complexity to Dyimini. The higher complexity for Dyimini is due to the blue category, which has a clear representation in Dyimini but appears at an earlier, lower consensus stage in Agarabi. *SI Appendix*, Movie S1 shows that low agreement around blue hues is predicted by IB for languages that operate around $1.026 \leq \beta_l \leq 1.033$, and this is consistent with several WCS languages (e.g., Aguacatec and Berik in *SI Appendix*, section 10; also ref. 29), as well as some other languages (9, 13).

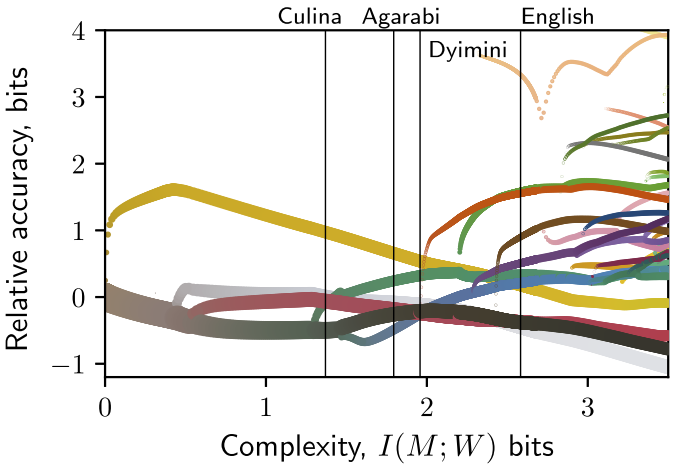

Fig. 5. Bifurcations of the IB color categories (Movie S1). The *y* axis shows the relative accuracy of each category *w* (defined in *Materials and Methods*). Colors correspond to centroids and width is proportional to the weight of each category, i.e., $q_\beta(w)$. Black vertical lines correspond to the IB systems in Fig. 4.

English is mapped to a relatively complex point in the IB hierarchy. The ability of IB to account in large part for English should not be taken for granted, since all IB encoders were evaluated according to a cognitive source that is heavily weighted toward the WCS languages, which have fewer categories than English. There are some differences between English and its corresponding IB system, including the pink category that appears later in the IB hierarchy. Such discrepancies may be explained by inaccuracies in the cognitive source, the perceptual model, or the estimation of β_l.

The main qualitative discrepancy between the IB predictions and the data appears at lower complexities. IB predicts that a yellow category emerges at the earliest stage, followed by black, white, and red. The main categories in low-complexity WCS languages correspond to black, white, and red, but these languages do not have the dominant yellow category predicted by IB. The early emergence of yellow in IB is consistent with the prominence of yellow in the irregular distribution of stimulus colors in CIELAB space (Fig. 2, *Lower Right*). One possible explanation for the yellow discrepancy is that the low-complexity WCS languages may reflect suboptimal yet reasonably efficient solutions, as they all lie close to the curve.

Discussion

We have shown that color-naming systems across languages achieve near-optimally efficient compression, as predicted by the IB principle. In addition, this principle provides a theoretical explanation for the efficiency of soft categories and inconsistent naming. Our analysis has also revealed that languages tend to exhibit only a slight preference for accuracy over complexity in color naming and that small changes in an efficiency trade-off parameter account to a large extent for the wide variation in color naming observed across languages.

The growth of new categories along the IB curve captures ideas associated with opposing theories of color term evolution (see also refs. 9 and 25). Apart from the yellow discrepancy, the successive refinement of the IB categories at critical points roughly recapitulates Berlin and Kay's (11) evolutionary sequence. However, the IB categories also evolve between phase transitions and new categories tend to appear gradually, which accounts for low-consensus regions (9, 12, 13). In addition, the IB sequence makes predictions about color-naming systems at complexities much higher than English and may thus account for the continuing evolution of high-complexity languages (25). This suggests a theory for the evolution of color terms in which semantic categories evolve through an annealing process. In this process, a trade-off parameter, analogous to inverse temperature in statistical physics, gradually increases and navigates languages toward more refined representations along the IB curve, capturing both discrete and continuous aspects of color-naming evolution in a single process.

The generality of the principles we invoke suggests that a drive for information-theoretic efficiency may not be unique to color naming. The only domain-specific component in our analysis is the structure of the meaning space. An important direction for future research is exploring the generality of these findings to other semantic domains.

Materials and Methods

Treatment of the Data. The WCS data are available online at www1.icsi.berkeley.edu/wcs. English data were provided upon request by Lindsey and Brown (25). Fifteen WCS languages were excluded from the LI source and from our quantitative evaluation, to ensure that naming probabilities for

each language were estimated from at least five responses per chip (*SI Appendix*, section 4.1).

LI Source. A source distribution can be defined from a prior over colors by setting $p(m_c) = p(c)$. For each language l, we constructed a LI source $p_l(c)$ by maximizing the entropy of c while also minimizing the expected surprisal of c given a color term w in that language (see *SI Appendix*, section 2 for more details). We obtained a single LI source by averaging the language-specific priors.

IB Curve. For each value of β the IB solution is evaluated using the IB method (14). IB is a nonconvex problem, and therefore only convergence to local optima is guaranteed. To mitigate this problem we fix $K = 330$ and use the method of reverse deterministic annealing to evaluate the IB curve (*SI Appendix*, section 1.4).

Dissimilarity Between Naming Distributions. Assume two speakers that independently describe m by $W_1 \sim q_1(w_1|m)$ and $W_2 \sim q_2(w_2|m)$. We define the dissimilarity between q_1 and q_2 by

$$\mathrm{gNID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\{I(W_1; W_1'), I(W_2; W_2')\}}, \quad [8]$$

where W_i' corresponds to another independent speaker that uses q_i. If q_1 and q_2 are deterministic, i.e., they induce hard partitions, then gNID reduces to NID (*SI Appendix*, section 3 for more details).

Relative Accuracy. We define the informativeness of a word w by

$$I_q(w) = D[\hat{m}_w \| m_0], \quad [9]$$

where $m_0(u) = \sum_m p(m)m(u)$ is the prior over u before knowing w. Note that the accuracy of a language can be written as $I_q(W; U) = \sum_w q(w)I_q(w)$, and therefore we define the relative accuracy of w (y axis in Fig. 5) by $I_q(w) - I_q(W; U)$.

ACKNOWLEDGMENTS. We thank Daniel Reichman for facilitating the initial stages of our collaboration, Delwin Lindsey and Angela Brown for kindly sharing their English color-naming data with us, Bevil Conway and Ted Gibson for kindly sharing their color-salience data with us, and Paul Kay for useful discussions. This study was supported by the Gatsby Charitable Foundation (N.T.), IBM PhD Fellowship Award (to N.Z.), and Defense Threat Reduction Agency (DTRA) Award HDTRA11710042 (to T.R.). Part of this work was done while N.Z. and N.T. were visiting the Simons Institute for the Theory of Computing at University of California, Berkeley.

1. Ferrer i Cancho R, Solé RV (2003) Least effort and the origins of scaling in human language. *Proc Natl Acad Sci USA* 100:788–791.
2. Levy RP, Jaeger TF (2007) Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, eds Schölkopf B, Platt JC, Hoffman T (MIT Press, Cambridge, MA), Vol 19, pp 849–856.
3. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108:3526–3529.
4. Gibson E, et al. (2013) A noisy-channel account of crosslinguistic word-order variation. *Psychol Sci* 24:1079–1088.
5. Jameson K, D'Andrade RG (1997) It's not really red, green, yellow, blue: An inquiry into perceptual color space. *Color Categories in Thought and Language*, eds Hardin CL, Maffi L (Cambridge Univ Press, Cambridge, UK), pp 295–319.
6. Regier T, Kay P, Khetarpal N (2007) Color naming reflects optimal partitions of color space. *Proc Natl Acad Sci USA* 104:1436–1441.
7. Baddeley R, Attewell D (2009) The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychol Sci* 20:1100–1107.
8. Regier T, Kemp C, Kay P (2015) Word meanings across languages support efficient communication. *The Handbook of Language Emergence*, eds MacWhinney B, O'Grady W (Wiley-Blackwell, Hoboken, NJ), pp 237–263.
9. Lindsey DT, Brown AM, Brainard DH, Apicella CL (2015) Hunter-gatherer color naming provides new insight into the evolution of color terms. *Curr Biol* 25:2441–2446.
10. Gibson E, et al. (2017) Color naming across languages reflects color use. *Proc Natl Acad Sci USA* 114:10785–10790.
11. Berlin B, Kay P (1969) *Basic Color Terms: Their Universality and Evolution* (Univ of California Press, Berkeley).
12. MacLaury RE (1997) *Color and Cognition in Mesoamerica: Constructing Categories as Vantages* (Univ of Texas Press, Austin, TX).
13. Levinson SC (2000) Yélî Dnye and the theory of basic color terms. *J Linguistic Anthropol* 10:3–55.
14. Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, eds Hajek B, Sreenivas RS (Univ of Illinois, Urbana, IL), pp 368–377.
15. Slonim N (2002) The information bottleneck: Theory and applications. PhD thesis (Hebrew Univ of Jerusalem, Jerusalem).
16. Shamir O, Sabato S, Tishby N (2010) Learning and generalization with the information bottleneck. *Theor Comput Sci* 411:2696–2711.
17. Palmer SE, Marre O, Berry MJ, Bialek W (2015) Predictive information in a sensory population. *Proc Natl Acad Sci USA* 112:6908–6913.
18. Harremoës P, Tishby N (2007) The information bottleneck revisited or how to choose a good distortion measure. *IEEE International Symposium on Information Theory*. Available at https://ieeexplore.ieee.org/document/4557285/. Accessed July 10, 2018.
19. Shannon CE (1959) Coding theorems for a discrete source with a fidelity criterion. *IRE Nat Conv Rec* 4:142–163.
20. Jaeger TF (2010) Redundancy and reduction: Speakers manage syntactic information density. *Cogn Psychol* 61:23–62.
21. Plotkin JB, Nowak MA (2000) Language evolution and information theory. *J Theor Biol* 205:147–159.
22. Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, ed Schubert LK (Association for Computational Linguistics, Stroudsburg, PA), pp 183–190.
23. Shannon C (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:623–656.
24. Cook RS, Kay P, Regier T (2005) The World Color Survey database: History and use. *Handbook of Categorization in Cognitive Science*, eds Cohen H, Lefebvre C (Elsevier, Amsterdam), pp 223–242.
25. Lindsey DT, Brown AM (2014) The color lexicon of American English. *J Vis* 14:17.
26. Csiszár I, Shields P (2004) Information theory and statistics: A tutorial. *Found Trends Commun Inf Theor* 1:417–528.
27. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR* 11:2837–2854.
28. Rose K (1998) Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE* 86:2210–2239.
29. Lindsey DT, Brown AM (2004) Color naming and color consensus: "Blue" is special. *J Vis* 4:55.

PNAS
www.pnas.org

Supplementary Information for

Efficient compression in color naming and its evolution

Noga Zaslavsky, Charles Kemp, Terry Regier and Naftali Tishby

Corresponding author: Noga Zaslavsky
E-mail: noga.zaslavsky@mail.huji.ac.il

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Contents

Movie Captions

Movie S1. Evolution of the IB color naming systems. Left panel: Bifurcation diagram, similar to Fig.5. This diagram shows the full range of IB solutions, whereas Fig.5 shows only the range relevant for the languages in our data. The black line indicates the location in the diagram that corresponds to the value of β. **Right panel:** Visualization (as in Fig.4) of the IB system that corresponds to β. The IB systems evolve as β gradually increases from $\beta = 1$, where there is only one category, to $\beta = 2^{13}$, where each color is mapped deterministically to its own unique category. In between these two extremes, the IB systems induce soft color categories. Structural phase transitions occur at critical values of β along this trajectory of efficient solutions, in which new categories appear. Low-consensus regions often appear in systems near these phase transitions.

Movie S2. Languages achieve near-optimal compression. Left panel: The red dot traces along the optimal systems on the IB curve (theoretical limit), while the blue dot follows nearby, indicating the position of selected languages just below the curve in the information plane. A total of 23 representative languages are shown, which were selected to demonstrate the range of empirical variation accommodated by the IB model and the relation of that variation to languages' positions near the IB curve. **Right panel:** Contour plots of the language's naming distribution (top) and the IB encoder (bottom) that correspond to the blue and red dots on the left panel, respectively. The IB systems captures much of the structural variability in the data, and even languages that are less similar to the IB systems are still highly efficient, as seen on the left panel.

Supporting Information Text

1. Theoretical framework

1.1. Summary of notation. We use capital letters to denote random variables (e.g. M and U), calligraphic letters to denote their support (e.g. $\mathcal{M}$ and $\mathcal{U}$), and lower case letters to denote a specific realization (e.g. m and u). In our formulation we consider a finite set of distributions $\mathcal{M}$. Each element in this set (i.e., each $m \in \mathcal{M}$) is a distribution over the set $\mathcal{U}$. In other words, m is a function that takes u as an argument. We use the notation $m(u)$ when we wish to make explicit that m is a function of u, or when we wish to denote the probability of a specific u according to m. It may be helpful to think of $m(u)$ in terms of conditional probabilities, i.e., $m(u) = p(u|m)$. Table S1 summarizes the notation used in the IB framework (1), in our current formulation of IB, in the framework of RKK (2) and in the adjusted RKK model (RKK+) which we constructed as a baseline for evaluation. A detailed description of RKK+ appears in section 4.

Table S1. Summary of notation

	Component	IB (1999)	IB (current)	RKK+ (current)	RKK (2015)
Communication model	Target variable / universe	$y \in \mathcal{Y}$	$u \in \mathcal{U}$	$u \in \mathcal{U}$	$t \in \mathcal{U}$
	Source variable	$x \in \mathcal{X}$	$m \in \mathcal{M}$	$m \in \mathcal{M}$	-
	Speaker's intended meaning	$p(y\|x)$	$m(u)$	$m(u)$	$s(t)$
	Source distribution / need	$p(x)$	$p(m)$	$p(m)$	$n(t)$
	Cluster / word	$\hat{x} \in \hat{\mathcal{X}}$	$w \in \mathcal{W}$	$w \in \mathcal{W}$	$w \in \mathcal{W}$
	Encoder / naming distribution	$q(\hat{x}\|x)$	$q(w\|m)$	$q(w\|m)$	$t \mapsto w$ if $t \in \text{cat}(w)$
	Decoder	$\hat{x} \mapsto q(y\|\hat{x})$	$q(\hat{m}\|w)$	$q(\hat{m}\|w)$	-
	Listener's interpreted meaning	$q(y\|\hat{x})$	$\hat{m}_w(u)$	$\hat{m}_w(u)$	$l(t)$
Optimization principle	Complexity	$I_q(X;\hat{X})$	$I_q(M;W)$	$\log K$	$K = \|\mathcal{W}\|$
	Distortion / communicative cost	$D[p(y\|x)\|\|q(y\|\hat{x})]$	$D[m\|\|\hat{m}]$	$D[m\|\|\hat{m}]$	$D[s\|\|l]$
	Accuracy	$I_q(\hat{X};Y)$	$I_q(W;U)$	$I_q(W;U)$	-
	Tradeoff parameter	β	β	-	-

1.2. Bayesian listener. We show that the ideal listener with respect to a given speaker is an optimal Bayesian decision maker. In our case, this means that we can assume an ideal listener that always decodes w deterministically by interpreting it as meaning $\hat{m}_w(u) = \sum_{m \in \mathcal{M}} q(m|w)m(u)$, where $q(m|w)$ is obtained by applying Bayes' rule,

$$q(m|w) = \frac{q(w|m)p(m)}{q(w)}\,, \qquad [S1]$$

where $q(w) = \sum_{m'} p(m')q(w|m')$. To show that this Bayesian listener is optimal, assume that the speaker's encoder is given by $q(w|m)$. The optimal listener for this speaker is defined by the decoder $q(\hat{m}|w)$ that minimizes

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) = I_q(M;W) - \beta \left(I(M;U) - \mathbb{E}_q \left[D[M\|\hat{M}] \right] \right) , \quad [S2]$$

where the second equality follows from Eq. (5). Note that $I(M;U)$ is constant in q and $I_q(M;W)$ depends on the encoder but not on the decoder. Only the last term depends on the decoder, and it holds that

$$\begin{aligned} \mathbb{E}_q \left[D[M\|\hat{M}] \right] &= \sum_{m,w,\hat{m}} p(m)q(w|m)q(\hat{m}|w)D\left[m\|\hat{m}\right] && [S3] \\ &= \sum_{m,w,\hat{m}} q(w)q(m|w)q(\hat{m}|w)D\left[m\|\hat{m}\right] && [S4] \\ &\geq \sum_{w} q(w) \operatorname*{argmin}_{\hat{m}'} \sum_{m} q(m|w)D\left[m\|\hat{m}'\right] && [S5] \end{aligned}$$

Therefore, there is a deterministic decoder $q(\hat{m}|w)$ that minimizes Eq. (S2),

$$q(\hat{m}|w) = \begin{cases} 1 & \text{if } \hat{m} = \operatorname*{argmin}_{\hat{m}'} \mathbb{E}_{q(m|w)} \left[D\left[m\|\hat{m}'\right] \right] \\ 0 & \text{otherwise} \end{cases} . \quad [S6]$$

Differentiating $\mathbb{E}_{q(m|w)} \left[D\left[m\|\hat{m}'\right] \right]$ with respect to $\hat{m}'$ and equating to 0 gives that the minimum is attained at $\hat{m}_w$. Since $\sum_u \hat{m}_w(u) = 1$ we did not need to impose this normalization constraint on the optimization, and because the KL divergence is convex in both arguments $\hat{m}_w$ is indeed the minimum.

1.3. Relation between IB and rate distortion theory. It has been shown that IB can be considered a special type of rate distortion (RD) with a variable distortion measure (3), and that the IB distortion measure has unique properties that distinguish IB from other RD problems (4). Furthermore, it was shown in (4) that IB can be considered a standard RD problem over probability measures, where the reconstruction alphabet is continuous. This view is closely related to the interpretation of IB as distributional clustering (5), in contrast to many applications of IB in the context of supervised learning (6). The setting we consider in this paper corresponds to a RD problem where M is compressed into $\hat{M}$. Although we are explicitly interested in the compression of M into a codeword W and in the reconstruction of $\hat{M}$ from W, it can be shown that the two problems are equivalent under mild assumptions.

A formal proof of this statement is beyond the scope of this work, but the main idea is that we can assume w.l.o.g. that the decoder is information lossless, i.e., $I_q(M;W) = I_q(M;\hat{M})$. In this case, minimizing $\mathcal{F}_\beta[q]$ is equivalent to minimizing the RD objective $I_q(M;\hat{M}) + \beta\mathbb{E}_q[D[M\|\hat{M}]]$, under the constraint $q(\hat{m}|m) = \sum_w q(w|m)\mathbf{1}_{[\hat{m}=\hat{m}_w]}$. It is possible to show that, under mild assumptions, this additional constraint on $q(\hat{m}|m)$ would not change the optimum of the RD problem. However, here we will only justify the assumption that the decoder is information lossless. Let $\varphi(w) = \hat{m}_w$ with respect to some encoder q. The decoder is information lossless if $\varphi(w)$ is a one-to-one mapping over the support of q (i.e., over $Sup(q) = \{w \in \mathcal{W} : q(w) > 0\}$). We can assume that this property holds, because otherwise it is possible to construct q' for which this property holds and $\mathcal{F}_\beta[q'] \leq \mathcal{F}_\beta[q]$. Assume there are $w_1, w_2 \in Sup(q)$ such that $w_1 \neq w_2$ and $\varphi(w_1) = \varphi(w_2)$. Define q' by merging them, namely for all m let $q'(w_1|m) = q(w_1|m) + q(w_2|m)$, $q'(w_2|m) = 0$, and for all $w \neq w_1, w_2$ let $q'(w|m) = q(w|m)$. This does not change the expected distortion; however, $I_{q'}(M;W) \leq I_q(M;W)$.

1.4. The IB method and deterministic annealing. Given a value of β, the IB method (1) iteratively updates the following IB equations until convergence,

$$\begin{aligned} q_\beta(w|m) &= \frac{q_\beta(w)}{Z(m;\beta)} \exp(-\beta D\left[m\|\hat{m}_w\right]) && [S7] \\ q_\beta(w) &= \sum_{m\in\mathcal{M}} q_\beta(w|m)p(m) && [S8] \\ \hat{m}_w(u) &= \sum_{m\in\mathcal{M}} m(u)q_\beta(m|w) , && [S9] \end{aligned}$$

where $Z(m;\beta)$ is the normalization factor. At the optimum, these equations are satisfied self-consistently. Because IB is a non-convex problem, the method of deterministic annealing (7) is often used to mitigate the problem of

converging to sub-optimal fixed points of the IB equations (e.g. 5, 8). Deterministic annealing starts at a low value of β ($\beta = 1$ in IB) where the solution is trivial, and then gradually increases β. For each β, the IB method is initialized with the solution found for the previous value of β. In practice, for better convergence, we evaluated the IB curve by reverse deterministic annealing (9); i.e., starting at a very high value of β, where the solution is given by a one-to-one mapping from M to W, and then gradually decreasing β. We repeated this process with 1500 values of β in $[1, 2^{13}]$.

2. Least informative source

How to accurately model a cognitive process that generates meanings for the speaker is an open question that is beyond the scope of this work. Instead, we wish to estimate a source distribution that is more realistic than the uniform distribution, but does not require prior knowledge. In this work we propose a general approach for doing so, based on the following observation: if a source distribution exists, it should be reflected somehow in the way people speak, i.e., in the naming distribution. Therefore, it makes sense to try to infer the source distribution directly from the naming data. We do so without making assumptions about the cognitive source by building on the notion of least informative priors. Our approach is domain-general; however, for simplicity we present it here in terms our color naming model. In section 7 we discuss other approaches for estimating the source distribution, and show that our conclusions also hold under these alternative source distributions.

2.1. Definition for a given language. We begin by defining a least informative prior over color chips, with respect to a given naming distribution $q_l(w|c)$. Because we assumed that each chip c is associated with a unique meaning m_c, any prior $p(c)$ induces a source distribution by setting $p(m_c) = p(c)$. One common approach for obtaining uninformative priors is by invoking the maximum entropy principle. However, in our case the maximum entropy distribution over color chips is simply the uniform distribution. Another natural approach in our setting is to find a distribution that maximizes the entropy of c while minimizing the expected uncertainty over c give a term w in the language. That is,

$$p_l(c) = \underset{p(c)}{\operatorname{argmax}}\ H(C) - H_q(C|W) \qquad \text{[S10]}$$

where $H_q(C|W) = -\sum_{c,w} p(c)q(w|c) \log \frac{q(c|w)}{p(c)}$ is the conditional entropy, and $q(c|w) = \frac{q(w|c)p(c)}{q(w)}$ is the posterior distribution of c given w.

This definition has two interesting interpretations, in addition to being a constrained maximum entropy distribution. First, note that

$$I_q(W;C) = \underset{p(c)}{\operatorname{argmax}}\ H(C) - H_q(C|W)\,, \qquad \text{[S11]}$$

which implies that $p_l(c)$ maximizes the mutual information between colors and words. This type of prior distribution is also called a capacity achieving prior, and can be evaluated using the Blahut-Arimoto algorithm (10, 11). Note that in the IB model, a language l would be maximally complex if the source distribution were defined from $p_l(c)$. This contrasts with the IB principle, which aims to minimize complexity. Second, $p_l(c)$ is considered the least informative prior over c in the sense that it minimizes information about the posterior $q(c|w)$ by maximizing the KL divergence between the prior and posterior. This interpretation follows from the identity

$$I_q(W;C) = \sum_w q(w) D[q(c|w) \| p(c)]\,, \qquad \text{[S12]}$$

and it is closely related to the notion of reference priors in Bayesian inference (12). Reference priors are considered objective priors in the sense that they depend solely on the given distribution $q(w|c)$, but not on other assumptions that may reflect subjective prior beliefs.

2.2. Estimation across languages. Our approach for estimating a LI source can be applied on a language-specific basis. However, we leave this language-specific analysis for future research and instead focus on estimating a single source distribution for all languages. We obtain this universal LI source by averaging across the language-specific LI priors, namely

$$p_{\mathrm{LI}}(m_c) = \frac{1}{L} \sum_{l=1}^{L} p_l(c)\,. \qquad \text{[S13]}$$

To control for overfitting and to test the ability of our model to generalize to languages which are not used for estimating the source, we performed 5-fold cross-validation over the languages that contribute to the average in Eq. (S13). Fifteen WCS languages were excluded from this process, to ensure that the naming probabilities for each

language were estimated from at least 5 responses for every chip. This regularization process is further explained in section 4.1, and the excluded 15 languages are listed in section 10. Section 10 contains the results for all 111 languages.

The full LI source, estimated by averaging over 96 languages, is shown in Fig.S1. This source distribution is non-uniform; however, it still has relatively high entropy, $H[p(m_c)] \approx 7.41$, compared to the maximal entropy $\log_2(330) \approx 8.36$. This means that the KL divergence between the LI source and the uniform source is roughly 1 bit.

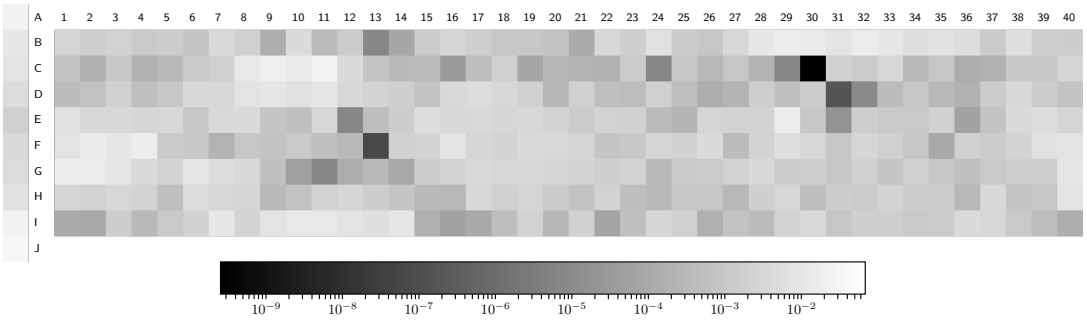

Fig. S1. The LI prior over the color chips obtained by averaging across the LI priors of 96 languages. Each chip is colored according to its probability mass, where black corresponds to probability 0 and white corresponds to probability 1. Gray colors are based on a logarithmic scale.

3. Dissimilarity measures

We compared different encoders, or color naming systems, by building on standard information-theoretic dissimilarity measures between clusterings (13). In our setting, these measures have an intuitive interpretation that relates them to the information between speakers of two languages.

Assume a language l_1 with lexicon $\mathcal{W}_1$ and an encoder $q_1(w_1|m)$, and a language l_2 with lexicon $\mathcal{W}_2$ and an encoder $q_2(w_1|m)$. In addition, assume that given a meaning $m \sim p(m)$, a speaker of l_1 produces a word $W_1 \sim q_1(w_1|m)$ and a speaker of l_2 independently produces a word $W_2 \sim q_2(w_2|m)$. The joint distribution of W_1 and W_2 is given by

$$q(w_1, w_2) = \sum_{m \in \mathcal{M}} p(m) q_1(w_1|m) q_2(w_2|m)\,. \qquad \text{[S14]}$$

Similarly, we can consider the joint distribution of two speakers of the same language that independently produce words W_i and W_i' given m,

$$q(w_i, w_i') = \sum_{m \in \mathcal{M}} p(m) q_i(w_i|m) q_i(w_i'|m)\,. \qquad \text{[S15]}$$

Intuitively, two languages are similar if the cross-language information $I(W_1; W_2)$ is large compared to the information within each language.

3.1. Normalized Information Distance (NID).

The normalized information distance (NID 13) is defined by

$$\mathrm{NID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\{H(W_1), H(W_2)\}}\,. \qquad \text{[S16]}$$

NID has been defined in (13) for hard partitions; i.e., in the case where $q(w|m)$ is a deterministic distribution. In this case NID has several desirable properties (13): it is a metric, it is bounded in the interval $[0, 1]$, and it was shown to outperform other methods for measuring similarity between hard clusterings. Therefore, we measured the distance between the mode maps that correspond to q_1 and q_2 by the NID between them.

3.2. Generalization of NID to soft partitions (gNID).

Although it is straightforward to apply the NID formula to soft partitions (soft-NID), we noticed that soft-NID is not sensitive enough to differences in the full probabilistic structure of the encoders. This can be seen in Fig.S2, which shows Dyimini for example. The soft-NID between Dyimini and different IB theoretical systems along the IB curve has a relatively flat part. This means that soft-NID can barely

distinguish between these different IB systems. We therefore slightly modified soft-NID in a way that also generalized NID to soft partitions. We define this generalization by

$$\mathrm{gNID}(W_1, W_2) = 1 - \frac{I(W_1; W_2)}{\max\{I(W_1, W_1'), I(W_2, W_2')\}} \,. \qquad [S17]$$

If $q_1(w_1|m)$ and $q_2(w_2|m)$ are both deterministic conditional distributions (i.e., W_1 and W_2 are selected deterministically given m), then gNID reduces to NID. To see this, notice that $I(W_i; W_i') = H(W_i) - H(W_i|W_i')$ and $H(W_i|W_i') = 0$ in the deterministic case.

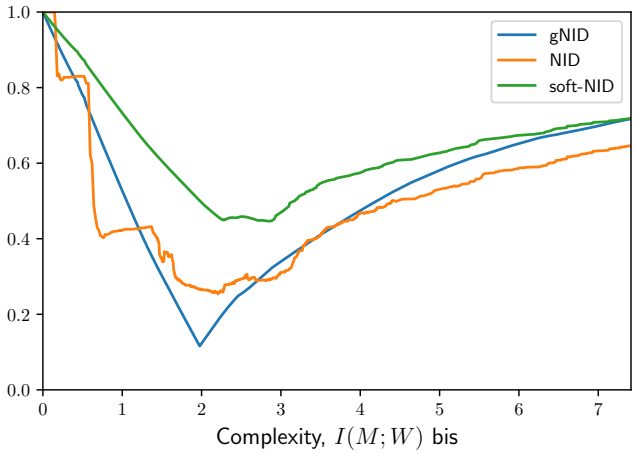

Fig. S2. Dissimilarity measures between the color naming system of Dyimini and the IB theoretical systems along the IB curve. gNID and soft-NID apply to the full distributions, whereas NID applies to their corresponding mode maps.

gNID has a few desirable properties. It holds that $\mathrm{gNID}(W_1, W_2) \leq 1$ because mutual information is non-negative, and $\mathrm{gNID}(W_1, W_2) = 1$ when W_1 and W_2 are independent because in that case $I(W_1; W_2) = 0$. When the two encoders are equivalent, then $\mathrm{gNID}(W_1, W_2) = 0$, as opposed to soft-NID which could be positive in this case. Although gNID in general is not necessarily non-negative, we did not encounter cases in which the gNID between a language's color naming distribution and an IB or RKK+ encoder was negative. In addition, for most languages gNID exhibits qualitatively similar behavior as seen for Dyimini (Fig.S2). That is, the gNID between the language and the IB systems follows a similar trend as NID and soft-NID; however, unlike NID and soft-NID, gNID is unimodal.

4. The RKK+ model

The RKK+ model is based on our communication model (Fig.1), but the definition of efficiency and the treatment of the data are derived from RKK's approach to color naming (2). Our communication model is very similar to RKK's communication model, although we relaxed a few assumptions made by RKK. In this section we discuss in detail the derivation of RKK+ from RKK's notion of efficiency, and explain the differences between the RKK+ model and RKK's color naming model. The mapping between our notation and RKK's notation is described in Table S1. For simplicity, we use here our notation for RKK+ and refer to the components of RKK's color naming model by the corresponding RKK+ notation.

4.1. Encoders based on major color terms. RKK's approach to the WCS color naming data relies on the notion of a major color term. According to RKK, w is a major color term in a given language, if it is the modal term for at least 10 color chips. Otherwise, w is considered a minor color term. For English, which was not included in RKK's color naming analysis, we set the threshold at 5 chips in order to obtain the 11 basic color terms in English. As in RKK's analysis, only data for major color terms is considered for the evaluation of the RKK+ model. That is, for each language l RKK+ considers a naming distribution $q_l^+(w|c)$ which is obtained from $q_l(w|c)$ by restricting it only to the major color terms in l. Restricting the data of a language to major terms may result in insufficient data for estimating the color naming distribution of that language. In 15 WCS languages some chips had fewer than 5 naming responses, and therefore we excluded these languages from the quantitative model evaluation and from the estimation of the LI source. These 15 languages are presented in section 10.

4.2. Relaxing RKK's assumptions.

Stochastic speaker. RKK made the simplifying assumption that the speaker chooses words deterministically, which induces hard partitions of the color space into named categories. For each color term w RKK defined $\text{cat}(w)$ by the set of colors that are named by w. This corresponds to a deterministic encoder: $q(w|c) = 1$ if $c \in \text{cat}(w)$ and 0 else. In RKK+ this assumption was relaxed because the encoder in our communication model can be stochastic.

Perceptual uncertainty. RKK assumed that the speaker has no perceptual uncertainty, which means that colors are represented by delta functions, i.e., $m_c(u) = \delta_{c,u}$. In our model we allow for perceptual uncertainty and instead assume that each color c is represented by the Gaussian m_c.

Bayesian listener. RKK assumed that the listener's interpreted meanings take the form

$$l_w(u) \propto \sum_{c \in \text{cat}(w)} \exp\left(-\frac{1}{2\sigma^2}\|u-c\|^2\right) . \qquad \text{[S18]}$$

Although this form is justified (2), we show in section 4.4 that a similar form can emerge directly from the need for efficiency. Therefore, we waive this assumption and consider a listener who is adapted to the speaker without additional constraints, as in the IB model.

4.3. Efficiency according to RKK. RKK argued that theoretically efficient languages minimize a communicative cost for a given level of complexity. We next present their definitions of complexity and communicative cost, and discuss the specific form these measures take in RKK+.

Complexity. RKK's notion of complexity is derived from the minimum description length principle on a domain-specific basis. In the domain of color, RKK defined the complexity of a language by the number of major terms in that language, denoted here by K. In RKK+ we slightly adjust this complexity measure and consider instead $\log K$. This does not change the essence of the measure nor the structure of the theoretically optimal systems, but allows us to measure complexity in bits, as in IB.

Communicative cost. RKK defined the error between the speaker's intended meaning and listener's interpreted meaning by the KL divergence between these two distributions. This definition coincides with the distortion measure in IB. RKK's communicative cost is the expected error, as it corresponds to the expected distortion in IB. Following the same argument as in section 1.2, we obtain that the ideal listener in RKK+ takes the same form as in IB; i.e., it is given by $\hat{m}_w$. Therefore, in RKK+ the communicative cost of an encoder $q(w|m)$ is given by

$$D[q] = \sum_{m,w} p(m)q(w|m)D\left[m\|\hat{m}_w\right] . \qquad \text{[S19]}$$

This definition is the same as Eq. (S3), but in RKK+ it applies to q_l^+ instead of q_l. We can therefore apply Eq. (5) to inversely relate the communicative cost $D[q_l^+]$ to the accuracy of the language according to RKK+. The complexity-accuracy pairs of the languages we considered, according to RKK+, are shown in Fig.S3.

4.4. Structure of the solution. An optimal speaker-listener pair in RKK+ jointly minimizes the expected distortion between them, for a given K. The hard constraint on the number of major terms is enforced by only considering encoders $q(w|m)$ over K terms. We have already seen that optimizing this distortion with respect to the speaker's interpreted meanings, while fixing the speaker's encoder, gives $\hat{m}_w$. Now, fix $\hat{m}_w$ and consider the encoder that minimizes Eq. (S19). Since this objective is linear in $q(w|m)$ the minimum is attained at

$$q(w|m) = \begin{cases} 1 & \text{if } w = w_m, \quad \text{where } w_m = \underset{w'}{\text{argmin}}\, D\left[m\|\hat{m}_{w'}\right] \\ 0 & \text{otherwise} \end{cases} . \qquad \text{[S20]}$$

Formally, this can be shown by following a similar argument as in section 1.2. This means that even though we relaxed the assumption that the speaker is deterministic, RKK+ does not predict any advantage for non-deterministic speakers that induce soft categories, and the theoretically optimal RKK+ systems can be characterized by hard partitions of color space. We can therefore define $\text{cat}(w)$ as in RKK's color naming model, namely $\text{cat}(w) = \{m \in \mathcal{M} : w_m = w\}$. Plugging back this encoder into the formula of $\hat{m}_w$ (i.e., Eq. (1)) and substituting the structure of m_c, gives a similar form as RKK's assumed listener in Eq. (S18).

4.5. Evaluation of the RKK+ bounds. The RKK+ fixed points are characterized by the self-consistent $\text{cat}(w)$ and $\hat{m}_w$. This suggests an iterative algorithm for finding these fixed points, which can be considered as K-Means over distributions where the KL divergence is used as the dissimilarity function. Denote by $C_w^{(t)}$ the set $\text{cat}(w)$ obtained at the t-th iteration of the algorithm. The algorithm can be described as follows:

- Initialize $C_w^{(0)}$ for $w \in \{1, \ldots, K\}$
- For $t = 1, \ldots$ (until convergence) update:

$$\hat{m}_w^{(t)}(u) = \frac{1}{|C_w^{(t-1)}|} \sum_{m \in C_w} m(u) \qquad [S21]$$

$$C_w^{(t)} = \left\{ m \in \mathcal{M} : w = \underset{w'}{\text{argmin}}\, D[m \| \hat{m}_{w'}^{(t)}] \right\} \qquad [S22]$$

This is a non-convex optimization problem, and only convergence to a local optimum is guaranteed. Therefore, for each K we repeated this algorithm 300 times with random initializations, and selected the best result. We evaluated the RKK+ bounds for $K = 2, \ldots, 11$. These bounds are shown in Fig.S3 (orange bars). The number of major terms in the languages we considered varies between 3-11.

4.6. Relation to IB. RKK+ is equivalent to IB when the lexicon size is restricted to K terms, and when $\beta \to \infty$. To see this, notice that taking $\beta \to \infty$ means that the speaker and listener only care about accuracy, and therefore minimizing $\mathcal{F}_\beta$ amounts to only minimizing the expected distortion. For every K we can evaluate the IB solution for $0 \leq \beta < \infty$, where the hard constraint on the lexicon size is imposed by considering only encoders $q(w|m)$ over a lexicon of size K. While the optimal IB curve is estimated for $K = |\mathcal{M}|$ (see 4), for smaller values of K we can obtain sub-optimal IB curves. This means that the IB curve upper bounds the RKK+ bounds. This relation between IB and RKK+ can be seen in Fig.S3.

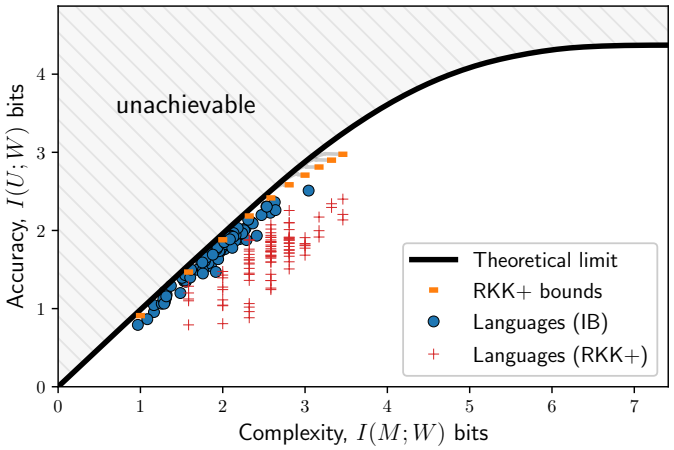

Fig. S3. Comparison between IB and RKK+. Complexity-accuracy values for all languages according to IB (blue dots) and RKK+ (red crosses). The IB curve (black) is evaluated for $K = 330$, and it defines the theoretical limit of achievable tradeoffs, including those achieved by the optimal systems according to RKK+. RKK+ bounds (orange bars) correspond to the deterministic limits of sub-optimal IB curves (gray curves) obtained by restricting the lexicon size to $K = 2, \ldots, 11$. The efficiency of the languages according to each model is evaluated with respect to the model's bounds.

5. Quantitative evaluation and variants of the IB model

Our goal in comparing IB with RKK+ is to test which principle can account better for the data, while holding all other elements of the model constant. Although IB and RKK+ are defined over the same communication model, there are two differences in the way these models treat the data: (1) RKK+ only considers major terms while IB considers the full set of naming responses, and (2) RKK+ evaluates each language against an optimal system with the same complexity, whereas in the IB model each language is evaluated against an optimal system at β_l which may have a different complexity than that of the language. We controlled for these differences by considering two variants of

the IB model that match how RKK+ treats the data. We show here that the results in both cases are similar to our main evaluation, which suggests that these two differences are mainly technical and do not impact our conclusions.

We evaluate RKK+ in the same way we evaluate IB. Namely, we are interested in (A) whether color naming systems across languages are near-optimally efficient according to RKK+; and (B) how well a theoretically optimal RKK+ encoder for a given K_l can explain the structure of the color naming distribution q_l^+ in languages with K_l major color terms. We use the same quantitative measures for evaluating IB and RKK+, namely ε_l, gNID and NID, where ε_l is defined with respect to the objective in each model. Although there is no tradeoff parameter in RKK+, the definition of ε_l coincides with the definition of ε_l in IB, because in RKK+ the complexity term cancels out. Recall that for IB we defined

$$\varepsilon_l = \frac{1}{\beta_l}\left(\mathcal{F}_{\beta_l}[q_l] - \mathcal{F}^*_{\beta_l}\right) = \frac{1}{\beta_l}\left(I_{q_l}(W;M) - I_{q_{\beta_l}}(W;M)\right) - \left(I_{q_l}(W;U) - I_{q_{\beta_l}}(W;U)\right). \quad [S23]$$

From Eq. (5) we get that

$$\varepsilon_l = \frac{1}{\beta_l}\left(I_{q_l}(W;M) - I_{q_{\beta_l}}(W;M)\right) + \left(D[q_l] - D[q_{\beta_l}]\right). \quad [S24]$$

If q_l and q_{β_l} have the same complexity then we get that $\varepsilon_l = D[q_l] - D[q_{\beta_l}]$. In RKK+ we have $\varepsilon_l = D[q_l^+] - D[q_{K_l}]$, where q_{K_l} is an optimal RKK+ encoder for K_l.

5.1. IB with constrained complexity. We considered a variant of the IB model in which β_l is determined such that the complexity at β_l matches the language's complexity (IB-C). Formally, this means that in IB-C β_l is selected such that $I_{q_l}(W;M) = I_{q_{\beta_l}}(W;M)$ and therefore $\varepsilon_l = D[q_l] - D[q_{\beta_l}]$. Table S2 shows the results for IB-C, together with the results for IB and RKK+ that are reported in main text (Table 1). The differences between IB and IB-C are not substantial, both for the LI source and for the uniform source. Therefore, our conclusions hold even for IB-C.

Table S2. Quantitative evaluation via fivefold cross-validation (including IB-C)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.18 (±0.07)	0.18 (±0.10)	0.31 (±0.07)	1.03 (±0.01)
	IB-C	0.18 (±0.07)	0.21 (±0.08)	0.31 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.47 (±0.10)	0.32 (±0.10)	
U	IB	0.24 (±0.09)	0.39 (±0.12)	0.56 (±0.07)	1.06 (±0.01)
	IB-C	0.24 (±0.09)	0.40 (±0.10)	0.56 (±0.08)	1.07 (±0.02)
	RKK+	0.95 (±0.22)	0.65 (±0.08)	0.50 (±0.10)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

5.2. IB for major color terms. Applying RKK+ to both major and minor terms can only increase the gap between the performance of RKK+ and IB. This is because in some languages there are many low frequency terms which do not much affect the partition of color space, however the optimal RKK+ encoders are very much affected by K. IB is more robust to low frequency terms, because the informational complexity in IB takes this into account by considering the frequency of each term. Therefore, we considered a variant of IB and a variant of IB-C in which they are applied to the color naming distributions restricted to major terms, i.e., to q_l^+ instead of q_l. Table S3 shows that the results in this case are not substantially different from the results in Table S2, which correspond to our main evaluation. Therefore, our conclusions hold whether or not the data are restricted to major color terms.

Table S3. Quantitative evaluation via fivefold cross-validation (based only on major color terms)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.14 (±0.06)	0.20 (±0.11)	0.31 (±0.07)	1.03 (±0.01)
	IB-C	0.14 (±0.06)	0.20 (±0.09)	0.31 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.47 (±0.10)	0.32 (±0.10)	
U	IB	0.19 (±0.07)	0.42 (±0.12)	0.57 (±0.07)	1.06 (±0.01)
	IB-C	0.19 (±0.07)	0.40 (±0.10)	0.56 (±0.08)	1.07 (±0.02)
	RKK+	0.95 (±0.22)	0.65 (±0.08)	0.50 (±0.10)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

6. Foundational assumptions

In this section we examine the foundational assumptions of our communication model more closely, and discuss the robustness of our results to these assumptions.

6.1. Choice of color space. Our model is based on the assumption that colors are represented in CIELAB space. To test the robustness of our results to this assumption, we repeated our full analysis with colors that are represented in the CIELUV color space (similarly to (14)) instead of CIELAB. Apart from this, all the other assumptions and methods were kept fixed. Table S4 shows quantitatively that this analysis yields similar results as the main analysis which is based on the CIELAB assumption. In particular, in both cases IB with the LI source provides the best account of the data. This conclusion is also supported by the qualitative results shown in Fig.S4 and in Fig.S5A, which are very similar to the corresponding results based on the CIELAB space. The main difference appears to be in the bifurcation diagram (Fig.S5B), where a red category appears much earlier compared to the results based on CIELAB.

Table S4. Quantitative evaluation via fivefold cross-validation (based on CIELUV)

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.14 (±0.06)	0.19 (±0.10)	0.30 (±0.09)	1.02 (±0.01)
	RKK+	0.71 (±0.23)	0.45 (±0.10)	0.29 (±0.10)	
U	IB	0.19 (±0.08)	0.36 (±0.12)	0.54 (±0.11)	1.03 (±0.01)
	RKK+	0.97 (±0.24)	0.66 (±0.07)	0.51 (±0.09)	

Averages over left-out languages ±1 SD for the least informative (LI) and uniform (U) source distributions. Lower values of ε_l, gNID and NID are better.

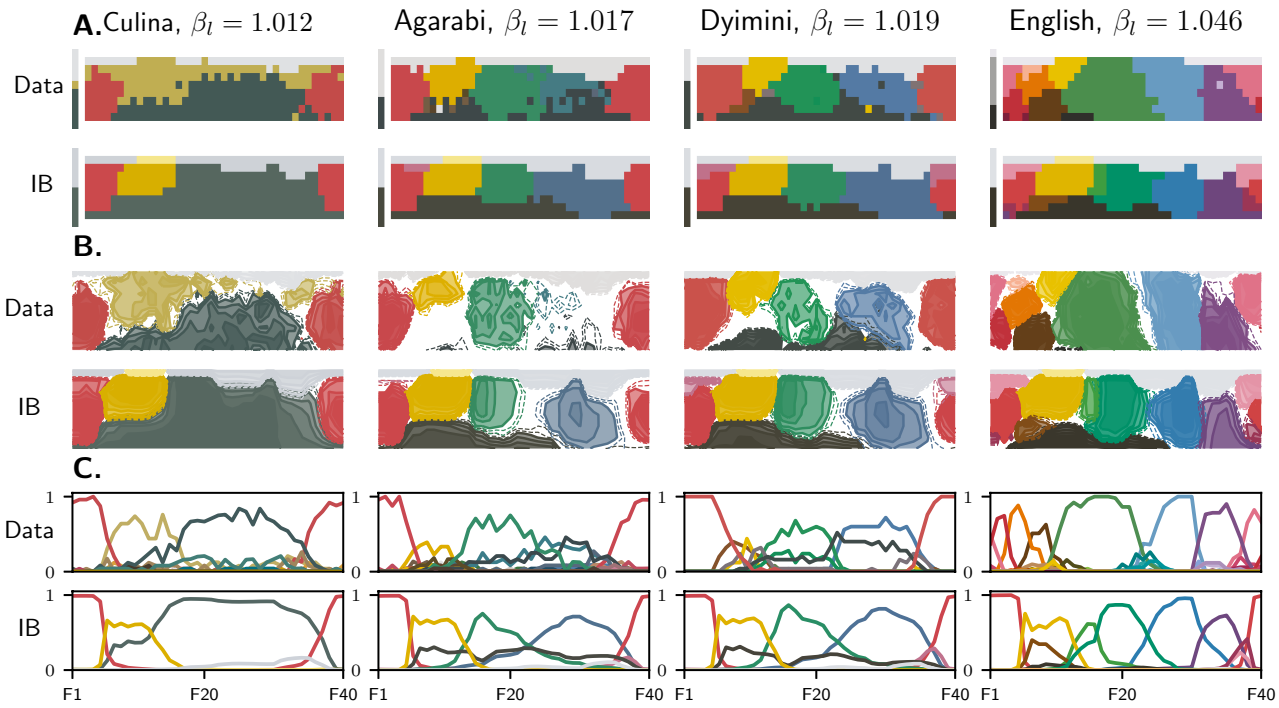

Fig. S4. CIELUV space. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F (**C**), similar to Fig.4 in main text but based on the results for CIELUV instead of CIELAB.

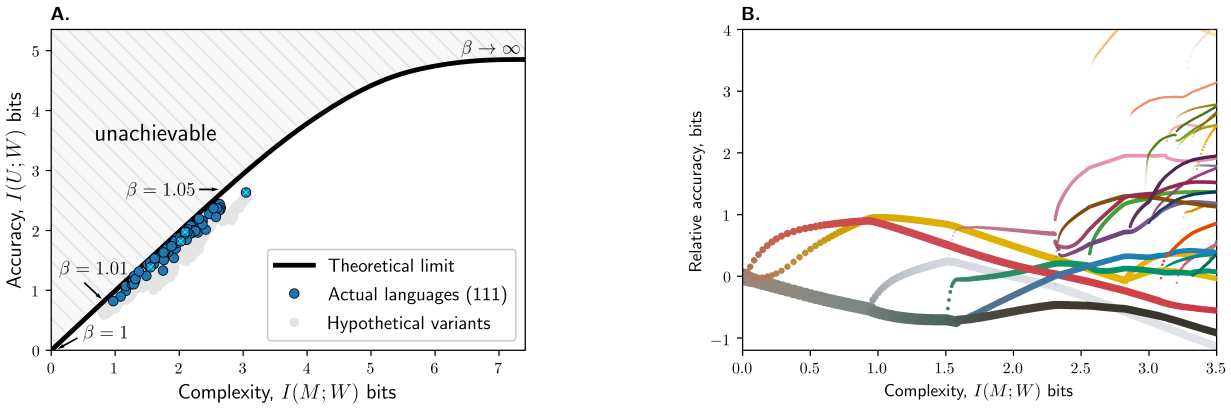

Fig. S5. CIELUV space. Information plane (**A**) and bifurcation diagram (**B**) for the full LI source. These figures are similar to Fig.3 and Fig.5 in main text, but they are based on the results for CIELUV instead of CIELAB.

6.2. Category effects and biological constraints. By grounding our model in a presumed universal perceptual color space such as CIELAB, we have implicitly assumed that this underlying representation is not affected by language. However, it is known that in fact there are lexical effects on the perceived similarity of colors (e.g. 15). While distances between colors in CIELAB may have been influenced to some extent by such category effects, we believe it is unlikely that this has introduced a substantial bias to our model. One reason for this belief is that our model is able to account for wide cross-language variation in color naming based on the same underlying perceptual space for all languages. Another reason is that category effects on color memory (e.g. 16, 17) have themselves been accounted for by assuming the same universal perceptual space, CIELAB, combined with knowledge of language-specific categories (18). These outcomes, which are consistent with a universal perceptual space, seem unlikely given a perceptual space that is instead strongly biased toward lexical categorization in one language, such as English.

It has recently been shown that pre-linguistic infants exhibit categorical distinctions that resemble common patterns in the WCS data (14), and this finding has been taken to suggest a pre-linguistic biological basis for color categorization. That conclusion is broadly consistent with our assumption of a universal color space, although our analysis is based solely on data from adults, and we do not attempt to directly engage the question of color categorization in infants.

6.3. Perceptual uncertainty. The color meaning space ($\mathcal{M}$) that we assumed has a free parameter, σ^2, that determines the speaker's level of perceptual uncertainty. We set $\sigma^2 = 64$ based on a result reported in (19) which suggested that this value corresponds to a distance over which two colors can be comfortably distinguished. To further justify this setting, we evaluated our IB model with a higher ($\sigma^2 = 500$) and lower ($\sigma^2 = 36$) level of perceptual uncertainty. The higher value, $\sigma^2 = 500$, corresponds to a level of perceptual uncertainty that has been used in previous studies (e.g. 2, 20). Table S5 shows the quantitative results for our IB model with different levels of perceptual uncertainty, and with respect to the full LI source. It can be seen that under higher uncertainty, the model is slightly worse on all three measures. Under lower uncertainty the model is slightly better in terms of ε_l but slightly worse in terms of gNID. This suggests that the value of σ^2 that we used is in a reasonable region; however, slightly lower values could perhaps improve the model. This remains a question for future work.

Table S5. Evaluation of IB with different levels of perceptual uncertainty.

	σ^2	ε_l	gNID	NID	β_l
Lower perceptual uncertainty	36	0.13 (±0.06)	0.23 (±0.11)	0.31 (±0.08)	1.01 (±0.01)
Baseline (main model)	64	0.18 (±0.07)	0.18 (±0.1)	0.31 (±0.07)	1.03 (±0.01)
Higher perceptual uncertainty	500	0.26 (±0.06)	0.31 (±0.12)	0.41 (±0.08)	1.77 (±0.20)

Numbers correspond to averages over languages ±1 SD. Lower values are better for ε_l, gNID and NID.

6.4. Validity of the WCS protocol. In the WCS protocol, field workers were instructed to encourage participants to provide short color terms. In practice, these instructions were not applied equally across languages, and in some languages this biased the free naming task towards frequently used terms. This raises a concern about the quality of

the WCS data and questions results based on these data. Gibson et al. (21) addressed this issue by comparing color naming data they collected in a free naming task and in a fixed naming task, and showing that their results were robust to these two conditions. To assure that our results were also not influenced by this issue, we applied a similar approach to our analysis.

Specifically, we considered the English color naming data that were collected by Lindsey and Brown (LB, 22) in a free naming task. LB used an improved experimental protocol for this task, and therefore the quality of their data is irrefutable. We also considered a modified version of these data which is based only on major terms (MT data), as described in section 4.1. Fig.S6D shows that the complexity and accuracy values evaluated from the LB data and the MT data are very similar. In addition, Fig.S6A-Fig.S6C show that the naming distribution estimated from the LB data is fairly similar to the naming distribution estimated from the MT data, and that the IB predictions are also similar in both cases. This suggests that our information-theoretic analysis is robust to restricting the naming responses to major terms, and thus the WCS data can be considered reliable in our setting.

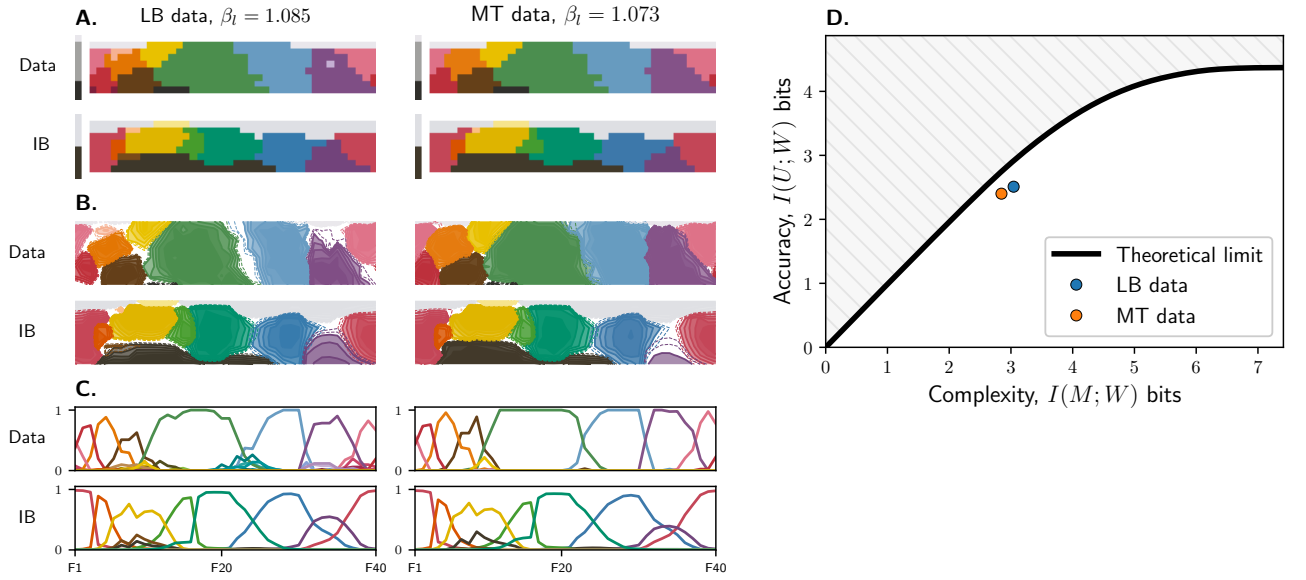

Fig. S6. English color naming data. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), as in Fig.4. Data rows correspond to the English color naming distribution estimated from the LB data (left), which considers all color terms, and from the modified MT data (right), which was restricted to major color terms. **D.** Complexity and accuracy evaluated based on the LB data the modified MT data.

7. Alternative source distributions

In this section we examine two alternatives to the LI source – the uniform distribution, which we used as a baseline for evaluation, and another approach based on image statistics.

7.1. Uniform distribution. The quantitative results for the uniform source are reported in the main text. We complete this picture by presenting Fig.S7A, Fig.S7B and Fig.S8, which are analogous to Fig.3, Fig.5 and Fig.4 in the main text, but were evaluated for the uniform source. In this case, the languages in our data also lie near the theoretical limit (Fig.S7), although not as close as they do with the LI source (this can be seen by comparing ε_l for IB under the uniform and LI source in Table 1). In addition, although both IB and RKK+ capture some of the structure in the data even with the uniform source (Fig.S8), this fit does not look as good as the fit based on the LI source (Fig.4 and section 10). This is consistent with Table 1, which quantitatively shows that the LI source improves the similarity between each model and the data.

Note that since the uniform source does not take into account communicative needs, the IB model with this source only reflects properties of the perceptual CIELAB space that are extracted by IB. The bifurcation diagram (Fig.S7B) in this case reveals a similar yellow discrepancy as observed for the LI source, in which a yellow category emerges at the earliest stage. This suggests that the yellow discrepancy is directly related to the irregular distribution of stimulus colors in CIELAB space.

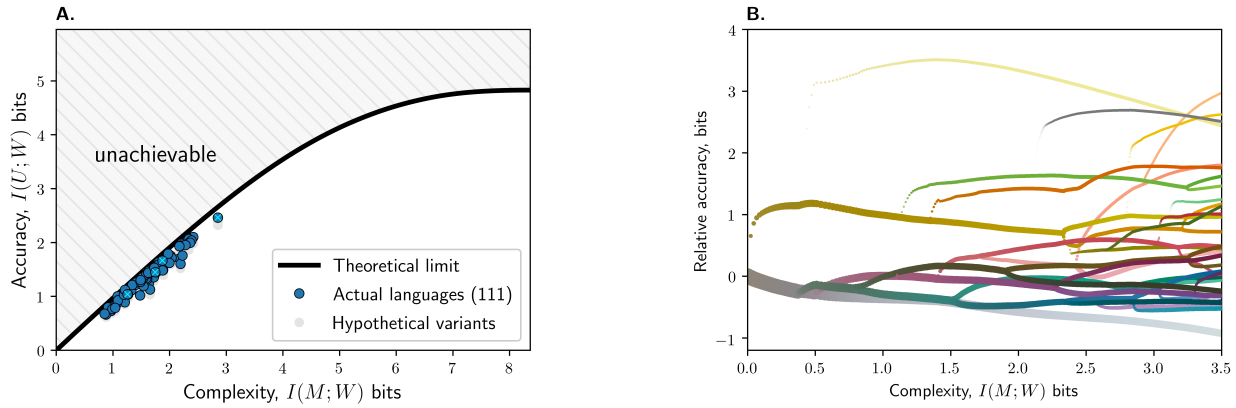

Fig. S7. Uniform source. Information plane (**A**) and bifurcation diagram (**B**) evaluated for the uniform source. For more details see captions of Fig.3 and Fig.5 in main text.

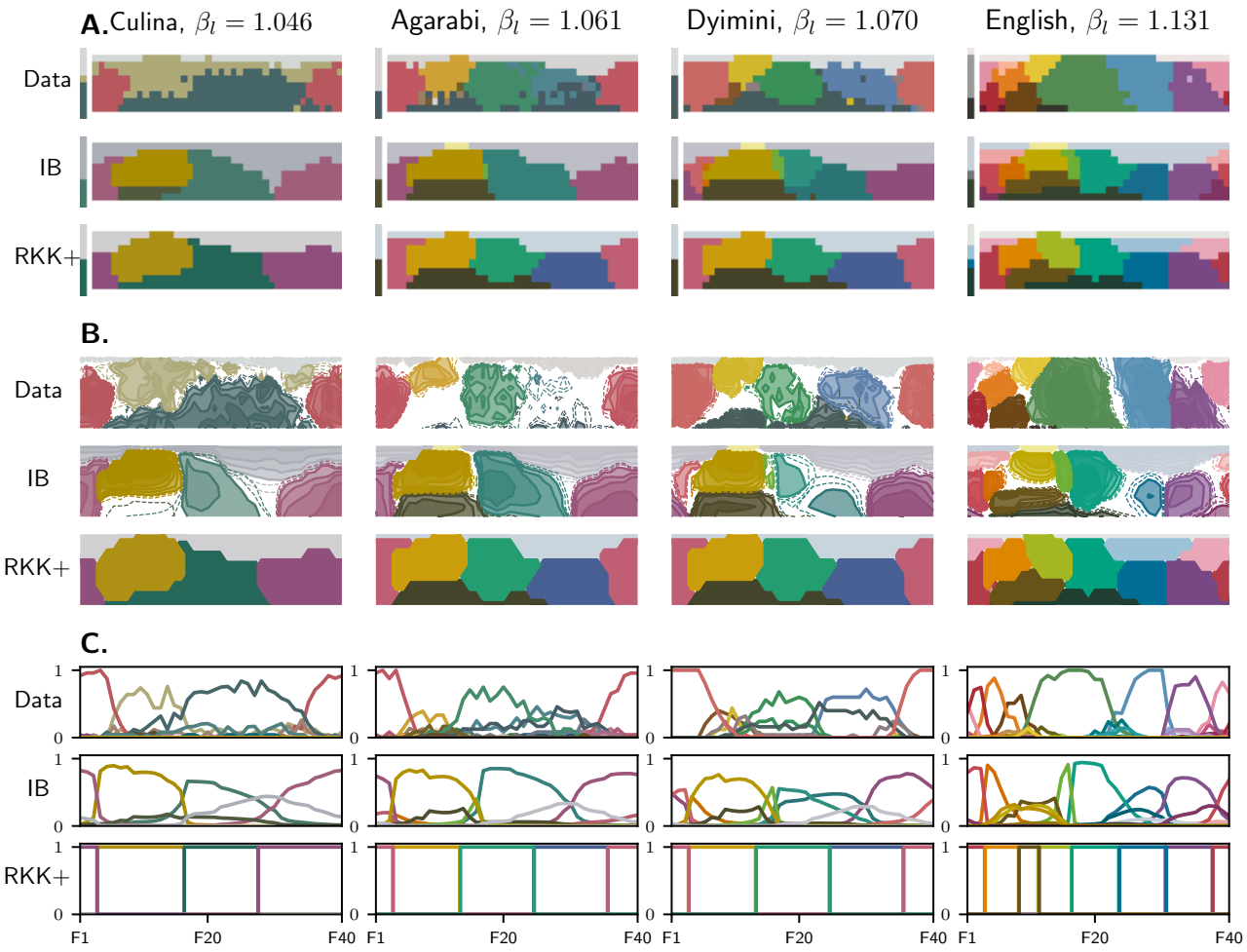

Fig. S8. Uniform source. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), for the color naming distributions (data) and for the IB and RKK+ models. These plots are similar to Fig.4 in main text, where the only difference is that they were evaluated with respect to the uniform source.

7.2. Salience-weighted distribution. Another possible approach for estimating the source distribution is based on the frequencies of colors in natural images. We used the color salience data of Gibson et al. (21), in which the salience of a color is defined by the frequency with which it appears in objects in a large set of images, relative to its frequency either in objects or in backgrounds, under the assumption that foreground objects are more likely to be spoken about than backgrounds are. Gibson et al. estimated the salience of 80 out of the 320 chromatic chips in the WCS palette, and obtained a salience-weighted (SW) prior by taking the probability of each chip to be proportional to its salience.

In order to apply the SW approach to our setting, we first constructed a salience function over CIELAB space by interpolating Gibson et al.'s salience data. We used RBF interpolation with basis functions $\phi(x - x_i) = \sqrt{\frac{\|x-x_i\|^2}{2\sigma^2} + 1}$ and $\sigma^2 = 64$ as in our main analysis. Based on this interpolated function, we estimated the salience of all 330 WCS chips and constructed a SW prior over them (see Fig.S9). This prior corresponds to a SW source.

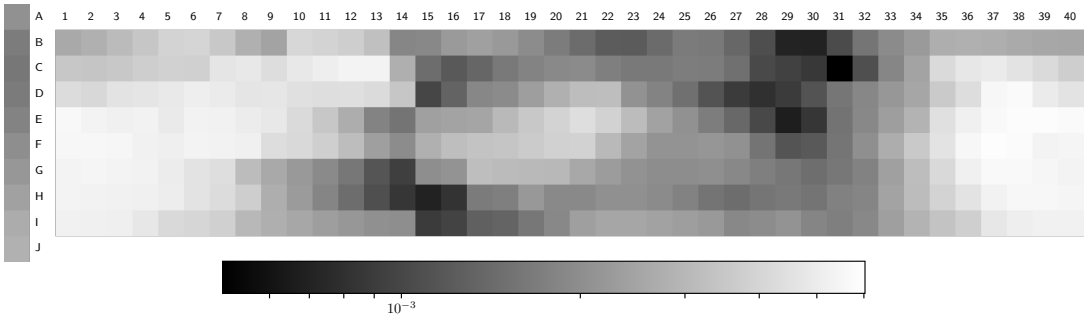

Fig. S9. The estimated salience-weighted (SW) prior over the 330 WCS chips. This prior was interpolated from the salience data of Gibson et al. (21).

We repeated our analysis exactly as described in the main text, but this time with the SW source. Our results show that in this case as well, naturally occurring color naming systems lie near the theoretical limit (Fig.S10A), and that IB achieves better scores than RKK+ (Table S6). Therefore, these results appear to be robust across the three reasonable source distribution we considered.

A comparison of Table S6 and Table 1 shows that the quantitative results with the SW source are similar to the results with the uniform source, and not as good as the results with the LI source. This can also be seen qualitatively by looking at Fig.S11 and Fig.S10B, which were evaluated for the SW source. Note that the effect of the SW source on the performance of the model is not specific to the IB principle — both IB and RKK+ do not fit the data well when evaluated with the SW source compared to the LI source or even to the uniform source. One possible explanation is that the SW source is strongly biased towards warm (reds/yellows) colors and does not weigh achromatic colors (in particular black and white) properly. This can clearly be seen in Fig.S9, and in Gibson et al.'s salience data before our interpolation. Although Gibson et al. argue that warm colors are more useful for communication than cool colors, and in that sense the SW source make sense, it seems unlikely that dark/light colors would have the low communicative need assigned to them by the SW prior.

Table S6. Quantitative evaluation (SW source)

Source	Model	ε_l	gNID	NID	β_l
SW	IB	0.24 (±0.09)	0.40 (±0.14)	0.54 (±0.12)	1.05 (±0.02)
	RKK+	0.96 (±0.22)	0.65 (±0.08)	0.51 (±0.10)	

Averages over left-out languages ±1 SD. Lower values of ε_l, gNID and NID are better.

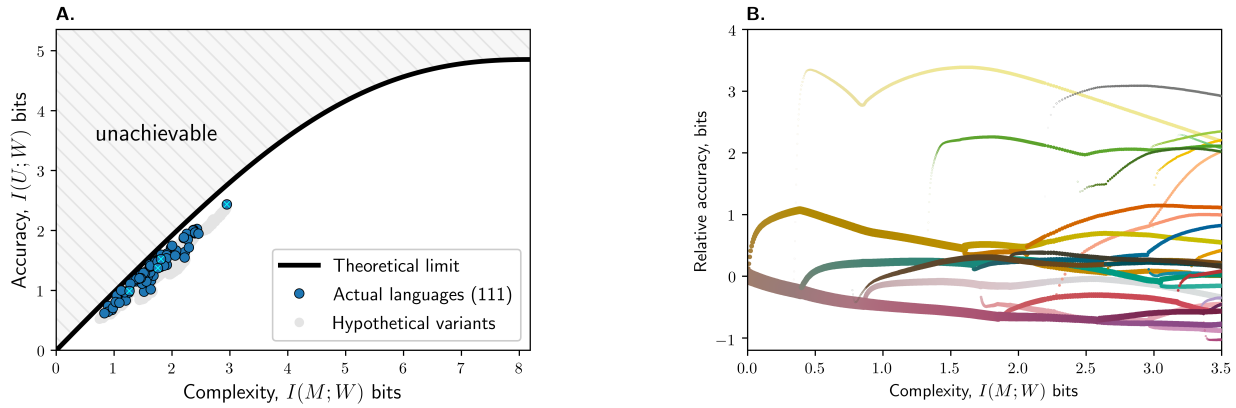

Fig. S10. SW source. Information plane (**A**) and bifurcation diagram (**B**) evaluated for the SW source. For more details see captions of Fig.3 and Fig.5 in the main text.

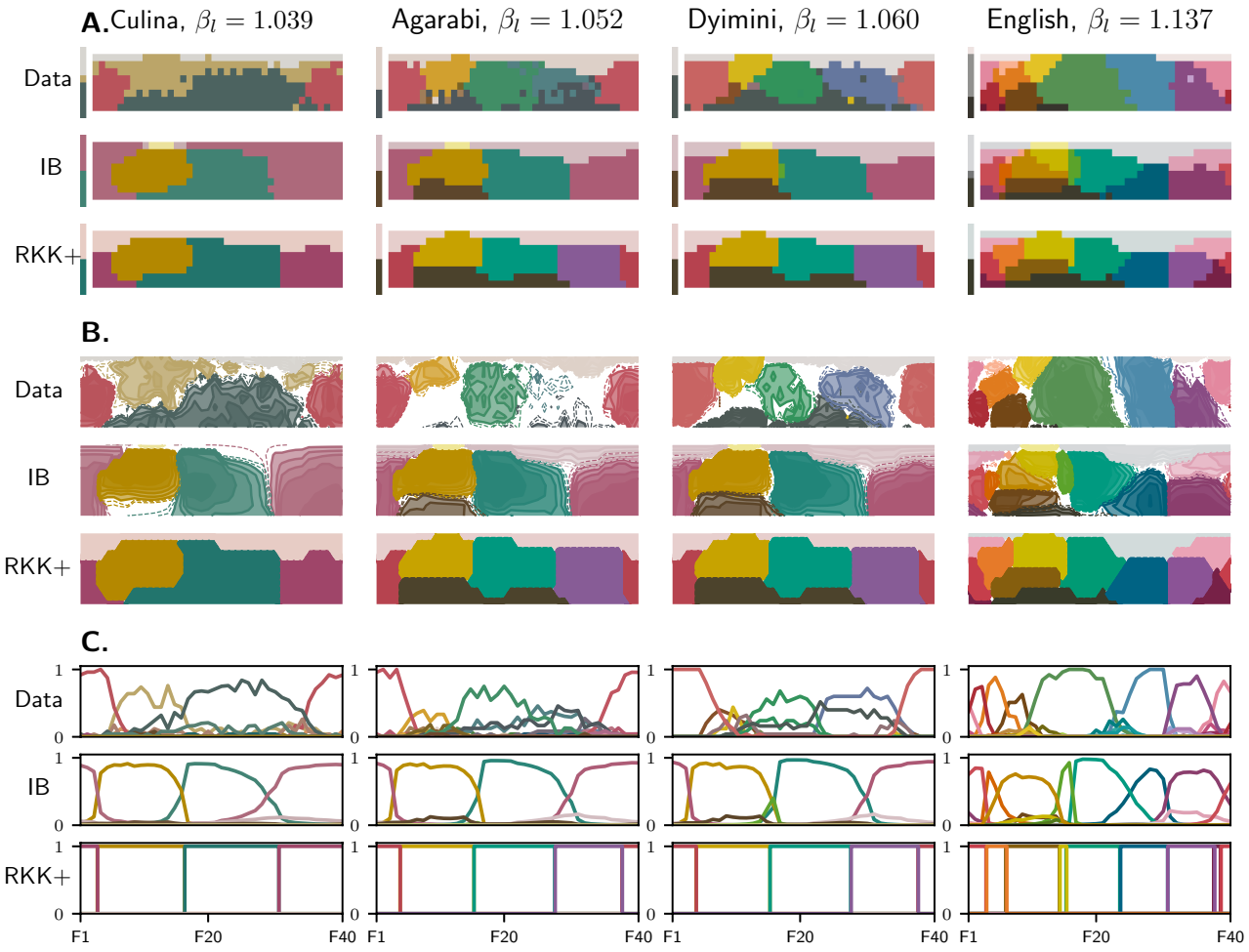

Fig. S11. SW source. Mode maps (**A**), contour plots (**B**) and naming probabilities along row F of the WCS palette (**C**), for the color naming distributions (data) and for the IB and RKK+ models. These plots are similar to Fig.4 in main text, where the only difference is that they were evaluated with respect to the SW source.

8. Hypothetical color naming systems

Is it a trivial result that naturally occurring color naming systems lie near the IB curve? Perhaps any 'reasonable-seeming' color naming system would lie near the curve, whether or not it is similar to naming systems found in the world's languages. Randomly generated color naming systems will typically lie close to the origin in the information plane. Such systems are non-informative and are thus not useful for color categorization. Therefore, in order to show that it is not trivial that naturally occurring color naming systems lie near the IB curve (and far from the origin), we considered two types of hypothetical color naming systems that maintain some informative structure about color space.

8.1. Rotation analysis. Following (20), we constructed a control set of 39 hypothetical variants for each language which were obtained by rotating its color naming distribution in the hue dimension across the columns of the WCS palette. Examples of a few hypothetical variants of Culina are shown in Fig.S12. $r = 0$ corresponds to the actual language, $r = 2$ corresponds to a shift of two columns to the right, and $r = -2$ corresponds to a shift of two columns to the left.

If languages are shaped by pressure for information-theoretic efficiency as defined by IB, we would expect that naturally occurring color naming systems would be more efficient than their hypothetical variants. To test this, for each rotated color naming system, $q_{l,r}$, we evaluated the deviation from optimality, or efficiency loss, in the same way we evaluated ε_l for the actual language, i.e. $\varepsilon_{l,r} = \min_\beta \frac{1}{\beta}(\mathcal{F}[q_{l,r}] - \mathcal{F}^*_\beta)$. We compared the efficiency of the language and the efficiency of its variants by considering $\varepsilon_{l,r} - \varepsilon_l$ (Δ efficiency loss) for IB with the full LI source. Fig.S13 shows that 93% of the languages are more efficient than all of their hypothetical variants. The remaining 7% are more efficient than most of their variants, and the preferred rotation is attained at a small $|r|$.

However, one could argue that these results are an outcome of the LI source, which was estimated with respect to the unrotated color naming systems. We therefore repeated this analysis with the uniform source. Fig.S14) shows that the results in this case are similar. This suggests that the actual languages are indeed more efficient than their hypothetical variants. The advantage of the actual languages can be explained by their alignment with the irregular structure of CIELAB space (20), which influences the accuracy of communication in the IB model. We also repeated this rotation analysis for colors that are represented in CIELUV space, and obtained similar results.

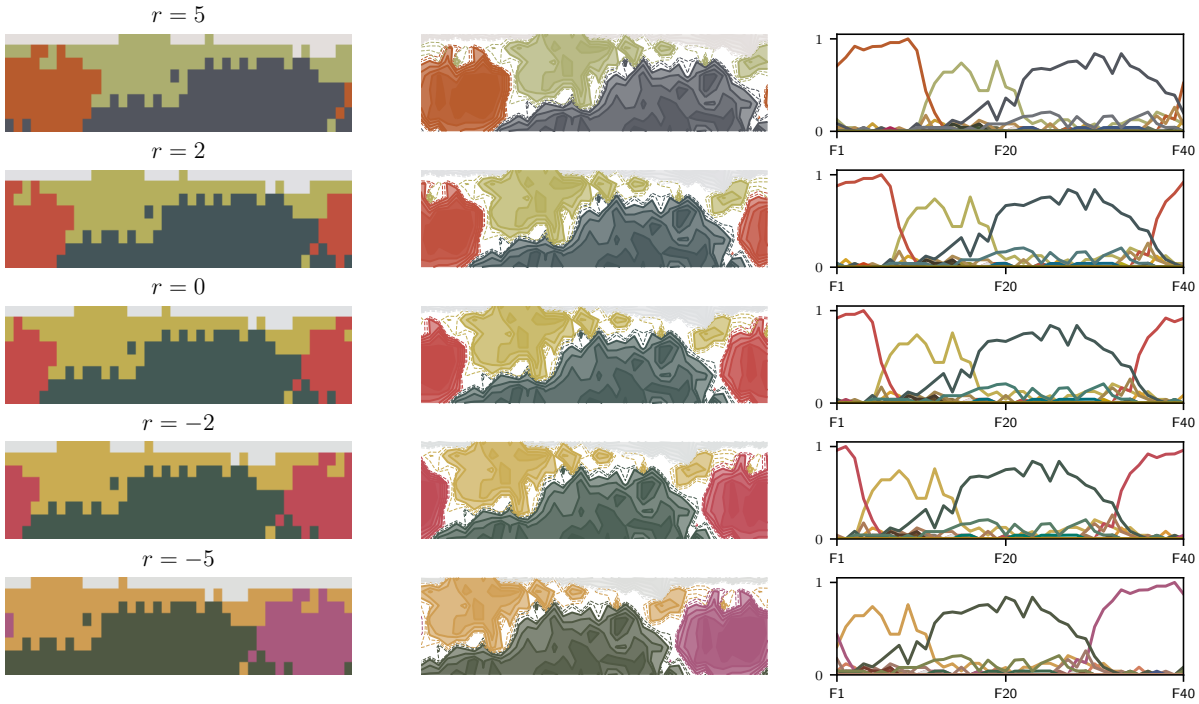

Fig. S12. Rotation example. Hypothetical variants for Culina obtained by rotating its color naming distribution in the hue dimension across the columns of the WCS palette. $r = 0$ corresponds to the actual language, $r = 2$ corresponds to a shift of two columns to the right, and $r = -2$ corresponds to a shift of two columns to the left. Colors correspond to the color centroid of each category, and columns correspond to mode maps (left), contour plots of the naming distribution (middle) and conditional probabilities along row F of the WCS palette (right).

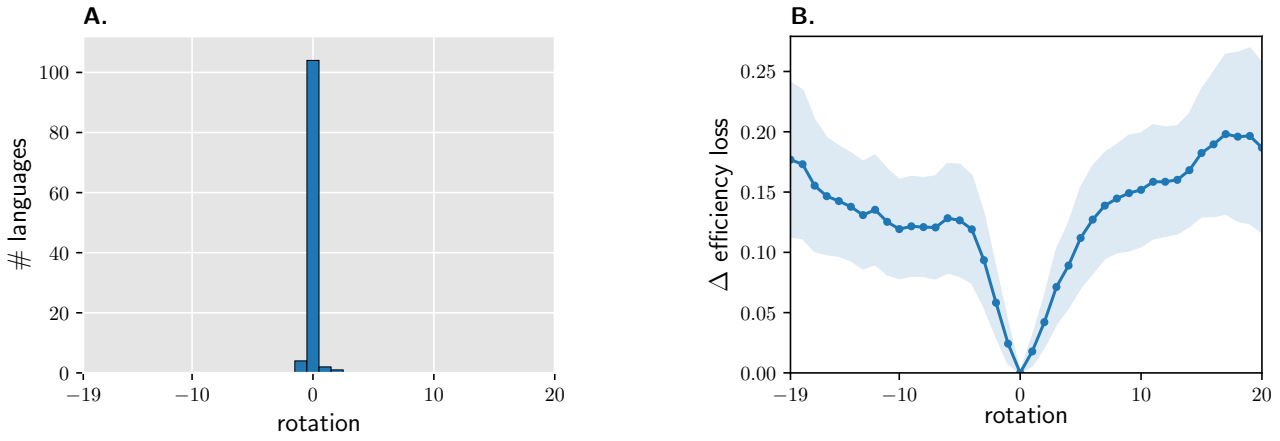

Fig. S13. Rotation analysis for the full LI source. A. Histogram of the most efficient rotation across languages. Rotation 0 corresponds to the actual language, and it is the most efficient for 93% of the languages in our data. **B.** Differences between the efficiency loss of the rotated language and the actual language, Δ efficiency loss $= \varepsilon_{l,r} - \varepsilon_l$. Lower values are better. Blue curve is the average across languages, and the colored region corresponds to ±1 SD across languages.

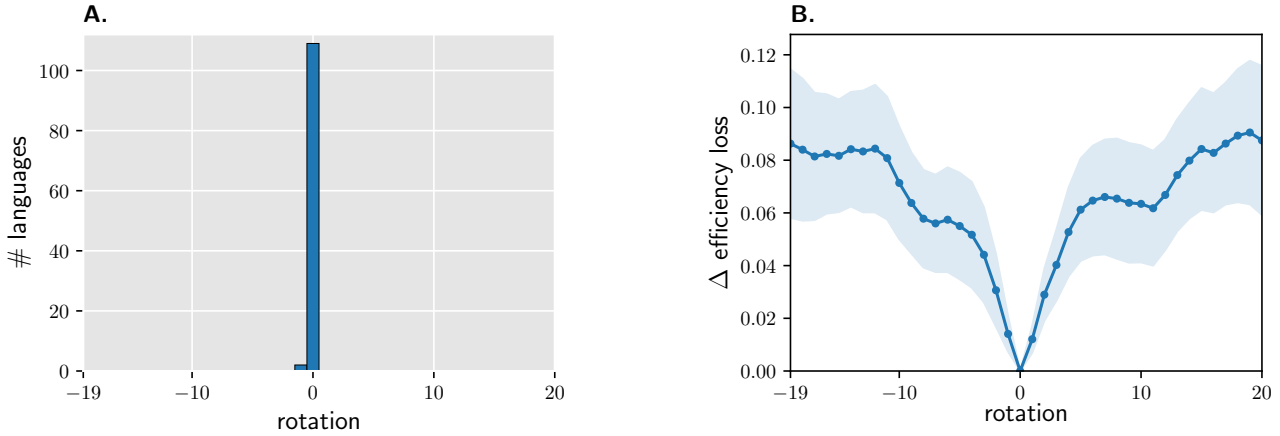

Fig. S14. Rotation analysis for the uniform source. A. Histogram of the most efficient rotation across languages. Rotation 0 corresponds to the actual language, and it is the most efficient for 98% of the languages in our data. **B.** Differences between the efficiency loss of the rotated language and the actual language, Δ efficiency loss $= \varepsilon_{l,r} - \varepsilon_l$. Lower values are better. Blue curve is the average across languages, and the colored region corresponds to ±1 SD across languages.

8.2. Structured control set based on random Gaussians. We considered another set of structured hypothetical systems in which the naming distribution is defined by random Gaussians over CIELAB space. We constructed a hypothetical system with K categories by (1) randomly selecting K chips c_w as representatives for categories $w = 1 \ldots, K$; (2) assigning to each category a random covariance matrix Σ_w; and (3) defining the color naming distribution by

$$q(w|m_c) \propto \exp\left(-\frac{1}{2}(c - c_w)^\top \Sigma_w^{-1}(c - c_w)\right) . \qquad [S25]$$

Σ_w induces a random transformation of CIELAB space and its eigenvalues are exponentially distributed with mean $\sigma^2 = 64$, which matches the level of perceptual uncertainty we used for constructing the color meaning space. We generated these random matrices as follows: a 3×3 diagonal matrix D was generated by sampling $D_{ii} \sim \text{Exp}(\frac{1}{\sigma^2 - 1}) + 1$, and a 3×3 matrix A was generated by sampling uniformly $A_{ij} \in [0, 1]$. The singular value decomposition of $A^\top A$ was evaluated, i.e. $A^\top A = U \Lambda V^\top$. Finally, $\Sigma_w = U D V^\top$.

We constructed these hypothetical systems with $K = 3, \ldots, 20$. For each K we sampled 100 systems, yielding a total of 1,800 hypothetical systems (see Fig.S15 for a few examples). We evaluated these systems with the IB

model based on the full LI source ($\varepsilon_l = 0.33 \pm 0.1$, gNID $= 0.39 \pm 0.16$, NID $= 0.44 \pm 0.13$) and the uniform source ($\varepsilon_l = 0.36 \pm 0.08$, gNID $= 0.47 \pm 0.15$,NID $= 0.5 \pm 0.13$). In both cases, these hypothetical systems are less efficient on average than the actual languages we considered.

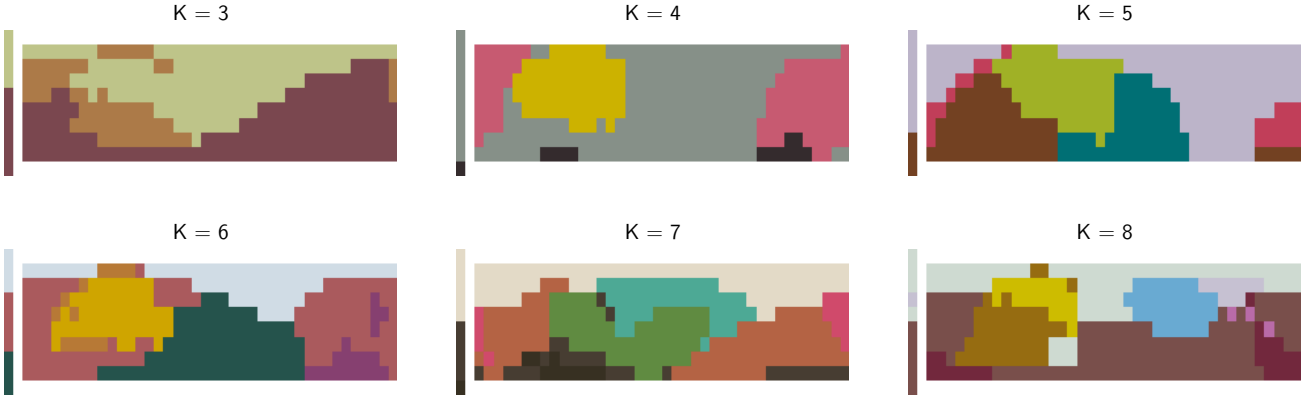

Fig. S15. Examples of hypothetical color naming systems based on K random Gaussians in CIELAB space.

9. Sensitivity analysis

In this section we test the sensitivity of our results to small errors in the structure of the meaning space, $\mathcal{M}$, that we assumed. To do so, we injected a small perturbation to each m_c and re-evaluated IB and RKK+ with the full LI source. We injected the perturbation by first drawing i.i.d. Gaussian variables $Z_{c,u} \sim \mathcal{N}(0, 0.01)$, and defining the perturbed model by $m'_c(u) \propto m_c(u)e^{Z_{c,u}}$. The results, summarized in table S7, are almost identical to the results without perturbation, which suggests that our analysis is robust to small amounts of noise in the perceptual model.

Table S7. Quantitative evaluation with the perturbed meaning space

Source	Model	ε_l	gNID	NID	β_l
LI	IB	0.18 (±0.07)	0.18 (±0.10)	0.31 (±0.07)	1.03 (±0.01)
	IB-C	0.18 (±0.07)	0.21 (±0.08)	0.30 (±0.08)	1.04 (±0.02)
	RKK+	0.70 (±0.23)	0.46 (±0.10)	0.31 (±0.10)	

Numbers correspond to averages over languages ±1 SD. Lower values are better for ε, gNID and NID.

Section 10 is omitted due to space limitation, and is available online:
www.pnas.org/content/suppl/2018/07/18/1800521115.DCSupplemental

References

1. Tishby N, Pereira FC, Bialek W (1999) The Information Bottleneck method in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.
2. Regier T, Kemp C, Kay P (2015) Word meanings across languages support efficient communication in *The Handbook of Language Emergence*, eds. MacWhinney B, O'Grady W. (Wiley-Blackwell, Hoboken, NJ), pp. 237–263.
3. Gilad-Bachrach R, Navot A, Tishby N (2003) An information theoretic tradeoff between complexity and accuracy in *Proceedings of the 16th Annual Conference on Learning Theory*.
4. Harremoës P, Tishby N (2007) The Information Bottleneck revisited or how to choose a good distortion measure in *IEEE International Symposium on Information Theory*. pp. 566–571.
5. Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. pp. 183–190.
6. Shamir O, Sabato S, Tishby N (2010) Learning and generalization with the Information Bottleneck. *Theoretical Computer Science* 411(29-30):2696–2711.
7. Rose K (1998) Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE* 86(11):2210–2239.
8. Elidan G, Friedman N (2005) Learning hidden variable networks: The Information Bottleneck approach. *Journal of Machine Learning Research* 6:81–127.
9. Slonim N, Tishby N (1999) Agglomerative Information Bottleneck in *Advances in Neural Information Processing Systems (NIPS)*. pp. 617–623.
10. Blahut R (1972) Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* 18(4):460–473.
11. Arimoto S (1972) An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory* 18(1):14–20.
12. Bernardo JM (2005) Reference analysis in *Bayesian Thinking Modeling and Computation*, Handbook of Statistics, eds. Dey D, Rao C. (Elsevier) Vol. 25, pp. 17 – 90.
13. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR* 11:2837–2854.
14. Abbott JT, Griffiths TL, Regier T (2016) Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences* 113(40):11178–11183.
15. Roberson D, Davies I, Corbett G, Vandervyver M (2005) Free-sorting of colors across cultures: Are there universal grounds for grouping? *Journal of Cognition and Culture* 5(3):349–386.
16. Roberson D, Davies I, Davidoff J (2000) Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* 129(3):369–398.
17. Roberson D, Davidoff J, Davies IR, Shapiro LR (2005) Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology* 50(4):378 – 411.
18. Cibelli E, Xu Y, Austerweil JL, Griffiths TL, Regier T (2016) The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PLOS ONE* 11(7):1–28.
19. Mokrzycki W, Tatol M (2012) Colour difference ΔE - a survey. *Machine Graphic and Vision* 8.
20. Regier T, Kay P, Khetarpal N (2007) Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* 104(4):1436–1441.
21. Gibson E, et al. (2017) Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences* 114(40):10785–10790.
22. Lindsey DT, Brown AM (2014) The color lexicon of American English. *Journal of Vision* 14(2):17.

Chapter 3

Direct Evidence that Color Naming Evolves Under Pressure for Efficient Coding

Direct evidence that color naming evolves under pressure for efficient coding

Noga Zaslavsky,[1,2,✉] Karee Garvin,[2] Charles Kemp,[3] Naftali Tishby,[1,4] and Terry Regier[2,5]

[1]Edmond and Lily Safra Centre for Brain Sciences, The Hebrew University of Jerusalem; [2]Department of Linguistics, University of California, Berkeley; [3]School of Psychological Sciences, The University of Melbourne; [4]Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem; [5]Cognitive Science Program, University of California, Berkeley.

Abstract

It has recently been hypothesized that semantic systems evolve under pressure to maintain efficient coding schemes. However, thus far, support for this hypothesis has been based largely on synchronic cross-language comparison, rather than on diachronic data. Here, we directly test the predictions of this efficient coding hypothesis in the domain of color naming, by analyzing recent diachronic data for a single language, Nafaanra. We show that color naming in Nafaanra has changed over the past four decades while remaining near-optimally efficient, and that this outcome would be unlikely under an alternative baseline process that does not incorporate pressure for efficiency. To our knowledge, this finding provides the first direct evidence in support of the view that color naming evolves under pressure for efficiency. The principle we invoke is general and has been applied to semantic domains other than color, thus it is possible that pressure for efficiency may shape the evolution of the lexicon more broadly.

1 Introduction

What forces shape the evolution of semantic systems? This general question has often been addressed in the specific case of color naming. Many theories hold that languages acquire new color terms with time, resulting in finer-grained color naming systems (e.g., Berlin and Kay, 1969; Kay and Maffi, 1999; MacLaury, 1997; Levinson, 2000). More recently, it has also been claimed (e.g., Lindsey et al., 2015; Regier et al., 2015; Gibson et al., 2017) that this historical evolutionary process, and color naming more generally, are shaped by the need for efficient communication — that is, the need to communicate accurately, with a simple lexicon. In particular, Zaslavsky et al. (2018) grounded this notion of efficiency in an independent

✉ noga.zaslavsky@mail.huji.ac.il

information-theoretic principle, the Information Bottleneck (IB: Tishby et al., 1999). IB can be formulated within rate–distortion theory (Shannon, 1948, 1959), the branch of information theory that addressed the problem of efficient coding under limited resources, also known as lossy data compression (Cover and Thomas, 1991). They showed that IB explains much of the observed cross-language variation in color naming, and hypothesized that languages evolve under pressure to remain near the IB theoretical limit of efficiency via an annealing-like process.

However, this hypothesis, as well as most theories concerning the evolution of color naming, has been supported by synchronic cross-language comparisons, rather than by direct evidence from diachronic data (but e.g. Biggam (2012) considered historical texts and Kay (1975) considered informant age as a proxy for change over time). In this work, we test directly the quantitative evolutionary predictions derived from the IB principle using diachronic color naming data. We do so by considering data for a single language, Nafaanra, that was collected first in 1978 as part of the World Color Survey (WCS: Kay et al., 2009), and again in 2018 by Garvin (in preparation). We show that color naming in Nafaanra has changed over the past four decades while remaining near the theoretical limit of efficiency, as predicted by IB, and that this result is unlikely to be explained by a baseline evolutionary process that is not influenced by pressure for efficiency. To our knowledge, this is the first finding that directly supports the idea that color naming, and possibly semantic systems more generally, evolve under pressure for efficiency.

2 Theoretical framework and predictions

We begin by reviewing Zaslavsky et al.'s (2018) IB color naming model and efficient coding hypothesis, on which the present study builds. Although this framework is presented here in the case of color naming, it is not specific to color, and can be applied to other semantic domains.

2.1 IB color naming model

The IB color naming model is based on a basic communication setting (Figure 1A) in which a speaker and a listener wish to communicate about colors. For simplicity, we consider only the set of colors shown in Figure 1B. The speaker obtains a mental color representation, M, drawn from a prior distribution $p(m)$. This prior was estimated by Zaslavsky et al. (2018) in a data-driven approach (see Zaslavsky et al., 2019a, for a systematic evaluation of several priors). The speaker's mental representations are grounded in color perception, following Regier et al. (2007, 2015), by assuming each color is represented by a Gaussian distribution over a standard perceptual color space (Figure 1C). The speaker then communicates their color representation by encoding it into a word W, using a stochastic encoder $q(w|m)$. The listener receives W and infers the speaker's representation by constructing another color representation, $\hat{M}$. An optimal Bayesian listener, as assumed here, infers from a given word, w, the representation

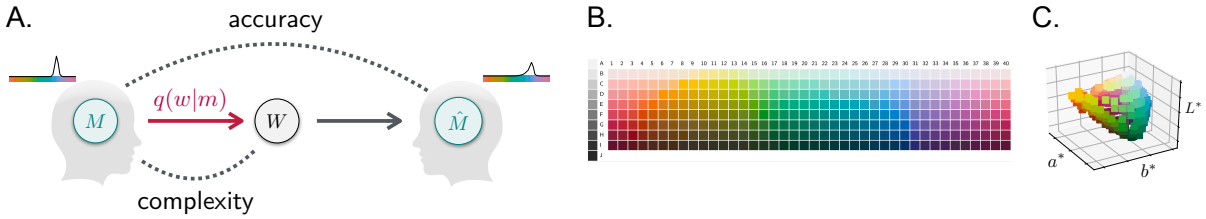

Figure 1. Color communication model (adapted from Zaslavsky et al., 2018). A. The speaker mentally represents a color as a Gaussian distribution, M, over a standard perceptual color space (shown in C), and communicates this representation by producing a word W. The listener receives W and infers the speaker's representation by reconstructing $\hat{M}$. B. The WCS color naming grid, which consists of 320 chromatic color chips and 10 achromatic color chips. C. Colors are represented in the 3-dimensional CIELAB space. Lightness is represented by the L^* dimension. Hue and saturation are represented in polar coordinates in the (a^*, b^*) plane.

$\hat{m}_w = \sum_m q(m|w)m$, where $q(m|w)$ is obtained by applying Bayes' rule with respect to the speaker's encoder and prior.

Zaslavsky et al. (2018) argued that human semantic systems, formulated as encoders, are pressured to optimize the IB tradeoff between the complexity and accuracy of the lexicon. IB can be formulated as a type of rate–distortion problem, where the messages that need to be compressed are defined by distributions over a set of relevant features. In our case, these messages are the mental representations described above, and a feature vector U is a points in the perceptual CIELAB color space (Figure 1C). According to IB, complexity roughly corresponds to the number of bits required for communication, and it is measured by the mutual information between M and W,

$$I_q(M;W) = \sum_{m,w} p(m)q(w|m) \log \frac{q(w|m)}{q(w)} \,. \tag{1}$$

Accuracy corresponds to the similarity between the speaker's and listener's representations, and it is measured by $I_q(W;U)$. Maximizing this term amounts to minimizing the expected Kullback–Leibler (KL) divergence between the two representations,

$$\mathbb{E}_q \left[D[m \| \hat{m}_w] \right] = \mathop{\mathbb{E}}_{\substack{m \sim p(m) \\ w \sim q(w|m)}} \left[\sum_u m(u) \log \frac{m(u)}{\hat{m}_w(u)} \right] \,. \tag{2}$$

Thus, high accuracy implies that the listener's inferred representation is similar to the speaker's representation. Achieving high accuracy requires a complex lexicon, while reducing complexity may result in accuracy loss. Optimal systems, according to the IB principle, minimize complexity while maximizing accuracy, for a tradeoff, $\beta \geq 0$, between these two competing objectives. Formally, an optimal encoder given β attains the minimum of the IB objective function,

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) \,, \tag{3}$$

across all possible encoders. Let $\mathcal{F}^*_\beta$ be the minimal value of this objective given β.

2.2 The efficient coding hypothesis

The theoretical limit of efficiency is determined by the set of encoders that attain $\mathcal{F}^*_\beta$ for different values of β. These encoders, denoted by $q_\beta(w|m)$, induce ideal color naming systems in the sense that they attain the maximal achievable accuracy given their complexity. Zaslavsky et al. (2018) evaluated this theoretical limit for color naming, shown in Figure 2. At the origin of the IB curve (black), the solution corresponds to $\beta \leq 1$, and it gives a minimally complex yet non-informative system, that can be implemented using a single word. As β increases from 1 to ∞, the ideal systems evolve by traveling on the IB curve, and becoming more complex and more accurate. Along this continuous trajectory, the systems undergo a sequence of structural phase transitions at critical values of β, in which new categories emerge. Zaslavsky et al. (2018) showed that (1) the actual color naming systems in the WCS+ dataset[1] lie near the theoretical limit; (2) the IB systems explain much of the observed cross-language variation, where β is the only language-dependent variable; and (3) the annealing-like process by which the IB systems evolve synthesizes the discrete aspects of Berlin and Kay's (1969) evolutionary sequence and continuous aspects of other accounts of color category evolution (MacLaury, 1997; Lyons, 1995; Levinson, 2000). On that basis, they hypothesized that languages evolve under pressure to remain near the theoretical limit. We refer to this hypothesis as the *efficient coding hypothesis* for semantic systems.

On this view, language change is driven to a large extent by shifts in the tradeoff parameter β. This parameter reflects language-specific factors, which may change over time, and determines the capacity resources (i.e., average number of bits) allocated for communication about the domain. β is closely related to Kemp et al.'s (2018) notion of *domain-level need*, although that notion refers to a probability distribution over semantic domains, rather than to the allocation of capacity resources. In both cases, however, specific social or cultural factors are not explicitly modeled, even though it is likely that such factors influence language change (e.g. Berlin and Kay, 1969). Explicitly taking into account such factors requires an extension of the current framework, which we leave for future work.

2.3 Quantitative diachronic predictions

The goal of this work is to directly test the efficient coding hypothesis by evaluating its predictions on diachronic color naming data, rather than synchronic data. Suppose we have access to the naming system of a given language, l, at a given point in time, t. Denote this system by $q_l^{(t)}(w|m)$. The evolutionary trajectory of l is defined by the trajectory of $q_l^{(t)}(w|m)$ over time. In practice, we may only obtain samples from this trajectory at discrete time points,

[1] The WCS+ dataset consist of the WCS data and color naming data from English (Lindsey and Brown, 2014).

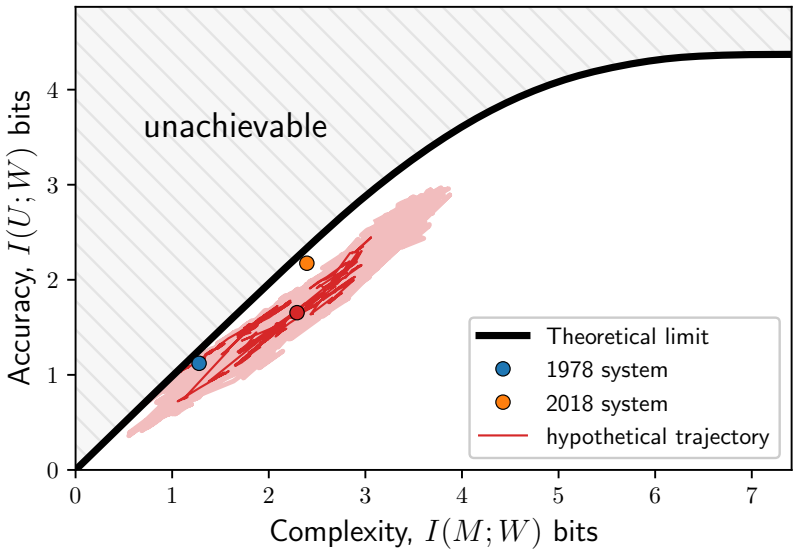

Figure 2. Information plane. The theoretical limit of efficiency (IB curve, black) is defined by the complexity–accuracy pairs of the optimal IB systems for different values of β (from Zaslavsky et al., 2018). The blue and orange dots show the complexity and accuracy of the Nafaanra system in 1978 and 2018. The light red area below the IB curve shows the area covered by 50 hypothetical trajectories, which were all initialized near the 1978 system. The red trajectory corresponds to the example in Figure 4B, and the red dot shows its final location.

$\mathcal{T} = \{t_0, \ldots, t_n\}$. Given these samples, the diachronic dataset for language l is defined by

$$Q_l(\mathcal{T}) = \left\{ q_l^{(t)}(w|m) : t \in \mathcal{T} \right\} . \tag{4}$$

Zaslavsky et al. (2018) derived two precise predictions for any given naming system, $q_l^{(t)}(w|m)$, which are extended here to $Q_l(\mathcal{T})$, while keeping the same quantitative measures for evaluation.

Inefficiency. The first prediction states that languages should be near-optimally efficient in the IB sense. Notice that $Q_l(\mathcal{T})$ induces a trajectory on the information plane (Figure 2), defined by the complexity–accuracy pairs of $q_l^{(t)}(w|m)$, for all $t \in \mathcal{T}$. If the efficient coding hypothesis is true, then there should be a sequence of tradeoffs, $\beta_l(t)$, such that the time-dependent *inefficiency*,

$$\varepsilon_l(t) = \tfrac{1}{\beta_l(t)} \left(\mathcal{F}_{\beta_l(t)} \left[q_l^{(t)} \right] - \mathcal{F}^*_{\beta_l(t)} \right) , \tag{5}$$

would be small. Notice that $\varepsilon_l(t) \in [0, H(M)]$, however in practice, achieving $\varepsilon_l(t) = 0$ is unlikely. A precise sense of "small" $\varepsilon_l(t)$ can be obtained by comparison to counterfactual outcomes. If for actual languages it holds that $\varepsilon_l(t) \ll H(M)$, and it is substantially lower compared to hypothetical systems that are not shaped by pressure for efficiency, then this would support the hypothesis. In section 3.2 we discuss in detail how this counterfactual data is generated. Finally, since $\beta_l(t)$ is unknown, a natural way to estimate it is by $\beta_l(t) = \operatorname{argmin}_\beta \{\mathcal{F}_\beta[q_l^{(t)}] - \mathcal{F}^*_\beta\}$, for every language l and time $t \in \mathcal{T}$.

Dissimilarity. The second prediction states that actual naming systems should be similar to their corresponding IB systems. This implies that the actual trajectory, $Q_l(\mathcal{T})$, is expected to be similar to the corresponding idealized trajectory along the IB curve, i.e., $Q_l^*(\mathcal{T}) = \{q_{\beta_l(t)}(w|m) : t \in \mathcal{T}\}$. The *dissimilarity* between any two naming systems is measured by the generalized Normalized Information Distance (gNID: Zaslavsky et al., 2018). To define gNID, consider the following process: Suppose that a speaker of language l_1 and a speaker of language l_2 obtain a representation $M \sim p(m)$. l_1 and l_2 could either be actual languages or hypothetical ones. In order to communicate M, the speaker of l_1 produces $W_1 \sim q_1(w|m)$, and the speaker of l_2 produces $W_2 \sim q_2(w|m)$. The cross-language information is defined by $I(W_1, W_2)$, and the within-language information is defined by $I(W_i, W_i')$, where W_i' is produced by another speaker of l_i, independently given M. gNID is based on the ratio between these informational terms, and it is defined by

$$\text{gNID}\left[q_1, q_2\right] = 1 - \frac{I(W_1, W_2)}{\max\left\{I(W_1, W_1'), I(W_2, W_2')\right\}}\,. \tag{6}$$

Thus, two naming systems are similar (low gNID) if they induce high cross-language information, normalized by their within-language information. We define the gNID between two trajectories by

$$\text{gNID}\left[Q_1(\mathcal{T}), Q_2(\mathcal{T})\right] = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{gNID}\left[q_1^{(t)}, q_2^{(t)}\right]\,. \tag{7}$$

If $\text{gNID}\left[Q_l(\mathcal{T}), Q_l^*(\mathcal{T})\right]$ is smaller than $\text{gNID}\left[Q_l(\mathcal{T}), Q_h(\mathcal{T})\right]$, where $Q_h(\mathcal{T})$ is a reasonable hypothetical trajectory not influenced by pressure for efficiency, in addition to $Q_l(\mathcal{T})$ being highly efficient, then that would provide converging evidence in support of the efficient coding hypothesis.

3 Diachronic data

In order to directly test the diachronic predictions discussed in section 2.3, we consider diachronic color naming data that exists for a single language, Nafaanra, spoken in West Ghana. We also consider a set of hypothetical color naming systems for our counterfactual analysis. We next describe our actual and counterfactual datasets.

3.1 Color naming in Nafaanra

Color naming data for Nafaanra was initially collected in 1978, as part of the WCS. Participants in the WCS experiment were asked to provide names for all 330 color chips in the WCS grid (Figure 1B). The 1978 Nafaanra system, $q_l'^{78}(w|m)$, was estimated by the proportion of participants who used the term w in reference to c_m, the chip associated with m (recall that each color is mapped to a unique mental representation). This system is shown in Figure 3A.

It was analyzed by Zaslavsky et al. (2018), together with all languages in the WCS+ dataset, and was shown to be near-optimally efficient (Figure 2, blue dot). Nafaanra data were collected again in 2018 by Garvin (in preparation), following the same WCS data collection protocol. We estimated the 2018 Nafaanra system, $q_l^{'18}(w|m)$, the same way the 1978 system was estimated. This system is shown in Figure 3B. Clearly, color naming in Nafaanra has changed substantially over the past four decades, by adding more color terms and becoming more fine-grained.[2] However, this observation alone does not indicate whether the system has changed in a way that is consistent with the diachronic predictions of the efficient coding hypothesis.

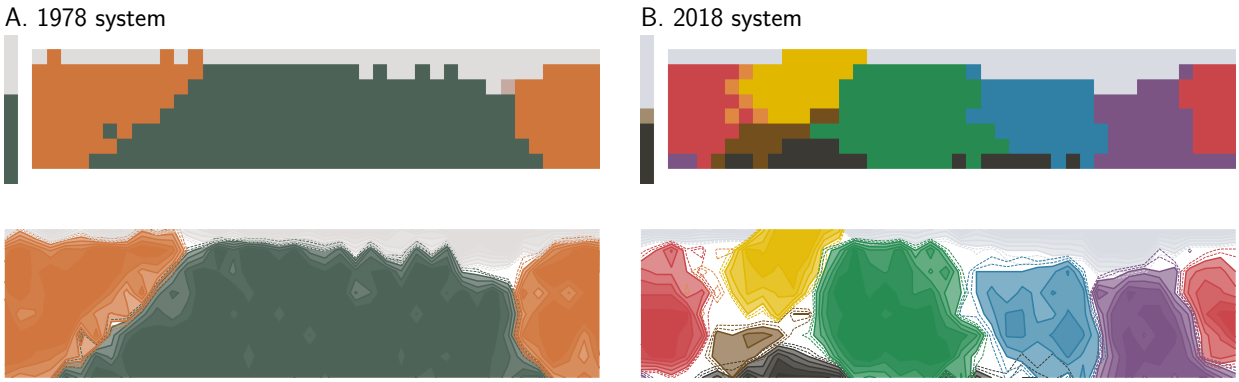

Figure 3. Color naming in Nafaanra as estimated from the 1978 and 2018 data. Upper panel (mode maps): Each chip in the WCS grid is colored according to its modal category. Colors correspond to the center of mass of the category. Lower panel: Contour plots of the distribution of color names across speakers. The frequency of use of each term is shown with the color of its category. Dashed lines correspond to agreement levels of 40%-45%, and solid lines correspond to agreement levels above 50%.

3.2 Counterfactual data

We generated two types of counterfactual datasets, for comparison with the actual Nafaanra data. First, we consider a standard approach for generating hypothetical variants of existing color naming systems (Regier et al., 2007). This approach is very useful for evaluating synchronic data, but not for evaluating diachronic data. Thus, we propose a method for generating hypothetical future trajectories of a given initial system — in this case, the 1978 Nafaanra system — which could then be compared with actual diachronic data.

Rotations. Regier et al. (2007) generated a set of hypothetical variants of actual color naming systems, by rotating each system along the hue dimension (columns) of the WCS grid. This method produces a set of 39 hypothetical system for each language, that preserve a similar category structure as the actual system. Therefore, they are more reasonable than randomly generated systems, which typically would not have a continuous category structure. Zaslavsky et al. (2018) applied this method to all 111 languages in the WCS+ dataset, including the 1978

[2] See Garvin (in preparation) for a detailed description of each color term and its origin.

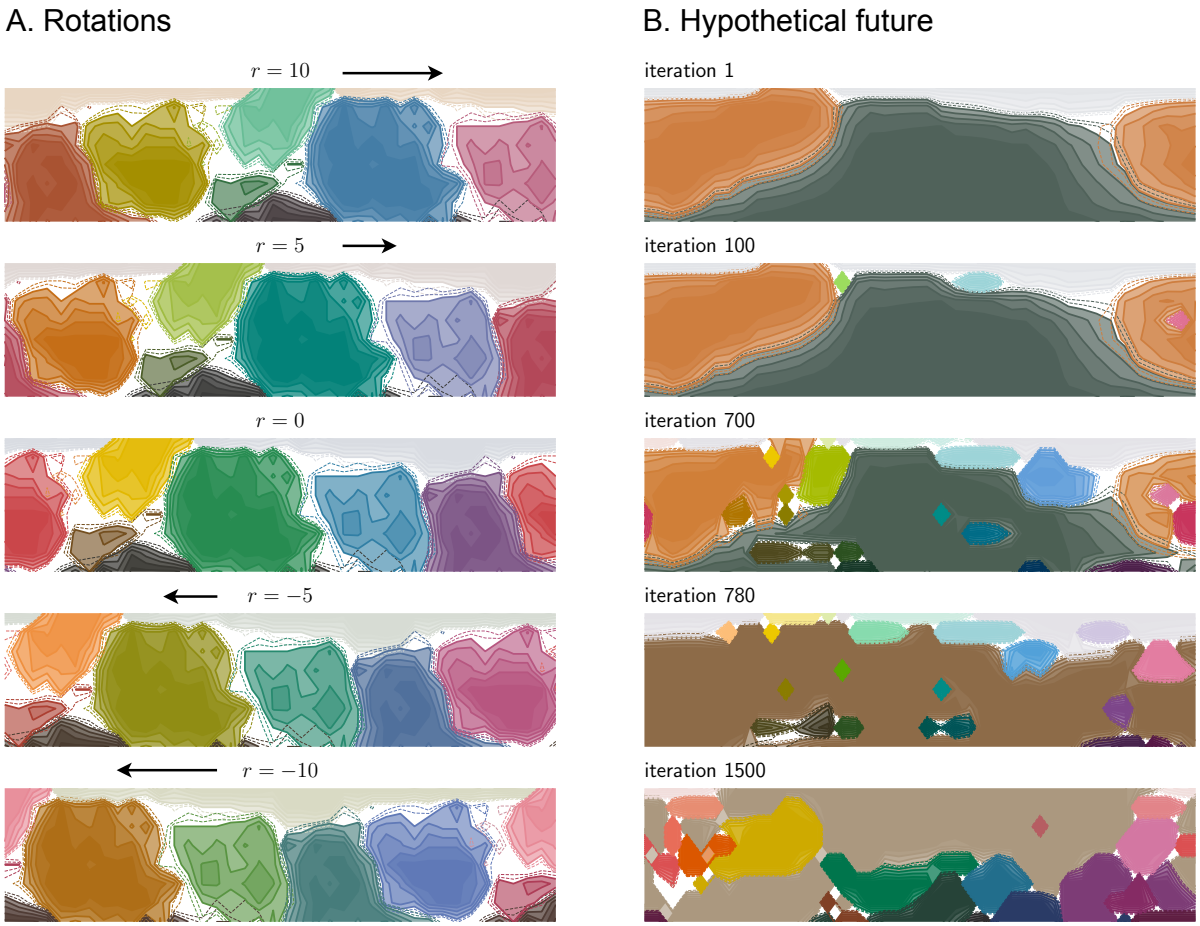

Figure 4. Counterfactual data. Contour plots of (A) rotated variants of the 2018 system, and (B) systems along one hypothetical trajectory. A. $r = 0$ corresponds to the actual system (same as Figure 3B). $r > 0$ corresponds to a shift of r columns to the right, and $r < 0$ corresponds to a shift of $|r|$ columns to the left. B. The initial system was fitted to the 1978 system. It evolves by simulating a stochastic process that allows new categories to emerge, drift, and occasionally vanish.

Nafaanra system, and showed that these languages are more efficient than their hypothetical variants, and more similar to the corresponding IB system. It is worth noting that β is fitted to each hypothetical system separately, in order to consider the best scores these hypothetical systems can achieve. Here, we apply the same method for constructing a set of hypothetical variants for the 2018 Nafaanra system (see Figure 4A for examples). While this counterfactual dataset is suitable for evaluating the efficiency of the 2018 system considered by itself, it is not suitable for evaluating diachronic data because it is intuitively unlikely that actual color naming systems evolve over time by rotations.

Hypothetical futures. In order to generate counterfactual diachronic data, we simulate language change via a stochastic process that preserves a continuous category structure, without any pressure to maintain efficient coding. To this end, we consider a class of artificial color naming systems, in which each category w induces a Gaussian distribution, $q(c|w) = \mathcal{N}(c; \mu_w, \Sigma_w)$, over CIELAB space. In practice, we discretized these Gaussians by restricting

them to the WCS grid. Abbott et al. (2016) considered a similar method of estimating color categories by Gaussian, and on that basis they accounted for focal colors. This suggests that the class of Gaussian color naming systems contains reasonably structured hypothetical systems. A system with k categories is defined by k Gaussians, and a k-dimensional probability vector $q(w)$. Given these parameters, the naming distribution is taken to be $q(w|m) \propto q(c_m|w)q(w)$. Our stochastic process takes an initial system from this class, and propagates it in time by allowing its parameters to change gradually.

Before we define the dynamics of this process, our parameterization requires further elaboration. First, to ensure that each Σ_w remains positive semi-definite, we parametrize it by another matrix, L_w, such that $\Sigma_w = L_w L_w^\top$. Second, to allow categories to emerge or vanish, we assume $K = 330$ potential categories, and keep a weight vector, π_w, for them. Only categories for which π_w is higher than a given threshold η are considered in the lexicon. For those categories, we define $q(w) \propto \pi(w)$. Therefore, η is a hyper-parameter that controls the tendency to add new categories. At the t-th iteration of the process, the system is defined by $\theta(t) = \{\mu_w^{(t)}, L_w^{(t)}, \pi_w^{(t)}\}_{w=1}^K$.

Given an initial system, $\theta(0)$, the dynamics of the process is defined as follows. At each iteration t, a potential category w_t is chosen at random. First, the weight vector is updated by randomly selecting whether to add or subtract η from $\pi_{w_t}^{(t-1)}$, and keeping the vector non-negative and normalized. Formally, the update equations for the weight vector are:

$$P_t = \max\{0, \pi_{w_t}^{(t-1)} + \eta s_t\}\,, \quad s_t \sim U(\{-1, 1\}) \tag{8}$$

$$\pi_w^{(t)} = \frac{1}{1 - \pi_{w_t}^{(t-1)} + P_t} \left(\delta_{w,w_t} P_t + (1 - \delta_{w,w_t})\pi_w^{(t-1)}\right)\,, \quad \forall w \in \{1, \ldots, K\}\,. \tag{9}$$

Next, if w_t is already in the lexicon, i.e. $\pi_{w_t}^{(t-1)} > \eta$, then with probability 0.5 its parameters are updated as follows:

$$\mu_{w_t}^{(t)} = \frac{1}{2}\left(\mu_{w_t}^{(t-1)} + c_t\right)\,, \quad c_t \sim q_{t-1}(c|w_t) \tag{10}$$

$$L_{w_t}^{(t)} = L_{w_t}^{(t-1)} + I + A^{(t)}\,, \quad A_{i,j}^{(t)} \sim \mathcal{N}(0, 1)\,. \tag{11}$$

The update rule for $\mu_{w_t}^{(t)}$ shifts it in the direction of c_t, which on average would be a small shift because c_t is sampled from $q_{t-1}(c|w_t)$. The update rule for $L_{w_t}^{(t)}$ adds to it a noise matrix, $A^{(t)}$, and the identity matrix, I, in order to encourage the category to grow over time.

Finally, it remains to set the initial set of parameters, $\theta(0)$, and threshold η. We set $\theta(0)$ such that the corresponding system will approximate the actual 1978 Nafaanra system. For each category in the 1978 system, we fit a Gaussian with a diagonal covariance matrix, and take $L_w^{(0)}$ to be its square root. For these categories, we take $\pi_w^{(0)}$ to be their proportion in the 1978 naming data. For the remaining potential categories, which are not in the lexicon, we set $\pi_w^{(0)} = 0$. For these categories, $\mu_w^{(0)}$ is initialized by randomly selecting a chip from the

WCS grid (with replacement). $L_w^{(0)}$ is initialized by $\sigma_w^{(0)} I + A_w^{(0)}$, where $A_w^{(0)}$ is sampled as in Eq. 11, and $\sigma_w^{(0)}$ is drawn uniformly from $[1, 5]$. We take $\eta = 0.01$, for which we observed a trend of gradual increase in the number of categories, reaching on average $k = 23.9$ after $1,500$ iterations.

We generated a set of 50 hypothetical trajectories by simulating this process for $1,500$ iterations. Examples of several systems along a trajectory from this sample are shown in Figure 4B. The initial system, which is the same for all trajectories, is indeed a good approximation of the 1978 system.

4 Results

We begin by analyzing the 2018 system in exactly the same way Zaslavsky et al. (2018) analyzed the 1978 system. It was already shown that color naming in Nafaanra was near-optimally efficient in 1978, and that it has changed substantially over the past 40 years. If the efficient coding hypothesis is true, then the 2018 system is also expected to be near-optimally efficient. The complexity and accuracy of the 2018 system are shown by the orange dot in Figure 2, and as expected, it lies near the theoretical limit. This observation is further validated by comparing the 2018 to its set of hypothetical variants. Figure 6C shows that the 2018 system is more efficient than all of its variants, and also more similar to the nearest optimal system. In this sense, the 2018 system is locally optimal. These findings alone, however, do not imply that the system had changed under persistent pressure for efficiency.

To find out whether it is likely that the observed change in Nafaanra was influenced by efficiency, we compare: (1) the actual trajectory, $Q_l(\mathcal{T})$, estimated from the 1978 and 2018 Nafaanra data; (2) the corresponding idealized IB trajectory, $Q_l^*(\mathcal{T})$; and (3) the set of hypothetical trajectories that the 1978 system could have followed, if it had changed without any pressure for efficiency. While the initial state of these hypothetical trajectories was fitted to the 1978 system, it is not clear which iteration corresponds to the 2018 system. Therefore, for each trajectory i and each iteration t, we ask how well $Q_h^i(\{0, t\})$ explains the actual trajectory.

Figure 5 shows that all the hypothetical trajectories diverge away from the curve over time, and become less efficient than the actual systems. This can also be seen by the light red area below the IB curve in Figure 2. This area is restricted because we forced the hypothetical systems to maintain Gaussian categories, which is a relatively good a-priori assumption for a color naming system. Interestingly, the initial system, which is a Gaussian approximation of the 1978 system, is more efficient than the actual systems. This suggests that the stochastic processes we simulated could in principle reach naturally-looking and highly efficient systems. However, we see that over time it does not tend to stay at such systems, even when initialized at a highly efficient system. Thus, we conclude that the fact that the 2018 system had remained near-optimally efficient over time is not a trivial property, because it is unlikely to be explained by a baseline process that is not pressured to remain near the theoretical limit.

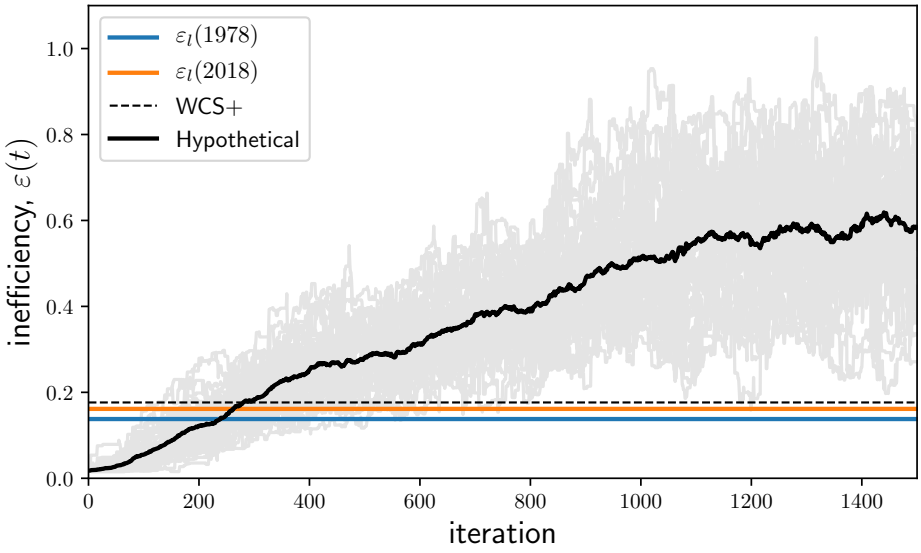

Figure 5. Inefficiency prediction. Gray curves show the inefficiency, $\varepsilon(t)$, of 50 hypothetical trajectories, as a function of the iteration t. The black curve is the average inefficiency across these trajectories. $\varepsilon_l(1978)$ and $\varepsilon_l(2918)$ correspond to the inefficiency of the 1978 and 2018 Nafaanra systems, and the dashed line corresponds to the average inefficiency across the WCS+ languages.

Next, we examine how well the IB trajectory explains the full structure of the actual trajectory. A qualitative comparison of the IB systems predicted for 1978 and 2018 (Figure 6A–B), and the actual systems (Figure 3), shows that the IB trajectory captures much of the structure seen in the data, including the observed stochastic category structure and the refinement of the system over time. However, this comparison also reveals how the actual systems deviate from the optimal ones: the 1978 does not exhibit the yellow category predicted by the model, and the 2018 exhibits purple and brown categories which are not predicted by the model.[3] These differences can potentially be explained by social or cultural factors that may influence language change, such as language contact, which are not taken into account in the model. Nonetheless, we argue that the idealized IB trajectory provides more insight about the actual trajectory, than the hypothetical trajectories. Qualitatively examining the systems along the hypothetical trajectories (Figure 4B for example), shows that although these systems maintain a reasonable category structure, over time they become intuitively unnatural and do not resemble the 2018 system (nor any of the WCS+ systems).

This observation is also supported by the quantitative analysis shown in Figure 7. Both the IB trajectory and the hypothetical trajectories start at a system that is more similar to the 1978 than to the 2018 system. For IB, as β increases, the IB systems along the curve become more similar to the 2018 system (lower gNID) and less similar to the 1978 system (higher gNID). The hypothetical trajectories, however, become less similar to the 1978 system but do

[3] At slightly higher values of β the model does predict brown and purple categories, resulting in a system that appears intuitively more similar to the 2018 Nafaanra system. An important direction for future work is to improve the estimation of β_l.

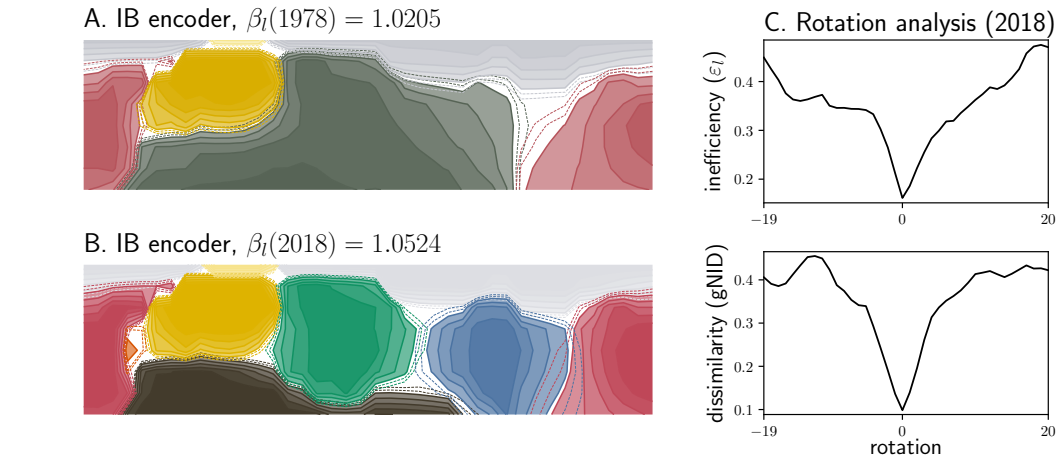

Figure 6. A–B. Contour plots of the IB systems that are predicted for the 1978 and 2918 Nafaanra systems. C. Rotation analysis for the 2018 Nafaanra system. This system is more efficient (lowest ε_l), and more similar to the nearest IB system (lowest gNID) than all of its hypothetical variants (see Figure 4A for examples).

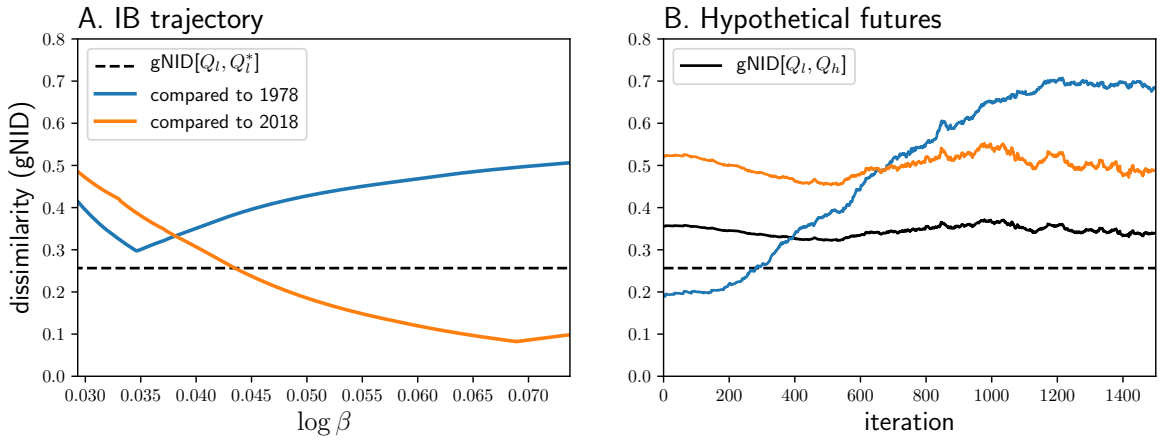

Figure 7. Dissimilarity prediction. A. Comparison between the IB systems along the theoretical limit, and the two Nafaanra systems. Dasshed line corresponds to the gNID between the actual trajectory $Q_l(\mathcal{T})$, and the IB trajectory, $Q_l^*(\mathcal{T})$, which is defined by the two IB systems at the initial and final values of β in this plot (these systems are shown in Figure 6A–B). B. Same as (A) for the hypothetical trajectories. Each curve shows the average across all trajectories.

not become similar to the 2018 system. This holds also for the gNID between the trajectories, i.e. $\text{gNID}[Q_l(\mathcal{T}), Q_l^*(\mathcal{T})] < \langle \text{gNID}[Q_l(\mathcal{T}), Q_h^i(\{0, t\})] \rangle_i$ for all t, as can be seen by comparing the dashed and solid black lines in Figure 7. This means that for every iteration t, if we stop the hypothetical process at that point and evaluate its similarity to the actual trajectory, it would perform worse compared to the IB trajectory.

Finally, we note that the hypothetical trajectories tend to add new categories over time and become more refined. This property is predicted by many theories of color category evolution (e.g., Berlin and Kay, 1969; Kay and Maffi, 1999; MacLaury, 1997; Levinson, 2000). However, our findings underscore that this alone is not enough for understanding the evolu-

tion of color naming, because many trajectories could exhibit this property while generating unnatural color naming systems.

5 Disucssion

In this work we have tested directly the hypothesis that color naming evolves under pressure to maintain efficient coding. We have done so by examining recent diachronic color naming data that exists for a single language, Nafaanra. We have shown that color naming in Nafaanra has changed over the past four decades in a way that is consistent with the predictions of the efficient coding hypothesis, and that this is not a trivial outcome because the observed change is unlikely to be explained by a baseline evolutionary process, that is not pressured by efficiency. To our knowledge, this is the first direct evidence in support of the general idea that the evolution of color naming, and possibly semantic systems more generally, is shaped by pressure for efficiency.

While we do not know what was the actual trajectory that the Nafaanra color naming system has undergone since 1978, our results suggest that the theoretically-motivated evolutionary trajectory derived from IB may be informative about the actual trajectory. At the same time, we do not argue that languages evolve by following exactly this idealized trajectory, because there are additional forces, such as language contact, that are likely to influence language change. An important direction for future research is to explore how these forces may operate in interaction with pressure for efficiency. Another important direction for future research is to test the extent to which our results extend to other languages, and to other semantic domains. The theoretical framework on which we build is not specific to color, and has been applied to other semantic domains (Zaslavsky et al., 2019b, and see also Kemp et al. (2018)), suggesting that efficient coding may be a fundamental principle in the evolution of the lexicon.

Acknowledgments

We thank Paul Kay for helpful discussions. This study was partially supported by the Gatsby Charitable Foundation (N.Z. and N.T.), and by the Defense Threat Reduction Agency (N.Z. and T.R.); the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

J. T. Abbott, T. L. Griffiths, and T. Regier. Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences*, 113(40): 11178–11183, 2016. doi: 10.1073/pnas.1513298113.

B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley and Los Angeles, 1969.

C. P. Biggam. *The semantics of colour: A historical approach*. Cambridge University Press, Cambridge, UK, 2012.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley New York, 1991.

K. Garvin. Nafaanra documentation project. In preparation.

E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017.

P. Kay. Synchronic variability and diachronic change in basic color terms. *Language in Society*, 4(3):257–270, 1975.

P. Kay and L. Maffi. Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101(4):743–760, 1999.

P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The World Color Survey*. Stanford: Center for the Study of Language and Information, 2009.

C. Kemp, Y. Xu, and T. Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), 2018. doi: 10.1146/annurev-linguistics-011817-045406.

S. C. Levinson. Yélî Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, 10(1):3–55, 2000.

D. T. Lindsey and A. M. Brown. The color lexicon of American English. *Journal of Vision*, 14 (2):17, 2014.

D. T. Lindsey, A. M. Brown, D. H. Brainard, and C. L. Apicella. Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, 25(18):2441–2446, 2015.

J. Lyons. Colour in language. In T. Lamb and J. Bourrieau, editors, *Colour: Art and science*, pages 194–224. University of Cambridge, Cambridge, UK, 1995.

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

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

T. Regier, C. Kemp, and P. Kay. Word meanings across languages support efficient communication. In B. MacWhinney and W. O'Grady, editors, *The Handbook of Language Emergence*, pages 237–263. Wiley-Blackwell, Hoboken, NJ, 2015.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948.

C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

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

N. Zaslavsky, C. Kemp, N. Tishby, and T. Regier. Communicative need in color naming. *Cognitive Neuropsychology*, 2019a. doi: 10.1080/02643294.2019.1604502.

N. Zaslavsky, T. Regier, N. Tishby, and C. Kemp. Semantic categories of artifacts and animals reflect efficient coding. In *41st Annual Conference of the Cognitive Science Society*, 2019b.

Chapter 4

Semantic Categories of Artifacts and Animals Reflect Efficient Coding

Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp (2019). Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the 41th Annual Conference of the Cognitive Science Society*.

Semantic categories of artifacts and animals reflect efficient coding

Noga Zaslavsky[1,2] **(noga.zaslavsky@mail.huji.ac.il)**
Terry Regier[2,3] **(terry.regier@berkeley.edu)**
Naftali Tishby[1,4] **(tishby@cs.huji.ac.il)**
Charles Kemp[5] **(c.kemp@unimelb.edu.au)**

[1]Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 9190401, Israel
[2]Department of Linguistics, University of California, Berkeley, CA 94720 USA
[3]Cognitive Science Program, University of California, Berkeley, CA 94720 USA
[4]Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 9190401, Israel
[5]School of Psychological Sciences, University of Melbourne, Parkville, Victoria 3010, Australia

Abstract

It has been argued that semantic categories across languages reflect pressure for efficient communication. Recently, this idea has been cast in terms of a general information-theoretic principle of efficiency, the Information Bottleneck (IB) principle, and it has been shown that this principle accounts for the emergence and evolution of named color categories across languages, including soft structure and patterns of inconsistent naming. However, it is not yet clear to what extent this account generalizes to semantic domains other than color. Here we show that it generalizes to two qualitatively different semantic domains: names for containers, and for animals. First, we show that container naming in Dutch and French is near-optimal in the IB sense, and that IB broadly accounts for soft categories and inconsistent naming patterns in both languages. Second, we show that a hierarchy of animal categories derived from IB captures cross-linguistic tendencies in the growth of animal taxonomies. Taken together, these findings suggest that fundamental information-theoretic principles of efficient coding may shape semantic categories across languages and across domains.

Keywords: information theory; language evolution; semantic typology; categories

Introduction

Cross-linguistic studies in several semantic domains, such as kinship, color, and numeral systems, suggest that word meanings are adapted for efficient communication (see Kemp, Xu, & Regier, 2018 for a review). However, until recently it had remained largely unknown to what extent this proposal can account for soft semantic categories and inconsistent naming, that could appear to pose a challenge to the notion of efficiency, and how pressure for efficiency may relate to language evolution. Recently Zaslavsky, Kemp, Regier, and Tishby (2018; henceforth ZKRT) addressed these open questions by grounding the notion of efficiency in a general information-theoretic principle, the Information Bottleneck (IB; Tishby, Pereira, & Bialek, 1999). ZKRT tested this formal approach in the domain of color naming and showed that the IB principle: (1) accounts to a large extent for cross-language variation in color naming; (2) provides a theoretical explanation for why observed patterns of inconsistent naming and soft semantic categories may be efficient; and (3) suggests a possible evolutionary process that roughly recapitulates Berlin and Kay's (1969) discrete implicational hierarchy while also accounting for continuous aspects of color category evolution. However, it is not yet clear to what extent these results may generalize to other semantic domains, especially those that are fundamentally unlike color.

Here we test the generality of this theoretical account by considering two additional semantic domains: artifacts and animals. These domains are of particular interest in this context because they are qualitatively different from color, they have not previously been comprehensively addressed in terms of efficient communication, and at the same time it is possible to apply to them the same communication model that has previously been used to account for color naming.

First, we consider naming patterns for household containers. This is a semantic domain in which categories are known to overlap and generate inconsistent naming patterns (Ameel, Storms, Malt, & Sloman, 2005; Ameel, Malt, Storms, & Assche, 2009). Although it has previously been shown that container naming in English, Spanish, and Chinese is efficient compared to a large set of hypothetical naming systems (Xu, Regier, & Malt, 2016), that demonstration did not consider the full probability distribution of names produced by different speakers, did not explicitly contrast monolingual and bilingual speakers, and was based on a smaller set of stimuli than we consider here. In this work we show that the full container-naming distribution in Dutch and French, including overlapping and inconsistent naming patterns, across a large set of stimuli, both in monolinguals and bilinguals, is near-optimally efficient in the IB sense.

Second, we test the evolutionary account of ZKRT in the case of animal categories. By analogy with Berlin and Kay's implicational hierarchy of color terms, Brown (1984) proposed an implicational hierarchy for the evolution of animal taxonomies based on cross-language comparison. We show that aspects of this hierarchy are captured by a sequence of efficient animal-naming systems along the IB theoretical limit. Our results also support the view that both perceptual and functional features shape animal categories across languages (Malt, 1995; Kemp et al., 2018).

The remainder of this paper proceeds as follows. First, we review the theoretical framework and formal predictions on which we build. We then present two studies that apply this approach to the aforementioned semantic domains.

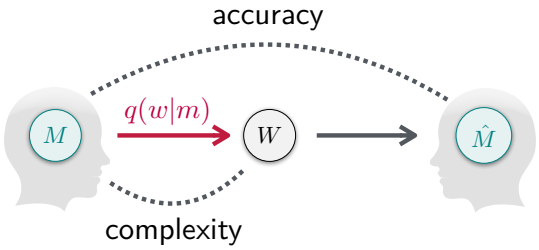

Figure 1: Communication model adapted from ZKRT. A speaker communicates a meaning M by encoding it into a word W according to a naming distribution $q(w|m)$. This word is then interpreted by the listener as $\hat{M}$. Complexity is a property of the mapping from meanings to words, and accuracy is determined by the similarity between M and $\hat{M}$.

Theoretical framework and predictions

We consider here the theoretical framework proposed by ZKRT, which is based on a simplified interaction between a speaker and a listener (Figure 1), formulated in terms of Shannon's (1948) communication model. The speaker communicates a meaning m, sampled from $p(m)$, by encoding it into a word w, generated from a naming (or encoder) distribution $q(w|m)$. The listener then tries to reconstruct from w the speaker's intended meaning. We denote the reconstruction by $\hat{m}_w$, and assume it is obtained by a Bayesian listener.[1] These meanings, m and $\hat{m}_w$, are taken to be mental representations of the environment, defined by distributions over a set $\mathcal{U}$ of relevant features. For example, if communication is about colors, then $\mathcal{U}$ may be grounded in a perceptual color space, and each color would be mentally represented as a distribution over this space.

Under these assumptions, efficient communication systems are those naming distributions that optimize the Information Bottleneck (IB; Tishby et al., 1999) tradeoff between the complexity and accuracy of the lexicon. Formally, complexity is measured by the mutual information between meanings and words, i.e.:

$$I_q(M;W) = \sum_{m,w} p(m)q(w|m)\log\frac{q(w|m)}{q(w)}, \qquad (1)$$

which roughly corresponds to the number of bits used to encode meanings into words. Accuracy is inversely related to the discrepancy between m and $\hat{m}_w$, measured by the expected Kullback–Leibler (KL) divergence between them:

$$\mathbb{E}_q[D[m\|\hat{m}_w]] = \mathop{\mathbb{E}}_{\substack{m\sim p(m)\\ w\sim q(w|m)}}\left[\sum_{u\in\mathcal{U}} m(u)\log\frac{m(u)}{\hat{m}_w(u)}\right]. \qquad (2)$$

Accuracy is defined by $I_q(W;U) = \mathbb{E}_q[D[\hat{m}_w\|m_0]]$, where

[1] The reconstruction of a Bayesian listener with respect to a given naming distribution is defined by $\hat{m}_w = \sum_m q(m|w)m$.

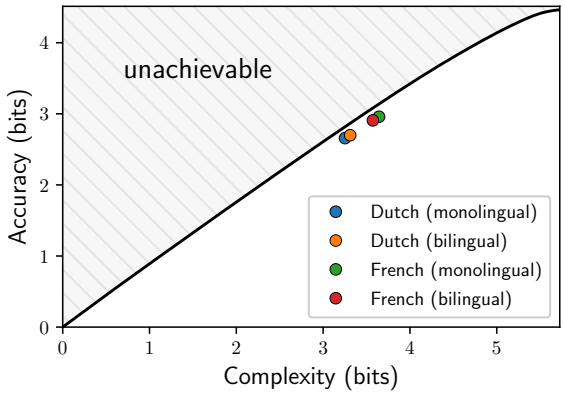

Figure 2: The black curve is the IB theoretical limit of efficiency for container naming, obtained by varying β. Points above this curve cannot be achieved. Complexity and accuracy tradeoffs in the four naming conditions are near-optimal.

m_0 is the prior representation before knowing w, and maximizing accuracy amounts to minimizing equation (2).[2]

Achieving maximal accuracy may require a highly complex system, while minimizing complexity will result in a non-informative system. Efficient systems are thus pressured to balance these two competing goals by minimizing the IB objective function,

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U), \qquad (3)$$

where $\beta \geq 0$ controls the efficiency tradeoff. The optimal systems, $q_\beta(w|m)$, achieve the minimal value of equation (3) given β, denoted by $\mathcal{F}^*_\beta$, and evolve as β gradually shifts from 0 to ∞. Along this trajectory they become more fine-grained and complex, while attaining the maximal achievable accuracy for their level of complexity. This set of optimal systems defines the theoretical limit of efficiency (see Figure 2).

If languages are pressured to be efficient in the IB sense, then for a given language l with naming system $q_l(w|m)$, two predictions are made. (1) Deviation from optimality, or *inefficiency*, should be small. This is measured by $\varepsilon_l = \frac{1}{\beta_l}(\mathcal{F}_{\beta_l}[q_l] - \mathcal{F}^*_{\beta_l})$, where β_l is estimated such that ε_l is minimized. (2) The *dissimilarity* between q_l and the corresponding IB system, q_{β_l}, should be small. This is evaluated by a dissimilarity measure (gNID) proposed by ZKRT. In addition, ZKRT suggested that languages evolve along a trajectory that is pressured to remain near the theoretical limit.

These predictions were previously supported by evidence from the domain of color naming. To apply this approach to other domains, i.e. to instantiate the general communication model, two components must be specified: a *meaning space*, which is the set of meanings the speaker may communicate; and a prior, $p(m)$, also referred to as a *need distribution* (Regier, Kemp, & Kay, 2015), since it determines the frequency with which each meaning needs to be communicated. In the following sections we present two studies that

[2] See (Zaslavsky et al., 2018) for detailed explanation.

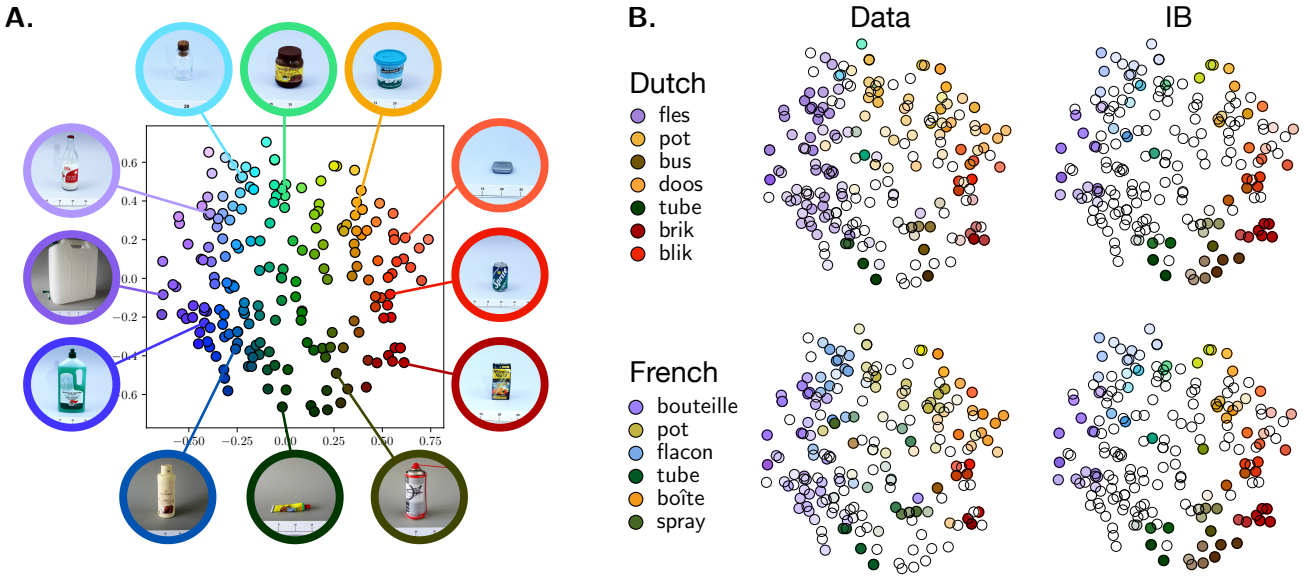

Figure 3: **A**. Two dimensional nMDS embedding and color coding of the containers stimulus set used by White et al. (2017). Images show a few examples. **B**. Monolingual naming distributions for Dutch (upper left) and French (lower left), together with their corresponding IB systems (right column), are visualized over the 2D embedding shown in (A). Each color corresponds to the color centroid of a container category, w, based on the color map in (A). Colors show category probabilities above 0.4, and color intensities reflect the values between 0.4 and 1. White dots correspond to containers for which no category is used with probability above 0.4. Legend for each language shows only major terms.

follow this approach and test its predictions in qualitatively different semantic domains.

Study I: Container names

The goal of this experiment is to test the theoretical predictions derived from IB in the case of container naming. It is not clear whether previous findings for color would generalize to this case for several reasons. First, the representation of artifacts is likely to involve more than just a few basic perceptual features, unlike color. Second, categories in this domain are believed to be strongly shaped by adaptation to changes in the environment (Malt, Sloman, Gennari, Shi, & Wang, 1999). At the same time, container categories tend to overlap, as in the case of color categories, posing a similar theoretical challenge to explain this observation in terms of communicative efficiency. Finally, the bilingual lexicon in this domain has been extensively studied, and it has been shown that bilingual naming patterns tend to converge (Ameel et al., 2005, 2009). However, it is not yet clear whether this convergence, or compromise, comes at a cost in communicative efficiency, or whether it may actually be formalized and explained in terms of efficiency.

Data. To address these open questions, we consider sorting and naming data collected by White et al. (2017), relative to a stimulus set of 192 images of household containers (see Figure 3A for examples). This set is substantially larger than those used in previous container-naming studies (e.g. Malt et al., 1999; Ameel et al., 2005), thus providing a better representation of this semantic domain. In the naming task, 32 Dutch and 30 French monolingual speakers, as well as 30 bilingual speakers, were asked to provide names for the containers in the stimulus set. Bilingual participants performed the task once in each language. The container-naming distribution in each of the four conditions (language $\times$ linguistic status) is defined by the proportion of participants in that condition that used the word w to describe a container c. A separate sorting task was performed by 65 Dutch speakers, who were asked to organize all containers into piles based on their overall qualities. Participants were also allowed to form higher-level clusters by grouping piles together. White et al. (2017) evaluated the similarity between two containers, denoted here by $\mathrm{sim}(c, c')$, based on the number of participants that placed them in the same pile or cluster (see White et al., 2017 for detail). In both tasks, participants were instructed not to take into account the content of the object (e.g., water).

Model. We ground the meaning space in the similarity data, following a related approach proposed by Regier et al. (2015) and Xu et al. (2016). While these data are from Dutch speakers, there are only minor differences in perceived similarities among speakers of different languages (Ameel et al., 2005). Therefore, we assume that these similarity judgments reflect a shared underlying perceptual representation of this domain. We take $\mathcal{U}$ to be the set of containers in the stimulus set, and define the mental representation of each container c by the similarity-based distribution it induces over the domain, $m_c(u) \propto \exp(\gamma \cdot \mathrm{sim}(c, u))$, where γ^{-1} is taken to be the

empirical standard deviation of $\mathrm{sim}(c, u)$. In contrast with the case of color, in which these mental representations were grounded in a standard perceptual space, here there is no standard perceptual space for containers, and so our assumed underlying perceptual representation requires further validation, which we leave for future work. We define the need distribution, $p(m_c)$, by averaging together the least informative (LI) priors for the different languages, as proposed by ZKRT. We used only the monolingual data for this purpose, and regularized the resulting prior by adding $\epsilon = 0.001$ to it and renormalizing.

Results

We estimated the theoretical limit of efficiency for container naming by applying the IB method (Tishby et al., 1999), as ZKRT did in the case of color naming, here with 1500 values of $\beta \in [0, 1024]$. We evaluated the empirical complexity and accuracy in the four naming conditions by entering the corresponding naming distributions in the equations for $I_q(M;W)$ and $I_q(W;U)$. The results are shown in Figure 2 and Table 1. It can be seen that container naming in Dutch and French lie near theoretical limit, both for monolinguals and bilinguals, and that bilinguals achieve similar levels of efficiency as monolinguals (Table 1). In all four cases, the corresponding IB solution is at $\beta_l \approx 1.2$, suggesting that there is only a weak preference for accuracy over complexity in this domain, as also found for color naming.

Consistent with the empirical observations of convergence in the bilingual lexicon, the complexity-accuracy tradeoffs in bilinguals are closer to each other (Figure 2, orange and red dots) compared to the monolingual tradeoffs (Figure 2, blue and green dots). This may be explained by a need to reduce the complexity of maintaining two naming systems simultaneously, while achieving monolingual-like levels of efficiency in each language. To test this possibility, we compared two joint French-Dutch systems that bilinguals may employ: one that randomly selects one of the two monolingual systems to name objects, and another that randomly selects one of the two bilingual systems. We found a 0.16% reduction in the complexity of the joint bilingual system compared to the joint monolingual system. Although this is a small effect, it may accumulate across domains to have a substantial impact. In addition, our simple calculation did not take into account similar word forms, which may also reduce complexity (Ameel et al., 2005). Thus, this finding suggests that the convergence in the bilingual lexicon may be shaped, at least in part, by pressure for efficiency.

The remainder of our analysis focuses on the monolingual systems, as they are more distinct and presumably more representative of each language. To get a precise sense of how challenging it may be to reach the observed levels of efficiency, we compared the actual naming systems to a set of hypothetical systems that preserve some of their statistical structure. This set was constructed by fixing the conditional distributions of words, while shifting how they are used by applying a random permutation of the containers. For each language we constructed $10,000$ such hypothetical systems. Table 1 shows that these hypothetical systems are substantially less efficient than the actual systems, and are also less similar to the IB systems. In fact, both languages achieve better (lower) scores than all of their hypothetical variants, providing a precise sense in which they are near-optimal according to IB. One possible concern is that this outcome may be a result of the LI prior, which was fitted to the naming data. To address this, we repeated this analysis with a uniform need distribution. The results in that case are similar (not shown), although as expected the fit to the actual systems is not as good compared to the LI prior.

Table 1: Evaluation of the IB container-naming model. Lower values indicate a better fit of the model. Values for hypothetical systems are averages ±SD over $10,000$ systems.

		Inefficiency	Dissimilarity
Dutch	monolingual	0.16	0.11
	bilingual	0.17	0.12
	hypothetical	0.29 (±0.02)	0.59 (±0.05)
French	monolingual	0.18	0.11
	bilingual	0.17	0.09
	hypothetical	0.31 (±0.01)	0.56 (±0.06)

The low dissimilarity scores for the actual languages, shown in Table 1, suggest that the observed soft category structure in this domain may also be accounted for by the IB systems. This is indeed supported by a fine-grained comparison between the naming distribution in both languages and their corresponding IB systems. To see this, we embedded the 192 containers in a 2-dimensional space by applying non-metric multidimensional scaling (nMDS) with respect to the similarity data, similar to Ameel et al. (2009). This was done using the scikit-learn package in Python. We initialized the nMDS procedure with a solution for the standard metric MDS that achieved the best fit to the similarity data out of 50 solutions generated with random initial conditions. For visualization purposes, we assigned a unique color to each container. The resulting 2D embedding and color coding of the containers stimulus set are shown in Figure 3A.

The monolingual systems in Dutch and French are shown in Figure 3B, together with their corresponding IB systems. These two IB systems are very similar, although not identical, which is not surprising given that the naming patterns in Dutch and French are fairly similar. Both the actual systems and the IB systems exhibit soft category structure and similar patters of inconsistent naming, as shown by the white dots. In addition, since each category is colored according to its centroid, similarity between the category colors together with their spatial distribution reflect the similarity between the full naming distributions. For example, the IB systems have a category that is similar to *fles* and *bouteille*, as well as a category that is similar to *doos* and *boîte* in Dutch and French respectively, although these categories in the IB systems are a bit narrower. The IB systems also capture the

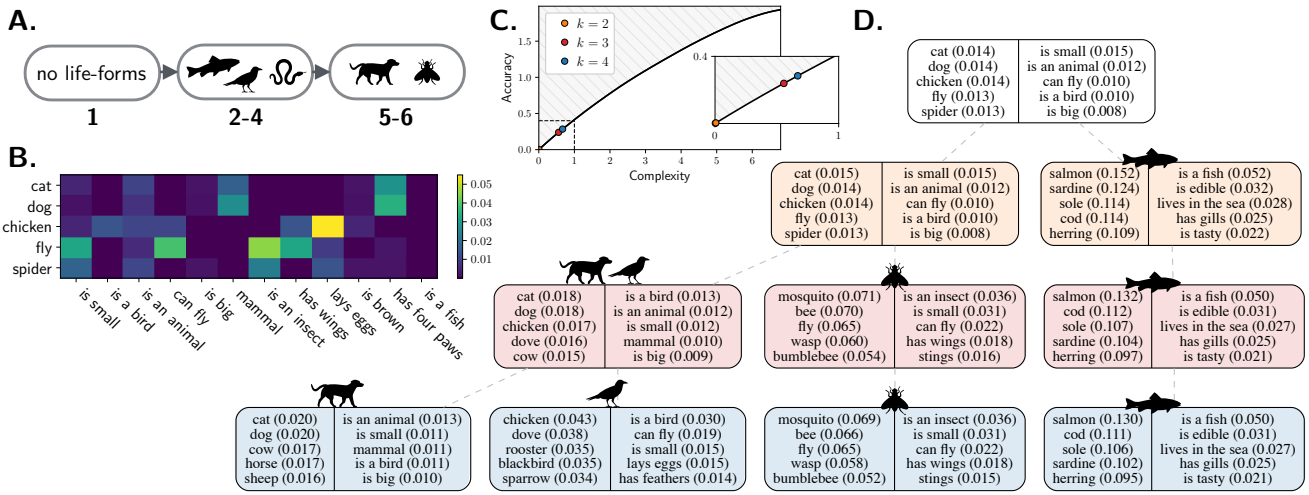

Figure 4: **A**. Brown's (1984) proposed hierarchy for animal categories. **B**. Subset of the conditional probabilities of features (columns) given animal classes (rows), for the 5 most familiar classes and 12 most frequently generated features. **C**. Theoretical limit for animal naming. Colored dots along the curve correspond to the systems shown in (D), with $k = 2, 3, 4$ categories. **D**. Animal category hierarchy derived from IB. Each level corresponds to an IB system. Each box corresponds to a category, which is represented by its top five classes (left) and features (right) and their probabilities given the category.

category *tube* quite well in both languages. However, there are also some apparent discrepancies. For example, the distinction between *bouteille* and *flacon* in French is reflected in both IB systems, although Dutch does not have the same pattern in this case (Ameel et al., 2005).

This analysis shows that efficiency constraints may to a substantial extent explain the container-naming distribution in Dutch and French, including soft category boundaries and inconsistent naming observed empirically, both in monolinguals and bilinguals. It thus supports the hypothesis that a drive for information-theoretic efficiency shapes word meanings across languages and across semantic domains. However, since this analysis is based only on two closely related languages, we were not able to test how well the results for this domain generalize across languages. Important directions for future research include testing whether these results generalize to other, preferably unrelated, languages, and further testing the extent to which the convergence in the bilingual lexicon is influenced by pressure for efficiency. The next section focuses on another semantic domain for which we are able to obtain broader cross-linguistic evidence.

Study II: Folk biology

Cross-language variation and universal patterns in animal taxonomies have been extensively documented and studied (Berlin, 1992), however this domain has not yet been approached in terms of efficient communication. By analogy with Berlin and Kay's theory, Brown (1984) proposed an implicational hierarchy for animal terms, based on data from 144 languages. Brown identified six stages for animal taxonomies, as illustrated in Figure 4A. Languages at the first stage do not have any lexical representation for life-forms. Languages at stages 2-4 add terms for *fish*, *bird* and *snake*, but Brown does not argue for any particular order for these categories. Terms for *mammal* and *wug* ("worm-bug", referring in addition to small insects) are added in stages 5 and 6, again with no implied order. Much of the data analyzed in this domain is not fine-grained, and Brown's proposal has been criticized (Randall & Hunn, 1984) mainly due to lack of sufficiently accurate data. Nonetheless, his observations can be considered as a rough approximation of cross-linguistic tendencies in this semantic domain. Therefore, in this work we aim at testing whether broad cross-linguistic patterns, as summarized by Brown's proposal, can be accounted for in terms of pressure for efficiency. More specifically, our goal is to derive from the IB principle a trajectory of efficient animal-naming systems, analogous to ZKRT's trajectory for color, and to compare this trajectory to the naming patterns reported by Brown. However, unlike previous comparisons to IB optima, due to the nature of available data, here we only attempt to make coarse comparisons.

To derive a trajectory of efficient animal-naming systems, we first need to specify the communication model in this domain. We ground the representations of animals in high-level, human-generated features. Specifically, we consider the Leuven Natural Concept Database (De Deyne et al., 2008), which contains feature data and familiarity ratings for animal classes (e.g., "cat", "chicken", etc.). These data were collected from Dutch speakers, and then translated to English. We follow Kemp, Chang, and Lombardi (2010), who considered 113 animal classes and 757 features from this database, and for each feature u and class c estimated the conditional probability $p(u|c)$ based on the number of participants who generated this feature for that class (see Figure 4B for exam-

ples). We take $\mathcal{U}$ to be the set of animal features, and assume each animal class is mentally represented by the distribution it induces over features, i.e. $m_c(u) = p(u|c)$, as estimated by Kemp et al. (2010). In addition, we follow Kemp et al. (2010) in using a familiarity-based prior over animal classes, in which the probability of a class is proportional to its familiarity score. We define the need distribution to be this prior.

Given these components, we estimated the theoretical limit for animal naming (Figure 4C) using the same method as before, this time with 3000 values of $\beta \in [0, 2^{13}]$. We then selected the most informative systems with $k = 2, 3, 4$ categories. The number of categories, k, was determined by considering categories w with probability mass $q_\beta(w) > 0.00001$. These systems are shown in Figure 4D, where each layer of the hierarchy corresponds to a system and each box corresponds to a category within that system. The top layer, with a single category, corresponds to a non-informative system that does not distinguish between different animal classes. This can be considered as a stage 1 system in Brown's sequence. The second layer (shown in orange) roughly corresponds to a stage 2 system. It consists of a *fish* category, as can be inferred from the distribution it induces over features and animals, and another category for all other animals. It lies very close to the origin in Figure 4C, as it maintains little information about most animals. The third layer (shown in red) corresponds to a system with categories for *fish* and *wug*, as well as a category that is dominated by birds and mammals. The *bird-mammal* category has greater probability mass (0.8) than the *wug* category (0.14), suggesting that it is more prominent even though these two categories appear together. This transition deviates from Brown's sequence in the early appearance of *wug* (although not strongly weighted here), and in lacking a *snake* category (although animals from that category do appear in the Leuven database). One possible explanation for this deviation is that the feature data on which we relied were obtained from Dutch participants, and are thus strongly biased toward Western societies. In the next layer (shown in blue), the 3-category system evolved to a 4-category system by refining the *bird-mammal* category, resulting in a system that roughly corresponds to a Brown stage 6 system, with the exception of *snake*.

These results suggest that animal naming systems may evolve under efficiency pressure much as color appears to, despite the qualitative difference between these domains. However, in order to test this proposal more comprehensively, fine-grained cross-linguistic animal naming data is required, comparable to the naming data for colors and containers. The fact that systems along the theoretical limit capture some cross-linguistic tendencies in animal taxonomies is notable, given that our characterization of the domain, in terms of features, was necessarily strongly biased toward animal representations in Western societies. This finding supports the idea that to some extent at least there is a shared underlying representation of animals across cultures (Mayr, 1969), while also raising the interesting possibility of some cross-language and cross-cultural differences in underlying representations. It is also worth noting that the salient features in the IB systems tend to be both perceptual (e.g., "is big") and functional (e.g., "is edible"), suggesting that both types of features may shape animal categories across languages, and that this may be consistent with pressure for efficiency (Kemp et al., 2018).

Although we introduced the hierarchy in Figure 4D as an account of category structure across languages, the same hierarchy could potentially serve as a model of hierarchical structure within a single language. This within-language interpretation resembles previous applications of the IB principle to language (Pereira, Tishby, & Lee, 1993), although these applications were based on corpus statistics. The within-language interpretation seems useful in the case of animal taxonomies, a semantic domain with strong hierarchical structure, as opposed to containers and even colors. A possible, yet speculative, reconciliation of the within-language and cross-language interpretations is that speakers may internally represent a hierarchy induced by an evolutionary sequence. For example, Boster (1986) showed that English speakers can recapitulate Berlin and Kay's implicational color hierarchy in a sequential pile-sorting task. Thus, it seems at least possible that a similar phenomenon may also hold for animal categories.

General discussion

Artifacts, animals, and colors are qualitatively different elements of human experience, yet our findings suggest that their semantic representations across languages is governed by the same general information-theoretic principle: efficient coding of meanings into words, as defined by the IB principle. We have shown that this theoretical account, which was previously tested only in the domain of color naming (ZKRT), generalizes to container names and animal taxonomies. This finding resonates with the proposal that word meanings may be shaped by pressure for efficient communication (Kemp et al., 2018). However, it goes beyond that proposal by explaining how pressure for efficiency may account for soft categories and inconsistent naming, both in monolinguals and bilinguals, and how it may relate to language evolution.

An important direction for future research is to test to what extent our results extend to other semantic domains, and ideally, to the lexicon as a whole. While it may not be possible to apply this approach to every aspect of the lexicon, we believe that the theoretical formulation considered here may be broadly applicable across semantic domains.

Acknowledgments

We thank Anne White, Gert Storms, and Barbara Malt for making their container naming and sorting data publicly available. The animal features and familiarity data we used were preprocessed by Kemp et al. (2010). We thank Simon De Deyne for initially sharing these data, and for useful discussions. This study was partially supported by the Gatsby Charitable Foundation (N.Z. and N.T.), and by the Defense

Threat Reduction Agency (N.Z. and T.R.); the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

Ameel, E., Malt, B. C., Storms, G., & Assche, F. V. (2009). Semantic convergence in the bilingual lexicon. *Journal of Memory and Language*, *60*(2), 270–290.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language*, *53*(1), 60–80.

Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton University Press.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley and Los Angeles: University of California Press.

Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, *25*(1), 61–74.

Brown, C. H. (1984). *Language and living things: Uniformities in folk classification and naming*. Rutgers University Press.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*(3), 216–233.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85–148.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230 - 262.

Mayr, E. (1969). The biological meaning of species. *Biological Journal of the Linnean Society*, *1*(3), 311-320.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics* (pp. 183–190).

Randall, R. A., & Hunn, E. S. (1984). Do life-forms evolve or do uses for life? Some doubts about Brown's universals hypotheses. *American Ethnologist*, *11*(2), 329-349.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck method. In *37th annual Allerton conference on communication, control and computing.*

White, A., Malt, B. C., & Storms, G. (2017). Convergence in the bilingual lexicon: A pre-registered replication of previous studies. *Frontiers in Psychology*, *7*.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS*, *115*(31), 7937–7942.

Part II

Information-Theoretic Approach to Communicative Need

Chapter 5

Color Naming Reflects both Perceptual Structure and Communicative Need

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier (2019). Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, 11(1):207–219. DOI: 10.1111/tops.12395.

TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2018) 1–13
© 2018 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8765 online
DOI: 10.1111/tops.12395

This article is part of the topic "Best Of Papers from the Cognitive Science Society Annual Conference," Wayne D. Gray (Topic Editor). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

Color Naming Reflects Both Perceptual Structure and Communicative Need

Noga Zaslavsky,[a,b] Charles Kemp,[c] Naftali Tishby,[a,d] Terry Regier[b,e]

[a]*Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem*
[b]*Department of Linguistics, University of California, Berkeley*
[c]*School of Psychological Sciences, The University of Melbourne*
[d]*Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem*
[e]*Cognitive Science Program, University of California, Berkeley*

Received 17 September 2018; accepted 7 October 2018

Abstract

Gibson et al. (2017) argued that color naming is shaped by patterns of communicative need. In support of this claim, they showed that color naming systems across languages support more precise communication about warm colors than cool colors, and that the objects we talk about tend to be warm-colored rather than cool-colored. Here, we present new analyses that alter this picture. We show that greater communicative precision for warm than for cool colors, and greater communicative need, may both be explained by perceptual structure. However, using an information-theoretic analysis, we also show that color naming across languages bears signs of communicative need beyond what would be predicted by perceptual structure alone. We conclude that color naming is shaped both by perceptual structure, as has traditionally been argued, and by patterns of communicative need, as argued by Gibson et al. —although for reasons other than those they advanced.

Keywords: Information theory; Color naming; Categorization

Correspondence should be sent to Noga Zaslavsky, Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 9190401, Israel. E-mail: noga.zaslavsky@mail.huji.ac.il

1. Introduction

Languages vary widely in the ways they partition colors into categories. At the same time, this variation is constrained, and similar color naming systems are often seen in unrelated languages (e.g., Berlin & Kay, 1969; Lindsey & Brown, 2006). The forces that give rise to this constrained variation have long been debated, and it is often held that a major role is played by perceptual structure (e.g., Kay & McDaniel, 1978). A variant of this view emphasizes in addition the importance of communicative forces, and it argues that languages divide perceptual color space into categories in ways that support efficient communication (Baddeley & Attewell, 2009; Jameson & D'Andrade, 1997; Lindsey et al., 2015; Regier et al., 2007; Regier et al., 2015; Zaslavsky et al., 2018).

Recently, Gibson et al. (2017) suggested an even greater role for communicative forces. They proposed that cross-language commonalities in color naming may reflect a human need to refer to particular colors more than others, and they presented this hypothesis as an alternative to one based on perceptual salience (p. 10785). They showed that color naming systems across languages support more precise communication about warm colors than cool colors, and that the objects we talk about tend to be warm-colored rather than cool-colored—suggesting that color naming systems may have adapted to a general human need to communicate preferentially about warm colors.

Here, we engage this argument and present results that suggest a somewhat different conclusion. We first present the core of Gibson et al.'s argument in detail and replicate their findings. We then consider an alternative explanation of their findings and show that greater communicative precision for warm than for cool colors, and greater need for warm colors, may both be explained by perceptual structure, without any additional communicative preference for warm colors. We next present a novel information-theoretic analysis of the link between need and communicative precision, and we use that analysis to infer need from color naming data. On that basis, we show that color naming across languages bears signs of communicative need beyond what would be predicted by perceptual structure alone. We conclude that color naming is shaped both by perceptual structure, as has traditionally been argued, and by patterns of communicative need, as argued by Gibson et al.—although our reasons for implicating need are different from theirs.

2. The argument of Gibson et al. (2017)

Gibson et al. found that across languages, warm colors tend to be communicated more precisely than cool colors. They also found that the objects we talk about tend to be warm-colored rather than cool-colored, and in that sense warm colors have higher communicative need. They concluded that the warm–cool asymmetry in communicative precision across languages "reflects colors of universal usefulness" and that the principle of color use "governs how color categories come about"

(p. 10785). They presented this idea as an alternative to proposals based on perceptual salience (p. 10785). Below we present the data they considered, and their definitions of communicative precision and communicative need, which inform our own analyses.

2.1. Data

Gibson et al. based their analysis primarily on color naming data from the World Color Survey (WCS: Cook, Kay, & Regier, 2005). The WCS dataset contains color naming data from 110 languages of non-industrialized societies. In the WCS, native speakers of each language were asked to provide a name for each of 330 color chips. Gibson et al. analyzed naming data for the subset of 80 color chips shown in Fig. 1, for all WCS languages and also for three languages for which they collected data: English, Spanish, and Tsimané. For each language l, each color term w in l, and each color chip c, they estimated the color naming distribution $p_l(w|c)$ as the proportion of speakers of l who used w rather than some other term to name c.

2.2. Communicative need

A need distribution, reflecting how often a given color c is used in communication, can be naturally considered a prior distribution $p(c)$ over colors (Regier et al., 2015). Gibson et al. considered two priors: a uniform prior and a "salience-weighted prior" (p. 27 of their SI). In the salience-weighted prior, the probability of each color was determined by the proportion of times that color appeared in a foreground object, rather than in the background, in their study of natural images. This prior was based on the assumption that foreground objects are more likely to be talked about than are backgrounds. This salience-weighted prior exhibits greater probability mass for warm colors than for cool colors (see Fig. 4C).

2.3. Communicative precision

Gibson et al. considered the expected surprisal of a given color c, with respect to a color naming distribution $p(w|c)$ and a prior $p(c)$, defined by

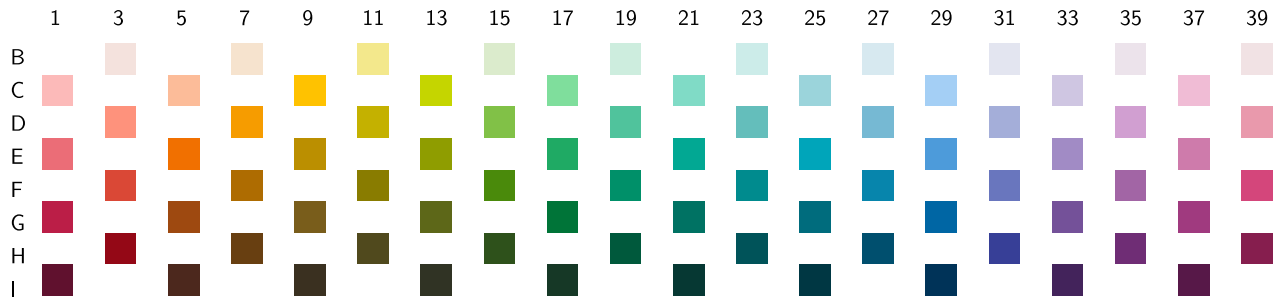

Fig. 1. The 80 color chips analyzed by Gibson et al. (2017), represented in the standard WCS palette. White spaces indicate WCS chips that were excluded from the analysis. The achromatic WCS color chips were also excluded.

$$S(c) = -\sum_{w} p(w|c) \log p(c|w), \tag{1}$$

where $p(c|w)$ is obtained by applying Bayes' rule:

$$p(c|w) = \frac{p(w|c)p(c)}{\sum_{c'} p(w|c')p(c')}. \tag{2}$$

Lower values of $S(c)$ correspond to higher communicative precision for a given color c. Gibson et al. found that across languages $S(c)$ tends to be lower for warm colors (reds/yellows) than for cool colors (blues/greens), when evaluated either with the uniform prior or with the salience-weighted prior. We replicated these results on very similar data (the WCS+ dataset; see below) for both priors, as shown in Fig. 3A and 3B.

Notice that $S(c)$ depends both on the prior $p(c)$ and on the naming system $p(w|c)$, and thus these results are an outcome of the combination of need and language. Here, we further explore the nature of this combination in two ways: first by using the same priors as Gibson et al. while considering new hypothetical color naming data, and second by keeping the color naming data fixed and considering new priors.

3. The role of perceptual structure

The crux of Gibson et al.'s argument is that the warm–cool asymmetry in precision may reflect the warm–cool asymmetry in need. Another possibility, however, is that both asymmetries may be produced by a common underlying cause, perhaps perceptual in nature. Fig. 2 re-plots the 80 colors from Fig. 1 in CIELAB color space, in which the Euclidean distance between nearby colors corresponds roughly to their perceptual dissimilarity (Brainard, 2003; but see also Komarova & Jameson, 2013). This visualization shows that there exist potentially relevant perceptual asymmetries of color—and in fact this perceptual structure has been used to explain patterns of color naming across languages (Jameson & D'Andrade, 1997; Regier et al., 2007, 2015; Zaslavsky et al., 2018). We wished to understand whether the structure of perceptual color space could also explain the asymmetry in precision documented by Gibson et al., or that in need, or both—a possibility acknowledged by Gibson et al. (p. 10789).

To test whether perceptual structure can account for the warm–cool precision asymmetry, we considered a set of hypothetical color naming systems that were derived solely from the structure of color space, without any additional element of communicative need. We began with the color naming data of the WCS, supplemented by data for English (Lindsey & Brown, 2014); we call this joint dataset WCS+. We considered the same 80 chips used by Gibson et al. Then for each actual language l, we constructed a corresponding hypothetical system by clustering the 80 color chips into k_l categories, using the k-

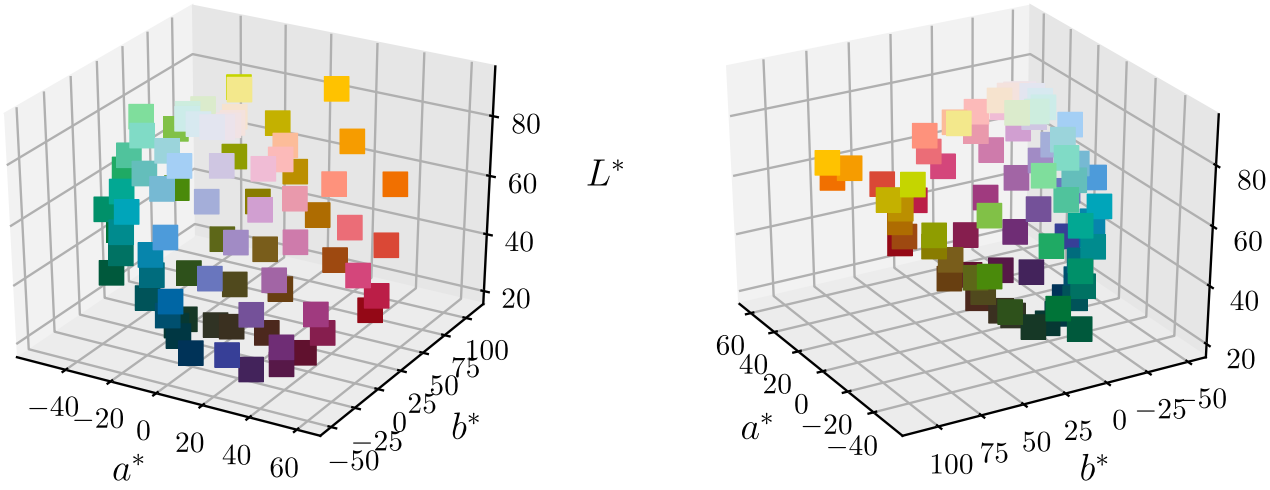

Fig. 2. The 80 color chips of Fig. 1, represented in CIELAB color space. L^* corresponds to lightness, and hue and saturation are represented in polar coordinates in the orthogonal plane defined by a^* and b^*. The irregular distribution of these colors reflects a perceptual asymmetry between warm and cool colors.

means algorithm with respect to the Euclidean distance between colors in CIELAB space. We took k_l to be the number of color terms in language l for which at least two speakers used that term to name the same color chip. In an attempt to avoid local optima, we ran the k-means algorithm 30 times for each language and retained the best solution. This procedure yielded a set of artificial color naming systems that are comparable in number of terms to those in our cross-language data but are determined only by the structure of perceptual color space, with no additional element of need.

The lower panels of Fig. 3 show that these k-means systems exhibit a warm–cool surprisal asymmetry broadly similar to that in the actual languages, both with the uniform prior (Fig. 3C) and with the salience-weighted prior (Fig. 3D). In support of this qualitative observation, with the salience-weighted prior, we found a strong correlation ($r = 0.73$, $p < 0.0001$) between $S(c)$ averaged across actual languages and $S(c)$ averaged across the corresponding k-means systems. With the uniform prior, although an overall warm–cool asymmetry is visually apparent, there is also a clear discrepancy between the actual languages and the k-means systems: Light colors tend to have relatively low surprisal in the actual languages, but high surprisal in the k-means systems. In this case we did not find a significant correlation between average surprisal across actual and k-means systems when considering all color chips, but we did find a significant correlation ($r = 0.57$, $p < 0.0001$) when focusing specifically on warm and cool colors by excluding the chips in rows ‘B’ and ‘I’ in Fig. 1A, which correspond roughly to light and dark. These results suggest that the warm–cool precision asymmetry found for actual languages under Gibson et al.’s priors may to some extent reflect perceptual structure.

Perceptual structure may also explain the pattern of color use or need that Gibson et al. reported and captured in their salience-weighted prior itself, according to which foreground objects (as opposed to their backgrounds) are more likely to be warm-colored rather than cool-colored. We found that their salience-weighted prior is correlated ($r = 0.49$, $p < 0.0001$) with the distance of each chip from central gray in CIELAB

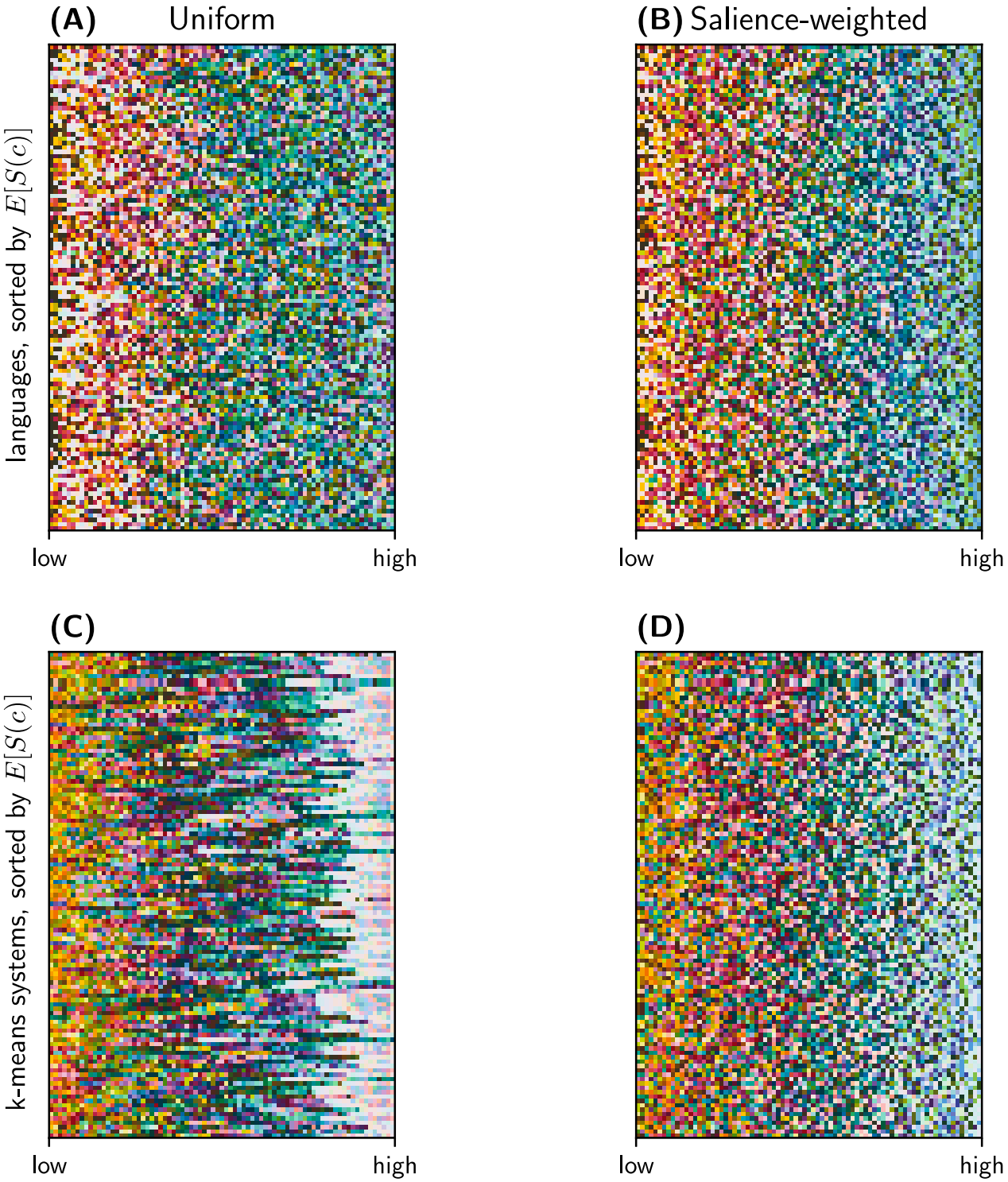

Fig. 3. (A and B) Replication of the results reported by Gibson et al. (2017) for the uniform prior and salience-weighted prior. Across languages, warm colors have lower expected surprisal than cool colors. (C and D) Analogous analyses in which each language's color naming system was replaced by a hypothetical color naming system obtained by k-means clustering of the color chips represented in CIELAB space. These perceptually derived hypothetical systems also exhibit a warm–cool surprisal asymmetry.

space,[1] suggesting that the salience-weighted prior reflects how "un-gray" and thus perceptually salient different colors are. It is possible that useful objects are often saliently (warmly) colored so as to attract human attention.

Taken together, these results suggest a possible perceptual common cause for both of the qualitative asymmetries in communicative precision and communicative need that Gibson et al. documented. However, these results still leave open the possibility that color naming across languages may be shaped by an element of need beyond what is predicted by perceptual structure. In the following sections we demonstrate an information-theoretic link between communicative need and precision, and use it to address this open question.

4. Information-theoretic link between need and precision

When viewing language in information-theoretic terms, one often considers a communication channel between a speaker and a listener (e.g., Baddeley & Attewell, 2009; Gibson et al., 2013; Plotkin & Nowak, 2000). However, this is not the only potentially relevant channel. From an information-theoretic perspective, any conditional distribution can be interpreted as a channel (Cover & Thomas, 2006), and in the present treatment, the lexicon is captured by the conditional distribution $p(w|c)$, which specifies the probability of using a color term w for a given color c. Therefore, the lexicon itself can be seen as a channel, and one may explore the capacity of that channel—that is, the maximal amount of information about color that can be conveyed by that lexicon.

Formally, the input to this channel is a color c, taken from a set $\mathcal{C}$ of colors, and the output is a word w, taken from a set $\mathcal{W}$ of possible words. Here we define $\mathcal{C}$ to be the 80 color chips shown in Fig. 1, and $\mathcal{W}$ to be an arbitrary set of K words, where K is determined by the number of color terms in the language. Shannon's channel coding theorem (Shannon, 1948) states that the maximal number of bits on average that can be transmitted per channel use is determined by the channel capacity, which is defined as the maximal mutual information between the input and output, namely by

$$\max_{p(c)} I(W;C), \tag{3}$$

where the maximization is over all possible choices of $p(c)$, and the mutual information is

$$I(W;C) = \sum_{c,w} p(c)p(w|c)\log\frac{p(c|w)}{p(c)}. \tag{4}$$

A distribution $p(c)$ over $\mathcal{C}$ that attains the channel capacity, that is, a maximizer of Eq. (3), is called a capacity-achieving prior (CAP). In our case, since $\mathcal{C}$ and $\mathcal{W}$ are finite sets, a capacity-achieving prior can be found via the Blahut–Arimoto algorithm (Arimoto, 1972; Blahut, 1972). This algorithm is based on the fact that by differentiating Eq. (4) with respect to $p(c)$ we get the following necessary and sufficient[2] condition for optimality:

$$p(c) \propto \exp(-S(c)). \tag{5}$$

We find it interesting that while Blahut and Arimoto derived the expression for $S(c)$ from the capacity achieving principle, the same expression has been used for different reasons by Gibson et al. and others (e.g., Piantadosi et al., 2011). Note that Eq. (5) defines a self-consistent condition for optimality, because $S(c)$ also depends on the prior. By taking the log on both sides of Eq. (5) we get that a prior is a CAP if and only if it satisfies

$$-\log p(c) = S(c) + \log Z, \tag{6}$$

where Z is the normalization factor of Eq. (5).

Thus, need and communicative precision are linked through the capacity achieving principle. Specifically, for a capacity-achieving prior, that is, a prior $p(c)$ that maximizes the information about color that is conveyed by a given lexicon, we should see a simple linear relationship, with slope 1, between $-\log p(c)$ and the expected surprisal (or communicative imprecision) $S(c)$. Notice that the link between $p(c)$ and $S(c)$ in Eq. (6) implies that, ideally, patterns in $p(c)$ would be mirrored in $S(c)$, and thus the link is consistent with Gibson et al.'s findings. However, this link makes a stronger claim in that it specifies more precisely what the relation between need and precision should be, and it does so on theoretically motivated grounds. In the next section we use this information-theoretic link to present new evidence that color naming across languages may indeed reflect universal patterns of communicative need, as well as perceptual structure.

5. Inferring need from naming data

The capacity achieving principle provides a basis for inferring a theoretically motivated need distribution from color naming data. Concretely, given a color naming system, this principle allows us to infer what the accompanying need distribution or prior should be in order to maximize the precision of the given lexicon.

We considered three different priors and assessed their effects in analyses of a single dataset, WCS+. We inferred a capacity-achieving prior from the WCS+ data itself (WCS-CAP, Fig. 4A): This is an idealized prior that is implicit in these actual color naming systems. We similarly inferred a capacity-achieving prior from the artificial naming data explored above that are derived from k-means clustering (KM-CAP, Fig. 4B): This is an idealized prior implicit in these artificial systems that are based on perceptual structure alone. In each case, following Zaslavsky et al. (2018), we evaluated the CAP $p_l(c)$ for each language l (real or artificial) with respect to its color naming distribution $p_l(w|c)$, and averaged together these language-specific priors in order to infer a universal need distribution.[3] That is, we defined

$$p(c) = \frac{1}{L}\sum_{l=1}^{L} p_l(c), \tag{7}$$

where $L = 111$ is the number of languages in the WCS+ dataset.

For comparison with these inferred priors, we also considered the salience-weighted prior of Gibson et al. (Fig. 4C), which is not inferred but is instead grounded directly in the frequency with which colors appear in foreground objects vs. backgrounds in natural images. For each of these three priors—WCS-CAP, KM-CAP, and salience-weighted—we entered it as $p(c)$ into Eq. (2), and then used Eq. (1) to obtain the expected surprisal $S(c)$ for each language in the WCS+ data given that prior. We then assessed each prior in two ways: first by asking whether we obtain the CAP-predicted linear relationship between $-\log p(c)$ and $S(c)$, and second by sorting chips by $S(c)$ and asking whether we observe the warm–cool surprisal asymmetry reported by Gibson et al. and also seen in our Fig. 3.

The results are shown in Fig. 5. Comparing first just the two inferred priors, WCS-CAP and KM-CAP, we see that the linear relation between $-\log p(c)$ and average $S(c)$ is dissociable from the warm–cool surprisal asymmetry: WCS-CAP shows a linear relation but not a clear warm–cool asymmetry, whereas KM-CAP shows a clear warm–cool

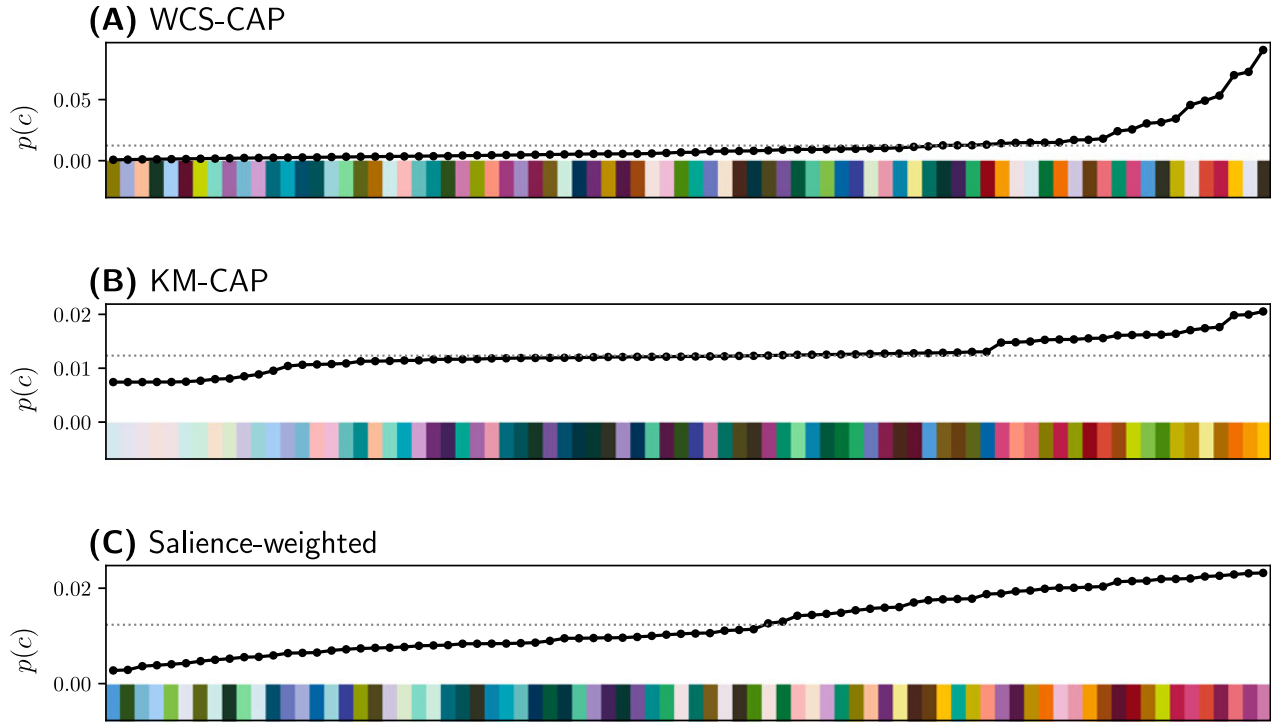

Fig. 4. Inferred (A: WCS-CAP, B: KM-CAP) and directly measured (C: salience-weighted) priors. Chips along the x-axis are rank ordered according to $p(c)$. Dashed line corresponds to a uniform prior. KM-CAP and salience-weighted exhibit a warm–cool asymmetry, whereas WCS-CAP exhibits a weaker tendency for warm colors and the two most needed colors according to this prior correspond to light and dark.

asymmetry but not a clear linear relation ($r = 0.32$, $p < 0.01$). The presence of a very clean linear relation for WCS-CAP reassures us that by averaging the language-specific CAPs, we inferred a universal need distribution largely consistent with Eq. (6).[4] It is perhaps more surprising that the warm–cool asymmetry vanishes under this well-motivated prior, given that it has persisted under others (recall Fig. 3). The absence of the warm–cool surprisal asymmetry under WCS-CAP demonstrates the sensitivity of this asymmetry to the assumed prior. At the same time, the lack of a clear linear relation between $-\log p(c)$ and average $S(c)$ under KM-CAP suggests that this prior is not well-suited for precise communication using the naturally occurring color naming systems of the WCS+ dataset. KM-CAP is ultimately derived from perceptual structure, whereas WCS-CAP is derived from the actual WCS+ languages, and both priors are derived using the same principle. Thus, the difference between them, seen in Figs. 4 and 5, can be attributed to features in the WCS+ data that are not simply a reflection of perceptual structure.

With this by way of stage-setting, consider now the results for the salience-weighted prior. It exhibits a warm–cool surprisal asymmetry on the WCS+ data (in fact, this panel simply replicates Fig. 3B) and also exhibits a roughly linear relation between $-\log p(c)$ and average $S(c)$, with slope close to 1 ($r = 0.83$, $p < 0.0001$). This linear relation is significant for two reasons. First, the fact that this relation is found for the salience-weighted prior but not for the perceptually based KM-CAP suggests that the salience-weighted prior (like WCS-CAP) exhibits signs of need beyond what is predicted by perceptual structure. Second, this roughly linear relation demonstrates an information-theoretic fit between cross-language color naming data and this prior, which was independently empirically obtained by Gibson et al.

6. Discussion

As stated in their title, Gibson et al. (2017) argued that "color naming across languages reflects color use." They presented this claim as an alternative to accounts of color naming based on perceptual salience. In support of this claim, they presented evidence of a warm–cool asymmetry in communicative need and a corresponding asymmetry in communicative precision in color naming across languages—suggesting that color naming systems may have adapted to a universal human tendency to communicate preferentially about warm colors. Here, we have cast this argument in a new light. We have shown that both qualitative asymmetries may be alternatively explained by a common cause: the structure of perceptual color space. Therefore, these two asymmetries are not an unambiguous sign that color naming reflects communicative need.

However, by invoking an information-theoretic principle that links need and precision, we have also presented a different form of evidence that color naming does in fact bear traces of universal patterns of communicative need beyond what perceptual factors would predict. Thus, we agree with Gibson et al. that communicative preferences appear to have left their imprint on color naming systems in the world's languages (see also Kemp & Regier, 2012 for a similar argument concerning kin terminologies). However, we differ with

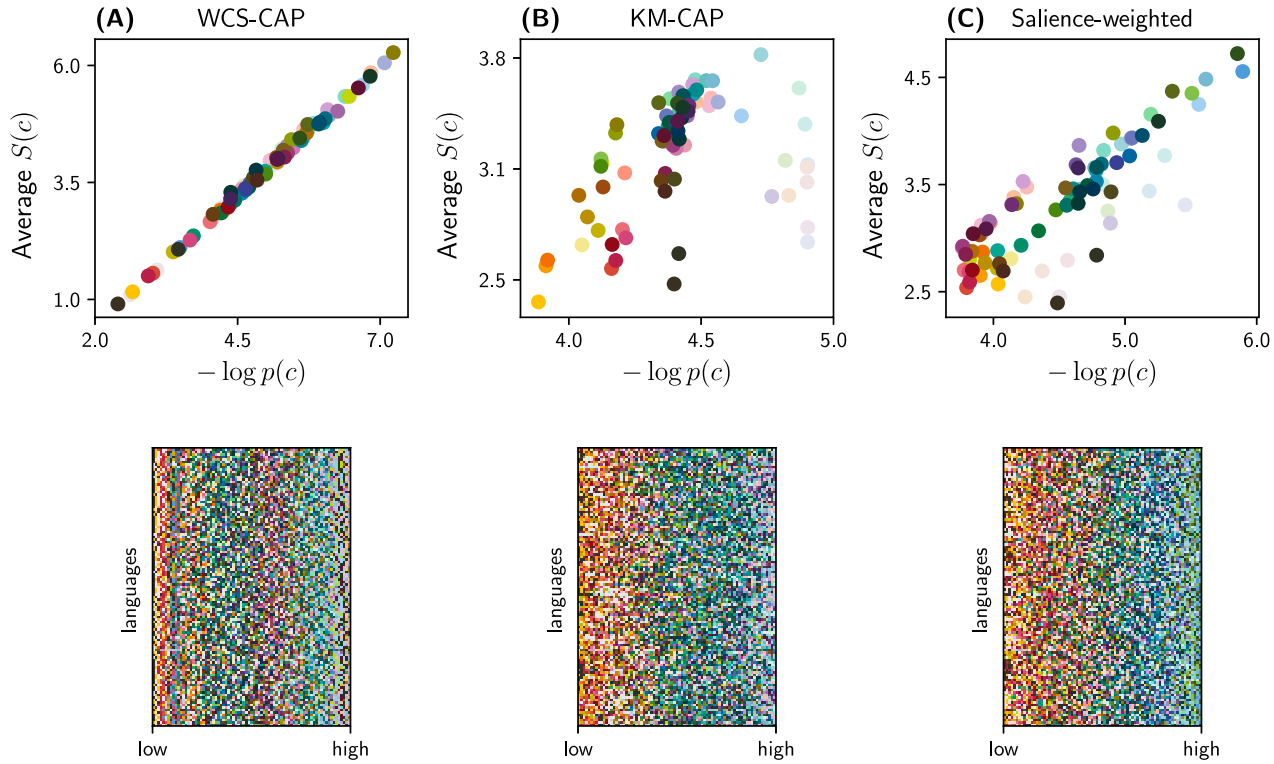

Fig. 5. Comparison between the priors of Fig. 4. Upper panels: Scatterplots of $-\log p(c)$ vs. average $S(c)$ across languages. Lower panels: Surprisal patterns for each prior, analogous to Fig. 3. See text for interpretation.

Gibson et al. in two respects: first, we reach this conclusion on different grounds, and second, we find that communicative need may operate in concert with, rather than as an alternative to, perceptual structure as a determinant of color naming.

More broadly, there is also another possible connection between perceptual structure and need. Although we have treated these two as independent factors, it may be the case that the structure of perceptual color space is itself adapted to the statistics of natural scenes (Shepard, 1994) and in that sense is influenced by need. Even in this case, however, the picture is not entirely straightforward. There is an important distinction in principle, and thus at least possibly in practice, between the frequency with which particular colors appear in the world and the frequency with which they must be communicated. It seems likely that our perceptual systems may have adapted to the former, and our languages to the latter.

Acknowledgments

We thank Bevil Conway and Ted Gibson for kindly sharing their salience-weighted prior with us, Delwin Lindsey and Angela Brown for kindly sharing their English color naming data with us, and Joshua Abbott for helpful discussions. This study was supported by the Gatsby Charitable Foundation (N.Z. and N.T.) and DTRA award

HDTRA11710042 (T.R.). Part of this work was done while N.Z. and N.T. were visiting the Simons Institute for the Theory of Computing at UC Berkeley.

Notes

1. We took central gray to be located at the midpoint between the CIELAB coordinates for the two achromatic chips that are most intermediate between black and white in the WCS palette, namely E0 and F0 (not shown in Fig. 1).
2. This follows from the concavity of $I(W;C)$ in $p(c)$. For more detail see Theorem 2.7.4 and section 10.8 in (Cover & Thomas, 2006).
3. We leave for later investigation the interesting question of language-specific need influences.
4. By substituting WCS-CAP into equation (6) we introduced a nonlinearity because the language-specific CAPs are averaged inside the log. In principle, this could have violated equation (6).

References

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20.

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*, *20*(9), 1100–1107.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Blahut, R. E. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473.

Brainard, D. H. (2003). Color appearance and color difference specification. In S. K. Shevell (Ed.), *The science of color, 2nd ed.* (pp. 191–216). Amsterdam: Elsevier.

Cook, R. S., Kay, P., & Regier, T. (2005). The World Color Survey database: History and use. In I. H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 223–242). Amsterdam: Elsevier.

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, E., Piantadosi, S.T., Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088.

Jameson, K., & D'Andrade, R. G. (1997). It's not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). Cambridge, UK: Cambridge University Press.

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610–646.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Komarova, N. L., & Jameson, K. A. (2013). A quantitative theory of human color choices. *PLOS One*, *8*(2), e55986.

Lindsey, D. T., & Brown, A. M. (2006). Universality of color names. *Proceedings of the National Academy of Sciences*, *103*(44), 16608–16613.

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of American English. *Journal of Vision*, *14*(2), 17.

Lindsey, D. T., Brown, A. M., Brainard, D. H., & Apicella, C. L. (2015). Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, *25*(18), 2441–2446.

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

Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, *205*(1), 147–159.

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

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28.

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

Chapter 6

Communicative Need in Color Naming

Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier (2019). Communicative need in color naming. *Cognitive Neuropsychology*. DOI: 10.1080/02643294.2019.1604502.

COGNITIVE NEUROPSYCHOLOGY
https://doi.org/10.1080/02643294.2019.1604502

Routledge
Taylor & Francis Group

Check for updates

Communicative need in colour naming

Noga Zaslavsky[a,b], Charles Kemp[c], Naftali Tishby[a,d] and Terry Regier[b,e]

[a]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel; [b]Department of Linguistics, University of California, Berkeley, CA, USA; [c]School of Psychological Sciences, The University of Melbourne, Parkville, Australia; [d]Benin School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel; [e]Cognitive Science Program, University of California, Berkeley, CA, USA

ABSTRACT

Colour naming across languages has traditionally been held to reflect the structure of colour perception. At the same time, it has often, and increasingly, been suggested that colour naming may be shaped by patterns of communicative need. However, much remains unknown about the factors involved in communicative need, how need interacts with perception, and how this interaction may shape colour naming. Here, we engage these open questions by building on general information-theoretic principles. We present a systematic evaluation of several factors that may reflect need, and that have been proposed in the literature: capacity constraints, linguistic usage, and the visual environment. Our analysis suggests that communicative need in colour naming is reflected more directly by capacity constraints and linguistic usage than it is by the statistics of the visual environment.

ARTICLE HISTORY
Received 2 October 2018
Revised 18 February 2019
Accepted 1 April 2019

KEYWORDS
Information theory; colour naming; categorization; semantic typology

1. Introduction

Colour naming varies widely across languages. At the same time, this variation is constrained, and certain universal tendencies of colour naming recur across unrelated languages (e.g., Berlin & Kay, 1969; Lindsey & Brown, 2006). Figure 1 shows the colour naming systems of four languages, illustrating this variation. As can be seen, both the number of terms and their extension vary across languages, but it is also the case that some cross-language commonalities can be found, such as the existence of terms roughly corresponding to English "red" and "yellow".

Why do the colour naming systems of the world's languages vary as they do? Why do we see these systems and not other logically possible ones? Broadly speaking, three classes of explanation have been proposed, emphasizing colour perception, communicative need, or both, as illustrated in Figure 2. Traditionally, cross-language variation has been explained largely in terms of perception (e.g., Kay & McDaniel, 1978; see Figure 2(a)). On this view, universal tendencies in colour naming are relatively direct reflections of universals in colour perception. Early work in this tradition did note in addition the apparent influence of cultural forces such as level of technological development, including dye technology, in determining the complexity of the colour lexicon, but these ideas were not pursued in depth and were instead presented as "plausible speculation" (Berlin & Kay, 1969, pp. 16–17). The influence of communicative forces was later explored via multi-agent simulations (e.g., Dowman, 2007; Loreto, Mukherjee, & Tria, 2012; Steels & Belpaeme, 2005), and a recent elaboration of these ideas has suggested concrete roles for both perception and communicative need: specifically it has been proposed that colour naming reflects perceptual structure as partitioned for communicative purposes (e.g., Jameson & D'Andrade, 1997; Komarova, Jameson, & Narens, 2007; Regier, Kay, & Khetarpal, 2007; see Figure 2(b)). In particular it has been proposed that colour naming across languages may be shaped by the need for *efficient communication*: the need to communicate about colour precisely, but at minimal cognitive cost (e.g., Lindsey, Brown, Brainard, & Apicella, 2015; Regier, Kemp, & Kay, 2015). Recently Gibson et al. (2017) pursued this progression of thinking to its logical extreme, proposing that communicative needs do not merely modulate an effect of perceptual structure—but rather that communicative needs themselves govern the character of colour categories (see Figure 2(c)), a hypothesis they situated as an "alternative" (p. 10785) to accounts based on

CONTACT Noga Zaslavsky noga.zaslavsky@mail.huji.ac.il
This article has been republished with minor changes. These changes do not impact the academic content of the article.
© 2019 Informa UK Limited, trading as Taylor & Francis Group

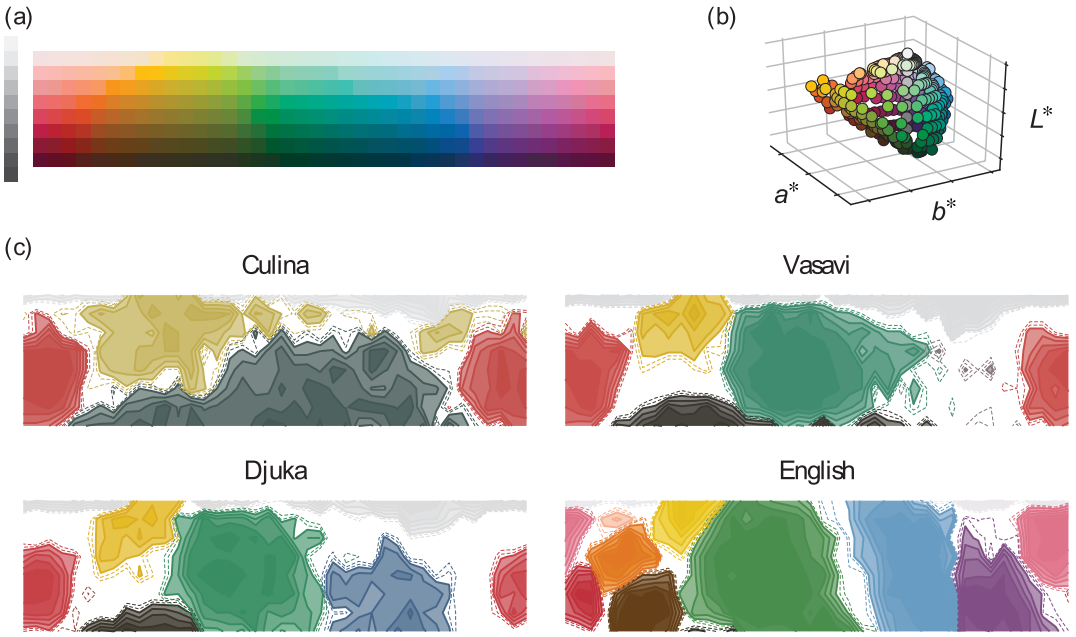

Figure 1. (Colour image online) (a) A standard colour naming stimulus grid, containing 320 colour chips and 10 achromatic colour chips (leftmost column). Columns correspond to equally spaced Munsell hues, rows correspond to equally spaced Munsell values (levels of lightness), and each chip is at the maximum available saturation (colourfulness) for that hue/lightness combination. (b) The 330 colour chips re-plotted in the CIELAB perceptual colour space, in which Euclidean distance between nearby colours is roughly correlated with perceptual dissimilarity. L^* corresponds to lightness, and hue and saturation are encoded in polar coordinates in the (a^*, b^*) plane. The chips are not evenly distributed in this space. For example, chips in the yellow region are exceptionally highly saturated (colourful) and therefore protrude farther outward away from other colours. This uneven distribution highlights presumably universal perceptual structure that may shape colour naming across languages. (c) Examples of colour naming systems from four languages in the colour naming dataset we consider here, plotted against the stimulus grid. Each plot shows the contours of the naming probabilities for each term in the language. The naming probabilities for each colour term are depicted in the colour corresponding to the centroid for that term. Solid lines correspond to level sets of 50% and above, and dashed lines correspond to level sets of 40% and 45%.

perception. Here, we argue that need and perception should both be taken into account, in line with earlier efficiency analyses (back to Figure 2(b)). We review a principled theoretical framework for achieving this integration of perceptual structure and communicative need (Zaslavsky, Kemp, Regier, & Tishby, 2018), and we use that framework to evaluate the character and role of communicative need in colour naming.

The empirical basis for our evaluation is a set of colour naming systems from 111 languages. These systems were drawn mainly from the World Color Survey (WCS: Kay, Berlin, Maffi, Merrifield, & Cook, 2009), which contains colour naming data from 110 languages of non-industrialized societies, with respect to the stimulus grid shown in Figure 1(a). In addition, we consider colour naming data from American English (Lindsey & Brown, 2014) which were collected with respect to the same stimulus grid. We refer to this joint dataset as the WCS+ dataset.

The remainder of this paper proceeds as follows. In section 2, we review recent evidence suggesting that communicative need—together with perceptual structure—plays an important role in shaping colour naming across languages. In section 3, we review a

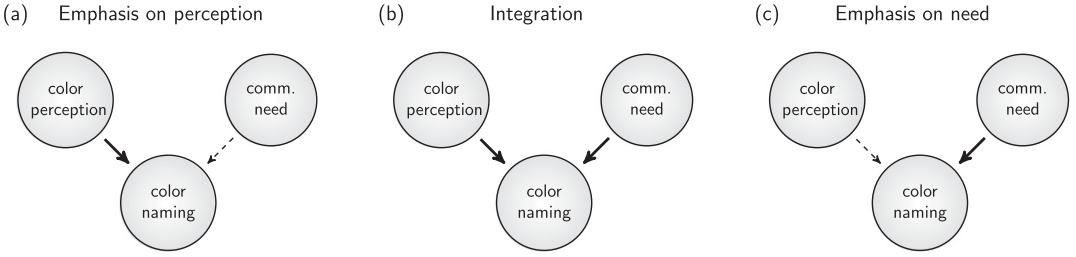

Figure 2. Colour naming may be shaped by colour perception, communicative need, or both.

recent computational model that integrates communicative need and perceptual structure, and that accounts for colour naming across languages in terms of an independent principle of efficiency. In this model, communicative need is formalized as a prior distribution over colours; however it is not yet clear how best to characterize this distribution. In section 4 we address this problem by presenting several estimation methods based on different factors that may reflect communicative need, specifically capacity constraints, linguistic usage, and the statistics of colours in the environment. Finally, in section 5, we evaluate these different factors by assessing how well the corresponding priors account for colour naming data across languages, in the context of the model mentioned above.

2. The importance of communicative need

When considering the possible role of communicative need in shaping colour naming, it is useful to distinguish two different kinds of need: *domain-level* need and *object-level* need (Kemp, Xu, & Regier, 2018). Domain-level need is the communicative importance of a given domain, such as colour, relative to other domains of human experience about which one may wish to communicate. For example, the observation that the introduction of dye technology may push a society or culture to develop a more fine-grained colour lexicon, mentioned above, is an observation about domain-level need: with the advent of dyes, colour as a domain presumably assumes greater cultural importance than it had previously, justifying greater complexity in this part of the lexicon. In section 3 we briefly discuss how domain-level need may be formalized in terms of the tradeoff between accuracy and complexity of the lexicon. Object-level need, in contrast, concerns how often one may need to communicate about particular objects within a domain—for example, within the domain of colour, one may need to talk about certain colours more than others. It is this sort of object-level need that is naturally captured as a prior distribution over colours, and that is the primary focus of this paper.

As noted above, early accounts of colour naming emphasized perception over (object-level) communicative need. Some justification for this stance is suggested by the fact that qualitative patterns of colour naming across languages can be accounted for fairly well based only on perceptual structure, assuming a uniform prior over colours (e.g., Regier et al., 2015). However this leaves open the possibility that a better account of the data might be obtained with a non-uniform need distribution.

In line with this possibility, Gibson et al. (2017) argued that some colours are more useful than others for human purposes, and that the usefulness of particular colours is a major determinant of colour naming across languages. Specifically, they argued for the greater usefulness of warm colours, relative to cool colours, and argued that this asymmetry in usefulness is reflected in patterns of colour naming. They showed that across languages, colour categories tend to support more precise communication for warm than for cool colours. They also examined colour statistics in a large dataset of natural images and found that objects (as opposed to their visual backgrounds) tend to be warm-coloured rather than cool-coloured, in a parallel to the warm-cool asymmetry in language. They suggested on this basis that colour naming across languages "reflects colors of universal usefulness" (p. 10785).

Zaslavsky, Kemp, Tishby, and Regier (2019) engaged this proposal, and argued for a somewhat different picture. They noted that the finding of a warm-cool asymmetry in language assumes a prior, and that the asymmetry vanishes under some well-motivated priors. They also found that the warm-cool naming asymmetry, when assessed using the same priors as Gibson et al. (2017), is present not only in natural colour naming systems, but also in a set of artificial colour naming systems that are based solely on perceptual structure, with no element of communicative need. These findings suggest that the warm-cool naming asymmetry, when it is found, cannot be taken as an unambiguous signature of communicative need. However this leaves open the possibility that there may be a different signature of need in the colour naming data of the world's languages. Zaslavsky et al. (2019) proposed such a signature, based on the notion of a capacity-achieving prior (treated below in section 4.1), and found that natural colour naming systems do indeed bear signs of communicative need beyond what would be predicted from perceptual structure alone. Thus, communicative need does appear to shape colour naming in the world's languages.

A natural conclusion from the work just reviewed is that patterns of colour naming may result from an integration of perceptual structure and communicative need. That conclusion leads to an important open question: what are the factors involved in communicative need, and how does need interact with perception in shaping colour naming? The remainder of this paper addresses that question.

3. Integration of communicative need and colour perception

The notion of efficient communication in colour naming was recently formalized by Zaslavsky et al. (2018), building on earlier work by Regier et al. (2015), in a way that integrates perceptual structure and communicative need. Zaslavsky et al.'s proposal grounded the notion of efficient communication in an independently motivated information-theoretic principle, the Information Bottleneck (IB) principle (Tishby, Pereira, & Bialek, 1999). On that basis, their proposal accounted to a large extent for the wide variation observed in colour naming across languages, provided a theoretical explanation for the existence of soft colour categories with graded membership, and synthesized previous accounts of colour category evolution. For these reasons, we adopt the IB colour naming model here as a framework within which different proposed sources of communicative need may be assessed.

The IB colour naming model is based on a simple communication scenario between a speaker and listener, illustrated in Figure 3(a). The speaker observes a colour c drawn from a prior distribution $p(c)$ over colours in the environment $\mathcal{U}$, and wishes to communicate this colour to the listener. The prior $p(c)$ reflects the communicative needs of the speaker, favoring certain colours over others (Kemp & Regier, 2012; Kemp et al., 2018). To account for perceptual uncertainty, it is assumed that the speaker does not have access to the exact colour but rather to a noisy mental representation of it, m_c,[1] formulated as a Gaussian distribution centred at c over colours in the CIELAB perceptual space (Figure 1 (b)). The speaker communicates this mental representation to the listener by producing a word w drawn from a shared lexicon $\mathcal{W}$, according to a naming distribution $q(w|c)$. The listener receives w and interprets this word by constructing a mental representation $\hat{m}_w$ that approximates the speaker's representation m_c.

According to the IB principle, the ideal speaker and listener are adapted to each other by jointly optimizing an information-theoretic tradeoff between the *complexity* of the lexicon and the *accuracy* of communication. This tradeoff is also illustrated in Figure 3(a). Below, we lay out the IB formulations of complexity, accuracy, and their tradeoff.

In IB terms, a colour naming distribution $q(w|c)$ is an encoder that compresses colours into words. As in rate-distortion theory (Shannon, 1959), the complexity of this encoder is measured by the information that the lexicon maintains about the speaker's representation, namely:

$$I(C;W) = \sum_{c\in\mathcal{U},w\in\mathcal{W}} p(c)q(w|c)\log\frac{q(w|c)}{q(w)}, \qquad (1)$$

where $q(w) = \sum_c p(c)q(w|c)$. This informational complexity roughly corresponds to the number of bits that are required to represent the lexicon on average. Similar informational costs have also been proposed as measures for cognitive effort in other contexts (e.g., Ferrer i Cancho & Solé, 2003; Genewein, Leibfried, Grau-Moya, & Braun, 2015; Marzen & DeDeo, 2017; Sims, 2016; Tkačik & Bialek, 2016).

The accuracy of communication is the extent to which the listener's interpreted representation is similar to the speaker's representation, or in other words the extent to which the distortion or discrepancy between these two representations is small. Since m_c and $\hat{m}_w$ are both distributions over colour space, a natural distortion measure (Harremoës & Tishby, 2007) is the expected Kullback-Leibler (KL) divergence between them:

$$\mathrm{E}[D[m_c \parallel \hat{m}_w]] = \sum_{c\in\mathcal{U},w\in\mathcal{W}} p(c)q(w|c)\sum_{u\in\mathcal{U}} m_c(u)\log\frac{m_c(u)}{\hat{m}_w(u)}. \qquad (2)$$

There is necessarily a tradeoff between accuracy and complexity. Maximizing accuracy amounts to minimizing the distortion given by Equation 2, which will be achieved when $D[m_c \parallel \hat{m}_w] = 0$, i.e., when $m_c \equiv \hat{m}_w$. This in turn will require a very complex lexicon, with a separate word for each colour, so that each colour can be communicated with perfect accuracy. On the other hand, minimizing complexity can be achieved by using a single word to describe all colours, but in this case accuracy will necessarily be low, i.e., communication will not be informative. The

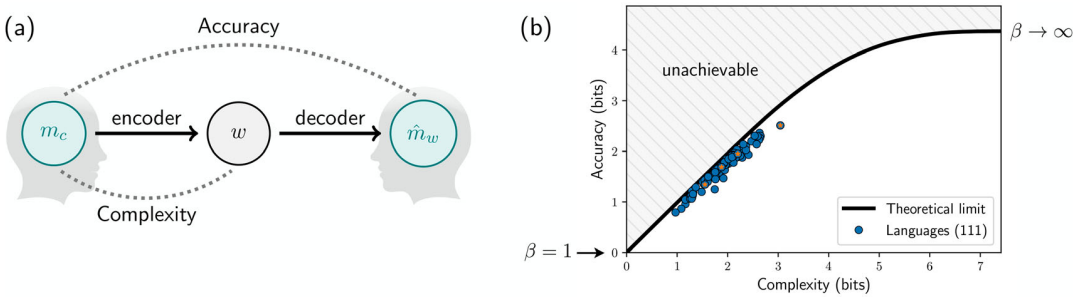

Figure 3. (Colour image online) (a) The basic communication model. A colour c is drawn from a prior distribution $p(c)$ that represents communicative need. The speaker observes c, mentally represents it by a distribution m_c, and communicates this representation to the listener by encoding it in a word w which is distributed according to an encoding naming distribution $q(w|c)$. The listener receives w and interprets (or decodes) it by constructing a mental representation $\hat{m}_w$. The complexity of the lexicon is determined by the encoder. The accuracy of the lexicon is determined by the similarity between the listener's and speaker's mental representations. (b) The theoretical limit of achievable complexity-accuracy tradeoffs, defined by the set of optimal IB systems, and the tradeoffs achieved by the colour naming systems of the WCS+ languages. Accuracy is inversely related to the expected distortion (Equation 2), such that maximal accuracy corresponds to zero distortion. All WCS+ languages achieve near-optimal tradeoffs. Orange stars correspond to the four languages shown in Figure 7, where they are ordered by complexity. Both figures are adapted from Zaslavsky et al. (2018).

tradeoff between these two competing forces is given by the following equation, which is equivalent to the standard IB objective function:

$$\min_{q(w|c),\ \hat{m}_w} I(C;W) + \beta \mathrm{E}[D[m_c \parallel \hat{m}_w]], \tag{3}$$

where the tradeoff parameter $\beta \geq 0$ controls how complexity and accuracy are balanced.

The optimal IB colour naming systems, i.e., the systems that optimize Equation 3 for different values of β, define the theoretical limit of achievable tradeoffs. Zaslavsky et al. (2018) evaluated this theoretical limit and found that the colour naming systems in the WCS+ data are near-optimal in that they lie near this theoretical limit (Figure 3(b)). This suggests that languages may have evolved under pressure for information-theoretic efficiency. It can be seen that variation in the tradeoff parameter β accounts for much of the cross-language variation in the WCS+ data—meaning that different languages navigate the tradeoff between accuracy and complexity in different ways, while remaining near the theoretical limit of efficiency. It is natural to interpret β as capturing domain-level need, or the cultural importance of colour as a domain in a given society (recall section 2): the more important it is to communicate accurately about colour, the more it is justified to allow greater complexity to achieve that accuracy—and this tradeoff is exactly what β controls. This notion is captured memorably in the title of a paper that described colour naming in a society for which colour is relatively unimportant: "We don't talk much about colour here" (Kuschel & Monberg, 1974); as would be expected, the colour system of this language was found to be very simple, having only three basic colour terms.

Importantly, the findings just reported were obtained with a specific non-uniform prior which is based on the notion of capacity-achieving priors (WCS-CAP, see section 4.1 for detail). It is not yet clear whether other well-motivated priors could provide a better account of the data.

In what follows, we systematically investigate the effect of different priors while keeping all the other components of the IB colour naming model fixed. Preparatory to doing so, it may be useful to note how both perceptual structure and the prior influence the IB objective function. The irregular distribution of colours in perceptual space (Figure 1(b)) influences the accuracy term (Equation 2), through m_c and $\hat{m}_w$. The prior $p(c)$ influences both terms of the IB objective function: complexity (Equation 1) and accuracy (Equation 2). Colours with higher communicative need, i.e., higher $p(c)$, will therefore be more dominant in the IB objective function (Equation 3), and thus there will be greater pressure to communicate those colours efficiently. Figure 4 illustrates this concretely, by showing how different priors we explore in the next sections emphasize different parts of perceptual colour space.

4. Characterizing communicative need

We explore three general classes of prior distribution, each derived from a different principle for inferring communicative need. First is the class of least

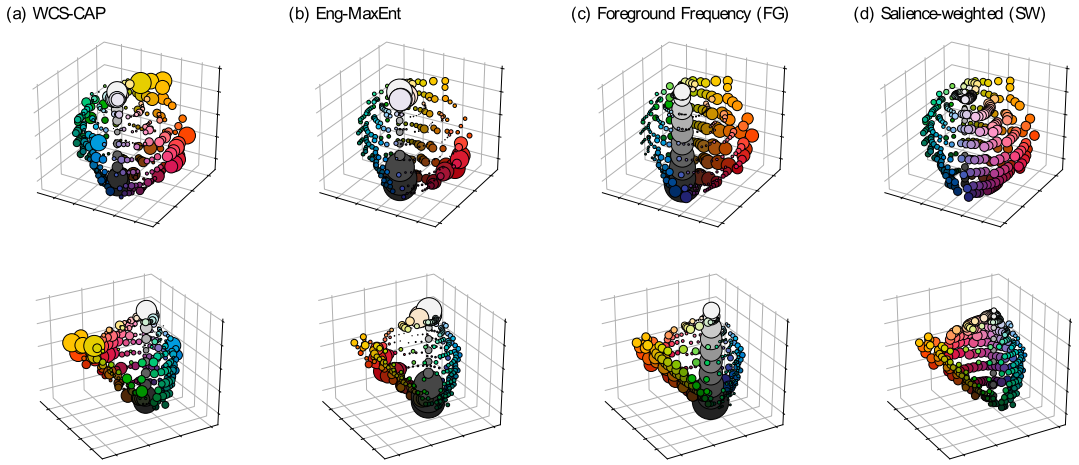

Figure 4. (Colour image online) Illustration of how communicative need may interact with perceptual structure in shaping colour naming. Each plot shows the 330 colour chips from Figure 1(b) as circles in CIELAB space, where the size of each circle is proportional to the colour's probability mass under four different priors defined in section 4. Each prior is a different distribution over perceptual space, which may give rise to different colour naming systems.

informative priors. This class aims to infer a prior without making any assumptions about external forces that may shape communicative need. The second class is based on the idea that communicative need is reflected in linguistic usage. The third class is based on the assumption that communicative need is reflected in colour statistics as encountered in the visual world, estimated from natural images.

4.1. Least informative priors

A natural approach to obtaining a prior distribution without any assumptions is by invoking the maximum entropy (MaxEnt) principle (Jaynes, 1982). The MaxEnt principle states that the most justified distribution is the one that maximizes uncertainty, measured in terms of entropy. In our setting, in its simplest form, this principle yields a uniform distribution over colour chips. A uniform prior has been used before to account for colour naming (e.g., Gibson et al., 2017; Regier et al., 2015), and thus we consider it here as a baseline. However, it is not clear whether in fact all colours are equally needed for communication in natural settings.

An alternative approach (Zaslavsky et al., 2018) aims to infer the prior directly from naming data, without making specific assumptions about the forces that may shape communicative need. This approach is based on the capacity-achieving principle (Shannon, 1948). In information theory, a channel is defined by a conditional distribution (Cover & Thomas, 2006). Thus, any colour naming distribution, $p(w|c)$, can be interpreted as a channel[2] that takes a colour c as its input and outputs a word w. The maximal amount of information that can be transmitted over a channel is the channel's capacity, and the ideal prior for that channel is called a capacity-achieving prior (CAP). In our setting, the CAP for a given naming distribution maximizes the amount of information the lexicon conveys about the observed colour. Formally, it is defined by

$$p_{\text{CAP}}(c) = \underset{p(c)}{\text{argmax}} \; I(C; W), \tag{4}$$

and can be obtained using the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972). Note that this CAP identifies a prior that maximizes complexity (Equation 1) for a given naming system, in contrast to the IB principle in which complexity is minimized over all possible naming systems for a given prior. Although these principles are related, they are also importantly different: the capacity-achieving principle is an optimality criterion for the prior whereas the IB principle is an optimality criterion for the naming system.

Given a colour naming distribution $p_l(w|c)$ for a specific language l, we can now obtain a CAP for that language, $p^{(l)}_{\text{CAP}}(c)$, which captures the pattern of communicative need for that language, inferred on the basis of the capacity-achieving principle. We then follow Zaslavsky et al. (2018) and average together the CAPs across languages l in the WCS+ dataset, to obtain a single universal prior.[3] The resulting prior, which we refer to as WCS-CAP, is shown in Figures 4(a) and 5(a). Because this prior is estimated from the WCS+ data, Zaslavsky et al. (2018) performed

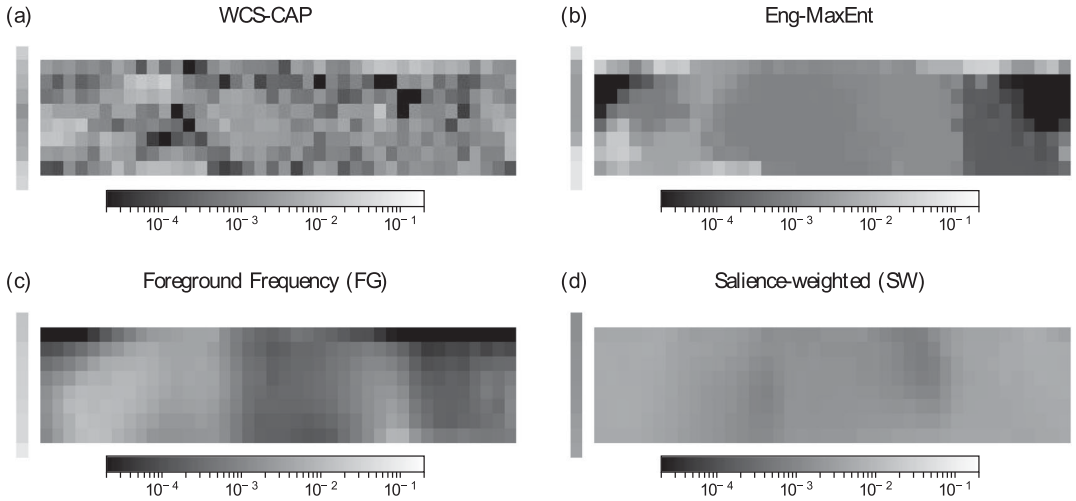

Figure 5. Prior distributions over the WCS grid. Each chip is coloured according to its probability mass (log-scale).

5-fold cross-validation and showed that WCS-CAP does not overfit the data (see Table 1).

4.2. Linguistic usage

It seems likely that the frequency of use of particular words in natural communication may reflect important aspects of communicative need, and priors estimated from corpus frequencies have been used to account for cross-linguistic variation in semantic domains other than colour (Kemp & Regier, 2012; Xu & Regier, 2014). However, a challenge for this approach is that it is not always clear how to infer a distribution over objects in the domain—colours, in our case—from corpus statistics, because corpus statistics provide frequencies only for words, and there are generally more objects in the domain (here, colour chips) than there are words (colour terms). Here we propose a general solution for this problem by applying the maximum entropy (MaxEnt) principle under constraints derived from corpus data.

Suppose we are given the naming distribution $p_l(w|c)$ for some language l, and we are also given word frequencies, $p_l(w)$, from a corpus for that language. For simplicity, assume that these word frequencies correspond only to cases in which these words are used for describing objects in the domain universe $\mathcal{U}$. Under this simplifying assumption, for $p_l(w|c)$ and $p_l(w)$ to be consistent with each other, it must hold that $\sum_c p_l(w|c)p(c) = p_l(w)$. This consistency requirement imposes a set of linear constraints on the prior, and of all the prior distributions that satisfy these constraints, we wish to select the one with maximal entropy, where entropy is defined by $H(C) = -\sum_c p(c)\log p(c)$. Formally, this gives the following optimization problem:

$$\begin{aligned} &\max_{p(c)} && H(C) \\ &\text{subject to} && \sum_{c\in\mathcal{U}} p_l(w|c)p(c) = p_l(w), \quad \forall w \in \mathcal{W}. \end{aligned} \tag{5}$$

This is a concave optimization problem, and can be solved using standard tools.[4]

In principle, this corpus-based MaxEnt approach can be applied on a language-specific basis, for every language for which $p_l(w)$ can be obtained. However, it is difficult to obtain such word frequencies for the WCS languages, because large representative corpora for these languages of non-industrialized societies do not exist. For English, in contrast, this approach is tractable because both naming data and corpus data exist. The English colour naming data collected by Lindsey and Brown (2014) contain over 100 words used across participants in their free-naming experiment. However, most of these words were used by only a few participants (see Lindsey & Brown, 2014). These words tend to be either rare, in which case their corpus frequencies may not be reliable, or words that are used metaphorically, in which case their frequencies are more likely to reflect usages other than describing colours. To mitigate this problem, we based this prior on only the 11 basic colour terms in English, which were used by all participants. We also obtained corpus frequencies for these 11 terms, as shown in Figure 6. The naming data and corpus frequency of each basic colour term define the constraints in Equation 5.

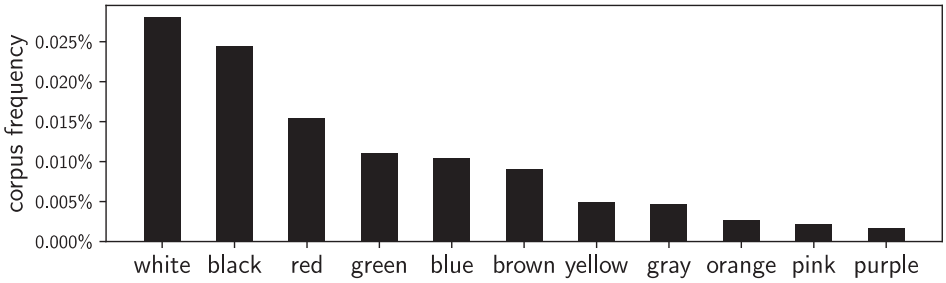

Figure 6. Frequencies of the 11 basic colour terms in English (case insensitive) from the Google n-gram (Michel et al., 2011) American English dataset for the year 2008 with a smoothing factor of 3 (average across the three preceding years). Since the English naming data from Lindsey and Brown (2014) were collected in the USA, this is a reasonably compatible corpus.

The resulting prior, Eng-MaxEnt, is shown in Figures 4(b) and 5(b). In contrast to WCS-CAP, this prior is estimated only from the English colour naming data and English corpus statistics, and thus it is independent of the WCS languages. We explore Eng-MaxEnt as a proposed approximation to a universal prior, on the assumption that corpus statistics in English may be shaped in part by universal communicative forces. We leave the interesting question of language-specific differences in usage and communicative need for future research (but see Regier, Carstensen, & Kemp, 2016, for treatment of this idea in another domain).

4.3. Visual environment

A natural possibility is that communicative need may be shaped largely by the statistics of colours in the world (e.g., Gibson et al., 2017; Yendrikhovskij, 2001). If this is the case, then a prior derived from the distribution of colours in the environment should provide a good account of colour naming. One way to approximate this distribution is from the statistics of colours in a large dataset of natural images. For example, Yendrikhovskij (2001) considered the total frequency of colours in a set of natural images. Gibson et al. (2017) also examined colour frequencies in natural images, but they noted that not all occurrences may be equally relevant for estimating need. Instead, they took as their measure of communicative need what they called the "salience" of particular colours: specifically, the frequency of a colour's appearance in objects that people tend to talk about, divided by the overall total frequency of that colour. Here we consider these two approaches, and another that is based only on a colour's frequency of appearance in foreground objects. This latter approach is based on the observation that if colours that appear in useful objects have greater communicative need, then this may hold regardless of the visual background of these objects.

To evaluate these different image-based approaches, we estimated (i) a prior based on the total frequency (TF) of colours; (ii) a prior based on the frequency of colours in foreground objects (FG); and (iii) a salience-weighted (SW) prior, similar to Gibson et al.'s approach but here based on the colours corresponding to all WCS chips whereas their analysis was based on a subset of these chips. Colour frequencies were estimated from Microsoft's COCO dataset (Lin et al., 2014), which contains over 80,000 annotated images.[5] The images were processed as follows. First, to filter out black-and-white images, only images with colourfulness index (Yendrikhovskij, Blommaert, & de Ridder, 1998) above 0.2 were considered. Approximately 3% of the images were excluded on this basis. Next, to avoid a bias toward large images, 50,000 pixels were randomly sampled from each image. These pixels were then converted to CIELAB coordinates (Figure 1(b)) and were classified as one of the WCS chips, or excluded if they were not close to any of the WCS chips.[6] Pixels with chroma less than the average chroma of pixels in the image were compared to the achromatic chips. Pixels with chroma above average were compared to the chromatic chips with closest lightness and hue values.

The resulting SW and FG priors are shown in Figures 4 and 5. The TF prior is not shown because it is fairly similar to the FG prior.[7]

5. Results

We assess the three classes of priors discussed above by entering each prior into the IB colour naming model, and evaluating how well the model with this

prior accounts for the WCS+ data. We follow the same quantitative evaluation method used by Zaslavsky et al. (2018), which is based on two goodness-of-fit scores:[8] (i) an inefficiency score, which measures the deviation from optimality of a given language's colour naming system; and (ii) a dissimilarity score, which measures the dissimilarity in extension between a given language's colour naming system and the corresponding optimal naming system predicted by the model. Lower values of these scores indicate a better fit to the data.

Table 1 shows the quantitative results based on these scores. WCS-CAP and Eng-MaxEnt achieve comparable scores, and outperform the other priors. A qualitative inspection of the results (Figure 7) shows that these priors predict slightly different solutions, but also agree to a large extent on the structure of the categories and resemble the actual systems. It is striking that Eng-MaxEnt—a prior that is derived only from English—is able to account so well for the WCS languages, which are from non-industrialized societies and the majority of which have fewer colour categories than English. This result suggests that there are general patterns of communicative need that are shared across cultures, and that these patterns can be inferred directly from linguistic data. While it is possible that Eng-MaxEnt and WCS-CAP also reflect perceptual structure, the influence of perception on these priors would be indirect, mediated via language use (Winter, Perlman, & Majid, 2018). For completeness, we compared these results with those obtained by using a capacity-achieving prior

Table 1. Evaluation of possible communicative need distributions.

Motivation	Data type	Prior	Inefficiency	Dissimilarity
Baseline	None	Uniform	0.24 (±0.09)	0.39 (±0.12)
Least informative	WCS+	WCS-CAP	0.18 (±0.07)	0.18 (±0.10)
Linguistic usage	English naming & corpus data	Eng-MaxEnt	0.19 (±0.09)	0.17 (±0.08)
Visual environment	Foreground freq.	FG	0.21 (±0.08)	0.31 (±0.12)
	Total freq.	TF	0.21 (±0.08)	0.34 (±0.14)
	Colour salience	SW	0.25 (±0.09)	0.40 (±0.12)

Notes: Inefficiency and dissimilarity scores are as defined by Zaslavsky et al. (2018). Reported scores correspond to averages across languages ± 1 SD. Lower values are better, and the best scores are in boldface. Results for the two uninformative priors are from Zaslavsky et al. (2018), where the scores for WCS-CAP are averages over left-out languages in 5-fold cross-validation.

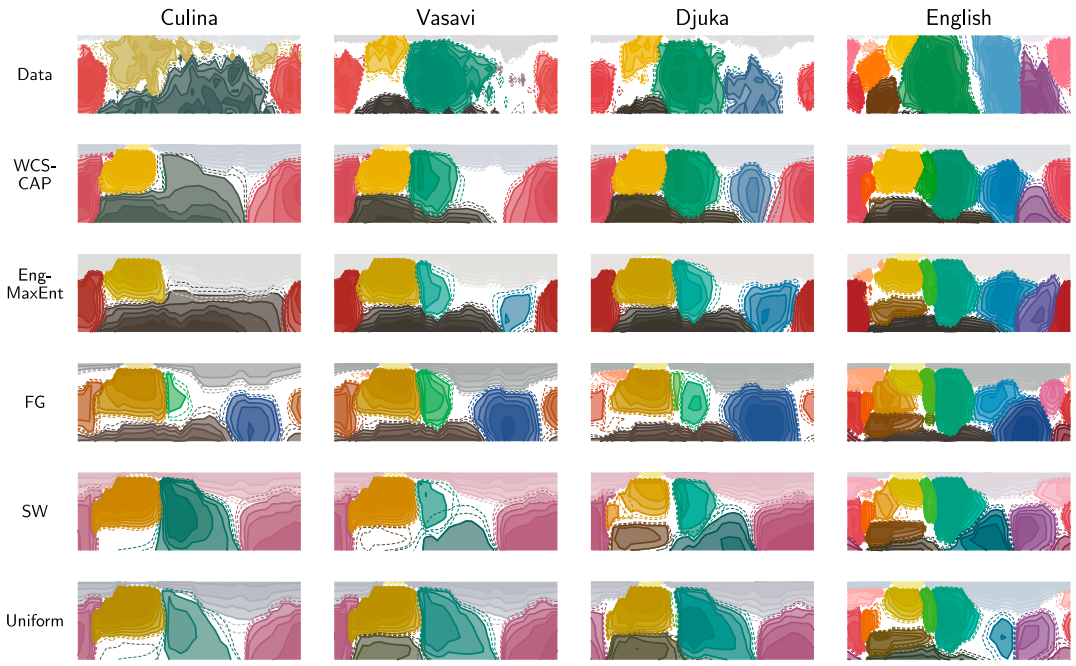

Figure 7. (Colour image online) Contour plots of colour naming systems from four languages (data row, same as Figure 1(c)) and the corresponding optimal systems which were predicted by the IB model under different priors. The variation shown for each model's prediction is caused by changes in the tradeoff parameter β that controls the location along the theoretical limit (see Figure 3(b) and section 3). Results for WCS-CAP and the uniform prior are from Zaslavsky et al. (2018).

estimated only from English naming data, and not those of any other language. This prior does not produce as good a fit to the actual data as do Eng-MaxEnt and WCS-CAP.

The relatively poor performance of the image-based priors is somewhat surprising, especially given that prior work (e.g., Gibson et al., 2017; Griffin, 2006; Yendrikhovskij, 2001) suggested that image statistics may play a central role in accounting for colour naming. Looking more closely at the results from the image-based priors may help to explain this seemingly inconsistent outcome.

Consider first the TF and FG image-based priors. They achieve similar scores and both perform better than the SW prior and the uniform prior, but not as well as the priors based on linguistic data (WCS-CAP and Eng-MaxEnt). These results seem inconsistent with the findings of Yendrikhovskij (2001), who found that colours sampled from 630 natural images form clusters in colour space that correspond roughly to known universal tendencies in colour naming. However, Steels and Belpaeme (2005) found that categories generated by Yendrikhovskij's method are correlated with human colour categories only slightly better than are categories derived from uniform sampling of colours.[9] In an attempt to more completely explore the apparent tension between our findings and those of Yendrikhovskij, we tried to replicate the findings of Yendrikhovskij (2001) using 1000 random images from the COCO dataset. Our analysis failed to replicate the qualitative results he obtained. This negative outcome could be due to the fact that we used a different set of images, or that the distribution of images in the COCO dataset is biased toward western cultures. However, there is also a further potential explanation for why the TF and FG frequencies do not perform well: they may not give good estimates of communicative need. Specifically, since most colours in natural images have low saturation (e.g., Hendley & Hecht, 1949; Steels & Belpaeme, 2005), the TF and FG frequencies are biased toward the achromatic chips. In our analyses, we excluded colours that were not sufficiently close to any of the WCS chips, but the bias toward the achromatics seems inherent to the statistics of colours in images in general, prior to any exclusion or filtering: the density near the achromatic chips is much higher than the density near the chromatic chips. This implies that the TF and FG priors predict greater communicative need for desaturated colours. Such a tendency seems unlikely given that consensus in colour naming, at least among English speakers, is positively correlated with chroma, such that highly saturated colours are named with highest consensus (Jraissati & Douven, 2018).

Consider now the SW prior. This prior is not biased toward desaturated colours. At the same time, it is closer to uniform than the other priors (Figure 5), it achieves scores similar to those of the uniform prior (Table 1), and it also predicts systems qualitatively similar to those predicted by the uniform prior (Figure 7). This suggests that the SW prior may be too close to uniform to accurately reflect communicative need.

6. Discussion

The possibility that both perceptual structure and communicative need may shape colour naming has long been discussed in the literature. However perception has traditionally been the focus of much more attention, and was incorporated first in computational accounts of colour naming, while communicative need remained an informal concept. Recently, this picture has started to change: the notion of communicative need has been cast formally as a prior over colours, and there is increasing evidence for the importance of this component. However, the factors that may characterize and shape communicative need have previously been only preliminarily explored. We approached this problem by exploring three major factors that may shape communicative need: capacity constraints, linguistic usage, and the visual environment. These factors were assessed within an independently motivated computational framework that integrates need and perception, and that predicts optimally efficient colour naming systems on that basis.

Our findings may be summarized in two main points. First, we found that different patterns of communicative need, instantiated as different priors, give rise to quite different efficient colour naming systems, given the same underlying perceptual structure. This finding further supports the idea that communicative need may have a substantial impact on colour naming, beyond the influence of perception. Second, we found that of the priors we considered, those based on capacity constraints and linguistic

usage provided the best fit to actual colour naming systems observed across languages. These best-performing priors were estimated from linguistic data, whereas other priors—uniform and image-based priors—did not account for the data as well. This suggests that communicative need may be well-estimated by the statistics of linguistic usage (Kemp & Regier, 2012; Regier et al., 2016; Xu & Regier, 2014), rather than by the statistics of the visual world to which language refers.

The corpus-based maximum entropy method for estimating need that we have presented here is novel, to our knowledge, and seems noteworthy for two reasons. First, it addresses the challenge of inferring communicative need from corpus statistics with minimal additional assumptions, and it can therefore in principle be applied widely across semantic domains. Second, while its performance is comparable to that of the capacity-achieving prior based on multiple languages in our dataset, it achieves this based on data from a single language. This suggests that there are important aspects of communicative need that are shared across languages, and that this method can be used to infer them. At the same time, we are not committed to the notion of an entirely universal prior. An important direction for future research is to test how well this corpus-based maximum entropy approach generalizes across languages and across domains, and to determine how and why communicative need varies across cultures, environments, and languages, beyond the simplifying assumption of a universal prior that we have made here.

Our findings do not imply that communicative need is uninfluenced by the statistics of the visual environment. Instead, they suggest that any influence of visual environment may be distal, and that language use may be a more direct reflection of need. This is broadly consistent with Boas' (1911, p. 26) view that cross-language variation in semantic categories "must to a certain extent depend upon the chief interests of a people": on this view, while the environment may shape a people's interests, it is those interests that directly shape the semantic categories of a given language—and those interests are presumably expressed through patterns of language use. This suggests two linked processes of adaptation. In the case of colour, colour naming may have adapted to communicative need and the structure of perceptual colour space—while need and perception may themselves have adapted to natural scene statistics (Shepard, 1994; Webster & Mollon, 1997), which may vary over time (Webster, Mizokami, & Webster, 2007) and space (McDermott & Webster, 2012). Although we have focused here on forces that shape colour naming, either directly or indirectly, it is also known that colour naming may in turn shape colour cognition and perception (Bae, Olkkonen, Allred, & Flombaum, 2015; Gilbert, Regier, Kay, & Ivry, 2006; Kay & Kempton, 1984; Roberson, Davies, & Davidoff, 2000; Winawer et al., 2007). Given the many moving parts in this overall picture, we find it striking that a universal perceptual colour space, and a universal prior based only on English usage, account for cross-language data as well as they do. Future research can usefully explore why this is the case, how far the universality extends, and when and under what circumstances language- and culture-specific forces dominate instead.

Notes

1. For simplicity, since it is assumed that each colour invokes a unique mental representation, we will treat c and m_c interchangeably when the distinction between them does not matter. For example, for any colour naming distribution $p(w|c)$ or prior $p(c)$, it holds that $q(w|m_c) = p(w|c)$ and $p(m_c) = p(c)$.
2. This *naming channel* is internal to the speaker, and it is distinct from the *communication channel* between the listener and speaker. The latter takes as input the word produced by the speaker and outputs the word perceived by the listener. The communication channel is left implicit in Figure 3(a) because this channel is assumed to be noiseless—i.e., the listener observes the speaker's word unaltered.
3. For compatibility with the analysis performed by Zaslavsky et al. (2018), we followed their regularization process and excluded fifteen languages from our quantitative evaluation (Table 1). We also repeated the evaluation process with all languages and obtained similar results; thus the regularization process does not influence our conclusions.
4. We used the python package cvxopt to solve this optimization problem. In general, it is possible that the feasible set would be empty, i.e. that there would be no prior that satisfies the constraints. However, this is not the case in our setting.
5. We considered the 2014 training dataset which contains 82,783 images. These images are annotated with object boundaries for objects from 80 different categories.
6. Conversion from RGB to CIELAB coordinates was done with the colorspacious python package, using illuminant

C. For the achromatic chips, only pixels with $\Delta E^2 = (L^*)^2 + (a^*)^2 + (b^*)^2 < 70$ were considered. For the chromatic chips, the comparison was based only on lightness and hue values, and pixels for which the square distance to the closest chromatic chip was greater than 400 were excluded. These thresholds were validated by manual inspection, to ensure that the converted pixels are indeed perceptually similar to the original ones.

7. The TF and the FG priors have similar structure and both give the highest probability mass to the achromatic colours. However, the FG prior gives less weight to the achromatic chips than the TF prior does. In addition, according to the FG prior, warm colours have higher probability than cool colours, similar to the SW prior we estimated, and consistent with the salience data of Gibson et al. (2017).
8. These two measures correspond to ε_I and gNID respectively. See (Zaslavsky et al., 2018) for more detail.
9. We thank Delwin Lindsey for drawing our attention to this connection.

Acknowledgments

We thank Joshua Abbott for helpful discussions, and Delwin Lindsey and Angela Brown for kindly sharing their English colour-naming data with us.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by DTRA (Defense Threat Reduction Agency) award HDTRA11710042 (NZ, TR) and the Gatsby Charitable Foundation (NZ, NT).

References

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, *18*(1), 14–20. doi:10.1109/TIT.1972.1054753

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*, 744–763. doi:10.1037/xge0000076

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Blahut, R. E. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, *18*(4), 460–473. doi:10.1109/TIT.1972.1054855

Boas, F. (1911). Introduction. In *Handbook of American Indian Languages, Vol.1* (pp. 1–83). Government Print Office (Smithsonian Institution, Bureau of American Ethnology, Bulletin 40).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, *31*(1) Blackwell Publishing Ltd, 99–132. doi:10.1080/03640210709336986

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788–791. doi:10.1073/pnas.0335980100

Genewein, T., Leibfried, F., Grau-Moya, J., & Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, *2*, 27. doi:10.3389/frobt.2015.00027

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., … Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790. doi:10.1073/pnas.1619666114

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, *103*, 489–494. doi:10.1073/pnas.0509868103

Griffin, L. D. (2006). Optimality of the basic colour categories for classification. *Journal of the Royal Society Interface*, *3*(6), 71–85. doi:10.1098/rsif.2005.0076

Harremoës, P., & Tishby, N. (2007). The Information Bottleneck revisited or how to choose a good distortion measure. *IEEE International Symposium on Information Theory*, 566–571. doi:10.1109/ISIT.2007.4557285

Hendley, C. D., & Hecht, S. (1949). The colors of natural objects and terrains, and their relation to visual color deficiency. *Journal of the Optical Society of America*, *39*(10), 870–873. doi:10.1364/JOSA.39.000870

Jameson, K., & D'Andrade, R. G. (1997). It's not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). Cambridge: Cambridge University Press.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952. doi:10.1109/PROC.1982.12425

Jraissati, Y., & Douven, I. (2018). Delving deeper into color space. *I-Perception*, *9*(4). doi:10.1177/2041669518792062

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The World Color Survey*. Stanford: Center for the Study of Language and Information.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*, 65–79. doi:10.1525/aa.1984.86.1.02a00050

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610–646. doi:10.1353/lan.1978.0035

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. doi:10.1126/science.1218811

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109–128. doi:10.1146/annurev-linguistics-011817-045406

Komarova, N. L., Jameson, K. A., & Narens, L. (2007). Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology*, *51*(6), 359–382. doi:10.1016/j.jmp.2007.06.001

Kuschel, R., & Monberg, T. (1974). 'We don't talk much about colour here': A study of colour semantics on Bellona Island. *Man, (New Series)*, *9*(2), 213–242. doi:10.2307/2800075

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Lawrence Zitnick, C. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *European Conference on Computer Vision* (pp. 740–755). Cham: Springer International Publishing.

Lindsey, D. T., & Brown, A. M. (2006). Universality of color names. *Proceedings of the National Academy of Sciences*, *103*(44), 16608–16613. doi:10.1073/pnas.0607708103

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of American English. *Journal of Vision*, *14*(2), 17. doi:10.1167/14.2.17

Lindsey, D. T., Brown, A. M., Brainard, D. H., & Apicella, C. L. (2015). Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, *25*(18), 2441–2446. doi:10.1016/j.cub.2015.08.006

Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, *109*(18), 6819–6824. doi:10.1073/pnas.1113347109

Marzen, S. E., & DeDeo, S. (2017). The evolution of lossy compression. *Journal of the Royal Society Interface*, *14*(130), 20170166. doi:10.1098/rsif.2017.0166

McDermott, K. C., & Webster, M. A. (2012). Uniform color spaces and natural image statistics. *Journal of the Optical Society of America A*, *29*(2), A182–A187. doi:10.1364/JOSAA.29.00A182

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L., (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. doi:10.1126/science.1199644

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLoS ONE*, *11*(4), e0151138. doi:10.1371/journal.pone.0151138

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441. doi:10.1073/pnas.0610341104

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney, & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*, 369–398. doi:10.1037/0096-3445.129.3.369

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

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, *4*(142–163), 1.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28. doi:10.3758/BF03200759

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198. doi:10.1016/j.cognition.2016.03.020

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489. doi:10.1017/S0140525X05000087

Tishby, N., Pereira, F. C., & Bialek, W. (1999). *The Information Bottleneck method*. Proceedings of the 37th Annual Allerton Conference on communication, Control and Computing.

Tkačik, G., & Bialek, W. (2016). Information processing in living systems. *Annual Review of Condensed Matter Physics*, *7*(1), 89–117. doi:10.1146/annurev-conmatphys-031214-014803

Webster, M. A., Mizokami, Y., & Webster, S. M. (2007). Seasonal variations in the color statistics of natural images. *Network: Computation in Neural Systems*, *18*(3), 213–233. doi:10.1080/09548980701654405

Webster, M. A., & Mollon, J. D. (1997). Adaptation and the color statistics of natural images. *Vision Research*, *37*(23), 3283–3298. doi:10.1016/S0042-6989(97)00125-9

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*, 7780–7785. doi:10.1073/pnas.0701644104

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary Is optimized for usage. *Cognition*, *179*, 213–220. doi:10.1016/j.cognition.2018.05.008

Xu, Y., & Regier, T. (2014). *Numeral systems across languages support efficient communication: From approximate numerosity to recursion*. Proceedings of the 36th Annual Meeting of the cognitive Science society.

Yendrikhovskij, S. N. (2001). A computational model of colour categorization. *Color Research & Application*, *26*, S235–S238. doi:10.1002/1520-6378(2001)26:1+<::AID-COL50>3.0.CO;2-O

Yendrikhovskij, S. N., Blommaert, F. J. J., & de Ridder, H. (1998). *Optimizing color reproduction of natural images*. The Sixth Color Imaging Conference: Color Science, Systems, and Applications.

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

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2019). Color naming reflects both perceptual structure and communicative need. *Topics in Cognitive Science*, *11*(1), 207–219. doi:10.1111/tops.12395

Part III

Evolution of Compressed Representations

Chapter 7

Deterministic Annealing and the Evolution of Information Bottleneck Representations

Deterministic annealing and the evolution of Information Bottleneck representations

Noga Zaslavsky[1] and Naftali Tishby[1,2]

[1]Edmond and Lily Safra Centre for Brain Sciences, The Hebrew University of Jerusalem
[2]Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem

Abstract

The Information Bottleneck (IB) framework provides a principled and broadly applicable approach for studying efficient compressed representations in artificial and biological systems. However, a comprehensive mathematical understanding of the optimal IB representations and the structural phase transitions they undergo via deterministic annealing exists only in a few limited cases. Here, we address the case of symbolic, or discrete, representations, which is particularly relevant to the emergence of language and abstract representations more generally. We characterize the structural changes in the IB representations as they evolve via a deterministic annealing process; derive an algorithm for finding critical points; and explore numerically the types of bifurcations and related phenomena that occur in IB. This work extends the theoretical grounds for understanding optimal representations within the IB framework.

1 Introduction

The Information Bottleneck (IB) framework [1] provides a principled approach for studying efficient compressed representations in artificial and biological systems. In this view, efficient representations should compress their inputs by maintaining the minimal amount of information on the input that is required for making accurate predictions about a target variable. In the past several years, there has been a surge of evidence for the wide applicability of IB in multiple fields, including deep learning [2, 3, 4, 5], and machine learning in general [6, 7], neuroscience [8, 9, 10], language [11, 12], and music [13]. However, a comprehensive mathematical understanding of the structure and evolution of the IB representations exists only in very few cases, usually when Gaussian assumptions are made [7, 14].

The goal of this work is to extend this understanding to the case of discrete random variables, which induce symbolic IB representations. This setting is particularly relevant to the emergence of language [12] and, more broadly, abstract representations. Structural phase

transitions[1] have previously been studied in related settings, such as clustering and classification [15, 16], and to some extent also in the case of IB [17]. The present work goes beyond these previous studies by (1) introducing order parameters that capture the evolution of the IB representations; (2) deriving a novel algorithm for finding critical points in which the representations undergo a phase transition; and (3) exploring numerically the types of phase transitions and related phenomena that occur in IB.

The remainder of this paper is structured as follows. In Section 2 we formulate the notion of efficient compressed representations and ground it in the IB principle. In Section 3 we characterize the evolutionary process of the IB representations and the structural phase transitions they undergo. In Section 4 we present numerical simulations that demonstrate these phenomena.

2 Efficient compressed representations

2.1 Setting

Let $X \in \mathcal{X}$ be a source random variable, $Y \in \mathcal{Y}$ a target variable, and $p(x, y)$ their joint distribution. We assume $p(x, y)$ is known, although in practical applications this distribution is often estimated from data (see [6] for confidence bounds). For simplicity, assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite sets with sizes m and n respectively. For any two random variables, denote by $\triangle(\mathcal{X})$ the $(m-1)$-dimensional simplex of distributions over the elements of $\mathcal{X}$, and by $\triangle(\mathcal{Y})^{\mathcal{X}}$ the set of conditional distributions of Y given X. That is, $\triangle(\mathcal{Y})^{\mathcal{X}} = \triangle(\mathcal{Y}) \times \cdots \times \triangle(\mathcal{Y})$ is the m-ary product of $\triangle(\mathcal{Y})$. We are interested in characterizing efficient representations of X.

Definition 1. *A representation $\hat{X} \in \hat{\mathcal{X}}$ is a stochastic function of X, defined by a conditional distribution $p(\hat{x}|x) \in \triangle(\hat{\mathcal{X}})^{\mathcal{X}}$. If $\hat{\mathcal{X}}$ is a discrete set of arbitrary symbols, then we say that $\hat{X}$ is a symbolic representation of X.*

In this work we consider symbolic representations, where $|\hat{\mathcal{X}}|$ is finite. From an information-theoretic perspective, $p(\hat{x}|x)$ is a stochastic *encoder* and $\hat{\mathcal{X}}$ is the *code alphabet*. In addition, Definition 1 implies that $\hat{X}$ obeys the Markov chain $Y - X - \hat{X}$.

This general setup is broadly applicable. For example, in supervised learning settings [e.g., 11, 6, 2], X would be an input of a classifier, Y would be a target label, and $\hat{X}$ would be an intermediate representation employed by the classifier. In unsupervised learning, this setting corresponds to distributional clustering [e.g., 18, 19], namely assignment of the points $p(y|x) \in \triangle(\mathcal{Y})$ to clusters $\hat{x} \in \hat{\mathcal{X}}$. In statistics, Y may be an unknown parameter of a distribution $p_y(x) = p(x|y)$, in which case X would be a sample from this distribution, and $\hat{X}$ would be a statistic of the sample. In the case of semantic systems [12], Y would be a set of relevant features in the environment, X would be a referent defined by a distribution over features,

[1] We use the term "phase transitions" a bit loosely. Strictly speaking, the phenomena we study are bifurcations, which are not necessary phase transitions in the physical sense.

i.e.$p(y|x)$, and $\hat{X}$ would be a word that is used to communicate the referent.

2.2 The Information Bottleneck method

In all of the settings mentioned above, we may ask: what would be an optimal representation? Intuitively, a good representation should require minimal resources, while achieving maximal predictive power. This intuition is formalized by the Information Bottleneck (IB) principle [1]. According to IB, the complexity of the representation is measured by $I_p(X;\hat{X})$, which is roughly the number of bits that are required for representing X using $\hat{X}$. The informativeness, or accuracy, of the representation is measured by $I_p(\hat{X};Y)$, which is the amount of *relevant information* about Y preserved by the representation. The optimal IB representations minimize $I_p(X;\hat{X})$, such that $I_p(\hat{X};Y)$ remains sufficiently high. Formally, this constrained optimization problem can be solved by minimizing the Lagrangian

$$\mathcal{F}_\beta[p(\hat{x}|x)] = I_p(X;\hat{X}) - \beta I_p(\hat{X};Y)\,, \tag{1}$$

where $\beta \geq 0$ is the Lagrange multiplier for the constraint on $I_p(\hat{X};Y)$. β can also be considered as a tradeoff parameter, or inverse-temperature in analogy to statistical mechanics [15]. Given β, denote the optimal value of the IB objective by $\mathcal{F}^*_\beta$, and the optimal complexity and accuracy by $I_\beta(X;\hat{X})$ and $I_\beta(\hat{X};Y)$ respectively. The IB theoretical limit is defined by the Pareto optimal tradeoffs $(I_\beta(X;\hat{X}), I_\beta(\hat{X};Y))$ as a function of β. This parametric curve [20] is called the *information curve* (see Figure 1A for example).

Tishby et al. [1] showed that a necessary condition for $p_\beta(\hat{x}|x)$ to be a stationary point of $\mathcal{F}_\beta$ is that it satisfies the following self-consistent equations:

$$\begin{cases} p_\beta(\hat{x}|x) &= \dfrac{p_\beta(\hat{x})}{Z_\beta(x)} \exp\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})]\right) \\ p_\beta(\hat{x}) &= \displaystyle\sum_{x\in\mathcal{X}} p(x)p_\beta(\hat{x}|x) \\ p_\beta(y|\hat{x}) &= \displaystyle\sum_{x\in\mathcal{X}} p(y|x)p_\beta(x|\hat{x}) \end{cases}, \tag{2}$$

where $Z_\beta(x)$ is the normalization factor, also known as the partition function, and $p_\beta(x|\hat{x})$ is obtained by applying Bayes' rule with respect to $p_\beta(\hat{x}|x)$ and $p(x)$. We refer to representations that satisfy (2) as *IB representations*. These representations can be found via the IB method (Algorithm 1), which is a variant of the Blahut–Arimoto algorithm [21, 22].

2.3 Effective cardinality

The cardinality of an IB representation $K(p_\beta)$ is defined by the cardinality of its support, $Supp(p_\beta) = \{\hat{x} \in \hat{\mathcal{X}} : p_\beta(\hat{x}) > 0\}$. That is, $K(p_\beta) = |Supp(p_\beta)|$. The following proposition

Algorithm 1: IB [Tishby et al., 1999]

Input: $p(x, y)$, initial mapping $p_0(\hat{x}|x)$, and tradeoff $\beta \geq 0$
Output: Fixed point of $\mathcal{F}_\beta$
$p(\hat{x}|x) \leftarrow p_0(\hat{x}|x)$
while $p(\hat{x}|x)$ *not converged* **do**
  $p(\hat{x}) \leftarrow \sum_x p(x)p(\hat{x}|x)$
  $p(y|\hat{x}) \leftarrow \sum_x p(y|x)p(x|\hat{x}(\hat{x}))$
  $p(\hat{x}|x) \leftarrow \frac{p(\hat{x})}{Z(x)} \exp\left(-\beta D[p(y|x)\|p(y|\hat{x})]\right)$
return $p(\hat{x}|x)$

shows that there may be a simple transformation that reduces the cardinality of the representation without compromising its optimality given β.

Proposition 1. *If $p_\beta(\hat{x}|x)$ is an IB representation with cardinality K, and there are $\hat{x}_1, \hat{x}_2 \in Supp(p_\beta)$ such that $p_\beta(y|\hat{x}_1) = p_\beta(y|\hat{x}_2)$, then there exists an IB representation $\tilde{p}_\beta(\hat{x}|x)$ with cardinality $K-1$ such that $\mathcal{F}_\beta[\tilde{p}_\beta] = \mathcal{F}_\beta[p_\beta]$.*

Proof. We construct a representation $\tilde{p}_\beta(\hat{x}|x)$ by merging $\hat{x}_1$ and $\hat{x}_2$. For all x and $\hat{x} \neq \hat{x}_1, \hat{x}_2$, let $\tilde{p}_\beta(\hat{x}|x) = p_\beta(\hat{x}|x)$. For $\hat{x}_2$ let $\tilde{p}_\beta(\hat{x}_2|x) = 0$, and for $\hat{x}_1$ let $\tilde{p}_\beta(\hat{x}_1|x) = p_\beta(\hat{x}_1|x) + p_\beta(\hat{x}_2|x)$. Given this construction, it is easy to verify that $\tilde{p}_\beta$ satisfies the IB equations (2), and that $\mathcal{F}_\beta[p_\beta] = \mathcal{F}_\beta[\tilde{p}_\beta]$. In addition, since $\tilde{p}_\beta(\hat{x}_2) = 0$, it holds that $Supp(\tilde{p}_\beta) = Supp(p_\beta)\backslash\{\hat{x}_2\}$, which implies that $K(\tilde{p}_\beta) = K - 1$, and this concludes the proof.

□

p_β and $\tilde{p}_\beta$ are equivalent representations in the sense that they keep the same information about X and Y. More generally, we define the equivalence class of p_β by the set of all representations $\tilde{p}_\beta$ that satisfy the IB equations (2) for the same value of β, and for which there exist mappings $\varphi : \hat{\mathcal{X}} \to \hat{\mathcal{X}}$ and $\psi : \hat{\mathcal{X}} \to \hat{\mathcal{X}}$ such that $\tilde{p}_\beta(y|\varphi(\hat{x})) \equiv p_\beta(y|\psi(\hat{x}))$. In other words, the equivalence class of p_β is determined by the set of distributions over $\mathcal{Y}$ that it induces, i.e.,

$$\{p(y) \in \triangle(\mathcal{Y}) : \exists \hat{x},\, p(y) \equiv p_\beta(y|\hat{x})\}\,. \tag{3}$$

Denote this equivalence class by $[p_\beta]$. Here, we focus on representations with minimal cardinality within their equivalence class.

Definition 2. *The effective cardinality of an IB representation p_β is*

$$k(p_\beta) = \min_{\tilde{p}_\beta \in [p_\beta]} K(\tilde{p}_\beta).$$

We say that $p_\beta(\hat{x}|x)$ is a canonical IB representation if $k(p_\beta) = K(p_\beta)$.

In the remainder of this paper we assume that the IB representations are canonical, unless stated otherwise. In particular, this implies that $p_\beta(y|\hat{x}_1) \neq p_\beta(y|\hat{x}_2)$ for all $\hat{x}_1 \neq \hat{x}_2$.

Algorithm 2: Reverse Deterministic Annealing for IB (RDA-IB)

Input: $p(x,y)$, scheduling $\beta_t > \beta_{t-1} > \cdots > \beta_1 \geq 0$
Output: Fixed points for all β_i
$p_0(\hat{x}|x) \leftarrow I_m$ (initialize)
for $i = t, t-1, \ldots, 1$ **do**
  $p_i(\hat{x}|x) \leftarrow \mathrm{IB}\left(p(x,y), p_{i-1}(\hat{x}|x), \beta_i\right)$ (initialize IB with the previous f.p.)
return $\{p_i(\hat{x}|x)\}_{i=1}^{t}$

Notice that for $\beta = 0$, the global optimum is trivial, and any $\hat{X}$ that is independent of X will attain the minimum $\mathcal{F}_0^* = 0$. In fact, this holds for all $\beta \in [0,1]$, because $I(X;\hat{X}) \geq I(\hat{X};Y)$ due to the Data Processing Inequality [23]. A canonical representation in this case is a constant $\hat{x}$, and so the effective cardinality is $k = 1$. As $\beta \to \infty$, the optimal mapping from X to $\hat{X}$ becomes deterministic, and the effective cardinality would be maximal. In particular, if $|\hat{\mathcal{X}}| \geq |\mathcal{X}|$, then the global optimum is attained by any one-to-one mapping from $\mathcal{X}$ to $\hat{\mathcal{X}}$.[2] In between these two extremes, as β gradually increases, the IB representations undergo a sequence of structural changes, also called phase transitions or bifurcations, in which the effective cardinality changes.

Intuitively, we can think of $I_\beta(X;\hat{X})$ as the logarithm of the effective cardinality because

$$k(p_\beta) \approx 2^{I_\beta(X;\hat{X})}. \tag{4}$$

This follows from the same typicality argument that Shannon applied in Rate-Distortion theory [24], which implies that $I_\beta(X;\hat{X})$ is roughly the minimal number of bits that are needed for encoding X using $\hat{X}$.

2.4 Reverse deterministic annealing

The IB optimization problem is non-convex, and thus Algorithm 1 is prone to converge to local minima of $\mathcal{F}_\beta$. A common approach for mitigating this problem is based on the notion of deterministic annealing [15, 16, and see also 25]. A deterministic annealing optimization procedure starts with an initial solution for a low value of β, e.g., $\beta = 0$, for which finding a globally optimal solution is trivial. Then, the solution is refined by invoking the iterative algorithm while gradually increasing β (cooling down the system) according to some annealing schedule. This process attempts to track the optimal solution as β increases from 0 to ∞.

Here, we are not only interested in the solution for $\beta \to \infty$, but rather in the whole trajectory which captures the evolution of the IB representations. In fact, if $|\hat{\mathcal{X}}| = |\mathcal{X}|$, then the solution for $\beta \to \infty$ is straight forward, as mentioned earlier. This suggests a *reverse deterministic annealing* procedure, which starts with a bijective representation and a large value of β, and

[2] We assume here that a non-trivial minimal sufficient statistics (MSS) of X for Y does not exists. If it does, then at the limit $\hat{X}$ would be isomorphic to the MSS.

then gradually decreases β. This procedure is summarized in Algorithm 2. The numerical simulations in Section 4 are based on reverse deterministic annealing because we found it to be more numerically stable than deterministic annealing, while yielding overall similar results.

3 Characterizing the evolution of the IB representations

In this section, we present several tools for characterizing the evolution of the IB representations and the structural phase transitions they undergo. More specifically, we propose several measures that reflect these structural changes as β varies, and introduce an algorithm for finding critical values of β. In addition, we analyze the structural phase transitions in the special case where Y is a deterministic function of X.

3.1 Bifurcations in IB

Bifurcation diagrams are a powerful method for observing qualitative changes in the fixed points of a dynamical system that occur when varying a bifurcation parameter [26]. In our case, the dynamics is defined by the iterative process of Algorithm 1, the fixed points of this process are the IB representations, and the bifurcation parameter is β. Typically, bifurcation diagrams show the fixed points of the system as a function of the bifurcation parameter. However, the IB fixed points are usually high-dimensional distributions, and so it is not always clear how to observe their bifurcations. Here we discuss how to address this issue in two cases: (1) when Y is binary, and (2) in the more general case of discrete variables.

3.1.1 Centroid bifurcations

We have shown in Section 2.3 that the set $\{p(y) \in \triangle(\mathcal{Y}) : \exists \hat{x},\, p(y) \equiv p_\beta(y|\hat{x})\}$ defines the equivalence class $[p_\beta]$. This implies that it is sufficient to consider $p_\beta(y|\hat{x})$ as a function of β, instead of $p_\beta(\hat{x}|x)$. In the case in which Y is binary, this reduces to a single parameter for each $\hat{x}$, namely $p_\beta(y=1|\hat{x})$. We refer to this type of bifurcation diagram as the *centroid bifurcation diagram*, because $p_\beta(y|\hat{x})$ can be viewed as the cluster centroids of the points $p(y|x) \in \triangle(\mathcal{Y})$ under the clustering $p_\beta(\hat{x}|x)$.

Figure 1D shows an example of this type of bifurcation diagram. For $\beta = 1$ there is only one possible value, $p_\beta(y=1|\hat{x}) = 0.5$, which corresponds to the prior distribution $p(y)$. This fixed point remains stable (and optimal) also for β greater than 1, but smaller from some critical value β_0. The first bifurcation occurs at β_0, when the prior centroid splits into two centroids and the effective cardinality increases. It is easy to verify that $p_\beta(y|\hat{x}) = p(y)$ remains a fixed point of the IB equations even for $\beta > \beta_0$, by simply substituting this solution in (2). However, this fixed point loses its stability at β_0 and is no longer globally optimal after that point. This type of phase transition is analogous to a pitchfork bifurcation [26]. A second critical point can

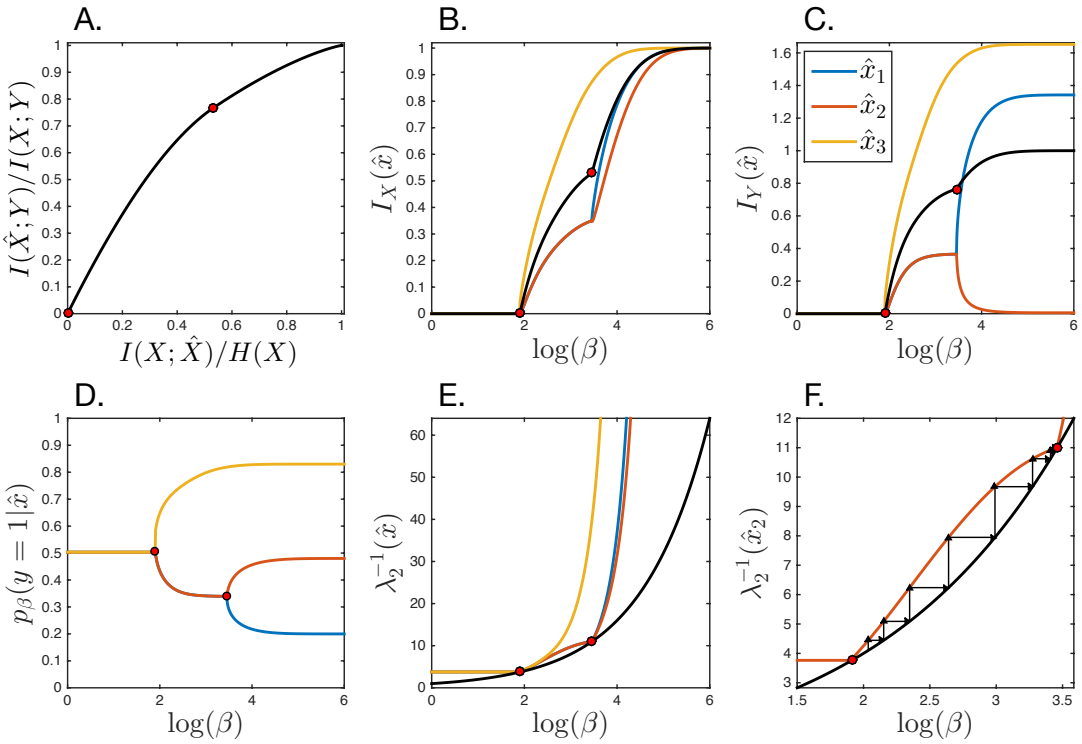

Figure 1. Characterization of the evolution of IB representations in an illustrative example. Here, Y is binary, X and $\hat{X}$ are trinary, $p(x)$ is uniform, and $p(y|x)$ is shown by the leafs of the bifurcation tree in panel D. The red points in all panels correspond to the two critical points that were found by Algorithm 3. **A.** Normalized information curve. **B-C.** Bifurcations of the (normalized) order parameters $I_X(\hat{x})$ and $I_Y(\hat{x})$ respectively. The expected values over $\hat{x}$, i.e. I_X and I_Y, are shown by the black curves. **D.** Centroid bifurcation diagram. **E.** Evolution of $\lambda_2^{-1}(\hat{x})$ for each $\hat{x}$, where $\lambda_2(\hat{x})$ is the second largest eigenvalue of $C_Y^{\beta,\hat{x}}$. The black curve shows β. **F.** Closer view of $\lambda_2^{-1}(x_2)$ (red curve in panel E) and β (black curve in panel E) near the two critical points. Black arrows show iterations of Algorithm 3 in which λ_2 is computed given β (vertical arrows) and then β is updated from λ_2 (horizontal arrows) until convergence. See main text for more detail.

also be seen in Figure 1D, in which another split occurs. The effective cardinality after this split is $K = 3$, and since in this case $|\hat{\mathcal{X}}| = 3$, another bifurcation after this point is impossible.

3.1.2 Informational bifurcations

Centroid bifurcation diagrams are useful when Y is binary, but are difficult to visualize when $|\mathcal{Y}| > 2$. Therefore, we propose an alternative approach that can be applied in the more general case of discrete variables. To this end, we define two informational measures that reflect the structural changes in IB as a function of β.

Definition 3. *Given an IB representation p_β for some $\beta \geq 0$, the point-wise information of*

$\hat{x} \in Supp(p_\beta)$ *about* X *or* Y*, denoted by* $I_X^\beta(\hat{x})$ *or* $I_Y^\beta(\hat{x})$ *respectively, are defined as*

$$I_X^\beta(\hat{x}) = D\left[p_\beta(x|\hat{x})\|p(x)\right] \tag{5}$$
$$I_Y^\beta(\hat{x}) = D\left[p_\beta(y|\hat{x})\|p(y)\right] , \tag{6}$$

where $D[\cdot\|\cdot]$ *is the Kullback-Leibler divergence. These measures are undefined for* $\hat{x} \notin Supp(p_\beta)$.

Notice that before the first phase transition, i.e. for $\beta < \beta_0$, if these measures are defined then necessarily $I_X^\beta(\hat{x}) = 0$ and $I_Y^\beta(\hat{x}) = 0$. At a critical point $\beta_{\hat{x}}$ after which $\hat{x} \in Supp(p_\beta)$, these two informational measures become non-negative. We refer to these measures as *order parameters*, as they are indicative of structural changes in the representation. Note that

$$I_\beta(\hat{X};Y) = \mathop{\mathbb{E}}_{\hat{x}\sim p_\beta(\hat{x})}[I_Y^\beta(\hat{x})] \tag{7}$$

and similarly, $I_\beta(X;\hat{X}) = \mathbb{E}[I_X^\beta(\hat{x})]$. In this sense, these two informational order parameters compose the information curve.

Figure 1B and Figure 1C show the changes of these order parameters as a function of $\log(\beta)$, for the same illustrative example considered in Section 3.1.1. We refer to these types of diagrams as *informational bifurcation diagrams*. The structural changes of the IB representations, directly observed in Figure 1D, are also reflected in the informational bifurcation diagrams, in which cardinality changes are accompanied by an emergence of an order parameter. This order parameter corresponds to the $\hat{x}$ that has been added to $Supp(p_\beta)$. These structural changes are also reflected in the expected values of the order parameters (black curves in Figure 1B-C), i.e., $I_\beta(X;\hat{X})$ and $I_\beta(\hat{X};Y)$, which have discontinuous derivatives with respect to β at the critical points. The following proposition shows that these discontinuities occur exactly at the same values of β.

Proposition 2. $\frac{\partial}{\partial\beta} I_\beta(X;\hat{X}) = \beta \frac{\partial}{\partial\beta} I_\beta(\hat{X};Y)$.

Proof. Substituting the explicit form of $p_\beta(\hat{x}|x)$, as given by (2), in $I_\beta(X;\hat{X})$ gives

$$\begin{aligned} I_\beta(X;\hat{X}) &= \sum_{x,\hat{x}} p(x)p_\beta(\hat{x}|x)\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})] - \log Z_\beta(x)\right) \\ &= -\mathop{\mathbb{E}}_x\left[\log Z_\beta(x)\right] - \beta\left(I(X;Y) - I_\beta(\hat{X};Y)\right) , \end{aligned}$$

where the second step follows from Lemma 1 in the Appendix. Therefore, the derivative with respect to β is

$$\frac{\partial}{\partial\beta} I_\beta(X;\hat{X}) = -\frac{\partial}{\partial\beta}\mathop{\mathbb{E}}_x\left[\log Z_\beta(x)\right] - \left(I(X;Y) - I_\beta(\hat{X};Y)\right) + \beta\frac{\partial}{\partial\beta} I_\beta(\hat{X};Y).$$

Lemma 2 in the Appendix shows that $\frac{\partial}{\partial\beta}\mathbb{E}_x\left[\log Z_\beta(x)\right] = I_\beta(\hat{X};Y) - I(X;Y)$. Substituting this in the equation above concludes the proof.

□

Another implication of Proposition 2 is that the discontinuities in $\frac{\partial}{\partial\beta}I_\beta(X;\hat{X})$ and $\frac{\partial}{\partial\beta}I_\beta(\hat{X};Y)$ coincide with Ehrenfest's definition of second-order phase transitions [27]. According to Ehrenfest, a second-order phase transition occurs if the second order derivative of the free energy $\mathcal{F}^*_\beta$ is discontinuous, but not the first order derivative. The following corollary shows that the n-th order derivative of $\mathcal{F}^*_\beta$ is given by the $(n-1)$-th order derivative of $-I_\beta(\hat{X};Y)$.

Corollary 1. $\frac{\partial}{\partial\beta}\mathcal{F}^*_\beta = -I_\beta(\hat{X};Y)$.

Proof. This follows directly from Proposition 2 because taking the derivative of $\mathcal{F}^*_\beta$ with respect to β gives

$$\frac{\partial}{\partial\beta}\mathcal{F}^*_\beta = \frac{\partial}{\partial\beta}I_\beta(X;\hat{X}) - \beta\frac{\partial}{\partial\beta}I_\beta(\hat{X};Y) - I_\beta(\hat{X};Y)\,.$$

□

Therefore, if the first-order derivative of $I_\beta(\hat{X};Y)$ is discontinuous, then so is the second-order derivative of $\mathcal{F}^*_\beta$. If $I_\beta(\hat{X};Y)$ is continuous in β, then this corresponds to Ehrenfest's second-order phase transition, and otherwise to a first-order phase-transition. Furthermore, proposition 2 and corollary 1 suggest that in practice it is sufficient to consider only $I^\beta_Y(\hat{x})$ as the order parameter. This conclusion is further supported by lemmas 3 and 4 in the Appendix, which show more precisely how the two order parameters and their derivatives are related.

3.2 Finding critical points

Thus far we have showed that the evolution of IB representations is reflected in a set of order parameters, $\mathcal{O} = \{I^\beta_Y(\hat{x}) : \hat{x} \in Supp(p_\beta), \beta \geq 0\}$. These parameters capture the evolutionary trajectory and the critical values of β in which second order phase transitions occur. A natural question is then: Given a joint distribution $p(x,y)$, what are the values of these critical points? To address this question, we propose an algorithm for finding such points. We refer to this algorithm as *Criticality Search* (Algorithm 3). First, we derive a necessary condition for a second-order phase transition, which will form the basis of the algorithm.

Following a similar argument as in [15], we consider small perturbations of the IB representation near a critical point. At a critical point in which a cluster splits continuously, there exist non-trivial perturbations $h_{\tilde{\beta}}(x,\hat{x})$ such that for all $\tilde{\beta} \geq \beta$, in a small vicinity of β, it holds that $\tilde{p}_\beta(\hat{x}|x) = p_\beta(\hat{x}|x) + h_{\tilde{\beta}}(x,\hat{x})$ satisfies the IB equations (2) for $\tilde{\beta}$. Assuming that the right derivatives of $h_{\tilde{\beta}}(x,\hat{x})$ and $p_{\tilde{\beta}}(\hat{x}|x)$ with respect to $\tilde{\beta}$ exist and are non-zero at $\tilde{\beta} = \beta$, then $\nabla_{h_{\tilde{\beta}}} \log p_{\tilde{\beta}}(\hat{x}|x)\big|_{\tilde{\beta}=\beta}$ is well-defined, and so are these derivatives for $\log p_{\tilde{\beta}}(x|\hat{x})$ and $\log p_{\tilde{\beta}}(y|\hat{x})$. As in [15], we neglect the influence of inter-cluster interactions, which yields

in our case the approximation

$$\mathbf{u}_{\hat{x},\beta}[x] \triangleq \sum_{x'} \frac{\partial \log p_\beta(x|\hat{x})}{\partial h_\beta(x',\hat{x})} \approx \beta \sum_y p(y|x) \sum_{x'} \frac{\partial \log p_\beta(y|\hat{x})}{\partial h_\beta(x',\hat{x})} \tag{8}$$

$$\mathbf{v}_{\hat{x},\beta}[y] \triangleq \sum_{x'} \frac{\partial \log p_\beta(y|\hat{x})}{\partial h_\beta(x',\hat{x})} = \sum_x \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \sum_{x'} \frac{\partial \log p_\beta(x|\hat{x})}{\partial h_\beta(x',\hat{x})} \,. \tag{9}$$

The coupled equations (8)-(9) can be re-organized and simplified as follows:

$$\mathbf{u}_{\hat{x},\beta}[x] \approx \beta \sum_y p(y|x) \sum_{x'} \frac{p(y|x')p_\beta(x'|\hat{x})}{p_\beta(y|\hat{x})} \mathbf{u}_{\hat{x},\beta}[x'] \tag{10}$$

$$\mathbf{v}_{\hat{x},\beta}[y] \approx \beta \sum_x \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \sum_{y'} p(y'|x)\mathbf{v}_{\hat{x},\beta}[y'] \,. \tag{11}$$

This gives two non-linear eigenvector conditions for a cluster split,

$$(\beta^{-1} I - C_X^{\beta,\hat{x}})\mathbf{u}_{\hat{x},\beta} = 0 \tag{12}$$

$$(\beta^{-1} I - C_Y^{\beta,\hat{x}})\mathbf{v}_{\hat{x},\beta} = 0 \,, \tag{13}$$

where $C_X^{\beta,\hat{x}}$ is a $|\mathcal{X}| \times |\mathcal{X}|$ matrix defined by

$$C_X^{\beta,\hat{x}}[x,x'] = \sum_y \frac{p(y|x)p(y|x')p_\beta(x'|\hat{x})}{p_\beta(y|\hat{x})} \,,$$

and $C_Y^{\beta,\hat{x}}$ is a $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix defined by

$$C_Y^{\beta,\hat{x}}[y,y'] = \sum_x \frac{p(y|x)p_\beta(x|\hat{x})p(y'|x)}{p_\beta(y|\hat{x})} \,.$$

For brevity, we simplify the notation by omitting the explicit reference to β and $\hat{x}$ when their actual values are implied or can be arbitrary. It follows that under our assumptions, a necessary (approximated) condition for a second-order phase transition that involves $\hat{x}$ is that β^{-1} is an eigenvalue of $C_X^{\beta,\hat{x}}$ and $C_Y^{\beta,\hat{x}}$. We note that the condition on C_X is closely related to the bifurcation analysis of [17]. Next, we show that both C_X and C_Y are stochastic matrices with the same non-zero eigenvalues.

Proposition 3. *C_Y and C_X have the same non-zero eigenvalues, and their largest eigenvalue is always 1 with $\mathbf{1}$ as an eigenvector.*

Proof. The first part follows from the fact that for any two $m \times n$ real matrices, A and B, it

holds that $AB^\top$ and $A^\top B$ have the same eigenvalues. For any given $\beta \geq 0$ and $\hat{x} \in \hat{\mathcal{X}}$, let

$$\begin{aligned} A[x,y] &= p(y|x) \\ B[x,y] &= \frac{p(y|x)p_\beta(x|\hat{x})}{p_\beta(y|\hat{x})} \,. \end{aligned}$$

It is easy to verify that $C_X = AB^\top$ and $C_Y = B^\top A$. Next, we will show that C_X and C_Y are stochastic matrices. All the values in these matrices are clearly positive, and so it remains to show that the rows sum up to 1. Notice that $B[x,y] = p_\beta(x|\hat{x},y)$, and thus

$$\begin{aligned} \sum_{x'} C_X[x,x'] &= \sum_{x'}\sum_{y} p(y|x)p_\beta(x'|\hat{x},y) = 1 \\ \sum_{y'} C_Y[y,y'] &= \sum_{y'}\sum_{x} p(y'|x)p_\beta(x|\hat{x},y) = 1 \,. \end{aligned}$$

It follows from the Perron–Frobenius Theorem that for both C_X and C_Y, the largest eigenvalue is always 1 with eigenvector $\mathbf{1}$. □

An immediate conclusion from Proposition 3 is that it is sufficient to find the eigenvalues only for the lower dimensional matrix, which is typically C_Y. Furthermore, this criticality condition becomes particularly simple when Y is binary.

Corollary 2. *Assume $|\mathcal{Y}| = 2$, then a necessary condition for a phase transition at β is that there is some $\hat{x} \in \hat{\mathcal{X}}$ for which $\beta = \det(C_Y^{\beta,\hat{x}})^{-1}$.*

Proof. For 2×2 stochastic matrices, the first eigenvalue is $\lambda_1 = 1$ and the second eigenvalue λ_2 is given by the determinant. Therefore, for a binary Y it holds that $\lambda_2(\hat{x}) = \det(C_Y^{\beta,\hat{x}})$, which implies that a necessary condition for (13) is $\beta = \det(C_Y^{\beta,\hat{x}})^{-1}$. □

Another conclusion from Proposition 3 is that the criticality condition cannot hold for $\beta < 1$, because the largest eigenvalue is always 1. This is consistent with the fact that the first critical point β_{c_0} is necessarily greater or equal than 1 (see Section 2.3). For $1 \leq \beta \leq \beta_{c_0}$, any trivial representation for which $p(x|\hat{x}) = p(x)$ and $p(y|\hat{x}) = p(y)$ is optimal, yielding $C_Y^0[y,y'] = \sum_x p(x|y)p(y'|x)$ which is independent of β and $\hat{x}$. Therefore, finding β_{c_0} amounts to finding the eigendecomposition of C_Y^0. For $\beta > \beta_{c_0}$, $C_Y^{\beta,\hat{x}}$ may vary with β resulting in the self-consistent condition $\beta^{-1} \in \mathrm{Eig}(C_Y^{\beta,\hat{x}})$ for criticality, where $\mathrm{Eig}(C_Y^{\beta,\hat{x}})$ is the set of eigenvalues of $C_Y^{\beta,\hat{x}}$. Therefore, finding critical points after β_{c_0} is no longer a simple eigendecomposition problem. To address this problem, we propose the Criticality Search algorithm.

3.2.1 Criticality Search

Criticality Search (Algorithm 3) is an iterative algorithm that finds candidate values of β that satisfy the self-consistent criticality condition. It starts with an initial guess $\beta_c(\hat{x}) = \beta_0$, computes the eigenvalues of C_Y (assuming Y is the lower-dimensional variable), and then checks

the criticality condition. If the condition is not met, the algorithm picks another candidate by making an educated guess:

$$\beta_c^{new}(\hat{x}) = \min\{\lambda^{-1}(\hat{x}) : \lambda \in \mathrm{Eig}(C_Y^{\beta_c,\hat{x}}), \lambda \neq 1\}\,. \tag{14}$$

When Y is binary, this guess simply becomes $\beta_c^{new}(\hat{x}) = \det(C_Y^{\beta,\hat{x}})^{-1}$. This process is repeated for each $\hat{x}$ until a point $\beta_c(\hat{x})$ that satisfies the condition is found, or when β is large enough such that a maximally-informative point is reached, i.e. when $I_\beta(\hat{X};Y) = I(X;Y)$.

The algorithm is demonstrated by the simulations of Figure 1. The red points in all panels correspond to the two critical points found by the algorithm. It can be seen that these points correspond to the structural phase transitions observed in the centroid bifurcation diagram (Figure 1D) and in the informational bifurcation diagrams (Figure 1B-C). The iterations of the algorithms are demonstrated in Figure 1F. This figure shows a run that was initialized with β_0 slightly larger than the first critical point. It converged to the second critical point for $\hat{x}_2$ by iterating between the red curve, which corresponds to $\lambda_2^{-1}(\hat{x})$, and the black curve, which corresponds to β. The fixed points of this iterated map are precisely the points in which the criticality condition is met. While our criticality condition only approximates a necessary condition for a phase transition, in all our numerical simulations the algorithm converged to actual critical points. This suggests that the condition we derived is a good approximation. In addition, we conjecture that while it is possible that the condition is met at non-critical points, these points might be unstable fixed points of the algorithm.

3.3 The deterministic case

To complete our characterization of the IB phase transitions, we discuss the special case in which Y is a deterministic function of X. This case exhibits qualitatively different behavior compared to cases in which $p(y|x) > 0$ for all x and y, and has recently been explored in the context of deep learning [28].

First, we argue that we can consider without loss of generality the case in which $p(y|x)$ is deterministic and defines a one-to-one mapping from X to Y. That is, for every x there is a unique value $y(x)$ such that $p(y'|x) = \delta_{y',y(x)}$. Otherwise, if there exist x_1, x_2 such that $y(x_1) = y(x_2)$, we can replace both of them by a single value $x_{1,2}$ such that $y(x_{1,2}) = y(x_1)$ and $p(x_{1,2}) = p(x_1) + p(x_2)$. This does not change the structure of the problem, that is, the IB clustering problem discussed in Section 3.1.1 remains the same. This also implies that we may assume without loss of generality that $|\mathcal{X}| = |\mathcal{Y}|$.

In this case, $I(X;\hat{X}) = I(\hat{X};Y)$ and the IB objective function becomes

$$\mathcal{F}_\beta[p] = (1-\beta)I_p(\hat{X};Y)\,.$$

There are three different regimes for β in this case: (i) when $\beta < 1$, the solution is the same

Algorithm 3: Criticality Search

Input: $p(x, y)$, initial $p_0(\hat{x}|x)$, and β_0
Output: Candidate critical points
for $\hat{x} \in \hat{\mathcal{X}}$ **do**
  $p(\hat{x}|x) \leftarrow p_0(\hat{x}|x)$ (initialize)
  $\beta_c(\hat{x}) \leftarrow \beta_0$
  $\lambda_c \leftarrow 0$
  while $\beta_c(\hat{x}) \neq \lambda_c^{-1}$ **do**
    $p(\hat{x}|x) \leftarrow \mathrm{IB}\left(p(x, y), p(\hat{x}|x), \beta_c(\hat{x})\right)$ (update encoder)
    $C_Y \leftarrow B^\top A$ (update C_Y)
    $U, D = \mathrm{EVD}(C_Y)$ (eigendecomposition of C_Y)
    $L \leftarrow \{\lambda_i : \lambda_i = D_{ii}, \forall i = 1, \ldots, n\} \setminus \{1\}$
    if $\exists \lambda \in L$ *such that* $\beta_c(\hat{x}) = \lambda^{-1}$ **then**
      $\lambda_c \leftarrow \lambda$ (found a candidate for $\hat{x}$)
    else if $I_p(\hat{X}; Y) = I(X; Y)$ **then**
      $\beta_c(\hat{x}) \leftarrow \infty$ (no candidates were found for $\hat{x}$)
      continue
    else
      $\beta_c(\hat{x}) \leftarrow \min_{\lambda \in L} \lambda^{-1}$ (educated guess for the next iteration)
return $\beta_c(\hat{x})$, $\forall t \in \hat{\mathcal{X}}$

as in the general case, i.e. it is the trivial solution for which $I(\hat{X}; Y) = 0$; (ii) when $\beta = 1$, $\mathcal{F}_\beta[p] = 0$ for all $p(\hat{x}|x)$, which means that any representation $p(\hat{x}|x)$ would be equally good; (iii) when $\beta > 1$, minimizing $\mathcal{F}_\beta[p]$ becomes equivalent to maximizing $I_p(\hat{X}; Y)$. The solution in this regime is equivalent to the solution when $\beta \to \infty$, and so the optimal representation would be a deterministic mapping from X to $\hat{X}$. Therefore, in this regime, the parameter that shapes the optimal representations is the hard constraint on $|\hat{\mathcal{X}}|$ rather than β.

Because β^{-1} is the slope of the information curve [20], the curve in the deterministic case is linear with slope 1 (or piecewise linear, as noted also in [28], if we relax the assumption that $y(x)$ is a bijective function, in which case the curve becomes flat once $H(Y)$ is reached). We identify, contra to [28], a sequence of structural phase transitions along this line, which are characterized by the solutions to the following optimization problems for $K = 1, \ldots, |\mathcal{X}|$:

$$\begin{aligned} \max_{p} \quad & I_p(\hat{X}; Y) \\ \text{such that} \quad & Supp(p) \leq K \,. \end{aligned}$$

These problems are NP-Hard, although in some cases (e.g., $K = |\hat{\mathcal{X}}|$) they are tractable.

4 Numerical examples

In this section we explore numerically (a) the types of structural phase transitions that may occur in IB; (b) related phenomena such as critical slowing down; and (c) the influence of $p(x, y)$ on the evolutionary trajectory of the representations. We do so by considering several numerical examples that are designed to be as simple as possible and at the same time convey important insight about the evolutionary trajectory of the IB representations.

4.1 Sensitivity to the source distribution

We begin by exploring the influence of the source distribution $p(x)$ on the evolution of the representations. To this end, we fix $p(y|x)$ and vary only $p(x)$. We take $Y \in \{0, 1\}$ and trinary X and $\hat{X}$. We define $p(y|x)$ by $p(y = 1|x_1) = 0.25$, $p(y = 1|x_2) = 0.48$, and $p(y = 1|x_3) = 0.75$. The choice of $p(y|x_2)$ is deliberately meant to break the symmetry in this example. The symmetric case will be explored in the next section. We consider four joint distributions defined by $p(y|x)$ and the following source distributions:

$$
\begin{aligned}
p_1(x) &= \begin{pmatrix} 0.45 & 0.1 & 0.45 \end{pmatrix} \\
p_2(x) &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \\
p_3(x) &= \begin{pmatrix} 0.18 & 0.64 & 0.18 \end{pmatrix} \\
p_4(x) &= \begin{pmatrix} 0.1 & 0.8 & 0.1 \end{pmatrix} .
\end{aligned}
$$

For each joint distribution $p_i(x, y) = p_i(x)p(y|x)$, we evaluated the evolutionary trajectory of the IB representations via Algorithm 2, the corresponding centroid bifurcation diagram, and the evolution of the second eigenvalue of $C_Y^{\beta,\hat{x}}$ for all $\hat{x}$. The results are shown in Figure 2. It can be seen that in all four cases there are two critical points. At these points, the effective cardinality increases, which is reflected in the emergence of a new distinct value in the centroid bifurcation diagrams (Figure 2A). Note that the effective cardinality corresponds to $Supp(p_\beta)$ only when the representation is canonical. For p_3 and p_4, all the representations found by Algorithm 2 are canonical, and therefore at the critical points $Supp(p_\beta)$ changes. This can be seen in Figure 2B, where $p_\beta(\hat{x})$ becomes positive for some $\hat{x}$.

Figure 2C shows that, as expected based on the theoretical analysis of Section 3, $\lambda_2(\hat{x})$ coincides with β^{-1} at critical points in which centroids splits continuously (e.g., the first phase transition for p_1). Interestingly, Figure 2A reveals that not all phase transitions correspond to continuous centroid splits (e.g., the second phase transition for p_1). However, even in these discontinuous cases $\lambda_2(\hat{x})$ seems to be indicative of the phase transition because it tends to reache β^{-1} at those critical points.

Finally, we observe a *critical slowing down* phenomenon near the phase transitions, in which the convergence time of the IB iterations diverges (Figure 2D). This phenomena has

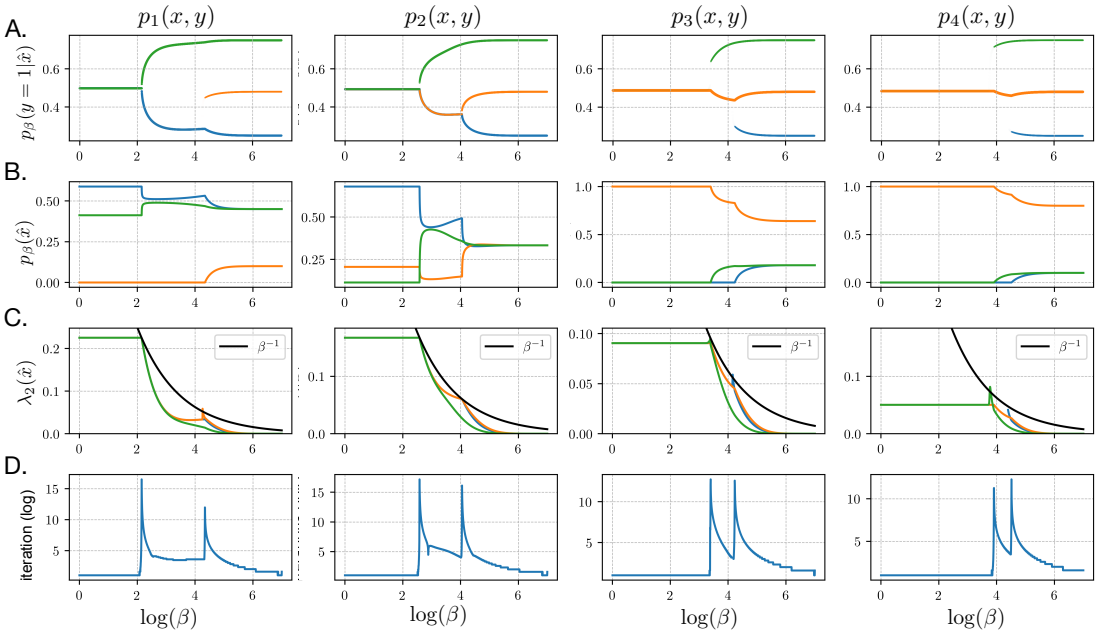

Figure 2. Numerical simulations with *asymmetric* distributions. The i-th column corresponds to the set of results for $p_i(x, y)$. Colored curves (blue, orange, green) in panels A-C correspond to different values of $\hat{x}$. A. Centroid bifurcation diagrams. B. $p_\beta(\hat{x})$ as a function of $\log(\beta)$. C. The evolution of the second eigenvalue $\lambda_2(\hat{x})$ as a function of $\log(\beta)$. D. Log convergence time of Algorithm 1, i.e., the number of IB iterations, as a function of $\log(\beta)$.

been known to happen near phase transitions in other settings [29, 30], and further analysis of this phenomena in the case of IB is left to future research.

This numerical exploration shows that the source distribution may have substantial influence on the location of the critical points, as well as their type. For example, bifurcations that appear as continuous splits, similar to pitchfork bifurcations, may change to what appears as a discontinuous emergence of a new centroid. In addition, our simulations suggest that the IB phase transitions may also be characterized by critical slowing down, in addition to the characterization of Section 3.

4.2 Symmetric distributions

Next, we repeat the same analysis with symmetric distributions. We constructed these distributions by taking the four asymmetric distributions from before and changing $p(y = 1|x_2) = 0.5$. Figure 3 shows the results in this case. Not surprisingly, the bifurcation diagrams are symmetric for these distributions (Figure 3A). In addition, these examples demonstrate that $p(x)$ may influence not only the type of bifurcations but also their number. For p_1 and p_2 there are two critical points, as before, however for p_3 and p_4 there is only one critical point. Furthermore, for p_3 and p_4 we observe a trinary split in which the effective cardinality jumps from $k = 1$ to $k = 3$. This appears to happen either via a continuous split (as in p_3) or via a discontinuous

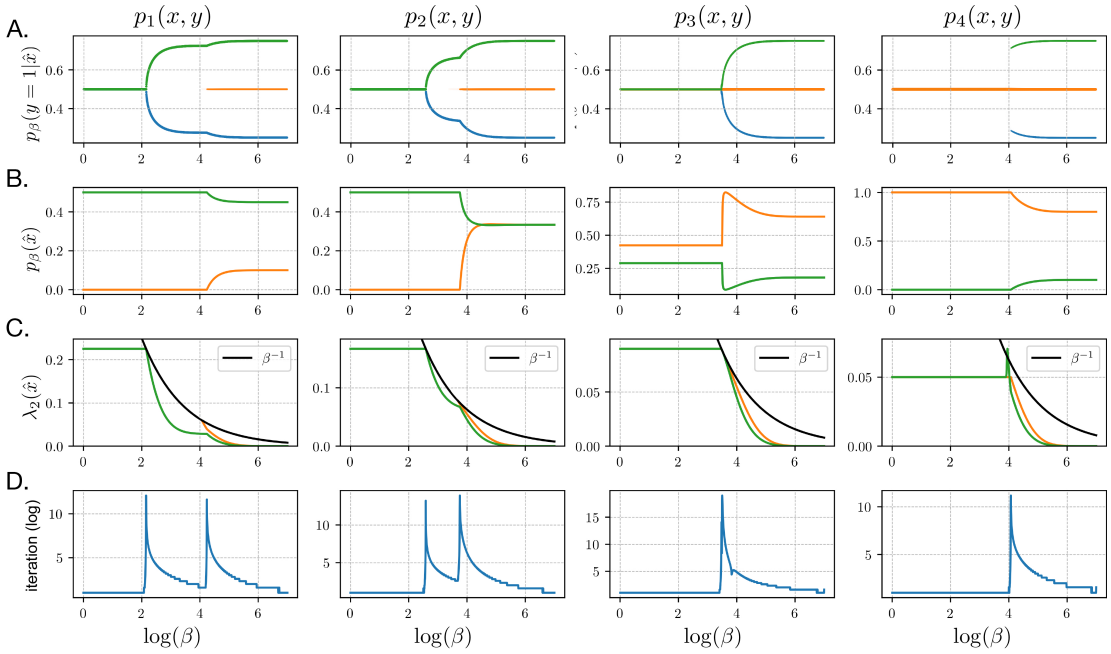

Figure 3. Numerical simulations with *symmetric* distributions. The i-th column corresponds to the set of results for $p_i(x, y)$. Colored curves (blue, orange, green) in panels A-C correspond to different values of $\hat{x}$. In some cases the blue and green curves overlap. A. Centroid bifurcation diagrams. B. $p_\beta(\hat{x})$ as a function of $\log(\beta)$. C. The evolution of the second eigenvalue $\lambda_2(\hat{x})$ as a function of $\log(\beta)$. D. Log convergence time of Algorithm 1, i.e., the number of IB iterations, as a function of $\log(\beta)$.

emergence of a new value (as in p_4). In the continuous case, which corresponds to the assumptions of our criticality condition, $\lambda_2(\hat{x}) = \beta_c^{-1}$ for all three clusters at the same critical point (intersection of the colored curves with the black curve in Figure 3C, p_3). This behavior is less clear is the discontinuous case (Figure 3C, p_4). In both cases, however, we observe critical slowing down near the phase transition (Figure 3D).

4.3 Water filling in Bayesian networks

In this final example, we extend our analysis to the multivariate case and illustrates a potential application of our approach to design principles for neural network architectures. Specifically, we use the methods of Section 3, but instead of the standard IB method we apply its multivariate extension [31, Multivariate IB (MVIB)]. MVIB takes the multi-information, which is defined for a set of random variables $\mathbf{Z} = (Z_1, \ldots, Z_n) \sim p(z_1, \ldots, z_n)$ by

$$\mathcal{I}(\mathbf{Z}) = D\left[p(z_1, \ldots, z_n) \middle\| \prod_{i=1}^{n} p_i(z_i)\right], \tag{15}$$

as a natural extension of mutual information in the multivariate case. The MVIB objective function is then $\mathcal{I}(\mathbf{X}, \hat{\mathbf{X}}) - \beta\mathcal{I}(\hat{\mathbf{X}}, \mathbf{Y})$, where $\mathbf{X}$, $\mathbf{Y}$ and $\hat{\mathbf{X}}$ are multivariate variables and the statistical dependencies between them are defined by a Bayesian network.

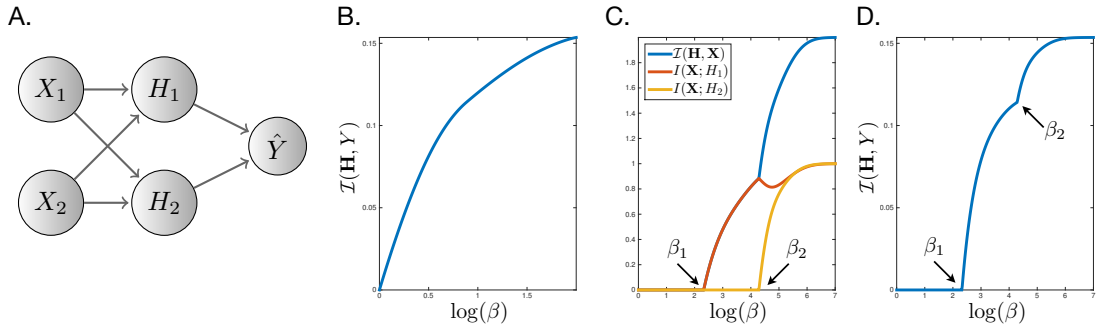

Figure 4. Numerical simulations in the multivariate case. A. The Bayesian network used in our simulations. B. The multivariate information curve. C. Information that the hidden representation maintains about the input. Note that for every β it holds that $\mathcal{I}(\mathbf{X},\hat{\mathbf{X}}) = I(\mathbf{X};H_1) + I(\mathbf{X};H_2)$. D. Information about the ground truth Y extracted by the hidden representation.

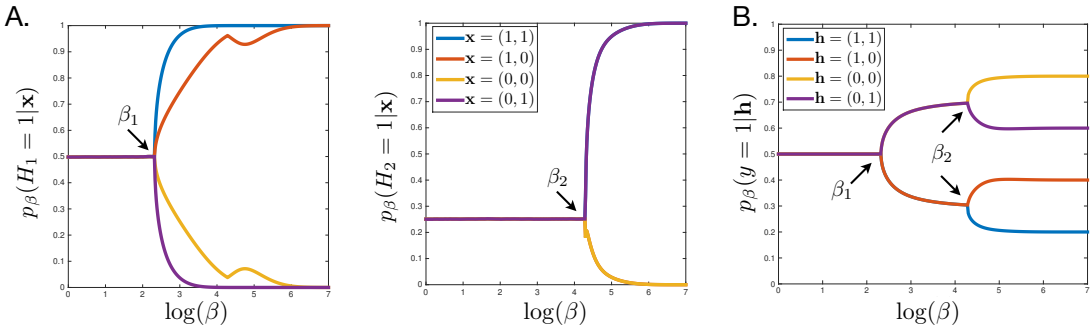

Figure 5. Evolution of the hidden representation. A. Bifurcation diagrams for the H_1 encoder (left) and H_2 encoder (right). B. Centroid bifurcation diagram.

To demonstrate our approach numerically, we consider the Bayesian network of Figure 4A, where $\mathbf{X} = (X_1, X_2)$ is the input, $\hat{\mathbf{X}} = \mathbf{H} = (H_1, H_2)$ is the hidden layer of the network, and $\hat{Y}$ is the network's prediction defined such that $p(\hat{Y} = y|\hat{\mathbf{x}}) = p(Y = y|\hat{\mathbf{x}})$. For simplicity, we assume that all variables — X_1, X_2, H_1, H_2, Y, and $\hat{Y}$ — are binary. We take $p(\mathbf{x})$ to be uniform, and define $p(y|\mathbf{x})$ by $p(y = 1|\mathbf{x} = (0,0)) = 0.8$, $p(y = 1|\mathbf{x} = (0,1)) = 0.6$, $p(y = 1|\mathbf{x} = (1,0)) = 04$, and $p(y = 1|\mathbf{x} = (1,1)) = 0.2$.

Figure 4 shows the multivariate information curve for this example, and the information that the hidden representation maintains about the input $\mathbf{X}$ and the desired output (or ground truth) Y. It is easy to verify that in this case $\mathcal{I}(\mathbf{X},\mathbf{H}) = I(\mathbf{X};H_1) + I(\mathbf{X};H_2)$. Figure 5 gives a more detailed view of the evolution of the hidden representation as a function of β.

These result demonstrate a water filling phenomenon for the hidden units of the networks, analogous to the water-filling phenomena in rate–distortion theory [23]. When $\beta < \beta_1$, both hidden units are independent of the input (Figure 5A), and do not maintain any information about $\mathbf{X}$ or Y (Figure 4C-D). The prediction of the network (Figure 5B) in this regime is based on the prior $p(y)$, which is uniform in this case. This means that the canonical hidden representation is constant, and thus both hidden units are redundant. When $\beta_1 < \beta < \beta_2$, only H_1 keeps

information about the input and output. In this case H_2 is redundant and can be eliminated from the network. When $\beta > \beta_2$, both units are informative, and their contribution is complementary. Namely, H_1 evolves to represent X_1 and H_2 evolves to represent X_2. Therefore, in this regime both units are necessary for the optimal architecture uses both of them.

5 Conclusions

In this work, we have cast the notion of efficient compressed representations in terms of IB, and characterized how these efficient representations evolve via a deterministic annealing process. The main contributions of this work are: (1) introduction of order parameters that capture the evolution of the IB representations and the structural phase transitions that they undergo; (2) derivation of an algorithm for finding critical points; and (3) numerical exploration of the phase transitions and related phenomena that occur in IB. Important directions for future research include an extension of our analysis to continuous variables; characterization of the critical slowing-down phenomenon in IB, and possibly methods for overcoming the computational problem this phenomenon raises. In addition, while the examples we considered here are merely illustrative, they demonstrate general principles that may apply to several fields. For example, some of these methods have already been applied to language evolution [12] and deep neural networks [2, 4]. This work lays out some of the theoretical grounds for extending these applications, as well as applying this approach more broadly.

Acknowledgments

This study was supported by the Gatsby Charitable Foundation. Part of this work was done while the authors were visiting the Simons Institute for the Theory of Computing.

References

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[2] N. Tishby and N. Zaslavsky, "Deep learning and the Information Bottleneck principle," in *IEEE Information Theory Workshop (ITW)*, 2015.

[3] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational Information Bottleneck," in *ICLR*, 2017.

[4] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[5] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *Journal of Machine Learning Research*, vol. 19, no. 50, pp. 1–34, 2018.

[6] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the Information Bottleneck," *Theoretical Computer Science*, vol. 411, no. 29-30, pp. 2696–2711, 2010.

[7] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information Bottleneck for Gaussian variables," *Journal of Machine Learning Research*, vol. 6, 2005.

[8] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, "Predictive information in a sensory population," *Proceedings of the National Academy of Sciences*, vol. 112, no. 22, pp. 6908–6913, 2015.

[9] J. Rubin, N. Ulanovsky, I. Nelken, and N. Tishby, "The representation of prediction error in auditory cortex," *PLOS Computational Biology*, vol. 12, no. 8, pp. 1–28, 2016.

[10] S. Wang, A. Borst, N. Zaslavsky, N. Tishby, and I. Segev, "Efficient encoding of motion is mediated by gap junctions in the fly visual system," *PLOS Computational Biology*, vol. 13, no. 12, pp. 1–22, 2017.

[11] N. Slonim and N. Tishby, "The power of word clusters for text classification," in *23rd European Colloquium on Information Retrieval Research*, 2001.

[12] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby, "Efficient compression in color naming and its evolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 31, pp. 7937–7942, 2018.

[13] N. Jacoby, N. Tishby, and D. Tymoczko, "An information theoretic approach to chord categorization and functional harmony," *Journal of New Music Research*, vol. 44, no. 3, pp. 219–244, 2015.

[14] M. Rey and V. Roth, "Meta-Gaussian Information Bottleneck," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1916–1924.

[15] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Phys. Rev. Lett.*, vol. 65, pp. 945–948, Aug 1990.

[16] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," in *Proceedings of the IEEE*, 1998, pp. 2210–2239.

[17] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The mathematical structure of information bottleneck methods," *Entropy*, vol. 14, no. 3, pp. 456–479, 2012.

[18] F. Pereira, N. Tishby, and L. Lee, "Distributional clustering of English words," in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 1993, pp. 183–190.

[19] N. Slonim and N. Tishby, "Document clustering using word clusters via the Information Bottleneck method," in *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, 2000, pp. 208–215.

[20] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proceedings of the 16th Annual Conference on Learning Theory (COLT)*, 2003.

[21] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[22] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[23] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley New York, 1991.

[24] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, 1948.

[25] G. Elidan and N. Friedman, "Learning hidden variable networks: The Information Bottleneck approach," *Journal of Machine Learning Research*, vol. 6, pp. 81–127, 2005.

[26] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview press, 1994.

[27] G. Jaeger, "The ehrenfest classification of phase transitions: Introduction and evolution," *Archive for history of exact sciences*, vol. 53, no. 1, pp. 51–81, 1998.

[28] A. Kolchinsky, B. D. Tracey, and S. V. Kuyk, "Caveats for Information Bottleneck in deterministic scenarios," in *International Conference on Learning Representations*, 2019.

[29] D. S. Fisher, "Scaling and critical slowing down in random-field Ising systems," *Physical Review Letter*, vol. 56, pp. 416–419, 1986.

[30] A. A. Middleton, "Critical slowing down in polynomial time algorithms," *Physical Review Letter*, vol. 88, p. 017202, 2001.

[31] N. Slonim, N. Friedman, and N. Tishby, "Multivariate Information Bottleneck," *Neural Computation*, vol. 18, no. 8, pp. 1739–1789, 2006.

Appendix

In this section we prove several technical lemmas that were used in our main analysis.

Lemma 1. *Let $Y - X - \hat{X}$ be a Markov chain such that $p(y, x, \hat{x}) = p(x, y)p(\hat{x}|x)$, and let $p(y|\hat{x})$ be the corresponding conditional distribution of Y given $\hat{X}$. Then*

$$\mathop{\mathbb{E}}_{x,\hat{x}} \left[D[p(y|x) \| p(y|\hat{x})] \right] = I(X;Y) - I(\hat{X};Y) \,.$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,\hat{x}} \left[D[p(y|x) \| p(y|\hat{x})] \right] &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x) \left[D[p(y|x) \| p(y|\hat{x})] \right] \\
&= \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|x)p(y)}{p(y|\hat{x})p(y)} \\
&= \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|x)}{p(y)} - \sum_{x,\hat{x},y} p(x)p(\hat{x}|x) \log \frac{p(y|\hat{x})}{p(y)} \\
&= I(X;Y) - I(\hat{X};Y) \,.
\end{aligned}$$

□

Lemma 2. $\frac{\partial}{\partial \beta} \mathbb{E}_x \left[\log Z_\beta(x) \right] = I_\beta(\hat{X};Y) - I(X;Y)$.

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \beta} \mathop{\mathbb{E}}_{x} \left[\log Z_\beta(x) \right] &= \sum_x p(x) \frac{\partial}{\partial \beta} \log Z_\beta(x) \\
&= \sum_x p(x) \frac{1}{Z_\beta(x)} \frac{\partial}{\partial \beta} \left(\sum_{\hat{x}} p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]} \right) \\
&= \sum_{x,\hat{x}} p(x) \frac{p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]}}{Z_\beta(x)} \left(\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) - \beta \frac{\partial}{\partial \beta} D[p(y|x) \| p_\beta(y|\hat{x})] \right) \\
&\quad - \sum_{x,\hat{x}} p(x) \frac{p_\beta(\hat{x}) e^{-\beta D[p(y|x) \| p_\beta(y|\hat{x})]}}{D[p(y|x) \| p_\beta(y|\hat{x})]} \\
&= \sum_{x,\hat{x}} p(x) p_\beta(\hat{x}|x) \left(\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) - \beta \frac{\partial}{\partial \beta} D[p(y|x) \| p_\beta(y|\hat{x})] \right) \\
&\quad - \sum_{x,\hat{x}} p(x) p_\beta(\hat{x}|x) D[p(y|x) \| p_\beta(y|\hat{x})] \,.
\end{aligned}$$

The first term is zero (assuming $p_\beta(\hat{x})$ is differentiable w.r.t. β) because

$$\mathop{\mathbb{E}}_{x,\hat{x}} \left[\frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) \right] = \sum_{\hat{x}} p_\beta(\hat{x}) \frac{\partial}{\partial \beta} \log p_\beta(\hat{x}) = \sum_{\hat{x}} \frac{\partial}{\partial \beta} p_\beta(\hat{x}) = 0 \,,$$

and so is the second term, for similar reasons:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,\hat{x}}\left[\frac{\partial}{\partial\beta}D[p(y|x)\|p_\beta(y|\hat{x})]\right] &= -\sum_{x,\hat{x},y} p(x)p_\beta(\hat{x}|x)p(y|x)\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) \\
&= -\sum_{\hat{x}} p_\beta(\hat{x})\sum_y\left(\sum_x p_\beta(x|\hat{x})p(y|x)\right)\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) \\
&= -\sum_{\hat{x}} p_\beta(\hat{x})\sum_y p_\beta(y|\hat{x})\frac{\partial}{\partial\beta}\log p_\beta(y|\hat{x}) = 0\,.
\end{aligned}$$

It follows that

$$\frac{\partial}{\partial\beta}\mathop{\mathbb{E}}_{x}\left[\log Z_\beta(x)\right] = -\mathop{\mathbb{E}}_{x,\hat{x}}\left[D[p(y|x)\|p_\beta(y|\hat{x})]\right]\,,$$

and applying Lemma 1 to the right hand side of this equation concludes the proof.

□

Lemma 3. *Let p_β be a canonical IB representation and $\hat{x} \in Supp(p_\beta)$, then*

$$I_X^\beta(\hat{x}) = \beta I_Y^\beta(\hat{x}) - \sum_x p_\beta(x|\hat{x})\left[\log Z_\beta(x) + \beta I_Y(x)\right]\,, \tag{16}$$

where $I_Y(x) = D[p(y|x)\|p(y)]$.

Proof. This follows from substituting (2) in the definition of $I_X^\beta(\hat{x})$, i.e.

$$\begin{aligned}
I_X^\beta(\hat{x}) &= \sum_x p_\beta(x|\hat{x})\left(-\beta D[p(y|x)\|p_\beta(y|\hat{x})] - \log Z_\beta(x)\right) \\
&= \sum_x p_\beta(x|\hat{x})\left[\beta\left(\sum_y p(y|x)\log\frac{p_\beta(y|\hat{x})}{p(y)} - I_Y(x)\right) - \log Z_\beta(x)\right] \\
&= \beta\sum_{x,y} p_\beta(x|\hat{x})p(y|x)\log\frac{p_\beta(y|\hat{x})}{p(y)} - \sum_x p_\beta(x|\hat{x})\left[\beta I_Y(x) + \log Z_\beta(x)\right] \\
&= \beta I_Y^\beta(\hat{x}) - \sum_x p_\beta(x|\hat{x})\left[\beta I_Y(x) + \log Z_\beta(x)\right]\,.
\end{aligned}$$

□

Lemma 4. *Let p_β be a canonical IB representation and $\hat{x} \in Supp(p_\beta)$, then*

$$\frac{\partial}{\partial\beta}I_X^\beta(\hat{x}) = \beta\frac{\partial}{\partial\beta}I_Y^\beta(\hat{x}) + g_\beta(\hat{x})\,,$$

where

$$g_\beta(\hat{x}) = I_Y^\beta(\hat{x}) - \frac{\partial}{\partial\beta}\left(\sum_x p_\beta(x|\hat{x})\left[\log Z_\beta(x) + \beta I_Y(x)\right]\right)$$

and $\mathbb{E}_{\hat{x}}\left[g_\beta(\hat{x})\right] = 0$.

Proof. The first part follows from differentiating (16) with respect to β. For the second part, notice that Proposition 2 implies that

$$\mathop{\mathbb{E}}_{\hat{x}}\left[\frac{\partial}{\partial\beta}I_X^\beta(\hat{x})\right] = \beta \mathop{\mathbb{E}}_{\hat{x}}\left[\frac{\partial}{\partial\beta}I_Y^\beta(\hat{x})\right] ,$$

and therefore $\mathbb{E}_{\hat{x}}\left[g_\beta(\hat{x})\right] = 0$.

□

Chapter 8

General Discussion

This thesis has identified several fundamental information-theoretic principles that may underlie human semantic systems and their evolution. The first principle is the Information Bottleneck (IB, Tishby et al., 1999), that arises as the link between semantic representations and Shannon's Rate–Distortion theory (Shannon, 1948, 1959). We have shown that the IB principle characterizes human semantic systems by first testing its theoretical predictions on cross-linguistic data in the domain of color naming, and then generalizing this account to two qualitatively different semantic domains, household containers and animal taxonomies. Furthermore, we have tested the evolutionary predictions of the IB principle on recent diachronic color naming data from one language, providing the first direct evidence, to our knowledge, that color naming may evolve under pressure to maintain efficient coding schemes. These findings suggest that efficient coding under limited resources, as defined by IB, may be a major force in the evolution of semantic systems.

Two additional information-theoretic principles that have been identified in this thesis are the capacity-achieving principle and the maximum-entropy (MaxEnt) principle, which aid in characterizing the forces that may shape communicative need and the influence of need on semantic systems. We have used the capacity-achieving principle to reveal new evidence that communicative need may shape color naming in interaction with perception, as opposed to traditional accounts that focused mainly on perception and recent accounts that focused mainly on need. This principle was also used in Part I, within the IB framework, as a theoretically-motivated method for estimating communicative need. We have proposed the MaxEnt principle with corpus constraints as another principled domain-general method for estimating communicative need, which considers the influence of linguistic usage rather than capacity constraints or the statistics of the visual environment. Our systematic evaluation of these factors in the domain of color naming suggests that linguistic usage may be the most relevant factor for characterizing the communicative need.

In the third part of this thesis, we have extended the mathematical foundations of the IB framework, which lie at the basis of our approach to semantic systems. We have characterized the structural changes in the IB representations as they evolve via a deterministic annealing

process; derived an algorithm for finding critical points; and explored numerically the types of bifurcations and related phenomena that occur in IB. This set of analytical results is central to our proposal that semantic systems may evolve along the IB theoretical limit via an annealing process, and many of the phenomena studied in Part III have been observed at a larger scale in the model simulations in Part I. Particularly noteworthy in this context is our simulation of the evolution of the IB color naming systems (see Chapter 2, Figure 5 and Movie S1), which demonstrates how color naming may evolve in an annealing process, undergoing a sequence of structural phase transitions. Additionally, in Part III we have began to explore how communicative need — that is, $p(x)$ in the standard IB formulation — may influence the structure and evolution of the optimal IB systems in carefully designed domains. This resonates with our systematic evaluation of several need distributions in Chapter 6, and suggests that studying small synthetic domains may help to gain a better understanding of the relation between communicative need, efficient compression, and the structure of semantic categories.

While our motivation in this thesis is to account for linguistic phenomena, the analytical tools we have developed in Part III apply to the IB framework in general. Therefore, these tools may be useful also in other important application of IB, for instance in deep learning (Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017).

The theoretical framework laid out in this thesis opens several important avenues for future research. A few examples are outlined below.

- **Additional semantic domains.** The generality of the principles we invoke suggests that they may apply more broadly across semantic domains. Therefore, an important direction for future research is to further test the extent to which these principles apply to domains beyond those explored in this thesis. This includes further testing our approach to communicative need across semantic domains, as well as further testing the predictions of the IB principle. For the latter, one major challenge is that it is not always clear how to specify the underlying representation of the domain. In the case of color, for example, this specification was derived from a well-established perceptual color space. However, in other semantic domains, such as containers or objects more generally, it is not always clear how to define such perceptual or conceptual spaces. Ideally, we would like to find a principled and domain-general approach for estimating this underlying representation.

- **Language evolution and efficient coding via multi-agent interactions.** The evolutionary process we have proposed describes how the lexicon may change while remaining near the IB theoretical limit. However, it does not provide a detailed explanation as to how these changes may occur via the dynamics of multi-agent interactions. This calls for studying the emergence of near-optimal semantic systems in more realistic settings of human communication and leaning. For example, it has been shown that empirically observed cross-linguistic patterns in color naming may emerge through an iterated language learning process (Xu et al., 2013, and see also Carstensen et al., 2014), however it is un-

known how this dynamics may be related to the IB principle and the evolutionary process we derived from it. In addition, recent deep learning approaches to emergent communication (e.g., Foerster et al., 2016) may provide a useful infrastructure for exploring the influence of other dynamical processes on the efficiency of the emerged lexicon.

- **Efficient compression in language development.** Our results suggest that pressure for efficient compression may also be an important force during language development. That is, our theoretical framework, which has been tested thus far on cross-linguistic data, also predicts that children should acquire new words by following a developmental trajectory that is pressured to remain near the IB theoretical limit. We are currently exploring this direction and testing these predictions on developmental naming data.

- **Informing machines with human-like semantics.** Current state-of-the-art methods for learning semantic representations in machines are based on training neural network language models on extremely large amounts of text (e.g., Devlin et al., 2019; Radford et al., 2019). These models are not grounded in a cognitively-motivated representation of the environment, and it is not yet clear to what extent these models reflect human semantic representations (Wang et al., 2019). This thesis has identified computational principles that characterize human semantic systems, and could potentially guide the development of artificial intelligence with human-like semantics.

More broadly, this thesis draws two intriguing connections between semantics and other lines of research. First, ideas from statistical physics, such as annealing and phase transitions, lie at the core of our evolutionary account of color naming, and may potentially apply to the lexicon more generally. While similar ideas have previously been applied to other aspects of language, such as word frequencies (e.g., Ferrer i Cancho and Solé, 2003) and semantic hierarchies (Pereira et al., 1993), the application of ideas from statistical physics to cross-language semantic variation and the evolution of the semantic categories is novel to our knowledge. We hope that this thesis will inspire more work along these lines.

Second, while most applications of information theory to language have focused on data transmission over a noisy channel (Gibson et al., 2019), this thesis focuses on the complementing information-theoretic problem, that is, lossy data compression. Rate–Distortion theory, the branch of information theory that characterizes optimal lossy compression, and to which the IB principle is closely related, has recently been applied to several cognitive abilities, such as decision making (Tishby and Polani, 2011; Genewein et al., 2015; Polanía et al., 2019), curiosity-driven learning (Still and Precup, 2012), visual working memory (Sims et al., 2012), and perception (Marzen and DeDeo, 2017; Sims, 2018). This thesis extends this body of literature with applications of Rate–Distortion theory to high-level semantic representations. Therefore, we believe that Rate-Distortion theory may complement current noisy channel approaches to language, and furthermore, provide a general theoretical framework for studying the interaction between semantic representations and other aspects of human cognition.

Bibliography

A. Alemi, I. Fischer, J. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

R. Baddeley and D. Attewell. The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*, 20(9): 1100–1107, 2009.

C. Bentz. *Adaptive Languages. An Information-Theoretic Account of Linguistic Diversity*. De Gruyter Mouton, Berlin, Boston, 2018.

B. Berlin. *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton University Press, 1992.

B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley and Los Angeles, 1969.

J. M. Bernardo. Reference analysis. In D. Dey and C. Rao, editors, *Bayesian Thinking Modeling and Computation*, volume 25 of *Handbook of Statistics*, pages 17 – 90. Elsevier, 2005.

R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

C. H. Brown. General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist*, 3(3):400–424, 1976.

C. H. Brown. *Language and living things: Uniformities in folk classification and naming*. Rutgers University Press, 1984.

L. Buesing and W. Maass. A spiking neuron as Information Bottleneck. *Neural Computation*, 22(8):1961–1992, 2010.

A. Carstensen, J. Xu, C. T. Smith, and T. Regier. Language evolution in the lab tends toward informative communication. In P. Bello, M. Guarini, M. McShane, and B. Scassellati, editors, *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 303–308. Austin TX: Cognitive Science Society, 2014.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information Bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6, 2005.

B. Corominas-Murtra, J. Fortuny, and R. V. Solé. Towards a mathematical theory of meaningful communication. *Scientific Reports*, 4:4587, 2014.

T. Cover and J. Thomas. *Elements of Information Theory (2nd Edition)*. Wiley-Interscience, Hoboken, NJ, 2006.

I. Csiszár and P. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

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

R. Ferrer i Cancho and R. V. Solé. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

J. R. Firth. A synopsis of linguistic theory 1930-55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, volume 1952-59, pages 1–32. The Philological Society, Oxford, 1957.

J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. 2016.

M. Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: Lossy source-channel communication revisited. *IEEE Transactions on Information Theory*, 49(5):1147–1158, 2003.

T. Genewein, F. Leibfried, J. Grau-Moya, and D. A. Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27, 2015.

E. Gibson, S. T. Piantadosi, K. Brink, L. Bergen, E. Lim, and R. Saxe. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088, 2013.

E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017.

E. Gibson, R. Futrell, S. Piantadosi, I. Dautriche, K. Mahowald, L. Bergen, and R. Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019.

R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT)*, 2003.

P. Harremoës and N. Tishby. The Information Bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory*, pages 566–571, 2007.

Z. S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

R. M. Hecht and N. Tishby. Extraction of relevant speech features using the Information Bottleneck method. In *INTERSPEECH-2005*, pages 353–356, 2005.

N. Jacoby, N. Tishby, and D. Tymoczko. An information theoretic approach to chord categorization and functional harmony. *Journal of New Music Research*, 44(3):219–244, 2015.

T. F. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23 – 62, 2010.

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, 1982.

P. Kay and C. K. McDaniel. The linguistic significance of the meanings of basic color terms. *Language*, 54:610–646, 1978.

P. Kay, B. Berlin, L. Maffi, W. R. Merrifield, and R. Cook. *The World Color Survey*. Stanford: Center for the Study of Language and Information, 2009.

C. Kemp and T. Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012.

C. Kemp, Y. Xu, and T. Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), 2018.

S. Kirby, M. Tamariz, H. Cornish, and K. Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87 – 102, 2015.

R. P. Levy and T. F. Jaeger. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, pages 849–856. 2007.

D. T. Lindsey and A. M. Brown. The color lexicon of American English. *Journal of Vision*, 14 (2):17, 2014.

D. T. Lindsey, A. M. Brown, D. H. Brainard, and C. L. Apicella. Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, 25(18):2441–2446, 2015.

V. Loreto, A. Mukherjee, and F. Tria. On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences*, 109(18):6819–6824, 2012.

R. D. Luce. Whatever happened to information theory in psychology? *Review of general psychology*, 7(2):183, 2003.

S. E. Marzen and S. DeDeo. The evolution of lossy compression. *Journal of The Royal Society Interface*, 14(130), 2017.

G. A. Miller. The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7(3):141 – 144, 2003.

G. P. Murdock. *Social structure*. Macmillan, Oxford, England, 1949.

S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

F. Pereira. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253, 2000.

F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1993.

S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

J. B. Plotkin and M. A. Nowak. Language evolution and information theory. *Journal of Theoretical Biology*, 205(1):147–159, 2000.

R. Polanía, M. Woodford, and C. C. Ruff. Efficient coding of subjective value. *Nature Neuroscience*, 22(1):134–142, 2019.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

T. Regier, C. Kemp, and P. Kay. Word meanings across languages support efficient communication. In B. MacWhinney and W. O'Grady, editors, *The Handbook of Language Emergence*, pages 237–263. Wiley-Blackwell, Hoboken, NJ, 2015.

E. Rosch. Principles of categorization. In E. Margolis and S. Laurence, editors, *Concepts: Core Readings*, pages 189–206. MIT Press, 1999.

K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of the IEEE*, pages 2210–2239, 1998.

K. Rose, E. Gurewitz, and G. C. Fox. Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.*, 65:945–948, Aug 1990.

O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the Information Bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948.

C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

C. R. Sims. Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389):652–656, 2018.

C. R. Sims, R. A. Jacobs, and D. C. Knill. An ideal observer analysis of visual working memory. *Psychological review*, 119(4):807–30, 2012.

N. Slonim. *The Information Bottleneck: Theory and applications*. PhD thesis, Hebrew University of Jerusalem, 2002.

N. Slonim and N. Tishby. Document clustering using word clusters via the Information Bottleneck method. In *Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 208–215, 2000.

N. Slonim and N. Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*, 2001.

L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489, 2005.

S. Still and D. Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

N. Tishby and D. Polani. Information theory of decisions and actions. In V. Cutsuridis, A. Hussain, and J. G. Taylor, editors, *Perception-Action Cycle: Models, Architectures, and Hardware*, pages 601–636. Springer New York, New York, NY, 2011.

N. Tishby and N. Zaslavsky. Deep learning and the Information Bottleneck principle. In *IEEE Information Theory Workshop*, 2015.

N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

K. von Finter and L. Matthewson. Universals in semantics. *The Linguistic Review*, 25(1-2): 139–201, 2008.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

S. Wang, A. Borst, N. Zaslavsky, N. Tishby, and I. Segev. Efficient encoding of motion is mediated by gap junctions in the fly visual system. *PLOS Computational Biology*, 13(12): 1–22, 2017.

L. Wittgenstein. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953.

J. Xu, M. Dowman, and T. L. Griffiths. Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1758), 2013.

L. Zhaoping. Colour categories, colour constancy, and lightness perception from information theory. In *ECVP '07 Abstracts*, page 201, 2007.

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.