Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

REGULAR ARTICLE

Assessing Pressures Shaping Natural Language Lexica

Anonymous author(s)

Jeanne Bruneau--Bongard, Emmanuel Chemla & Thomas Brochhagen.
To appear in Cognitive Science

[1]Department Name, Institution Name, State Name, Country Name

Correspondence
Anonymous corresponding author.
Email: email

Abstract

Human languages balance communicative informativity with complexity, conveying as much as needed through the simplest means required to do so. Yet, these concepts — informativity and complexity — have been operationalized in various ways, and it remains unclear which definitions best capture empirical linguistic patterns. A particularly successful operationalization is that offered by the Information Bottleneck framework, which suggests a balance between complexity and informativity across domains like color, kinship, and number. However, we show that the notion of complexity employed by this framework has some counterintuitive consequences. Focusing on color terms, we then study to what extent this and other notions of complexity play a role in explaining cross-linguistic regularity. We propose a method to assess their explanatory contributions; and to probe whether they enter in a joint optimization or in a trade-off competition. This offers a more general framework to study language change and the forces that shape it, where instead of showing that a given model is compatible with existing data, the data is used to adjudicate between candidate measures.

KEYWORDS
information theory, language complexity, semantic typology, computational linguistics, language change

CURRENT MODELS OF LANGUAGE CHANGE

Languages across the world carve up reality in different ways. For instance, some languages distinguish between colors that other languages amalgamate into a single word. Such lexical discrepancies have been studied in the domain of colors [Zaslavsky et al., 2018], as well as in many others such as person [Zaslavsky et al., 2021], number [Denić and Szymanik, 2023], tense [Mollica et al., 2021] and pronominal systems [Saldana and Maldonado, 2024]. Beyond such variation, languages also follow similar abstract organizational patterns in their vocabularies [Kemp and Regier, 2012, Zaslavsky et al., 2018, Kemp et al., 2018, Xu et al., 2020,

Abbreviations: IB, Information Bottleneck; WCS, World Color Survey.

Brochhagen and Boleda, 2022, Jackson et al., 2019, Carlsson et al., 2023]. These patterns have been argued to reflect the result of an interplay between common but non-deterministic desiderata for these languages [Kemp and Regier, 2012, Zaslavsky et al., 2018]. Under this view, recurring universal patterns and their variations can be explained as the outcome of conflicting pressures for languages to help convey as much information as possible, while being as simple as possible.

The Information Bottleneck (IB) framework [Tishby et al., 1999] has been used as an operationalization of this trade-off. Rooted in information theory, this framework provides a characterization of compression of *meanings* into *words* as an optimization of a trade-off between informativity and a notion of complexity which derives from analytic principles. This approach has gained popularity in recent years, notably because it has been shown to fit the organization of meaning across languages in numerous semantic domains remarkably well [Zaslavsky et al., 2019, Zaslavsky et al., 2021, Zaslavsky et al., 2022, Tucker et al., 2022, Carlsson et al., 2023, Mollica et al., 2020].

Here, we first identify aspects of the framework that could be different. For instance, the IB complexity measure that is typically used comes directly from abstract theories of information, and it has not been put in competition with alternatives that could be relevant for direct linguistic processes, such as the borrowing of new words across languages or the existence of synonyms. Similarly, the proposed form of a trade-off is only one of several ways that different constraints may influence language change. Having identified these sources of alternative models, we thus propose a method to *compare* between various potential models of semantic evolution. Very recently, [Tucker et al., 2025] have demonstrated how incorporating pressures for utility in addition to complexity and informativity provides a better fit to the World Color Survey data than the classic IB model. Our study provides new empirical and systematic methods to explore the space of theoretical options of pressures that shape language and the way they interact, including but also beyond the IB framework. Our results in the domain of color suggest a nuanced and multifaceted picture of language change, in which the cross-linguistic organization of lexica is explained by a more diverse notion of complexity, and in which interactions between two or more evolutionary pressures may be involved.

Below we first introduce the IB framework, its associated notions of informativity and complexity, and the way these are measured in general. In Experiment 1 we explore more practically how the IB complexity measure behaves in situations at stake in language change. In Experiment 2, we present other plausible complexity measures, and their combinations, to eventually test which may be involved in the process of language change. Lastly, in Experiment 3, we relax the usual assumption of a trade-off between informativity and complexity and ask whether a linear combination of pressures can better explain real-world data.

TECHNICAL BACKGROUND: THE INFORMATION BOTTLENECK FRAMEWORK

The IB framework describes optimal ways for a language q to compress meanings ($\mathcal{M}$) into words ($\mathcal{W}$) while capturing as much information about the environment ($\mathcal{U}$) as possible. Specifically, [Tishby et al., 1999] shows that in an information-theoretic sense, this amounts to minimizing:

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(U;W) \tag{1}$$

where U, M and W are random variables that represent possible states of the environment, meanings and words, respectively. Meanings are distributions over the environment $\mathcal{U}$, as further developed in Supplementary Material A. I stands for Shannon's mutual information (see Supplementary Material A), and β is a trade-off parameter that controls for the relative importance of the two terms. The first term $I_q(M;W)$ is a measure of compression, also referred to as *complexity* in [Zaslavsky et al., 2018]: it should be minimized. In the second term, $I_q(U;W)$ can be understood as a measure of *informativity*,[1] which should thus be maximized ($\beta > 0$).

Informativity

More precisely, the informativity of a language is defined as the negative of the expected Kullback-Leibler divergence (D) between the speaker's intended meanings, represented by the random variable M, and the listener's interpretations, represented by the random variable $\hat{M}$. It is linked to the mutual information of W and U through a constant independent of the language itself, as follows (see Supplementary Material A.2.2 for derivation):

Informativity

The informativity of a language q is

$$\text{Informativity} = -\mathbb{E}[D[M||\hat{M}]] = I_q(U;W) - I(M;U) \tag{2}$$

This notion of informativity has been used in a variety of contexts and is well-established [Regier et al., 2015, Xu et al., 2020, Mollica et al., 2020, Chen et al., 2023]. Intuitively, a more informative language is one in which the information that its signals carry can be recovered more faithfully (as measured by the expected divergence between speaker meaning M and listener interpretation $\hat{M}$).

Complexity

The IB complexity of a language is defined as the mutual information of M and W, interpreted as the amount of information (in bits) required to represent the intended meaning [Zaslavsky et al., 2018]:

IB Complexity

The IB complexity of a language q is

$$\text{IB complexity} = I_q(M;W)$$

This measure of complexity emerges canonically from the formulation of the Information Bottleneck: it is $I(M;W)$, the complementary term to informativity in Eq. 1. For reasons of parsimony, this emergent notion is at an advantage compared to other notions of complexity that may need to be independently motivated. Notwithstanding, there are many other ways complexity can be operationalized. For instance, through minimal description length [Xu et al., 2020, Steinert-Threlkeld, 2021]; the

[1] [Zaslavsky et al., 2018] uses the term *accuracy*. We call it *informativity* here to distinguish this notion from individual-level task success, as accuracy is understood, a.o., in the machine learning / emergent communication literature [Graesser et al., 2020, Ren et al., 2020]. See [Brochhagen, 2018] for discussion.

number of semantic features a word represents [Denić et al., 2022]; or the complexity of the meanings encoded [Brochhagen et al., 2018]. These word-level complexity measures are then aggregated to obtain a language-level complexity measure. The advantage of these measures is that they behave as expected in the sense that additions to the lexicon lead to increases in complexity. The IB framework, on the other hand, offers a language-level description of complexity whose naturalness comes from analytical principles[2].

Our first goal (Exp. 1) will be to assess the extent to which the analytical IB complexity measure can also be understood in intuitive terms. This is important as we do not only want to propose a way to adjudicate between measures through their empirical fit but to also do so in terms of their explanatory capability. The latter, construed as the conceptual vantage point afforded by a given measure, crucially depends on first understanding what it quantifies in linguistic terms.

IB curve, Optimality, and Efficiency

The IB framework is analytically elegant, and it also has a lot of empirical traction. It has been shown that languages strike a rather close optimal trade-off between informativity and IB complexity. In the domain of colors for instance, [Zaslavsky et al., 2018] showed that languages from the World Color Survey (WCS, see [Kay, 2011]) are very close to the Pareto front, also called the *IB-curve* in this framework (see Fig. 1). The Pareto front represents the optimum, in the sense that hypothetical languages lying on this curve are optimally trading-off informativity and complexity: it is impossible to increase informativity without also increasing complexity; inversely, a decrease in complexity necessarily implies a decrease in informativity. Other possible but unattested languages (green, orange and blue dots) are further away from this optimum than attested languages from the WCS (purple dots).

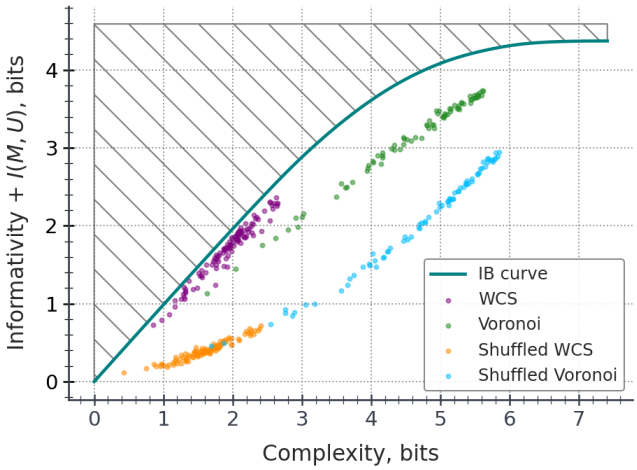


FIGURE 1 IB curve (teal line), representing the optimally efficient languages. The area above this curve is unachievable. Languages from the World Color Survey are plotted in purple. The corresponding shuffled languages (see Experiment 1) are plotted in orange. Voronoi languages (see Experiment 1) are shown in green, and their corresponding shuffled languages in blue.

[2] See [Mollica, 2024] for a discussion on the use of different formalizations of complexity in efficient communication analyses of semantic typology.

EXPERIMENT 1: BEHAVIOR OF THE IB COMPLEXITY MEASURE

While the notion of IB complexity naturally emerges from analytical considerations, we are interested in understanding how it behaves in various controlled situations. Specifically, we explore three ways in which languages could have more words (ways that have some evolutionary plausibility, see [Thouzeau et al., 2024]): adding a synonym to an already existing word, borrowing a word from another language, and splitting a word of a language into two words. Each of these processes should, even if marginally, increase the complexity of a language as they add material to the lexicon. Below, we ask whether IB complexity reflects this.

Languages

We tested the effect of these three vocabulary extensions on the IB complexity of four types of languages, illustrated in Fig. 2 (see detail in Supplementary Material C). First, we used actual languages from the WCS [Kay, 2011] (Fig. 2b). These data consist in names for the colors of the stimulus palette's chips (Fig. 2a) as given by 2,618 speakers of 110 spoken languages. Second, we tested artificial, so-called Voronoi languages [Jäger et al., 2011]. These languages divide the color space into connected partitions (see Fig. 2d). Like attested WCS languages, they also fall close to the IB-curve (see Fig. 1). As shown below, this is a consequence of the way IB optimality is set up. Third, we tested randomly shuffled WCS languages (Fig. 2c) and, fourth, randomly shuffled Voronoi languages (Fig. 2e), where each word distribution $[q(w|m)]_{w\in\mathcal{W}}$ gets randomly assigned to a new meaning. As shown in Fig. 1, these shuffled languages fall further away from the IB curve. This allows us to also study the behavior of the IB complexity measure for languages that are further away from optimality.

Experiment 1A: Adding synonyms

Adding a synonym to a language should increase its complexity. We consider two ways of adding a synonym to a language: (i) spread the probability of the original word equally between two synonyms (*divide* method, described in Table 1), and (ii) duplicate a word, with the same probability, and re-normalize the probabilities over all words (*add* method, described in Table 2). In the case of *divide*, the other words have the same probabilities as in the original language. In the case of *add* they relinquish some mass to the synonyms. (For the results in the main text, the two synonyms themselves always receive the same probabilities, the Supplementary Material D.2 shows that similar results hold when relaxing this constraint).

For the *divide* method, one can prove analytically that the complexity of the new language q' is the same as the complexity of the initial language q (see Supplementary Material D.1).

For the *add* method, Fig. 3a shows simulation results: adding a synonym to a language does not systematically increase or decrease the complexity of the language (11 % of the WCS and 10% of the shuffled WCS languages show a strictly decreased complexity, no change is observed for the Voronoi and shuffled Voronoi languages – see Supplementary Material D for details).

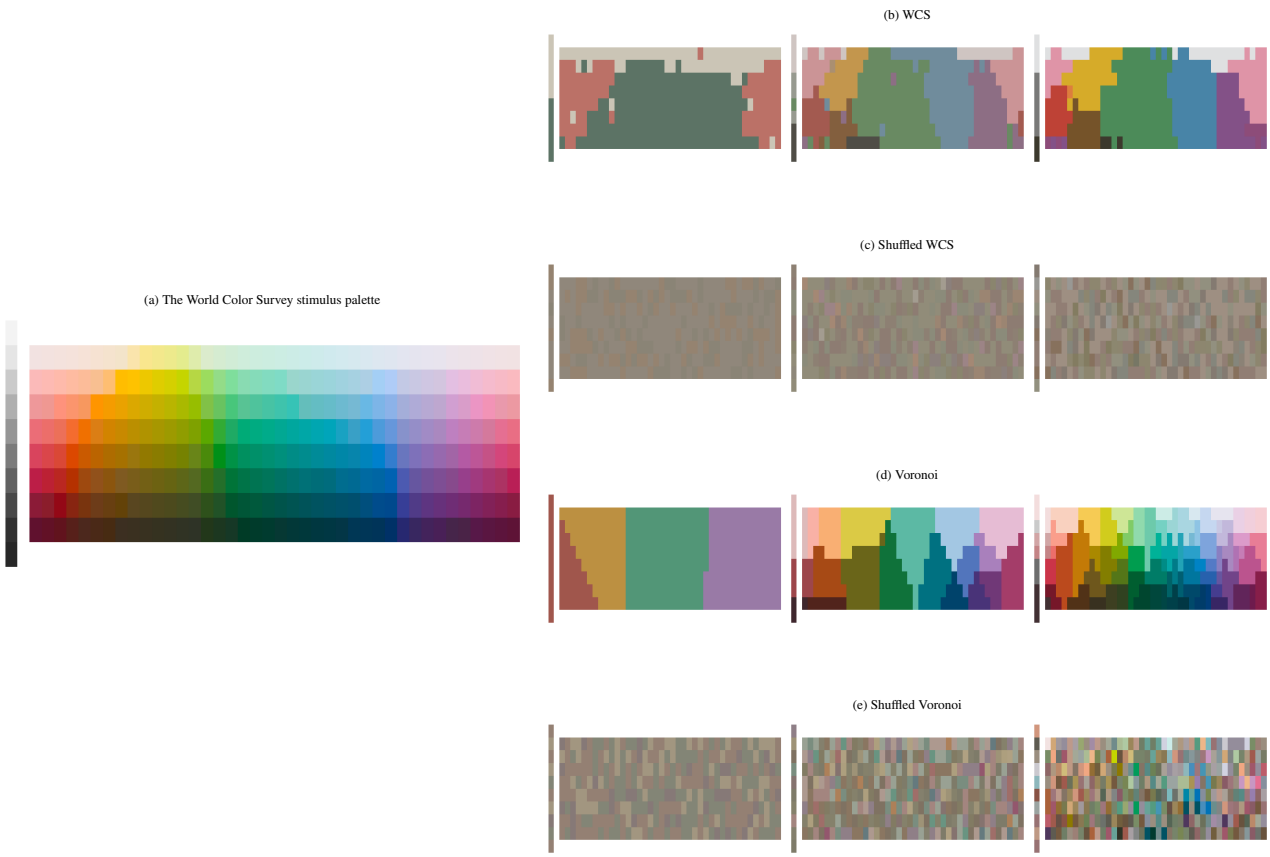


FIGURE 2 **(a)** Colors as organized in the WCS stimulus palette. In (b-e), the colors instead indicate which part of the palette are grouped together in a single word (based on the most frequent word for each cell), with examples of **(b)** WCS (from left to right: Yacouba, Garìfuna (Black Carib), Camsa), **(c)** the corresponding shuffled WCS, **(d)** Voronoi languages (from left to right, with divisions of the color space into 4, 20 and 80 connected zones), and **(e)** the corresponding shuffled Voronoi languages.

Divide			
$\forall k \in \{1, ..., K-1\}$	$q'(w_k\|m)$	$=$	$q(w_k\|m)$
$\text{syn}_1, \text{syn}_2$	$q'(\text{syn}_i\|m)$	$=$	$\frac{q(w_K\|m)}{2}$

TABLE 1 Adding a synonym with the *divide* method. From q a source language, with a lexicon $\mathcal{W} = \{w_1, w_2, ..., w_K\}$, we obtain q' a new language with lexicon $\mathcal{W}' = \{w_1, ..., w_{K-1}\} \cup \{\text{syn}_1, \text{syn}_2\}$, by replacing w_K with two identical synonyms syn_1 and syn_2 and keeping the initial probabilities of the other words.

In both cases, for synthetic as well as natural languages, adding synonyms does not systematically increase the complexity of a language.

In sum, these results show that the IB complexity of a language does not necessarily increase with the addition of synonyms. Instead, depending on the method and the synonyms added, IB complexity often also stays the same or even decreases.

Add			
$\forall k \in \{1, ..., K-1\}$	$q'(w_k \mid m)$	$=$	$\frac{q(w_k \mid m)}{1 + q(w_K \mid m)}$
$\text{syn}_1, \text{syn}_2$	$q'(\text{syn}_i \mid m)$	$=$	$\frac{q(w_K \mid m)}{1 + q(w_K \mid m)}$

TABLE 2 Adding a synonym with the *add* method. From q a source language with lexicon $\mathcal{W} = \{w_1, w_2, ..., w_K\}$, we obtain q' a new language with lexicon $\mathcal{W}' = \{w_1, ..., w_{K-1}\} \cup \{\text{syn}_1, \text{syn}_2\}$, by replacing w_K with two identical synonyms syn_1 and syn_2 and re-normalizing the probabilities over all words.

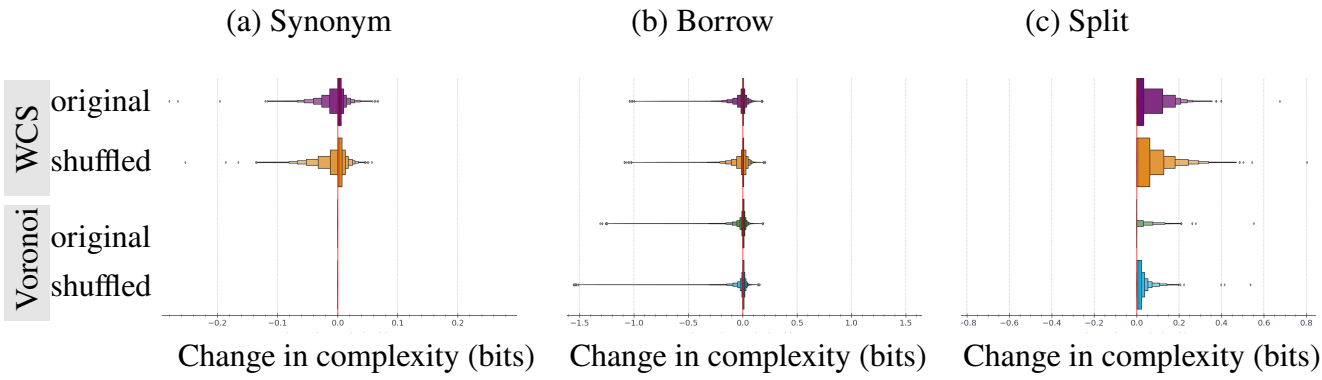


FIGURE 3 Difference in complexity when adding a new word to WCS (purple), shuffled WCS (orange), Voronoi (green) and shuffled Voronoi (blue) languages, using three different methods: **(a)** by adding a synonym of an existing word (*add*), **(b)** by borrowing from another language of the same type, **(c)** by splitting a word into 2. Positive values correspond to an increase in complexity. The red line corresponds to an absence of change. There is no change in complexity in **(a)** for Voronoi as well as shuffled Voronoi languages because the word distributions for each chip are diracs (see Supplementary Material D for more details).

Experiment 1B: Borrowing words from other languages

Another natural way for a language to gain a word is to borrow one from another language. We thus examine the behavior of IB complexity when a word is added from another language of its type (WCS to WCS, Voronoi to Voronoi, shuffled WCS to shuffled WCS, and shuffled Voronoi to shuffled Voronoi), as described in Table 3. For simplicity, we consider borrowings from words with unchanged semantics.

Borrowing a word from another language should increase the overall complexity of a lexicon as it adds new material to the lexicon. A measure of complexity should reflect this. Fig. 3b shows that it is not the case: 18 % of the WCS, 16 % of the shuffled WCS, 23 % of the Voronoi and 19 % of the shuffled Voronoi languages show a strict decrease in complexity instead.

Experiment 1C: Splitting words into two

One may also consider what happens when a word is split into two, thereby gaining informativity and increasing complexity (see [Fuller, 2014] for discussion of the relevance of this in evolution). To do so, as

Borrow			
$\forall k \in \{1, ..., K\}$	$q'(w_k \vert m)$	$=$	$\frac{q(w_k \vert m)}{1 + q_{\text{other}}(\text{other} \vert m)}$
other	$q'(\text{other} \vert m)$	$=$	$\frac{q_{\text{other}}(\text{other} \vert m)}{1 + q_{\text{other}}(\text{other} \vert m)}$

TABLE 3 Borrowing a word from another language. From q a source language with a lexicon $\mathcal{W} = \{w_1, w_2, ..., w_K\}$, we obtain q' a new language with lexicon $\mathcal{W}' = \mathcal{W} \cup \{\text{other}\}$, by borrowing a word other from another language q_{other}. The first row describes how the probabilities of the initial words are modified. The second row specifies which probabilities get assigned to the new word other. For each language, we test this for every word of every other language in the same dataset.

described in Table 4, starting from a word w_{split}, we split its extension into two subsets $\mathcal{M}_1$ and $\mathcal{M}_2$, and then replace the original word with two words, each covering one of these two subsets (see Supplementary Material E.1 for details).

Split			
	$q'(\text{split}_1 \vert m)$	$=$	$\mathbb{1}_{\mathcal{M}_1}(m) q(w_K \vert m)$
	$q'(\text{split}_2 \vert m)$	$=$	$\mathbb{1}_{\mathcal{M}_2}(m) q(w_K \vert m)$
	$q'(w \vert m)$	$=$	$q(w \vert m)$ otherwise

TABLE 4 Splitting a word into two. The key step consists in dividing the set of meanings $\mathcal{M}$ into two subsets: $\mathcal{M}_1$ and $\mathcal{M}_2$, as described in Supplementary Material E.1. From a source language q with a lexicon $\mathcal{W} = \{w_1, w_2, ..., w_K\}$, we obtain a new language q' with lexicon $\mathcal{W}' = \{w_1, ..., w_{K-1}\} \cup \{\text{split}_1, \text{split}_2\}$, by replacing w_K with two new words split_1 and split_2 over $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively.

One can prove analytically that the complexity of the new language q' is greater than the complexity of the initial language q (see Supplementary Material E.2 for a detailed derivation): $C(q') \geq C(q)$.

A simulation of this process on the natural and synthetic languages confirms that splitting a word into two systematically leads to an increase in complexity, as shown in Fig. 3c. In this case, the IB complexity measure accordingly behaves as expected, with additions to the lexicon leading to increases in complexity.

Interim discussion

We investigated several ways to increase the vocabulary of a language by a single word. Each of these processes increases the complexity of a language in the sense that new material is added to the lexicon. Our analytical and simulation results over four types of languages show that this expectation is not always borne out by the IB complexity measure. Only the split of a word into two leads to a consistent increase in IB complexity.

EXPERIMENT 2: COMPARING COMPLEXITY MEASURES

General Framework

Our results so far show that IB complexity captures some but not all facets of what, conceptually, makes a language complex. Moreover, the IB framework raises another empirical issue: What accounts for the gap between theoretically achievable optimality and the near-optimality of attested languages? As shown in Fig. 1, WCS languages are close to the IB curve but they are not on it. This gap is noticeable across semantic domains (see [Zaslavsky et al., 2021, Denić and Szymanik, 2023, Mollica et al., 2021, Saldana and Maldonado, 2024]). If languages optimize the IB trade-off then this difference between theory and data needs to be explained. And this opens another important possibility: since the data is not perfectly accounted for (which will never be the case), we must ask how other plausible models can account for it. Can other complexity measures provide a reasonable approach and a better fit? Can we consider optimizations other than a trade-off? We explore these matters in the following two experiments.

From a conceptual perspective, a good complexity measure is one that follows general principles such as changes in complexity reflecting changes in the lexicon that make its vocabulary more complex. For instance, the addition of words through synonyms, borrowings, or splits as explored here (see Exp. 1). From an empirical and functionalist perspective, a good complexity measure is one for which languages are optimized. Taking this perspective, our aim is to provide a method that allows for the adjudication of different complexity measures and the way they interact.

Methods

To evaluate a given complexity measure from the empirical point of view, we thus want to know how optimal the WCS languages are when measured by it.

In Experiment 2, we evaluate and compare complexity measures under the assumption that they stand in a trade-off relationship with informativity, as defined in the IB framework. We therefore only manipulate the complexity part of the trade-off.

Our method consists of the following steps, explained below:

1. Select complexity measures to be compared
2. Select a set of variants to benchmark attested languages against
3. Derive an empirical Pareto Front
4. View each complexity measure as a classifier (a language is optimal or not)
5. Compare complexity measures via the area under the Receiver Operating Characteristic (ROC) curve.

Selection of Complexity Measures

First, we need to select the complexity measures whose empirical fit we want to compare. We here use IB complexity, degree of convexity (as defined in [Steinert-Threlkeld and Szymanik, 2020] and used in [Koshevoy and Szymanik, 2024]), vocabulary size, and combinations of the three.

The degree of convexity of a language is a candidate complexity measure because humans have been argued to prefer convex concepts ([Gärdenfors, 2000]). Convexity is defined as follows:

Degree of Convexity

Let q be a language and P_q denote its corresponding palette, assigning the most frequently used word to each color chip (see Fig. 2b). Let $\{P_{q,1} \dots P_{q,K}\}$ be the zones that correspond to each word in q. The degree of convexity $\mathcal{C}_{\text{convexity}}$ of q is

$$\mathcal{C}_{\text{convexity}}(q) = \frac{\sum_{i=1}^{K} |P_{q,i}| \times \frac{|P_{q,i}|}{|\text{ConvexHull}(P_{q,i})|}}{\sum_{i=1}^{K} |P_{q,i}|}$$

where $|\cdot|$ denotes the number of chips in a zone, and $\text{ConvexHull}(\cdot)$ is the smallest convex set that contains the zone.

Intuitively, the degree of convexity of a language reflects to what extent the zones corresponding to each of its words are connected in the semantic space, leaving no gaps between any two chips to which a word refers to. Note that this measure, similarly to the IB complexity measure and contrary to vocabulary size, assumes a measure on the semantic space.

In contrast, vocabulary size may seem to be a crude proxy of complexity. We are interested in testing it to assess whether such a simple complexity measure can be relevant, particularly when compared to more sophisticated ones like IB complexity or degree of convexity.

The space of possible languages used to estimate optimality

Second, for each complexity measure, we want to know whether attested languages are optimal compared to other alternative languages. Since we cannot compare the WCS languages to the (infinite) set of all possible languages, we need to select a subsample. To this end, we generate sets of alternative languages in different ways, to also study the influence of such choices on the outcome. As candidate sets of hypothetical language variants, we used the variants from [Zaslavsky et al., 2018] (*Z18*), used for the same purpose in that study, as well as those from Experiment 1: shuffled WCS languages (*Shuffled WCS*), variants obtained by adding synonyms to the WCS languages (*Added Synonyms*), Voronoi languages (*Voronoi*) and IB-optimal languages (*IB Curve*). Although implausible, these languages offer a baseline within the infinite space of possible languages to benchmark against.

Derivation of the Empirical Pareto Front and ε-Pareto Front

Thirdly, we need a way to determine how well empirically attested lexica (the color systems from the WCS) are optimized for a given complexity measure (e.g., vocabulary size) when compared to a given variant set (e.g., shuffled WCS languages). We use Pareto optimality to this end.

A language is Pareto optimal within a set of languages if it is not dominated by any other language in this set in terms of complexity and informativity: there is no language in this set that is at the same time more informative and less complex. Formally:

Pareto optimal languages and Pareto front

A language q is Pareto optimal for a set $\mathcal{Q}$ of languages, a given complexity measure $\mathcal{C}$ and a given informativity measure $\mathcal{I}$ iff

$$\forall q' \in \mathcal{Q} : \mathcal{I}(q) > \mathcal{I}(q) \text{ or } \mathcal{C}(q) < \mathcal{C}(q') \text{ or } q \text{ and } q' \text{ are equal in both dimensions.}$$

The inferred Pareto front $P_{\mathcal{Q},\mathcal{C},\mathcal{I}}$ is the piecewise affine function connecting Pareto optimal languages.

We can thus ask how many of the WCS languages are Pareto optimal given the free parameters above: a complexity measure $\mathcal{C}$ and a set of languages $\mathcal{Q}$. It is important to choose $\mathcal{Q}$ carefully if it is not exhaustive since this is what a given measure is benchmarked against. Hypothetical languages that are expected to be optimal or near optimal given a complexity measure $\mathcal{C}$ accordingly provide a conservative benchmark for that measure. For example, the Voronoi languages from Experiment 1 are convex. They therefore provide a hard benchmark to surpass if degree of convexity is part of how optimality is assessed.

Pareto optimality, as defined above, does not leave space for languages that could have evolved to be nearly optimal. This is why we introduce the notion of *near* Pareto optimality. We formalize this as a language being no more than at a distance of ε to the Pareto front.

ε-optimal languages

Let $\mathcal{Q}$ be a set of languages, $\mathcal{C}$ a complexity measure and $\mathcal{I}$ an informativity measure. Let $P_{\mathcal{Q},\mathcal{C},\mathcal{I}}$ be the corresponding estimated Pareto front (see above).
Then a language q is ε-optimal iff

$$d\Big(\big(\mathcal{C}(q)\,,\,\mathcal{I}(q)\big),\, P_{\mathcal{Q},\mathcal{C},\mathcal{I}} \Big) \leq \varepsilon$$

where d is the Euclidian distance.

Comparing Complexity Measures

Now that we have a notion of (ε-)optimality, a complexity measure can be seen as a classifier: some languages are (ε-)optimal and others are not, for this measure. If we assume that the WCS languages have been optimized based on some complexity measure, then for a complexity measure to be empirically plausible, WCS languages should appear to be optimal or nearly optimal with respect to it. We can therefore evaluate a complexity measure as we would evaluate a classifier: a complexity measure provides a good empirical fit if a large proportion of the ε-optimal languages are WCS languages. To abstract away from the choice of ε, we study this proportion across the whole range of possible values for ε, in a so-called Receiver Operating Characteristic (ROC) curve. We use the Area Under the ROC Curve (AUC) as a metric to compare classifiers. AUC ranges from 0 to 1, with 1 indicating that the measure perfectly identifies WCS languages as optimal without classifying any variant as optimal; and, reversely, 0 reflecting a measure that classifies all variants as more optimal than any of the WCS languages. The higher the AUC, the better suited the complexity measure.

Cross-validation

To preclude over-fitting, estimate sampling variation, and ensure a robust comparison between complexity measures, we performed a 10-fold cross-validation across complexity measures for a given set of variants. That is, the results below correspond to AUCs of held out WCS languages that were not used to estimate the Pareto front. Since this process of holding out data is repeated 10 times, multiple AUCs are obtained for each complexity measure and variant set (see Supplementary Material F for details).

Results

Figures 4 **A**, **B** and **C** illustrate where the WCS languages and the variants lie in the (complexity, informativity) space for the complexity measures we evaluated. They also highlight how the Pareto fronts were inferred. Figure 4 **D** shows an example of ROC curves, obtained combining all variants.

Figure 5 shows the behavior of our complexity-fitting metric for different complexity measures and different hypothetical languages to compare against empirically attested WCS languages.

The results highlight that optimality greatly depends on the sets of hypothetical variants one compares attested lexica against. For instance, using the IB Curve variants (i.e. variants exclusively on the IB Pareto front), the AUC of the corresponding IB complexity measure becomes 0, since no natural language can, by definition, be more optimal than these languages. This is a limitation that cannot be overcome in a large domain like color. Assessing optimality against the largest set of variants is the next best option we have. Here this corresponds to using the union of all variants, as in the right most column of Figure 5. Against this set of variants, IB complexity performs best (AUC$= 0.75 \pm 0.06$), followed by degree of convexity (AUC$= 0.66 \pm 0.05$) and vocabulary size (AUC$= 0.64 \pm 0.05$). Performing an uncorrected paired t-test, we find that the difference in AUC between IB complexity and vocabulary size is significant ($p = 0.02$), as well as between IB complexity and degree of convexity ($p = 0.03$). See Tables G6 and G7 in the Supplementary Material for detailed results.

So far, we described the comparison of individual complexity measures against each other. However, the world's languages may be sensitive to several types of complexity. We therefore investigated composite complexity measures as well. To do so, we first computed z-scores associated to each complexity measure (to make their range comparable), then we identify the best combinations of these z-scored measures. We did so using an exhaustive grid search over all composite measures of the form $\alpha_{\text{IB}} Z(\mathcal{C}_{\text{IB}}) + \alpha_{\text{convexity}} Z(\mathcal{C}_{\text{convexity}}) + \alpha_{\text{voc size}} Z(\mathcal{C}_{\text{voc size}})$, with α_i taking 21 equally spaced values between 0 and 1, and $\alpha_{\text{IB}} + \alpha_{\text{convexity}} + \alpha_{\text{voc size}} = 1$. The AUC scores of the best composite measures for each set of variants are shown in Figure 5 in dark gray. As before, the best composites are evaluated on held-out data using 10-fold cross-validation to avoid overfitting (see Supplementary Material F for details).

Focusing on the largest set of variants (*all*), the best composite measure is a mix of IB complexity ($\alpha_{\text{IB}} = 0.9$), vocabulary size ($\alpha_{\text{voc size}} = 0.05$), and degree of convexity ($\alpha_{\text{convexity}} = 0.05$). Still considering all variants, we find that the best composite measure performs better than all three monolithic measures, with significant differences to vocabulary size ($p = 0.002$) and degree of convexity ($p < 0.001$). However, the difference is not significant for the IB complexity ($p = 0.1$).

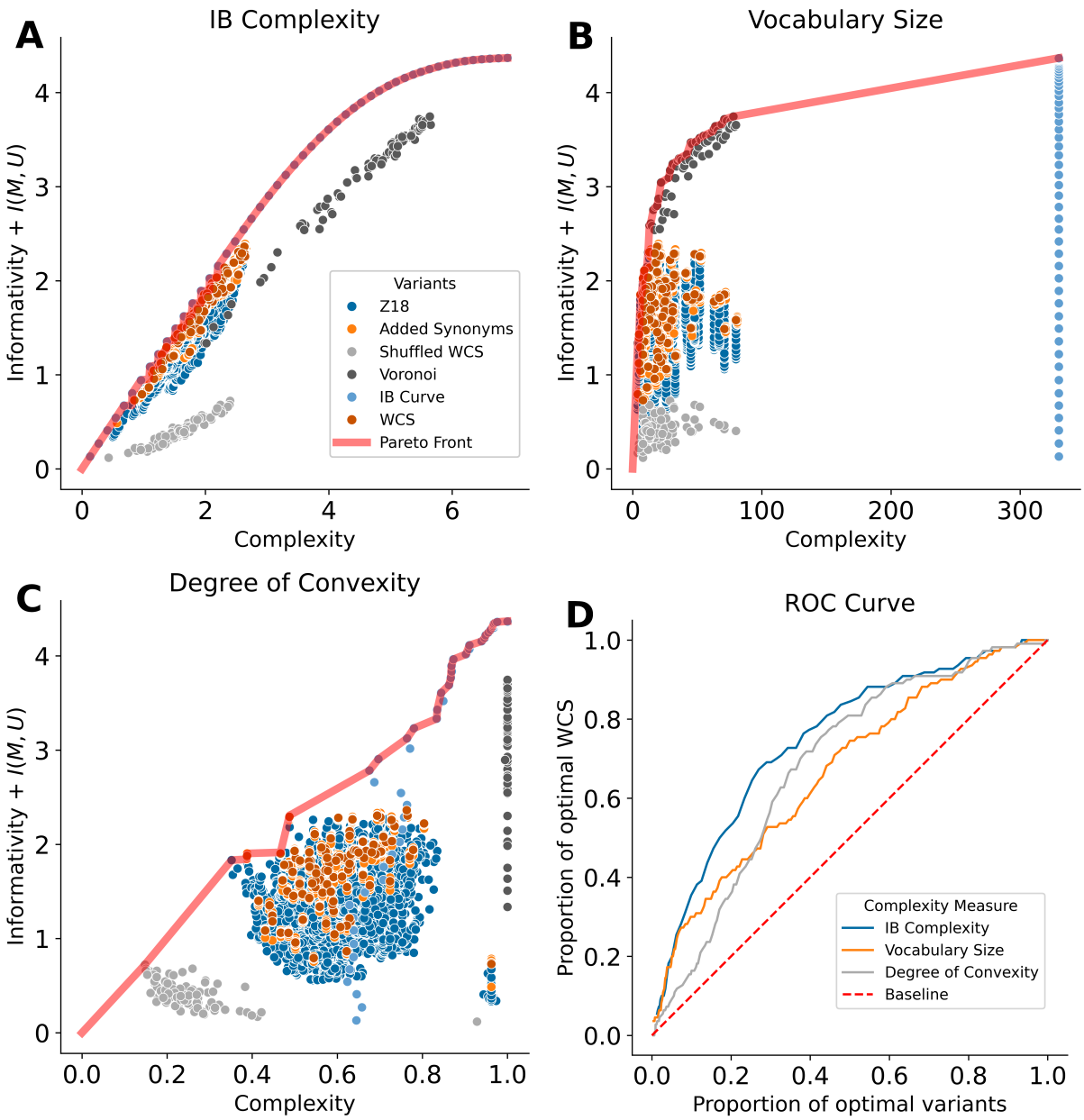


FIGURE 4 **A**, **B** and **C**: complexity (x-axis) and informativity (y-axis) of the WCS languages and variants for our three monolithic complexity measures (IB complexity – **A**, Vocabulary size – **B**, Degree of convexity – **C**). The red lines represent the inferred Pareto fronts. **D**: ROC curves obtained when combining all variants, for each monolithic complexity measure. It shows the proportion of ϵ-optimal WCS languages (y-axis) against the proportion of ϵ-optimal variants (x-axis), aggregated over values of ϵ.

Discussion

So far, these results show that (1) the choice of the set of variants has a strong influence on the results, (2) the best composite measure for the largest set of variants involves a combination of IB complexity, vocabulary size and degree of convexity.

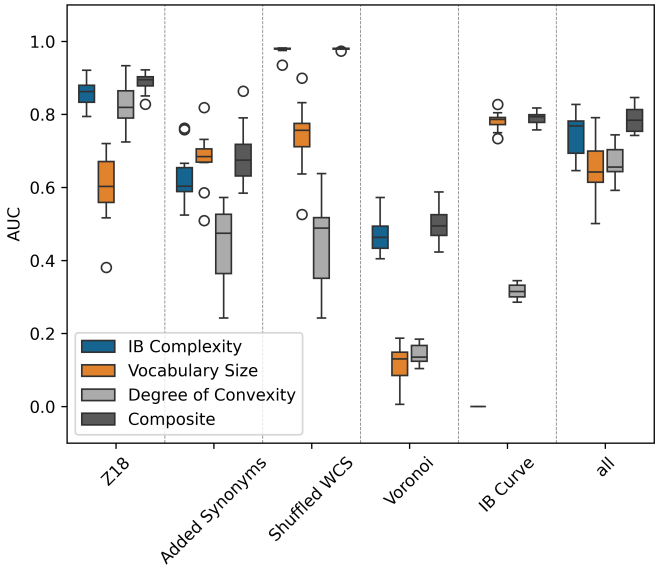


FIGURE 5 EXPERIMENT 2 Areas under the curve (y-axis) for each complexity measure when benchmarked against a given set of variants (x-axis). AUCs are obtained using 10-fold cross-validation. Descriptive statistics of these results can be found in Supplementary Material, Table G6.

EXPERIMENT 3: COMPARING HOW PRESSURES INTERACT

We asked above what complexity measure(s) are empirically most plausible. To do so, we assumed, as is usually done (e.g., in Eq. 1), a trade-off between informativity and complexity. Here, we go one step further and ask whether data is better explained when complexity enters in a trade-off with informativity, or whether all constraints are balanced through an additive competition. That is, whether the relationship between the pressures that different measures instantiate is better explained by other types of relationship than the trade-off that is standardly assumed in the literature. To do so, we investigated whether a linear loss function led to a better fit to attested languages than the assumption of a trade-off does. We used a linear loss function that includes all four parameters (IB complexity, vocabulary size, degree of convexity and informativity). As before, we computed the z-scores of these parameters so that values fall on a comparable range. The loss function is:

$$\mathrm{LF}_{\alpha_{\mathrm{IB}},\alpha_{\text{voc size}},\alpha_{\text{convexity}},\alpha_{\text{info}}} = \alpha_{\mathrm{IB}} Z(\mathcal{C}_{\mathrm{IB}}) + \alpha_{\text{convexity}} Z(\mathcal{C}_{\text{convexity}}) + \alpha_{\text{voc size}} Z(\mathcal{C}_{\text{voc size}}) - \alpha_{\text{info}} Z(\mathcal{I})$$

where $0 \leq \alpha_i \leq 1$ and $\alpha_{\mathrm{IB}} + \alpha_{\text{convexity}} + \alpha_{\text{voc size}} + \alpha_{\text{info}} = 1$.

We adopt analogous notions of optimality to those used in Experiment 2. This allows us to evaluate optimality using the same method as before.

Optimal languages – Linear loss function

A language q is optimal for a set $\mathcal{Q}$ of languages and a given quadruplet $(\alpha_{\mathrm{IB}}, \alpha_{\text{voc size}}, \alpha_{\text{convexity}}, \alpha_{\text{info}})$ iff

$$\forall q' \in \mathcal{Q}, \mathrm{LF}_{\alpha_{\mathrm{IB}},\alpha_{\text{voc size}},\alpha_{\text{convexity}},\alpha_{\text{info}}}(q') \geq \mathrm{LF}_{\alpha_{\mathrm{IB}},\alpha_{\text{voc size}},\alpha_{\text{convexity}},\alpha_{\text{info}}}(q)$$

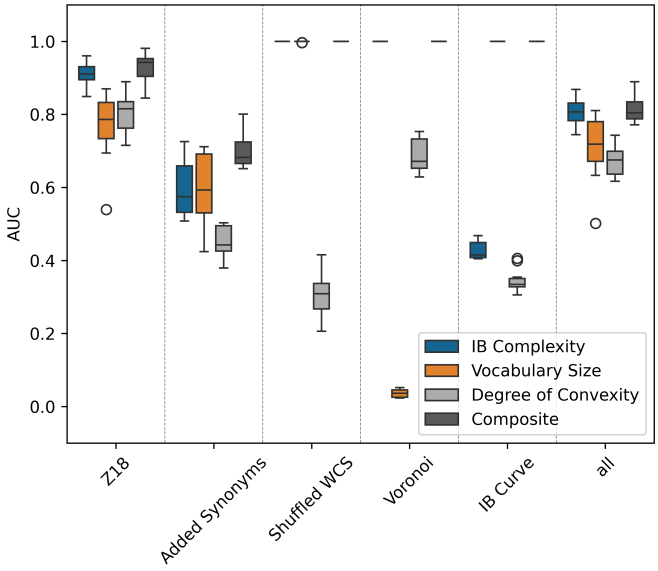


FIGURE 6 EXPERIMENT 3 Area under the ROC-like curves (y-axis) for each linear loss function when benchmarked against a given set of variants (x-axis). AUCs are obtained using 10-fold cross-validation. Descriptive statistics of these results can be found in Supplementary Material, Table G8.

ε-optimal languages – Linear loss function

Let $\mathcal{Q}$ and $(\alpha_{\text{IB}}, \alpha_{\text{voc size}}, \alpha_{\text{convexity}}, \alpha_{\text{info}})$ be a set of languages and a given quadruplet. Then a language q is ε-optimal iff

$$\left| \text{LF}_{\alpha_{\text{IB}},\alpha_{\text{voc size}},\alpha_{\text{convexity}},\alpha_{\text{info}}}(q) - \min_{q' \in \mathcal{Q}} \text{LF}_{\alpha_{\text{IB}},\alpha_{\text{voc size}},\alpha_{\text{convexity}},\alpha_{\text{info}}}(q') \right| \leq \varepsilon$$

Figure 6 shows the results obtained using an additive linear loss function for each complexity measure $\mathcal{C}_i$ alone, with the same weight put on informativity as on complexity (that is $\alpha_i = 0.5$, $\alpha_{\text{info}} = 0.5$ and all other αs equal to zero). It also shows the AUCs obtained by the best composite measures, now represented by a given quadruplet $\alpha_{\text{IB}}, \alpha_{\text{voc size}}, \alpha_{\text{convexity}}, \alpha_{\text{info}}$.

Results

As before, we find that the choice of the set of variants influences the results. When considering all variants (right-most panel in Fig. 6), we find that IB complexity performs better with a significant difference than vocabulary size ($p = 0.006$) and degree of convexity ($p < 0.001$). The best composite measure ($\alpha_{\text{IB}} = 0.42, \alpha_{\text{convexity}} = 0.07, \alpha_{\text{voc size}} = 0.045, \alpha_{\text{info}} = 0.465$) leads to similar results: the mean AUC is higher than the one for vocabulary size and degree of convexity (see Table G8 in Supplementary Material), and the difference is significant ($p = 0.02$ and $p < 0.001$ respectively, see Table G9 in Supplementary Material). However, we find no significant difference between IB complexity and the composite measure.

Comparing these results to the ones of Experiment 2, we find that the linear loss function leads to better results than the trade-off assumption for all measures, and that these differences are statistically significant for IB complexity and vocabulary size ($p = 0.01$, see Table G10 in Supplementary Material).

Discussion

In sum, these results show that using a simple linear loss function can lead to an overall better fit to attested color systems than alternative trade-off formulations. This suggests that the data can be better explained when (multiple facets of) complexity and informativity are balanced through an additive competition. For monolithic complexity measures, the best fit is found for IB complexity. A comparably good fit can be reached using a composite measure involving a mixture of all considered complexity measures and informativity.

GENERAL DISCUSSION

Our results show that IB complexity is sensitive to some but not all of the processes that make lexica more complex (Exp. 1). However, they also highlight the robustness of IB complexity in the sense of providing a good fit for the world's lexica, both when assuming a trade-off with informativity (Exp. 2) and in terms of a linear competition between pressures (Exp. 3).

Beyond providing a more detailed characterization of IB complexity, we also showed how the question of which pressures shape the organization of meaning can be approached in a data-driven manner. Under the assumption that languages evolve to reflect the forces that shape them, one can adjudicate between candidate pressures. Our method not only allows for the evaluation of the best single or composite operationalization of, say, complexity (e.g., IB complexity vs. vocabulary size vs. degree of convexity), but also for the evaluation of different formulations of the way pressures are combined (e.g., a trade-off vs. a linear loss function). We find, for the pressures we compared, that the assumption of a linear competition between pressures better explains the data compared to the assumption of a trade-off for two measures, and does equally good for the two other measures.

Our analysis highlights a central, yet often overlooked, challenge in this line of research: the choice of variants is paramount. Whenever possible, all variants should be considered exhaustively.[3] However, in many cases, such as that of color, this is intractable. This creates a problem for the empirical measure of optimality, because it is relative to a set of variants. This is the case for previous studies which ask whether, given a set of constraints, the observed languages are optimal; and also the current study which asks what constraints may be at stake for the observed languages to be optimal. Nonetheless, we can offer two practical recommendations when choosing these variants. First, one can seek *adversarial variants*, which maximally challenge the final conclusions. For instance, one may consider variants that are artificially optimized for the constraints under investigation. When the set of constraints is the (moving) target of the study, this set may have to be constructed dynamically, and tractability should again be kept under control. However, while this may protect against false positives, this can also hide true positives: the observed languages may be on their way to optimality but still exhibit some margin for further optimization. Choosing only adversarial variants could thus exaggerate the distance between empirically observed languages and optimality. Second, one may create *prior variants* using domain knowledge. For instance, as discussed in Experiment 2 in evaluating the pressure for convexity in color naming, Voronoi-tessellated languages

[3] Even in domains where all variants can be generated one may also consider whether they ought to be included with differential likelihoods. But, in the long run, this could also be part of an accompanying acquisition model, for instance.

serve as a more informative benchmark than shuffled WCS languages. Injection of domain knowledge is a justifiable compromise on a pure data-driven approach: when pure bottom-up methods are not tractable, priors must play a stronger role. But domain knowledge may be hard to translate into constraints on variants, and this choice for variants is also dynamic, because domain knowledge may evolve with time. Our general recommendation then is that the choice of variants be chosen with empirical and theoretical care, and most importantly that result interpretation be made in relation to these choices.

There are four main takeaways of these findings. The first is that even though IB complexity provides a good fit for attested languages, it sometimes falls short in behaving as one would expect a notion of linguistic complexity to behave. The second is that the usual assumption of a trade-off between informativity and complexity is not necessary to explain real-world data: a linear combination of the different pressures at stake does equally good or better. The third takeaway is that (near-)optimality is a matter of comparison, making the choice of the comparative set of variants and their justification key. The fourth takeaway concerns ways forward. Here we have focused only on a single domain (namely colors), and a set of pressures limited to four. Moreover, we studied only two types of interactions between them: a trade-off and a linear loss function. Thus, while these results showcase a clear and scalable way forward in the study of efficient communication, more pressures and their combination should be investigated in the future; ideally across multiple semantic domains.

Moving forward, one may also want to consider a more proximal justification of the possible principles at stake: one could propose such constraints based on a direct understanding of the local dynamics of language use and acquisition, which are ultimately the drivers for global and larger scale language changes. We hope that our work, by allowing flexibility in the models tested, will allow the two relevant strands of work, behavioral and computational, to converge into one.

AUTHOR CONTRIBUTIONS

Redacted.

ACKNOWLEDGMENTS

Redacted.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AND CODE

Relevant data and code used to obtain the results of this research are provided here (anonymous link).

References

[Brochhagen, 2018] Brochhagen, T. (2018). *Signaling under uncertainty*. PhD thesis, University of Amsterdam.

[Brochhagen and Boleda, 2022] Brochhagen, T. and Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. *Cognition*, 226:105179.

[Brochhagen et al., 2018] Brochhagen, T., Franke, M., and van Rooij, R. (2018). Coevolution of Lexical Meaning and Pragmatic Use. *Cognitive Science*, 42(8):2757–2789. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12681.

[Carlsson et al., 2023] Carlsson, E., Dubhashi, D., and Regier, T. (2023). Iterated learning and communication jointly explain efficient color naming systems. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

[Chen et al., 2023] Chen, S., Futrell, R., and Mahowald, K. (2023). An information-theoretic approach to the typology of spatial demonstratives. *Cognition*, 240:105505.

[Denić et al., 2022] Denić, M., Steinert-Threlkeld, S., and Szymanik, J. (2022). Indefinite pronouns optimize the simplicity/informativeness trade-off. *Cognitive Science*, 46(5).

[Denić and Szymanik, 2023] Denić, M. and Szymanik, J. (2023). Languages optimize the trade-off between lexicon size and average utterance length: A case study of numeral systems. LingBuzz Published In:.

[Fuller, 2014] Fuller, J. L. (2014). The vocal repertoire of adult male blue monkeys (cercopithecus mitis stulmanni): a quantitative analysis of acoustic structure. *American journal of primatology*, 76(3):203–216.

[Graesser et al., 2020] Graesser, L., Cho, K., and Kiela, D. (2020). Emergent linguistic phenomena in multi-agent communication games.

[Gärdenfors, 2000] Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought*. MIT Press, Cambridge London.

[Jackson et al., 2019] Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., and Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

[Jäger et al., 2011] Jäger, G., Metzger, L. P., and Riedel, F. (2011). Voronoi languages. *Games and Economic Behavior*, 73(2):517–537.

[Kay, 2011] Kay, P. (2011). The world color survey.

[Kemp and Regier, 2012] Kemp, C. and Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, 336(6084):1049–1054.

[Kemp et al., 2018] Kemp, C., Xu, Y., and Regier, T. (2018). Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, 4(1):109–128. _eprint: https://doi.org/10.1146/annurev-linguistics-011817-045406.

[Koshevoy and Szymanik, 2024] Koshevoy, A. and Szymanik, J. (2024). Convexity bias makes languages efficient.

[Mollica, 2024] Mollica, F. (2024). A note on complexity in efficient communication analyses of semantic typology. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

[Mollica et al., 2020] Mollica, F., Bacon, G., Xu, Y., Regier, T., and Kemp, C. (2020). Grammatical marking and the tradeoff between code length and informativeness.

[Mollica et al., 2021] Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., and Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49):e2025993118. Publisher: Proceedings of the National Academy of Sciences.

[Regier et al., 2015] Regier, T., Kemp, C., and Kay, P. (2015). Word Meanings across Languages Support Efficient Communication. In MacWhinney, B. and O'Grady, W., editors, *The Handbook of Language Emergence*, pages 237–263. Wiley, 1 edition.

[Ren et al., 2020] Ren, Y., Guo, S., Labeau, M., Cohen, S. B., and Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model.

[Saldana and Maldonado, 2024] Saldana, C. and Maldonado, M. (2024). Exploring horizontal homophony in pronominal paradigms: A case study where cross-linguistic regularities defy individual learning biases.

[Steinert-Threlkeld, 2021] Steinert-Threlkeld, S. (2021). Quantifiers in Natural Language: Efficient Communication and Degrees of Semantic Universals. *Entropy*, 23(10):1335. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[Steinert-Threlkeld and Szymanik, 2020] Steinert-Threlkeld, S. and Szymanik, J. (2020). Ease of learning explains semantic universals. *Cognition*, 195:104076.

[Thouzeau et al., 2024] Thouzeau, V., Dezecache, G., Schlenker, P., Dunbar, E., Chemla, E., and Ryder, R. J. (2024). Phylogenetics of primate call functions: a methodological proposal and case study. under revision.

[Tishby et al., 1999] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. arXiv:physics/0004057.

[Tucker et al., 2025] Tucker, M., Shah, J., Levy, R., and Zaslavsky, N. (2025). Towards human-like emergent communication via utility, informativeness, and complexity. *Open Mind*, 9:418–451.

[Tucker et al., 2022] Tucker, M., Shah, J., Levy, R. P., and Zaslavsky, N. (2022). Towards human-agent communication via the information bottleneck principle. In *RSS Workshop on Social Intelligence in Humans and Robots*.

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

[Zaslavsky et al., 2022] Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., and Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, 7(2):184–199.

[Zaslavsky et al., 2018] Zaslavsky, N., Kemp, C., Regier, T., and Tishby, N. (2018). Efficient compression

in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

[Zaslavsky et al., 2021] Zaslavsky, N., Maldonado, M., and Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression.

[Zaslavsky et al., 2019] Zaslavsky, N., Regier, T., Tishby, N., and Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. arXiv:1905.04562 [cs].

SUPPLEMENTARY MATERIAL

A THE IB FRAMEWORK

A.1 General framework

In all generality, the Information Bottleneck framework [Tishby et al., 1999] describes optimal ways, in an information-theoretic sense, of maximally compressing a signal X into a signal $\hat{X}$ while keeping as much information as possible about a relevant signal Y. In the case of semantic typology, X would represent *meanings*, $\hat{X}$ *words* and Y the universe. [Tishby et al., 1999] shows that finding an optimal compression given by $\left\{p(\hat{x}|x)\right\}_{x,\hat{x}}$ amounts to minimizing the functional

$$\mathcal{F}_\beta[p] = I(X;\hat{X}) - \beta I(Y;\hat{X}) \tag{A1}$$

where I stands for Shannon's mutual information (see below), and β is a trade-off parameter that controls for the importance of preserved meaningful information about Y compared to high compression.

Let X and Y be two discrete random variables variables with values over the space $\mathcal{X}$ and $\mathcal{Y}$ respectively. Let $P_{(X,Y)}$ be their joint distribution P_X and P_Y their respective marginal distributions. We repeat the definition of standard information theory metrics (log stands for $\log_2$):

Shannon's entropy	**Shannon's joint entropy**	**Shannon's mutual information**
H(X)	H(X;Y)	I(X;Y)
$= -\sum_{x\in X} P_X(x)\log P_X(x)$	$= -\sum_{x\in X}\sum_{y\in\mathcal{Y}} P_{(X,Y)}(x,y)\log P_{(X,Y)}(x,y)$	$= \sum_{x\in X}\sum_{y\in\mathcal{Y}} P_{(X,Y)}(x,y)\log\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}$
		$= H(X) + H(Y) - H(X;Y)$

Details of the setup and notations from [Zaslavsky et al., 2018]

Let us consider a speaker and a listener in an environment $\mathcal{U}$, with possible states denoted as $u \in \mathcal{U}$. The speaker has an intended meaning $m \in \mathcal{M}$, which is a distribution over $\mathcal{U}$. This intended meaning is subject to a prior distribution p over the meanings. The speaker wishes to communicate m by producing a word w taken from a shared lexicon $\mathcal{W}$. This word w is chosen according to the policy of an encoder q, representing the language that the speaker and listener both speak (we later identify a language with its corresponding encoder q). The listener then interprets w as an inferred meaning $\hat{m}$.

Summing up (see also Table A1):

- Meanings are distributions over the environment $\mathcal{U}$: $m : \begin{array}{l} \mathcal{U} \longrightarrow [0,1] \\ u \longmapsto m(u) \end{array}$
- The intended meaning m of the speaker is determined by the prior distribution p: $P_{\mathcal{M}}(m) = p(m)$
- The speaker produces a word w to communicate m according to the language q: $P_{\mathcal{W}}(w|m) = q(w|m)$
- The listener interprets w to mean $\hat{m}$. [Zaslavsky et al., 2018] assumes an ideal Bayesian listener, who always decodes w 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:

$$q(m|w) = \frac{q(w|m)p(m)}{q(w)}$$

We then have:

$$\hat{m}_w(u) = \frac{1}{q(w)} \sum_{m \in \mathcal{M}} q(w|m)p(m)m(u)$$

Notation	Description	Type
$u \in \mathcal{U}$	State of the environment	Object
$m \in \mathcal{M}$	Intended meaning of the speaker	Distribution over $\mathcal{U}$
p	Prior distribution	Distribution over $\mathcal{M}$
$w \in \mathcal{W}$	Word produced by the speaker	Object
$q(\cdot\|m)$	Encoder / Language	Distribution over $\mathcal{W}$
$\hat{m}$	Interpretation of the meaning by the listener	Distribution over $\mathcal{U}$

TABLE A1 Summary of the notations used in [Zaslavsky et al., 2018].

Prior distribution

We use the same prior as [Zaslavsky et al., 2018], namely a universal source for all languages that is derived from the notion of least informative priors (see Supplementary Material of [Zaslavsky et al., 2018]).

Environment

Following [Zaslavsky et al., 2018], we define the environment as the set of the chips of the WCS stimulus palette.

Meanings

We use the same meaning space as [Zaslavsky et al., 2018]: each chip c of the WCS stimulus palette is associated to a meaning m_c, an isotropic Gaussian centered around c in the 3D CIELAB color space (see Supplementary Material of [Zaslavsky et al., 2018] for details).

$$\begin{aligned} m_c : \ \mathcal{U} &\longrightarrow [0,1] \\ u &\longmapsto \lambda_c \exp(-\tfrac{\|u-c\|^2}{\sigma^2}) \end{aligned}$$

Because of the way the meanings are defined, there is a one-to-one mapping between meanings $m \in \mathcal{M}$ and chips $c \in \mathcal{U}$. We may thus in what follows refer to a meaning by referring to a chip.

A.2 Applying the IB framework

That is where the IB framework comes into play: it describes optimal ways of compressing meanings ($\mathcal{M}$) into words ($\mathcal{W}$) while capturing as much information about the environment ($\mathcal{U}$) as possible. Adapting Eq. A1, this amounts to finding a language q that minimizes the functional

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(U;W) \tag{A2}$$

A.2.1 Obtaining the IB curve

Such optimal languages can be approached by minimizing Eq. A2 for different values of β, using the IB method described in [Tishby et al., 1999]. Given a value of β, this method iteratively updates the following equations until convergence:

$$q_\beta(w|m) = \frac{q_\beta(w)}{Z(m,\beta)} \exp(-\beta D\left[m||\hat{m}_w\right])$$
$$q_\beta(w) = \sum_{m\in\mathcal{M}} q_\beta(w|m)p(m)$$
$$\hat{m}_w(u) = \sum_{m\in\mathcal{M}} m(u)q_\beta(w|m)$$

where $Z(m,\beta)$ is a normalization factor. Following [Zaslavsky et al., 2018], we used reverse deterministic annealing to mitigate the problem of converging to sub-optimal fixed points of the equations above. Reverse deterministic annealing starts at a very high value of β, where the solution is given by a one-to-one mapping from $\mathcal{M}$ to $\mathcal{W}$, and then gradually decreases β. For each β, the IB method is initialized with the solution found for the previous value of β. The process is repeated with 1,500 values of β in $[1, 2^{13}]$. These β values and the prior distribution over $\mathcal{M}$ were taken from [Zaslavsky et al., 2018]. The implementation is available here (`information-bottleneck-method` folder).

A.2.2 Informativity and Complexity

In [Zaslavsky et al., 2018], the two mutual information terms from Eq. A2, $I_q(M;W)$ and $I_q(U;W)$, are interpreted as crucial features of the languages, namely *complexity* and *informativity*.[¶]

Informativity

The informativity of a language is defined as the opposite of the expected Kullback-Leibler divergence (D) between the speaker's intended meanings represented by the random variable M and the listener's interpretations represented by the random variable $\hat{M}$:

$$\text{Informativity} = -\mathbb{E}[D[M||\hat{M}]]$$

[¶] [Zaslavsky et al., 2018] uses the word *accuracy*, but here we stick to *informativity* to avoid any confusion with Machine Learning / Emergent Communication vocabulary [Graesser et al., 2020, Ren et al., 2020]. See also [Brochhagen, 2018] for discussion.

The Kullback-Leibler divergence between a speaker's intended meaning m and its corresponding interpreted meaning $\hat{m}$ is defined as:

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

One can then derive Eq. 2:

$$\begin{aligned}
\mathbb{E}\left[D[M||\hat{M}]\right] &= \sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{M}} p(m) q(w|m) D[m||\hat{m}_w] \\
&= \sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{M}} p(m) q(w|m) \sum_{u \in \mathcal{U}} m(u) \log \left[\frac{m(u) q(w)}{\sum_{m' \in \mathcal{M}} p(m') m'(u) q(w|m')} \right] \\
&= \sum_{w \in \mathcal{W}} \left(\sum_{m \in \mathcal{M}} p(m) q(w|m) \right) \left(\sum_{u \in \mathcal{U}} m(u) \right) \log[q(w)] \\
&\quad + \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} p(m) \left[\sum_{w \in \mathcal{W}} q(w|m) \right] m(u) \log[p(m) m(u)] \\
&\quad - \sum_{m \in \mathcal{M}} p(m) \left(\sum_{w \in \mathcal{W}} q(w|m) \right) \left(\sum_{u \in \mathcal{U}} m(u) \right) \log[p(m)] \\
&\quad - \sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} p(m) q(w|m) m(u) \log \left[\sum_{m' \in \mathcal{M}} p(m') m'(u) q(w|m') \right] \\
&= \sum_{w \in \mathcal{W}} q(w) \log[q(w)] + \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} P_{(\mathcal{U},\mathcal{M})}(u,m) \log \left[P_{(\mathcal{U},\mathcal{M})}(u,m) \right] \\
&\quad - \sum_{m \in \mathcal{M}} p(m) \log[p(m)] - \sum_{u \in \mathcal{U}} \sum_{w \in \mathcal{W}} P^q_{(\mathcal{U},\mathcal{W})}(u,w) \log \left[P^q_{(\mathcal{U},\mathcal{W})}(u,w) \right] \\
&= -H_q(W) - H(U;M) + H(M) + H_q(U;W) \\
&= [H(M) + H(U) - H(U;M)] - [H(U) + H_q(W) - H_q(U;W)] \\
&= I(M;U) - I_q(U;W)
\end{aligned}$$

Complexity

The complexity $\mathcal{C}$ of a language is defined in [Zaslavsky et al., 2018] as being the mutual information of M and W, interpreted as the amount of information (in bits) required to represent the intended meaning. It can be simplified as:

$$\mathcal{C}(q) = I_q(M;W) = \sum_{w \in \mathcal{W}} \sum_{m \in \mathcal{M}} p(m) q(w|m) \log \frac{q(w|m)}{q(w)}$$

B MINOR DEFINITIONS

Definition 1 (Extension of a word)**.** The *extension* of a word w in a language q is the set of meanings for which the probability of that word is strictly positive:

$$E_q(w) = \{m \mid q(w|m) > 0\}$$

Definition 2 (overlap between words)**.** Two words w and w' of a language q *do not overlap* if their extensions are disjoint:

$$E_q(w) \cap E_q(w') = \emptyset$$

Definition 3 (isolated words)**.** A word is *isolated* in a language q if it does not overlap with any other:

$$w \text{ isolated } \Leftrightarrow \forall w' \in \mathcal{W}_q \, (w' \neq w \Rightarrow E_q(w) \cap E_q(w') = \emptyset)$$

C SYNTHETIC LANGUAGES

The code and necessary data to run the different tests is available here (`test-IB-complexity` folder). The cleaned and augmented WCS data were obtained from the supplementary materials of [Carlsson et al., 2023], accessible from here.

C.1 Voronoi languages

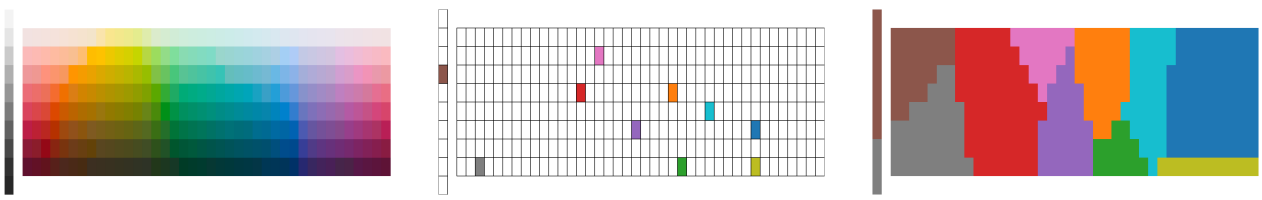
A Voronoi language partitions the color space into K connected regions $\{R_k\}_{k=1\ldots K}$. It is obtained from the following process (see illustration in Fig. C1):

1. Randomly select K points $\{(x_i, y_i)\}_{i=1\ldots K}$ in the color space (removing black and white, so in a rectangle of size 8×41, see Fig. C1, Step 1). For simplicity, we consider the gray scale as an extension of the color space.
2. Compute a Voronoi tessellation with the K points as centroids: each chip of coordinates (x, y) gets assigned to the region R_j such that

   $$\forall i = 1 \ldots K, (x - x_j)^2 + (y - y_j)^2 \leq (x - x_i)^2 + (y - y_i)^2$$

   In other words, each region of centroid (x_i, y_i) consists of all points of the plane closer to that centroid than to any other. This is illustrated in Fig. C1, Step 2.
3. Create a language q that reflects this partition of the space.

   $$q(w_k|m) = \begin{cases} 1 \text{ if } c_m \in R_k \\ 0 \text{ otherwise} \end{cases}$$



The WCS color space | Step 1: Randomly select K points in the color space | Step 2: Create Voronoi tesselation from the previous points

FIGURE C1 Schematic illustration of the steps to sample a Voronoi language.

We created 77 Voronoi languages by applying the process above with each value of $K \in [4, 80]$ (80 is the maximal lexicon size of the WCS languages).

C.2 Shuffled languages

Each language q can be seen as a matrix $L = [q(w|m)]_{m,w} \in \mathscr{M}_{|\mathcal{M}|,|\mathcal{W}|}([0,1])$. Obtaining a shuffled language thus simply consists in permuting the rows of this matrix:

$$L_{\text{shuffled}} = [q(w|\sigma(m))]_{m,w}\,, \text{ where } \sigma : \mathcal{M} \to \mathcal{M} \text{ is a permutation}$$

Shuffling a language corresponds to creating new extensions for each word. In particular, we lose the connectedness of the Voronoi languages.

D ADDING SYNONYMS: METHODS AND ANALYTICAL RESULTS

We review here several methods to add synonymy to a language. We demonstrate in a series of analytical results that in most occasions this does not impact complexity.

D.1 Adding synonyms: The *Divide* method does not affect complexity

Let q denote a language with a vocabulary of size K given by $\mathcal{W} = \{w_1, w_2, ..., w_K\}$. Consider a language q' over the words $\mathcal{W}' = \{w'_1, ..., w'_{K-1}\} \cup \{\text{syn}_1, ..., \text{syn}_N\}$. The idea is that the words w'_k correspond to the words w_k, and the words syn_n are N duplicates of w_K, in the precise sense of Table D2.

Divide			
$\forall k \in \{1, ..., K-1\}$	$q'(w'_k\|m)$	$=$	$q(w_k\|m)$
$\forall n \in \{1, ..., N\}$	$q'(\text{syn}_n\|m)$	$=$	$\dfrac{q(w_K\|m)}{N}$

TABLE D2 *Divide* method: Create a new language q' starting from a language q, by replacing a given word w_K with N identical synonyms syn_n.

Theorem 1. *The Divide method does not affect the complexity of a language.*

Proof. This will follow from Theorem 2, demonstrated below, with $\alpha_n = \frac{1}{N}$ for all n. □

D.2 Adding synonyms: The *Share* method does not affect complexity

Consider the more general *Share* method which creates synonyms by 'duplicating' a word, as in the *Divide* method, with the added flexibility that the different synonyms may have different probabilities. More precisely, with $\alpha_1, ..., \alpha_N$ that sum to 1:

Theorem 2. *The Share method does not affect the complexity of a language.*

Share			
$\forall k \in \{1,...,K-1\}$	$q'(w'_k\|m)$	$=$	$q(w_k\|m)$
$\forall n \in \{1,...,N\}$	$q'(\text{syn}_n\|m)$	$=$	$\alpha_n q(w_K\|m)$

TABLE D3 *Share* method: Create a new language q' by adding N unbalanced synonyms syn_n of a given word w_K from a language q.

Proof. The complexity of the new language can be computed as:

$$\begin{aligned}
\mathcal{C}(q') &= \sum_{w'\in\mathcal{W}'}\sum_{m\in\mathcal{M}} p(m)q'(w'|m)\log\left[\frac{q'(w'|m)}{q'(w')}\right] \\
&= \sum_{k=1}^{K-1}\sum_{m\in\mathcal{M}} p(m)q'(w'_k|m)\log\left[\frac{q'(w'_k|m)}{q'(w'_k)}\right] + \sum_{n=1}^{N}\sum_{m\in\mathcal{M}} p(m)q'(\text{syn}_n|m)\log\left[\frac{q'(\text{syn}_n|m)}{q'(\text{syn}_n)}\right] \\
&= \sum_{k=1}^{K-1}\sum_{m\in\mathcal{M}} p(m)q(w_k|m)\log\left[\frac{q(w_k|m)}{q(w_k)}\right] + \sum_{n=1}^{N}\sum_{m\in\mathcal{M}} p(m)\alpha_n q(w_K|m)\log\left[\frac{\alpha_n q(w_K|m)}{\alpha_n q(w_K)}\right] \\
&= \sum_{k=1}^{K-1}\sum_{m\in\mathcal{M}} p(m)q(w_k|m)\log\left[\frac{q(w_k|m)}{q(w_k)}\right] + \sum_{m\in\mathcal{M}} p(m)\left(\sum_{n=1}^{N}\alpha_n\right) q(w_K|m)\log\left[\frac{q(w_K|m)}{q(w_K)}\right] \\
&= \sum_{k=1}^{K}\sum_{m\in\mathcal{M}} p(m)q(w_k|m)\log\left[\frac{q(w_k|m)}{q(w_k)}\right] = \mathcal{C}(q)
\end{aligned}$$

□

D.3 Adding synonyms: The *Add* method is similar to the *Divide* method for isolated words

Consider the *Add* method which duplicates a word into N versions of itself, and then normalizes the probabilities through all words in the vocabulary (not only within these N synonyms). After having duplicated the word w_K, the probabilities of picking a word given m (before correction) would sum to $1+(N-1)q(w_K|m)$ (the $N-1$ new copies create an excess of probability). The probabilities must thus be normalized as made precise in Table D4.

Add			
$\forall k \in \{1,...,K-1\}$	$q'(w'_k\|m)$	$=$	$\dfrac{q(w_k\|m)}{1+(N-1)q(w_K\|m)}$
$\forall n \in \{1,...,N\}$	$q'(\text{syn}_n\|m)$	$=$	$\dfrac{q(w_K\|m)}{1+(N-1)q(w_K\|m)}$

TABLE D4 The *Add* method creates a new language q' by replacing a word w_K with N synonyms syn_n sharing its probability.

We will now consider the effect of the *Add* method when words do not overlap with one another:

Theorem 3. *The Add method applied to an isolated word is the same as the Divide method.*

Proof. Consider a language where w_K is isolated, that is its extension does not overlap with the extension of any other word. Then, for $k<K$: $\dfrac{q(w_k|m)}{1+(N-1)q(w_K|m)} = q(w_k|m)$, because whenever the numerator is not

null (for $m \in E_q(w_k)$), the denominator is 1 (m is outside of $E_q(w_K)$). Now notice that $q(w_K|m) = \mathbb{1}_{E_q(w_K)}$, because either w_K is not applicable, or w_K is the *only* applicable word. It follows that $\frac{q(w_K|m)}{1+(N-1)q(w_K|m)} = \frac{\mathbb{1}_{E_q(w_K)}}{1+(N-1)\mathbb{1}_{E_q(w_K)}} = \frac{q(w_K|m)}{N}$, because whenever the numerator is not null, the denominator is N.

With these two results ($\frac{q(w_k|m)}{1+(N-1)q(w_K|m)} = q(w_k|m)$ and $\frac{q(w_K|m)}{1+(N-1)q(w_K|m)} = \frac{q(w_K|m)}{N}$), it is visible that Table D4 is the same as Table D2. □

Corollary 1. *For Voronoi and shuffled Voronoi languages, adding synonyms with the Add method does not affect complexity.*

Proof. Voronoi and shuffled Voronoi languages form partitions of the world, so that *all* of their words are isolated. □

E SPLITTING A WORD INTO TWO

E.1 Process to split a word into two

Let q be a language with a vocabulary given by $\mathcal{W} = \{w_1, ..., w_K\}$. We want to split w_K into two new words split_1 and split_2. We will do so as follows:

1. Retrieve the extension of w_K: $E_q(w_K) = \{m \in \mathcal{M} \mid q(w_K|m) > 0\}$
2. Find the corresponding chips delimiting a region in the WCS palette: $(x, y) \in \text{Region}(w_K)$
3. Compute the bounding box of the region:

$$x_{\max} = \max\{x \mid (x,y) \in \text{Region}(w_K)\} \quad | \quad y_{\max} = \max\{y \mid (x,y) \in \text{Region}(w_K)\}$$
$$x_{\min} = \min\{x \mid (x,y) \in \text{Region}(w_K)\} \quad | \quad y_{\min} = \min\{y \mid (x,y) \in \text{Region}(w_K)\}$$

4. Compute the coordinates of the chips that will be the centroids of the two sub-regions:

$$(x_0, y_0) = \left(x_{\min} + \left\lfloor \tfrac{x_{\max}-x_{\min}}{4} \right\rfloor, y_{\min} + \left\lfloor \tfrac{y_{\max}-y_{\min}}{2} \right\rfloor\right)$$
$$(x_1, y_1) = \left(x_{\max} - \left\lfloor \tfrac{x_{\max}-x_{\min}}{4} \right\rfloor, y_{\min} + \left\lfloor \tfrac{y_{\max}-y_{\min}}{2} \right\rfloor\right)$$

   If (x_0, y_0) and/or (x_1, y_1) are not in $\text{Region}(w_K)$, we take their closest points in $\text{Region}(w_K)$.
5. Partition the color space into two, by taking a Voronoi tessellation with (x_0, y_0) and (x_1, y_1) as centroids.
6. Translate this to obtain a partition of the meaning space: $\mathcal{M} = \mathcal{M}_1 \sqcup \mathcal{M}_2$
7. Create a new language q' over $\mathcal{W}' = \{\text{split}_1, \text{split}_2\} \cup \mathcal{W} \setminus \{w_K\}$ with two new words split_1 and split_2 instead of w_K, as described in Table E5.

E.2 Analytical result: Splitting a word into two increases complexity

Theorem 4. *The split method (if not vacuous) increases the complexity of a language.*

Proof. Let q be a language with a vocabulary given by $\mathcal{W}$. Let us take $w_K \in \mathcal{W}$ and apply the split method (see Table E5) to obtain a language q'. To calculate the difference between $\mathcal{C}(q')$ and $\mathcal{C}(q)$, we first note

Split		
$q'(\text{split}_1\|m)$	=	$\mathbb{1}_{\mathcal{M}_1}(m)q(w_K\|m)$
$q'(\text{split}_2\|m)$	=	$\mathbb{1}_{\mathcal{M}_2}(m)q(w_K\|m)$
$q'(w\|m)$	=	$q(w\|m)$ otherwise

TABLE E5 *Split* method: Create a new language q' from a language q, by replacing a given word w_K with two new words split_1 and split_2.

that the terms of these sums which do not involve w_K, split_1 or split_1 will cancel each other out. We are left with:

$$\begin{aligned}
\mathcal{C}(q') - \mathcal{C}(q) = & \sum_{m\in\mathcal{M}} p(m)q'(\text{split}_1|m)\log\tfrac{q'(\text{split}_1|m)}{q'(\text{split}_1)} + \sum_{m\in\mathcal{M}} p(m)q'(\text{split}_2|m)\log\tfrac{q'(\text{split}_2|m)}{q'(\text{split}_2)} \\
& - \sum_{m\in\mathcal{M}} p(m)q(w_K|m)\log\tfrac{q(w_K|m)}{q(w_K)} \\
= & \sum_{m\in\mathcal{M}_1} p(m)q(w_K|m)\log\tfrac{q(w_K|m)}{q'(\text{split}_1)} + \sum_{m\in\mathcal{M}_2} p(m)q(w_K|m)\log\tfrac{q(w_K|m)}{q'(\text{split}_2)} \\
& - \sum_{m\in\mathcal{M}_1} p(m)q(w_K|m)\log\tfrac{q(w_K|m)}{q(w_K)} - \sum_{m\in\mathcal{M}_2} p(m)q(w_K|m)\log\tfrac{q(w_K|m)}{q(w_K)} \\
= & \sum_{m\in\mathcal{M}_1} p(m)q(w_K|m)\log\tfrac{q(w_K)}{q'(\text{split}_1)} + \sum_{m\in\mathcal{M}_2} p(m)q(w_K|m)\log\tfrac{q(w_K)}{q'(\text{split}_2)}
\end{aligned}$$

When the split is not degenerated, $q(w_K)$ is greater than both $q'(\text{split}_1)$ and $q'(\text{split}_2)$, and both terms above are therefore positive. Hence, $\mathcal{C}(q') > \mathcal{C}(q)$. □

F CROSS-VALIDATION

In this section, we explain how we used cross-validation to ensure a robust comparison between our different complexity measures in Experiments 2 and 3.

For defined complexity measures (IB complexity, vocabulary size, degree of convexity), instead of evaluating the full set of WCS languages, we:

1. Randomly split the WCS languages into 10 folds.
2. For each fold i, we:
   (a) Define the Pareto frontier with the chosen variants and folds $1 \dots i-1$ and $i+1 \dots 10$
   (b) Compute the AUC score with the held-out WCS fold i and the chosen variants

We then obtain 10 different AUC scores for each (complexity measure, variants) couple, whose repartition we show in Fig. 5.

To find the best composite complexity measure or linear function, what we do is:

1. Randomly split the WCS languages into 10 folds
2. Generate a grid of composite complexity measures / linear functions. For each composite complexity measure / linear function:

- For each fold i, define the Pareto frontier with the chosen variants and folds $1 \ldots i-1$ and $i+1 \ldots 10$.
- Compute the AUC score with folds $1 \ldots i-1$ and $i+1 \ldots 10$ and the chosen variants

3. Derive the best complexity measure / linear function by taking the mean coefficients that lead to the higher AUC score for each fold.
4. Define the Pareto frontier with all the data, using the best complexity measure / linear function.
5. For each fold i,
   (a) Compute the AUC score using the chosen variants and fold i.
6. Compute the mean AUC score across folds

G DETAILED RESULTS

	all	Z18	Shuffled WCS	Added Synonyms	Voronoi	IB Curve
IB Complexity	0.75 ± 0.06	0.86 ± 0.04	0.98 ± 0.01	0.63 ± 0.08	0.47 ± 0.05	0.0 ± 0.0
Vocabulary Size	0.64 ± 0.09	0.6 ± 0.1	0.74 ± 0.1	0.68 ± 0.08	0.12 ± 0.05	0.78 ± 0.03
Degree of Convexity	0.66 ± 0.05	0.83 ± 0.06	0.45 ± 0.12	0.44 ± 0.11	0.14 ± 0.03	0.32 ± 0.02
Composite	0.79 ± 0.04	0.89 ± 0.03	0.98 ± 0.0	0.69 ± 0.09	0.5 ± 0.05	0.79 ± 0.02

TABLE G6 EXPERIMENT 2 Mean and standard deviation of the areas under the ROC-like curves (AUC) for each (complexity measure, variants) pair, obtained using cross-validation.

	Composite	Degree of Convexity	Vocabulary Size
Degree of Convexity	$p < 0.001$		
Vocabulary Size	$p = 0.002$	$p = 0.5$	
IB Complexity	$p = 0.1$	$p = 0.03$	$p = 0.02$

TABLE G7 EXPERIMENT 2 *p*-values obtained when performing an uncorrected paired t-test to test a significant difference in the means in AUC between two different complexity measures, for the *all variants* set.

	all	Z18	Shuffled WCS	Added Synonyms	Voronoi	IB Curve
IB Complexity	0.81 ± 0.04	0.91 ± 0.03	1.0 ± 0.0	0.6 ± 0.08	1.0 ± 0.0	0.43 ± 0.02
Vocabulary Size	0.71 ± 0.09	0.77 ± 0.1	1.0 ± 0.0	0.6 ± 0.1	0.04 ± 0.01	1.0 ± 0.0
Degree of Convexity	0.67 ± 0.05	0.81 ± 0.06	0.31 ± 0.06	0.45 ± 0.04	0.69 ± 0.05	0.34 ± 0.03
Composite	0.81 ± 0.04	0.93 ± 0.04	1.0 ± 0.0	0.7 ± 0.05	1.0 ± 0.0	1.0 ± 0.0

TABLE G8 EXPERIMENT 3 Mean and standard deviation of the areas under the ROC-like curves (AUC) for each (complexity measure, variants) pair, obtained using cross-validation.

	Composite	Degree of Convexity	Vocabulary Size
Degree of Convexity	$p < 0.001$		
Vocabulary Size	$p = 0.02$	$p = 0.3$	
IB Complexity	$p = 0.7$	$p < 0.001$	$p = 0.006$

TABLE G9 **EXPERIMENT 3** *p*-values obtained when performing an uncorrected paired t-test to test a significant difference in the means in AUC between two different complexity measures, for the *all variants* set.

	Added Synonyms	IB Curve	Shuffled WCS	Voronoi	Z18	all
Composite	$p = 0.6$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.02$	$p = 0.1$
Degree of Convexity	$p = 0.8$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.1$	$p = 0.3$
IB Complexity	$p = 0.03$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.01$
Vocabulary Size	$p = 0.007$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.01$

TABLE G10 **COMPARING EXPERIMENT 2 AND EXPERIMENT 3** *p*-values obtained when performing a paired t-test to test a significant difference in the means in AUC between the two different experiments (trade-off v. linear loss function), for each complexity measure (row) and each variant set (column).

References

[Brochhagen, 2018] Brochhagen, T. (2018). *Signaling under uncertainty*. PhD thesis, University of Amsterdam.

[Carlsson et al., 2023] Carlsson, E., Dubhashi, D., and Regier, T. (2023). Iterated learning and communication jointly explain efficient color naming systems. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

[Graesser et al., 2020] Graesser, L., Cho, K., and Kiela, D. (2020). Emergent linguistic phenomena in multi-agent communication games.

[Ren et al., 2020] Ren, Y., Guo, S., Labeau, M., Cohen, S. B., and Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model.

[Tishby et al., 1999] Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. arXiv:physics/0004057.

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