A note on complexity in efficient communication analyses of semantic typology

Francis Mollica (mollicaf@gmail.com)
University of Melbourne
Melbourne, VIC Australia

Abstract

Recently the principles of efficient communication have provided useful characterizations of semantic typology: the diversity of attested languages can be described by competing pressures for simplicity and informativeness. While this approach has achieved success in several semantic domains, the formalizations used to define complexity across domains vary. In this note, we list the conditions under which the two main approaches of defining complexity: channel rate and description length, unify and, thus, conclusions about near-optimal communicative efficiency generalize across formalizations. We illustrate this equivalence using simulations of communicative efficiency for Boolean concepts. We round out this note discussing the (un)importance of description languages and the limits on generalizing this equivalence for other behavioral targets for explanation.

Keywords: semantic typology; ideal learner; efficient communication; (algorithmic) information theory

Introduction

There is considerable diversity in how languages of the world carve up semantic domains. For example, a Mexican Spanish speaker would make a distinction between *celeste* (sky blue) and *azúl* (blue); whereas, an English speaker cannot make this distinction with a single basic color term. In the past decade, the diversity in semantic typology has been characterized as efficient communication (for review see, Kemp, Xu, & Regier, 2018). In evolving communicatively efficient linguistic systems, languages must trade-off a pressure to be simple yet successful for communication—i.e., informative. By virtue of this trade-off, there are several optimal communication systems that depend on the extent to which a language weights simplicity vs informativity. Attested linguistic systems are spread across the optimal frontier of efficient languages. Thus, the diversity in semantic typology is explained as efficient communication with each language differently weighting pressures for simplicity vs informativeness.

The success of the communicative efficiency account has been demonstrated in multiple semantic domains, with attested languages being near optimal with respect to this trade-off (Kemp & Regier, 2012; Y. Xu, Liu, & Regier, 2020; Zaslavsky, Kemp, Regier, & Tishby, 2018; Steinert-Threlkeld, 2021); however, the precise formalization of the simplicity-informativity trade-off has varied across analyses, (for review see, Mollica & Zaslavsky, in press), with some analyses operationalizing complexity in terms of description lengths (e.g., Kemp & Regier, 2012) and others adopting the well studied information-theoretic framework, Information Bottleneck (IB; Tishby, Pereira, & Bialek, 1999; Zaslavsky et al., 2018). This begs a question, do the results of one approach hold for the other approach? That is, are the optimal languages in the IB sense, optimal in the description length approach?

In this note, we present the conditions under which there is a sufficient equivalence for claims of near-optimality to hold between both description length and information-theoretic notions of the simplicity-informativeness trade-off. We note that the Information Bottleneck is a form of compression, a Rate-Distortion trade-off under Shannon (1948)'s information theory. We show that under reasonable assumptions and with few limitations, the description length approach is also a form of compression under Kolmogorov (1963)'s algorithmic information theory. Specifically, the description length complexity is equivalent to Kolomogrov's structure function, which in expectation has an asymptotic equivalence to the Information Bottleneck (Grunwald & Vitányi, 2004; Vereshchagin & Vitányi, 2004). The upshot is that the results of communicative efficiency analyses are preserved across both approaches. We will illustrate this using simulations of communicative efficiency for Boolean concepts. Additionally, we will address the un-importance of the choice of description language for communicative efficiency analyses and contrast it to the clear importance of description language in anthropological explanations (e.g., Goodenough, 1956), concept learning/generalization (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008) and ideal learning models (e.g., Mollica & Piantadosi, 2022).

Equivalence of Formalization

Let us start with the intuitions.

Information theory concerns the transmission of information across a channel. Under Shannon (1948)'s theory, communication is the the encoding of information from a source distribution into a code that is transmitted across a (possibly noisy) channel and decoded at a destination (Figure 1a; Shannon, 1948). Information is defined as the amount of surprise a source message conveys relative to other candidate messages[1]. Efficient communication balances the complexity of encodings–i.e., the complexity of word-meaning mappings and the vocabulary size, against the amount of communication error–i.e., the distortion between source and destination messages. Efficient communication is, thus, a com-

[1] As a result, the content and even the representation of source messages are irrelevant to Shannon's information theory.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

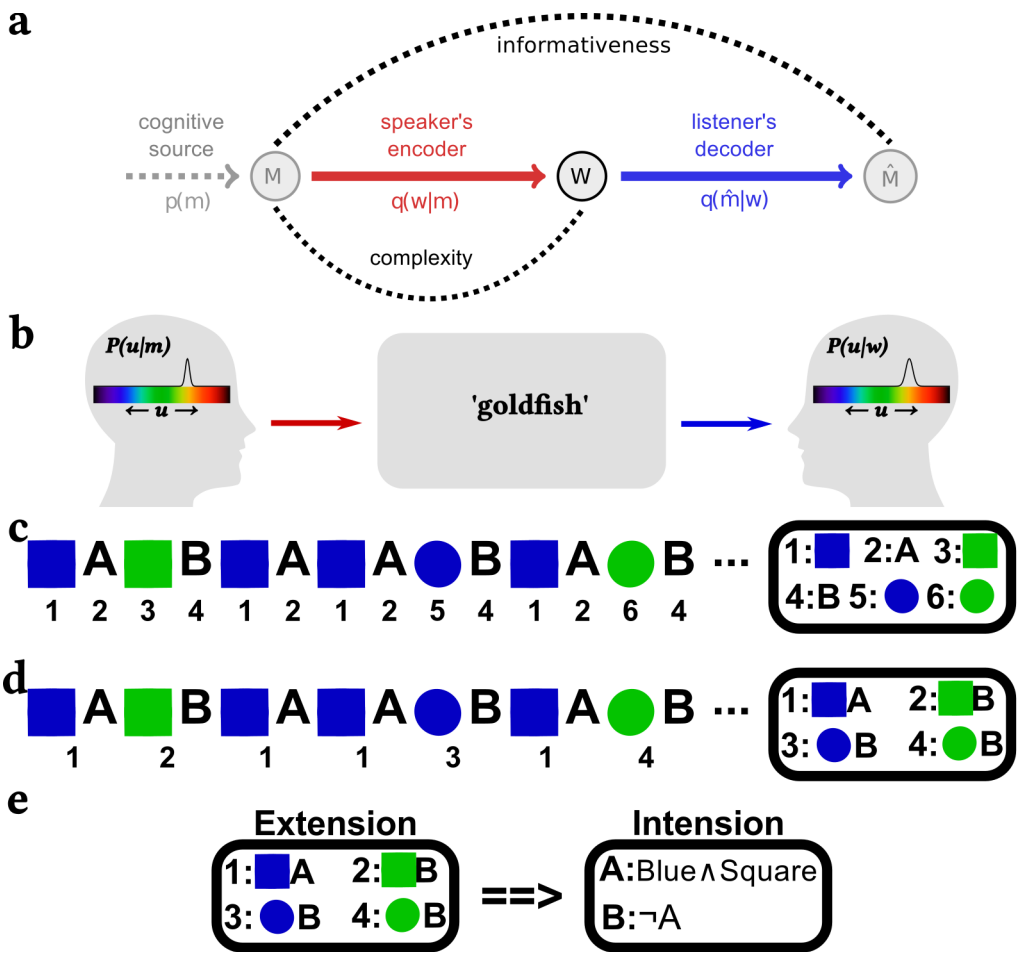

Figure 1: a) Shannon (1948)'s model of communication. Dashed lines emphasize the relationship between the model and the pressures for simplicity and informativeness. b) Shannon's model maps to human communication, where a speaker has source meanings (represented as probability distributions over the semantic domain) encoded into words, which allow listeners to decode an intended meaning. c) Consider an infinite sequence of pairs of meanings (colored shapes) and words (A or B). The description of this sequence is given by a model (in black box) that assigns a number to each *unique* element in the sequence, and an encoding of the sequence that indexes each element (number under the sequence). d) We can consider a different model that captures the regularities between word-meaning pairs and encode the sequence under this model. e) Contrast a model where indexing is determined by explicitly enumerating the sequence (e.g., word's extensions) with a model that generates indices following a description language (akin to a word's intension).

pression of source meanings onto words that minimizes the expected distortion with respect to the source distribution. As the source distribution is often chosen as the communicative need of the message, efficient communication is ecologically rational. Shannon's theory is a useful model of human communication (Figure 1b).

Algorithmic information theory concerns the encoding of information *within* a mathematical object. Imagine a source distribution and encoder generates an infinite sequence of meaning-word pairs. The algorithmic information, or Kolmogorov complexity, of this sequence is the length of the shortest program that generates exactly this sequence. As a worst case scenario, there is no way to describe the data that is more compact than simply listing the data itself in full (Figure 1c). However, if there are regularities in the sequence (e.g., a consistent relationship between words and meanings), then we can often find a program that generates the sequence without having to fully specify the sequence. As a trivial reduction in the program size, we can list all possible word-meaning pairs and then encode the sequence in terms of the index of those combinations (Figure 1d). If the set of all possible word-meaning pairs is small, this may be sufficient; however, if the set of word-meaning pairs are large and word-meaning mappings are non-arbitrary, we might be able to produce a shorter program that describes their generation–i.e., the intensional semantics (Figure 1e). Efficient encoding of the meaning-word sequence generated by a source and encoder is compression of the regularities between words and meanings to minimize both program length and encoding costs. Algorithmic information theory is a useful model of human learning. In fact, it has been applied directly to language acquisition (Chater & Vitányi, 2007; Hsu, Chater, & Vitányi, 2013) and underlies ideal learning models of cognitive development (Ullman & Tenenbaum, 2020).

Now, both approaches are interested in finding optimal encodings of the semantic domain that are simple and informative. Nonetheless, the target of optimization in both approaches is different. IB models optimize the encoding of word-meaning mappings, where meanings are fixed distributions over the semantic domain. In contrast, the algorithmic Rate-Distortion fixes word-meaning mappings as one-to-one and optimizes the encoding of sets of elements in the semantic domain, i.e., a word's extension, into meanings.

A more formal treatment might help clarify these issues: First, let's lay out the IB objective function. Then, we will outline the description length approach currently used in efficient communication approaches to semantic typology. Finally, we will demonstrate how a few assumptions can transform the description length approach to the Kolmogorov structure function.

Information Bottleneck. Efficient communication for semantic typology is primarily concerned with source-coding (Zaslavsky et al., 2018). That is, efficient languages should be simple codes that retain all the relevant information. As noted above, simplicity and informativeness will trade-off as simpler codes necessarily lose information. In the IB approach for semantic typology, speaker meanings m are encoded into words w which are recovered as listener meanings $\hat{m}$. Meanings, themselves, are probability distributions over elements u in a semantic domain. Complexity is measured as the amount of information shared between meanings and words $I(M;W)$. Informativeness is measured as the expected distortion between speaker and listener meanings, where distortion is defined as the KL Divergence $E[KL(M|\hat{M})]$. The IB objective function is:

$$\mathcal{F}_\beta[q(w|m)] = I(M;W) - \beta E[KL(M|\hat{M})], \tag{1}$$

where $q(w|m)$ is the encoder of meanings into words and β is the trade-off parameter determining the relative importance of simplicity vs informativeness.

One difference between the description length and the IB approach is that IB allows for a source meaning to map to multiple words; whereas, the description length approach forces a one-to-one mapping between words and meanings. We can integrate this deterministic word-meaning mapping assumption into the IB model by changing our complexity term (Strouse & Schwab, 2017). The IB complexity can be decomposed into two terms $I(M;W) = H(M) - H(M|W)$, the expected information content of meanings $H(M)$ and the conditional entropy $H(M|W)$, which accounts for non-determinism in word-meaning mappings. Assuming that words and meanings are one-to-one mapped, the conditional entropy is 0 and thus, we have the deterministic IB objective function:

$$\mathcal{F}_\beta[q(w|m)] = H(M) - \beta E[KL(M|\hat{M})]. \tag{2}$$

Description Length. The description length approach defines the meaning of a word M_w as a description of the word's extension $M_w \subseteq U$. We can think of this description as an encoding and denote the encoding function as $\mathcal{E}$. For example, the kinship domain can be encoded using a first-order logic description language with predicates for genealogical relationships (Kemp & Regier, 2012). Complexity is defined as the length ℓ of the shortest description of a word's extension $\ell(\mathcal{E}(M_w))$. For example, BROTHER could be encoded as $\exists.z\, female(z) \wedge parent(z,x) \wedge parent(z,y) \wedge male(y)$; however, this is not minimal as the $female(z)$ could be removed. Similar to the IB approach, informativeness is defined as the communicative cost between a speaker and listener $KL[M|\hat{M}]$. Thus, the description length objective function is:

$$\arg\min_{\mathcal{E}(M_w)} \ \ell(\mathcal{E}(M_w)) + KL[M|M_w]. \tag{3}$$

Now, there are three assumptions that we need to make in order to show equivalence between Equation 3 and the Kolmogorov structure function. First, we need to assume that in a given communicative interaction, speakers have a unique intention u in mind (see Regier, Kemp, and Kay (2015) for a similar assumption). In this case, the informativeness term reduces to $-\log P(u|\hat{m})$ and Equation 3 is identical to the well studied, two-part code minimum description length objective function (Rissanen, 1978):

$$\arg\min_{\mathcal{E}(M_w)} \ \ell(\mathcal{E}(M_w)) - \log P(u|\hat{m}), \tag{4}$$

where the first terms corresponds to the encoding of the meaning and the second term corresponds to the encoding of the data under the meaning.

Second, we need to assume that the description language of the model is expressively complete for the domain—i.e., the description language can encode all possible patterns of data. This assumption is easily met by and, often, a desideratum of analyses using description languages (e.g., Goodenough, 1956). Still, we include it for the sake of completeness. If the description language is complete and we have found the minimum description length, then $\ell(\mathcal{E}(M_w))$ is a relaxed Kolmogorov Complexity $K(M_w)$.

Finally, we need to assume that the recovered meaning distribution over the extension of a word is uniform $p(u|\hat{m}) = |M_w|^{-1}$. From the perspective of a communication system analysis, this assumption is probably the least well motivated. Usually, we have some idea about the communicative need of individual elements of the universe and would incorporate that information into our analysis. In this regard, assuming a uniform distribution is akin to ignoring what we know about communicative need. However, from a learning perspective this assumption is well motivated as a size-principle likelihood, which tracks how humans use data when learning (Tenenbaum, 1999; F. Xu & Tenenbaum, 2007; Gweon, Tenenbaum, & Schulz, 2010).

If we make these three assumptions, then Equation 3 can be rewritten as the Kolmogorov structure function (Grunwald & Vitányi, 2004):

$$h(\{u \in M_w\}) = K(M_w) + \log|M_w|. \tag{5}$$

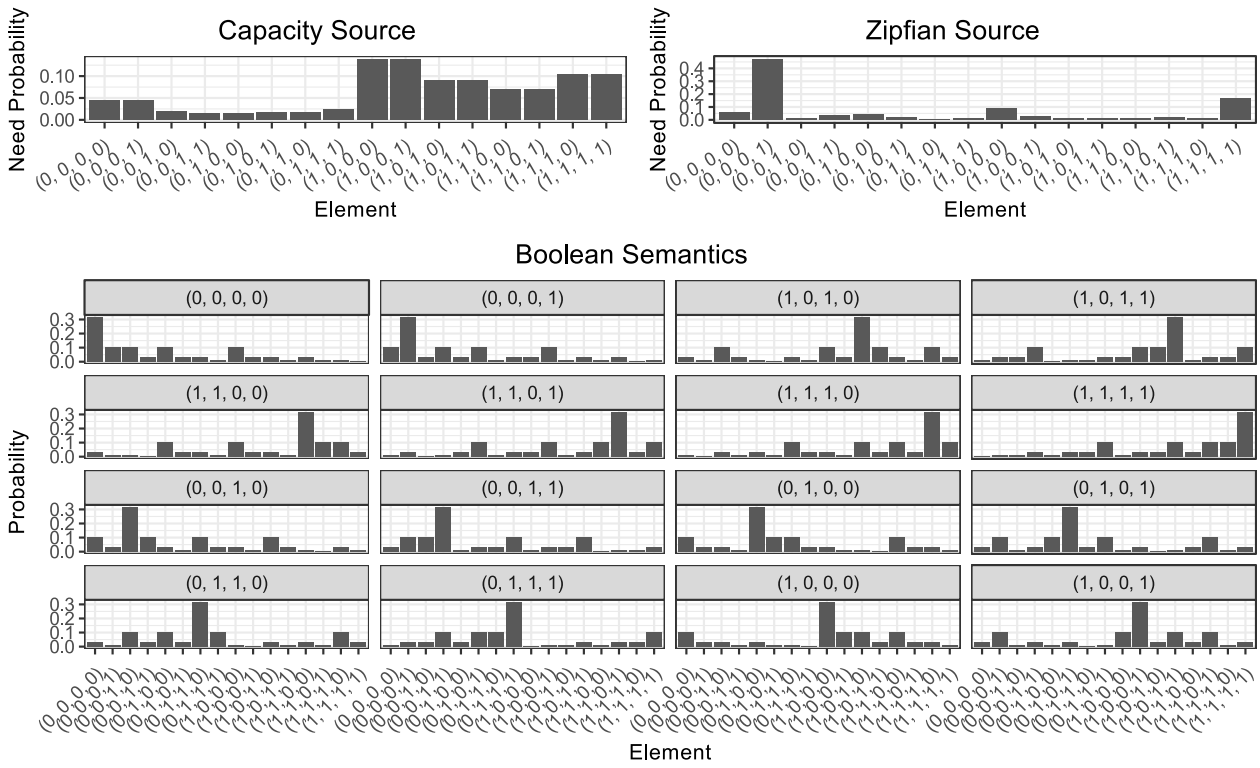

Figure 2: Assumptions for the Shannon communication model for our Boolean universe. The top panels show the source probability for the meaning corresponding to each element in the domain, represented as Boolean feature vectors. The bottom plot shows the meaning distributions $p(u|m)$, where each facet represents a different meaning.

1	Word-meaning mappings are deterministic.
2	Speakers have unique intentions.
3	The description language is expressively complete.
4	The recovered meaning distribution is uniform.

Table 1: Assumptions for Equivalence of IB and Description Length Approaches

When taken in expectation over the source domain, the Kolmogorov Structure Function 5 is asymptotically equivalent to Shannon's Rate-Distortion Theory (Grunwald & Vitányi, 2004). To summarise, the assumptions under which this equivalence hold are enumerated in Table 1. At face value, the equivalence between the two approaches is clearer by noticing that the informativeness term in Equation 2 and Equation 3 are identical and that under a Shannon-Fano code– i.e., a description language with $\ell(\mathcal{E}(m)) = -\log p(m)$, the entropy of a source distribution is equivalent up to a constant with the expected Kolmogorov complexity:

$$H(M) = -\sum p(m)\log p(m) \stackrel{+}{=} \sum p(m)K(m). \qquad (6)$$

Case Study: Communicating Boolean Concepts

To illustrate this equivalence, we will show that the optimal deterministic IB frontier is replicated under a description length analysis. Let's consider a communication system for the semantic domain of Boolean concepts with four features. To use the IB to calculate the frontier of optimal languages, we need to specify the source distribution $P(m)$ and the meaning distributions $P(u|m)$. For illustration, we will cross two assumptions about source and meaning distributions (Figure 2). We will consider a Zipfian biased source and a capacity-achieving source, similar to Zaslavsky et al. (2018)[2]. We will also consider meanings that are governed by the semantics $P(u|m_{u^*}) \propto (0.75)^n(0.25)^{4-n}$, where n denotes the number of shared features between u and u^*, and meanings where we have permuted the domain to flout the semantics (not shown in Figure). Using Equation 2 we will compute the optimal deterministic communication systems for each constellation of our assumptions[3]. Additionally, we generated 10,000 random partitions to demonstrate how the near-optimality of non-optimal communications systems is preserved across approaches.

For each of these systems, we look up the minimal descrip-

[2]As there are no attested languages for this fictional domain we're making up, we constructed a fictional evolutionary trajectory of encoders and used them instead of attested languages. Importantly, this source retains the near uniform properties of capacity achieving sources.

[3]Following standard practice, we use reverse deterministic annealing to compute the IB frontiers. For the capacity source, the β schedule was 2^x for x from 4 to 0 by 0.005 increments. For the Zipfian source, the β schedule was 2^x for x from 8 to 0 by 0.005 increments.

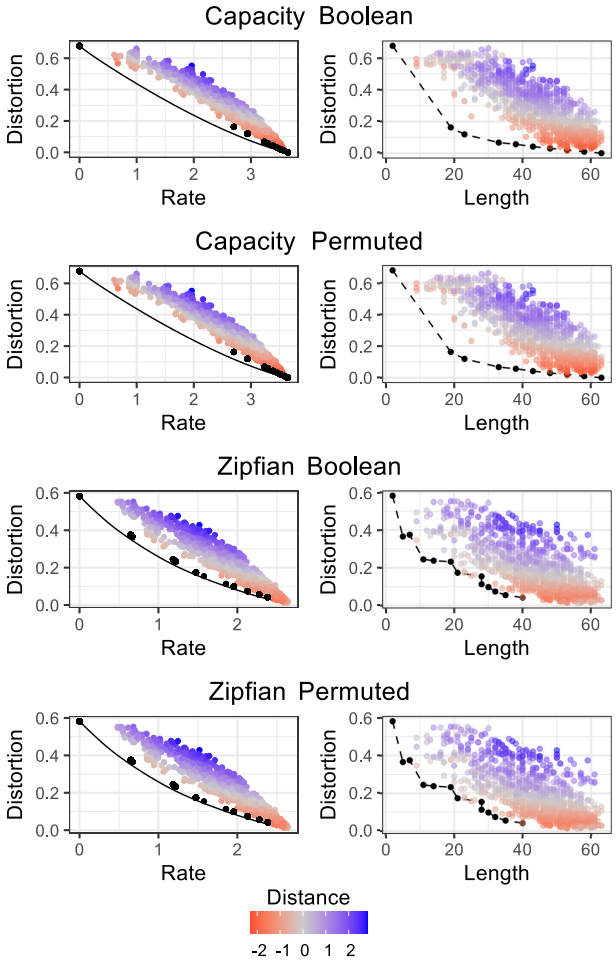

Figure 3: Left panels contain IB frontiers (solid lines), optimal deterministic encoders (black dots) and random partition encoders shaded based on their distance from the optimal IB frontier. Right panels depict the description length trade-off with the same deterministic encoders and random partition encoders still shaded based on their distance from the IB frontier. The dashed line is illustrative of the description length frontier.

tion length[4] from Carcassi and Szymanik (2023) for each partition in the Boolean logic description language: conjunction, disjunction and negation $\{\wedge, \vee, \neg\}$ and calculate the communicative success. Figure 3 plots both the IB pareto-frontier (left panels; solid line) and the translated description length frontier[5] (right panels; dashed line) with optimal deterministic languages plotted as black points and random partitions plotted with colors reflecting their distance to the IB frontier (red is closer; blue further). As can be seen, optimal deterministic IB languages generally lie along the frontier under either formalism.

Lacuna: Will efficient communication analyses of semantic typology generalize across description languages?

In cognitive science, there is a long history of strongly motivating description languages for theory building. For example, Feldman (2000) used Boolean Logic as a description language for concept learning; whereas, Goodman et al. (2008) and Piantadosi, Tenenbaum, and Goodman (2016) used first order logic. Kemp (2012) has also argued for first-order logic being an important description language for concept learning. Kemp and Regier (2012)'s analysis of kinship used first-order logic with relationship primitives; whereas, Mollica and Piantadosi (2022) uses compositional functions as the description language in their developmental account of kinship. While description languages tend to be domain general, others have motivated specific description languages grounded in cognitive abilities, notably for describing numeral systems (Y. Xu et al., 2020) and counting routines (Piantadosi, Tenenbaum, & Goodman, 2012). It is generally held that choice of description language reflects important hypotheses about the underlying features to the semantic domain. For example, anthropology has long been interested in what the minimal set of features an anthropologist must collect to understand kinship systems (Goodenough, 1956). Therefore, there is good reason to question whether results about near optimal communication under one description language will generalize to other description languages.

The answer depends on the description language. If the description language is not capable of capturing any possible input pattern[6]—i.e., not expressive, then we have no guarantees that the analysis will hold. Surprisingly, if the description language is expressive, then judgements of near-optimality will generalize to any description language that is also expressive, following Kolmogorov's invariance theorem (for similar argument see Chater & Vitányi, 2007). Sparing technical details, the encoding of an extension can under any expressive description language be translated into any other expressive description language by writing a translation program and appending it to every encoding. As the length of the translation program is constant, it can be ignored in the optimization[7]. As an illustration, Figure 4 shows the description length frontier for the capacity-achieving source, Boolean se-

[4]Carcassi and Szymanik (2023) calculated the minimal description length in terms of the number of logical operators used. We will use those values here. Further, we will add the lengths of different concepts rather than maximally compressing the logical formulas, as under Kolmogorov complexity. To ensure robustness, we conducted the analysis where we used Lempel-Ziv-Welch compression (Welch, 1984) on the formulas and our illustration holds, likely because there is minimal overlap in most description languages.

[5]NB This frontier is purely illustrative. This is not the Kolmogorov structure function as we are not searching over all possible descriptions. As a result, it is possible for some of the random languages to be below this line.

[6]Any possible input pattern does not necessarily mean every logically possible input pattern. More frequently analyses are only concerned with a theoretically constrained subset of all logically possible inputs. For example, most analyses of kinship are not concerned with uniquely identifying every individual on a family tree. It will be sufficient to demonstrate that description languages are equally expressive on the set of all possible relevant inputs.

[7]It should be noted that the invariance theorem does not mean that any description language can achieve the shortest description for every concept, but rather no other description language could compress every concept any better than a constant amount.

mantics analysis for four expressive Boolean logical description languages. As Carcassi and Szymanik (2023) explain, there are 420 unique description languages of Boolean operators and this illustration holds for all of them.

While this is great news for communicative efficiency studies, where we are interested in the asymptotic efficiency of the system, explanations of other behavioral phenomena (e.g., concept learning/generalization) can depend on the description language. For example, in ideal learner models, the description language is used as a simplicity bias for induction. Thus, in the absence of data the preference of one concept over another is the relative simplicity of the concepts' descriptions in a fixed description language. The relative complexity of two concepts can flip under different description languages (e.g., see Table 2). In fact, Carcassi and Szymanik (2023) studied exactly this issue when evaluating different experimental designs for assessing the recoverability of a description language, or Language of Thought (LoT). Their take homes are: 1) experimental designs need to measure the learning trajectory to have modest success recovering description languages from data because trajectories emphasize the relative order of descriptions; and 2) if descriptive languages preserve the relative orderings of descriptions across languages, the two languages are non-identifiable. This is a long-acknowledged sticking point for using behavioral data to distinguish between description languages as desired by anthropology (Burling, 1964). The consequence is we need to

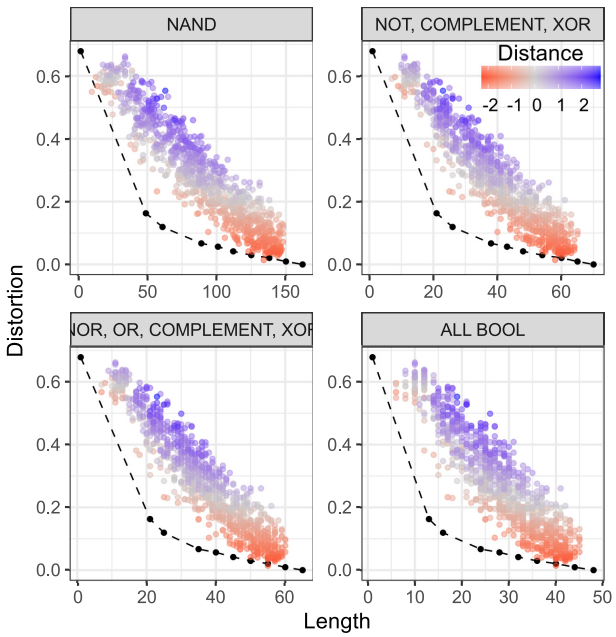

Figure 4: Description length trade-off plots for the capacity source, Boolean semantics communication systems in four additional description languages. The dots represents the same encoders in the top row of Figure 3; thus, the shading is the same across Figures/panels. While there are minor variations in the location of the encoders across description languages, the general patterns of near optimality are preserved. ALL BOOL uses all of the Boolean operators as well as the unary operator negation.

LoT	Formula	Length
$\{\neg, \vee\}$	A= $\neg(p \vee (\neg r \vee \neg s))$	8
	B= $\neg(p \vee \neg(q \vee r)) \vee \neg(q \vee (r \vee \neg p))$	15
$\{\vee, \nleftrightarrow, \barwedge\}$	A= $p \nleftrightarrow ((p \barwedge p) \barwedge (r \barwedge s))$	9
	B= $p \nleftrightarrow (q \vee r)$	5

Table 2: Different description languages (LoT) can change the relative complexity of concepts. Both A and B are concepts with equivalent extension; however, in the disjunction and negation LoT $\{\neg, \vee\}$, A is simpler than B. Whereas in the disjunction, negative biconditional and negative conjunction LoT $\{\vee, \nleftrightarrow, \barwedge\}$, B is simpler than A. The Boolean features in the universe are denoted as *p, q, r* and *s*.

continue to strongly motivate our description languages and to qualify our findings in these models to account for limitations of the description languages.

Discussion

The goal of this note was to show that, under reasonable assumptions (Table 1), efficient communication analyses of semantic typology will generalize across formalizations. Therefore, it's fine to conduct whichever analysis is easier for the available data. For example, when well motivated assumptions about communicative need and meanings are available, IB is a solid formalization. Whereas when well-motivated and easily searchable description languages for the domain are available, the description length approach is appropriate. That said, optimization is generally easier in the IB formalization than in the description length one.

While choice of description language is not an important factor in characterizing the asymptotic efficiency of communicative system, it is prudent to reiterate that description language will matter for other behavioral targets, including explaining evolutionary trajectories. Modelling communicative pressures with information theory is akin to assuming a Shannon-Fano code as the description language. In contrast, developmental/learning models often motivate description languages appealing to cognitive constraints on the semantic domain (e.g., core cognition primitives; Piantadosi et al., 2012) or on the process of searching for descriptions/concepts (Bramley, Dayan, Griffiths, & Lagnado, 2017; Fränken, Theodoropoulos, & Bramley, 2022; Gong, Gerstenberg, Mayrhofer, & Bramley, 2023). As a result, the source of the evolutionary pressures can make different predictions for the evolutionary trajectory; though they can agree (e.g., in the domain of color; Gyevnar, Dagan, Haley, Guo, & Mollica, 2022).

To those hoping to use communicative efficiency analyses of semantic universals to inform description languages as theories of mental representations, this note might be disappointing, yet not condemning. Comparing description languages is still possible and is best approached by modelling the developmental/learning trajectory following Piantadosi et al. (2016) and Carcassi and Szymanik (2023).

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Burling, R. (1964). Cognition and componential analysis: God's truth or hocus' pocus? *American anthropologist*, *66*(1), 20–28.

Carcassi, F., & Szymanik, J. (2023). The boolean language of thought is recoverable from learning data. *Cognition*, *239*, 105541.

Chater, N., & Vitányi, P. (2007). 'ideal learning'of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, *137*, 101506.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

Goodenough, W. H. (1956). Componential analysis and the study of meaning. *Language*, *32*(1), 195–216.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Grunwald, P., & Vitányi, P. (2004). Shannon information and kolmogorov complexity. *arXiv preprint cs/0410002*.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Gyevnar, B., Dagan, G., Haley, C., Guo, S., & Mollica, F. (2022). Communicative efficiency or iconic learning: Do acquisition and communicative pressures interact to shape colour-naming systems? *Entropy*, *24*(11), 1542.

Hsu, A. S., Chater, N., & Vitányi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in cognitive science*, *5*(1), 35–55.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological review*, *119*(4), 685.

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

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 369–376.

Mollica, F., & Piantadosi, S. T. (2022). Logical word learning: The case of kinship. *Psychonomic Bulletin & Review*, 1–34.

Mollica, F., & Zaslavsky, N. (in press). *Information-theoretic and machine learning methods for semantic categorization.* Oxford University Press.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, *123*(4), 392.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, 237–263.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Steinert-Threlkeld, S. (2021). Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, *23*(10), 1335.

Strouse, D., & Schwab, D. J. (2017). The deterministic information bottleneck. *Neural computation*, *29*(6), 1611–1630.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning* (Unpublished doctoral dissertation). MIT.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th annual allerton conference on communication, control and computing*.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Vereshchagin, N. K., & Vitányi, P. M. (2004). Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, *50*(12), 3265–3290.

Welch, T. A. (1984). A technique for high-performance data compression. *Computer*, *17*(06), 8–19.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

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

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.