How Efficiency Shapes Human Language

Edward Gibson, Richard Futrell, Steven Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, Roger Levy

How Efficiency Shapes Human Language

Edward Gibson[1,*], Richard Futrell[2], Steven T. Piantadosi[3], Isabelle Dautriche[4], Kyle Mahowald, Leon Bergen[5], & Roger Levy[1,*]

[1]Massachusetts Institute of Technology, [2]University of California, Irvine, [3]University of California, Berkeley, [4]University of Edinburgh, [5]University of California, San Diego,

[*]Corresponding authors: egibson@mit.edu; rplevy@mit.edu

Abstract

Cognitive science applies diverse tools and perspectives to study human language. Recently, an exciting body of work has examined linguistic phenomena through the lens of efficiency in usage: what otherwise puzzling features of language find explanation in formal accounts of how language might be optimized for communication and learning? Here, we review studies that deploy formal tools from probability and information theory to understand how and why language works the way that it does, focusing on phenomena ranging from the lexicon through syntax. These studies show how a pervasive pressure for efficiency guides the forms of natural language and indicate that a rich future for language research lies in connecting linguistics to cognitive psychology and mathematical theories of communication and inference.

Keywords: language evolution, communication, language efficiency, cross-linguistic universals, language learnability, language complexity

How Efficiency Shapes Human Language

Why do languages look the way they do?

Depending on how you count, there are 6000-8000 distinct languages on earth. The differences among languages seem pretty salient when you have to learn a new language: you have to learn a new set of sounds, a new set of words, and a new way of putting the words together; nevertheless, linguists have documented a wealth of strong recurring patterns across languages [1, 2]. These language universals, which may be exceptionless or instead strong statistical tendencies [3], are key desiderata for explanatory theories of language structure. Here we review recent progress made in explaining some of these universals through a theoretical framework in which languages must offer communicative efficiency under information processing and learning constraints.

Because language is used for transferring information in social environments, a likely influence on language might be how its structure affects **efficiency** of use: languages may be under pressure to be structured so as to facilitate easy, rapid, and robust communication [4, 5, 6, 7, 8, 9, 10, 11, 12]. Traditional theories de-emphasized efficiency of use as a source for similarities of structure across languages, most prominently because of the existence of ambiguity at all levels of linguistic structure: if a component of language is ambiguous, that would arguably make it difficult to use [13]. But recent information-theoretic analyses have shown that the existence of ambiguity is actually not a problem in language use, because context usually disambiguates [14]. Furthermore, the recent availability of large corpora across languages, and natural language processing tools to analyze them, has enabled quantitative evaluation of the possibility that communicative efficiency is a force shaping language structure. We can therefore conceive of a **language's utility** in terms of not only its learnability (or complexity) but also its efficiency of use, such that more efficient languages might be more useful overall [15].

This article reviews the convergent picture that such studies provide: across levels of linguistic analysis—from words to syntax—the form of human language exhibits a strong tendency to be structured for efficient use. In what follows below, we first define efficiency and other concepts from information theory that will be used in this review [16]. Then we review the evidence that the existence of ambiguity in linguistic structures out of context is actually an argument for the information-theoretic approach and not against it. In the main bulk of the review, we next summarize recent results about cross-linguistic universals that can be explained by concepts related to efficiency. Most of this evidence comes from analyses of large texts across languages (see Box 1 for a summary of this and other methods). We then provide evidence that communicative efficiency must often be balanced against the complexity and learnability of languages.

While it is often possible to formulate these other constraints in information-theoretic terms [e.g. as in 17], it remains to be seen whether this is possible in all cases. Finally, we present some challenges for the efficiency / learnability framework.

What is Communicative Efficiency?

Our notions of 'communication' and 'efficiency' are drawn from information theory, which provides a general yet mathematically precise picture of these concepts. The information-theoretic view of communication is summarized in Figure 1. First, an information source selects a message to be transmitted. Next, the message is encoded into a signal, and that signal is sent to a receiver through some medium called a channel. The receiver then decodes the signal to recover the intended message. Successful communication takes place when the message recovered at the destination is equal to the message selected at the source, or diverges from it only slightly. Efficiency in communication means that successful communication can be achieved with minimal effort on average by the sender and receiver. Typically, effort is quantified using the length of messages, so efficient communication means that signals are short on average while maximizing the rate of communicative success. The task of choosing a code that makes signals short on average, according to the distribution over messages coming out of the information source, is called **source coding**.

In the case of human language, the speaker is the information source and sender, the utterance is the signal, and the listener is the receiver and destination. Under this view, human languages are codes that enable information to be transmitted through the channel of the acoustic or visual environment. This information-theoretic picture of language is fundamentally usage-based, because communicative efficiency requires that the messages that we often want to send can be transmitted with minimal effort.

The transmission process may be noisy, meaning that there exist errors in transmission. For example, a speaker and listener may be speaking in a noisy room where the listener cannot perceive all the words the speaker says, or the speaker may be making speech errors, or the listener may not be paying full attention: these and all phenomena that introduce error during communication are called noise. Information theory describes codes which enable communicative efficiency even in the presence of noise of various kinds. When a code enables communication despite noise, it is called robust. The task of choosing a code to enable communication in the presence of noise, according to the characteristics of the channel, is called **channel coding**.

The overall communicative efficiency of a language in this framework boils down to a simple intuition: an efficient language should enable a speaker to transmit many different messages successfully with minimal effort. Note that information theory as specified so far is entirely agnostic to the form and content of messages—it makes no assumptions about the meaning ultimately communicated by an utterance of language.

The underlying messages could be model-theoretic statements about the world, or they could be imperative instructions, or they could contain information about the relative social status of speaker and listener; all that matters from the perspective of information theory is that messages are transmitted accurately, regardless of their content.

Despite this extreme generality, it is nevertheless possible to develop a rich mathematical apparatus for describing communicative efficiency, which is summarized in Box 2. The theory was originally developed in applied settings for the development of telecommunications and cryptography systems [16, 18]. However, it has seen wide application in fields related to cognitive science, including theoretical neuroscience [19], statistical complexity theory [20, 21], and models of rational action under information processing constraints [22, 23]. Fruitful applications of information theory to language are now possible due to large datasets and computational power that make it possible to estimate information-theoretic quantities such as entropy from language data [24].

The existence of ambiguity out of context

One domain where theories of efficient use have been argued to have trouble is the existence of ambiguity, a pervasive phenomenon in language that can be observed in lexical, morphological, syntactic, and semantic systems. Perhaps most famously, Chomsky argued that ambiguity is a hallmark of an inefficient communication system because it permits the possibility of confusion [13]. Chomsky has used this to argue that, in fact, language is not "designed" for communication at all, but rather for some other functions (perhaps, e.g., for thinking). Indeed it is unclear why one would ever design an ambiguous communication system or what role ambiguity might serve [25]. This puzzle is resolved by recognizing the role of the context of usage: to a first approximation, context resolves all communicatively relevant ambiguity [26] (even for children learning their native language [27]), and in natural conversation the participants can easily query each other to rapidly clarify any residual uncertainties [28].

In this light, ambiguity becomes a communicatively desirable feature of language: by leaving out information inferrable from context, we can speak more concisely. This fact can be proven rigorously in information theory [14], and can be appreciated intuitively by looking at places where we attempt to communicate with zero ambiguity, such as legal contracts and computer programming languages. Both of these often end up being wordy or pedantic to an extent that feels unnecessary. The extra effort it would take to make utterances unambiguous (even out of context) would simply be wasted in ordinary language use settings.

Evidence of Communicative Efficiency in Human Language

Evidence of communicative efficiency is provided below, at the levels of the lexicon, syntax, and morphology.

The Lexicon

The lexicon (or dictionary) is the set of words or lexemes that a speaker of a human language knows. Each lexical entry consists of a sequence of sounds paired with a meaning. It is estimated that the average American adult knows about 40,000 words [29].

Word Length. One of the simplest domains in which communicative efficiency has been studied is in word lengths. A well-known statistical law of linguistics, popularized by George K. Zipf [30], is that more frequent words tends to be shorter (e.g. "the" vs "accordion"). This makes intuitive sense: if signals that are sent most frequently are shorter, then we can decrease the average length of a message. For this reason, communication systems like Morse Code also have this principle (e.g. the dot sequence for a frequent letter like "e",".", is shorter than for an infrequent one like"q","- - . -"). However, Zipf worked before information theory provided a mathematical framework for understanding optimal codes. In an optimal code, the length of a signal will depend on its probability in context, not its overall frequency [31]. For instance, if context tells us that there are only two ways for the British to land, Paul Revere can get by with just a single-bit message (one vs. two lights in the Old North Church) instead of having to signal out an entire sentence. Likewise, since language processing mechanisms use contextual information and previous linguistic input to make predictions [32], a more information-theoretically refined version of Zipf predicts that, if language is functioning in a communicatively efficient manner, word length should depend on predictability in context rather than frequency: words that tend to be predicted by context should be even shorter than their frequency predicts. This was shown to be true in a variety of languages by showing that a word's probability in context in a corpus was a stronger determinant of its length than the word's overall frequency [33] (see Figure 2). This pattern likely reflects lexicalization of processes that shorten predictable words, observable both in language production choices and in historical change. The existence of information-theoretically efficient shortening processes in production choice is demonstrated in [34], who show that speakers are more likely to choose the short form of a near synonymous pair like "chimp"/"chimpanzee" in a predictive context than in a neutral context. Together, these results indicate that the processes shaping both the lexicon and word choice are influenced by information-theoretic considerations of predictability–over and above frequency, and exactly as should be expected for an efficient communication system.

The partitioning of semantic space in the lexicon. Beyond the form of words, there has been a great deal of research over the past 70 years on why a language has the words that it does, within a variety of semantic domains, including kinship relations [e.g., 35, 36], color [37, 38], spatial relations [39, 40, 41, 42], and numeral systems [43, 44].

Principles of efficient communication play an increasing role in our understanding of the evolution of word meanings, exemplified by the influential work of [45, 46, 17]. The idea is that perhaps lexica are optimized to balance (i) informativeness—that is, how well words allow us to make communicatively relevant distinctions about states of the world; against (ii) complexity—that is, how well the system of meaning distinctions represented by the lexicon can be mentally represented and learned. In one influential example, it is shown that the kinship systems of all the world's languages lie near the **Pareto frontier**, or the range of optimal possible tradeoffs, between informativeness and complexity [47] (Figure 3A). For example, in some languages, such as Swedish, there are unique words for all 4 grandparents. While this improves the kinship system's ability to uniquely identify people by removing the linguistic ambiguity of "grandmother" and "grandfather" (which could refer either to the maternal or paternal side), it comes at the cost of increased complexity. According to this analysis, the kinship systems of human languages could not be improved in terms of communicative usefulness without increasing complexity, nor made simpler without incurring communicative cost.

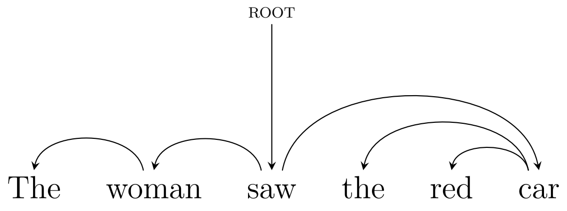
A similar idea has also been applied to the domain of color words. It is well known that color term inventories vary a great deal across languages, from as few as two or three (corresponding to English black, white, and red) to as many as 12 that everyone in the culture knows, as in many modern industrialized communities today [37, 38, 48, 49, 50]. It turns out that the **communicativity** of the set of color words within a language is partially optimized according the perceptual properties of color [51, 52, 53], as defined by the **CIELAB color space** [54], across the **World Color Survey**, a large database of how people name colors from all over the world. The idea here is that when culture necessitates the use of a new color term, both the previous and the new color word partition are driven by the shape of the perceptual color space. This approach has the advantage of explaining why unrelated languages seem to get similar color term partitions. Such an analysis seems ripe for other sense domains, when adequate domain models can be figured out [e.g., 55, 56].

An alternative motivation for the set of color words in a language is usefulness in describing the objects in the world [57, 58]. To evaluate this idea, one can first compute a **color score** associated with each color chip in an array of colors, which reflects how easy it is to communicate that color to a listener, using a word. Consider an arbitrary color, say a light blue. And now consider a word that one might use to label this color, say "blue". We can compute a color score associated with this color and label as a product of of two factors: (a) a production factor: how likely a person would be to use that label for that color (if it's a really good blue, then that probability may be greater than .5); and a comprehension factor: how difficult it would be for a listener to pick the correct physical color from the set of choices, given the label "blue". If there are a lot of other colors that might also be labeled with the same word, then this difficulty is relatively high. Formally, this difficulty is computed as the **surprisal** of the color, given the word. We can then compute the average color score for a color over all words, by summing all the scores for each label for that color. The resulting average color score is the optimal number of yes-no guesses it takes a listener to guess which chip the speaker meant: the number of bits it takes to convey that information. Given this simple information-theoretic score, it turns out that all languages convey warm colors (yellow, red, orange, brown) using fewer bits of information than cool colors (blue, green) (Figure 3B) [58]. Furthermore, objects in the world tend to be warm-colored, whereas backgrounds tend to be cool-colored (blues and greens) (Figure 3C). This suggests a causal explanation for the warm/cool ease of communication generalization that we observe across languages: people invent words to describe objects, not backgrounds. So perhaps color words are introduced into the space according to their usefulness in discriminating among objects [cf. 59].

Under the usefulness hypothesis, all languages have the same underlying perceptual color structure. What differs is the extent to which these perceptual categories are useful in a culture, which can evolve over time, depending on what needs to be labeled. Evidence for this idea comes from analyses of Hadza [60] and Tsimane' [58], both of which have similar color term structure to industrialized languages, despite having relatively low total information in the World Color Survey [61, 62].

Syntax

Syntax is the way words are combined **compositionally** to form sentences. The most general notation for describing syntactic dependencies among words is called **dependency syntax** [63]. Dependency syntax considers only the order in which individual words must be combined together in order to derive the meaning of a sentence. For example, in Example (1) below, the adjective "red" modifies the noun "car", not the noun "woman", and this is indicated by drawing a dependency link from the **dependent**, "car", to its **head**, "red". Headship is a key element of all contemporary syntactic frameworks [64, 65, 66] and can (for the most part) be determined in a relatively theory-neutral way [67], rendering the work described here robust to a range of particular assumptions about finer details of syntactic structure.



Language understanding requires recovering the dependency structure given the words of a sentence. To study efficiency in syntax, researchers ask whether the word order and morphology of languages are optimized to make this process easy and accurate.

Order of subject, verb, and object. One main focus for efficiency-based explanations of word order has been the order of subject, verb, and object. For example, in the sentence "Sam ate oranges", the subject is "Sam", the verb is "ate", and the object is "oranges". English word order is Subject-Verb-Object (SVO); across languages, the most common orders are (in order) SOV, SVO, and VSO, a distant third [68]. There are a number of proposed explanations for the prevalence of SOV and SVO word order: some based in dependency locality (see below) [69], and some based in the theory of Uniform Information Density: that the information content of an utterance per unit time should be roughly constant [70, 71, 72]. Based on co-occurrence statistics for a small set of subjects, verbs, and objects, [73] argue that SVO order produces the most uniform information profile, followed by SOV and VSO.

Another theory for the distribution of SOV and SVO order arises from principles of robust sentence understanding through a noisy communicative channel [74, 75, 76]. In this noisy-channel theory, a transmitter sends a message of the form SVO or SOV, and then the receiver might receive a corrupted version where a word has been deleted. Under these circumstances, SVO order more robustly communicates which noun is the subject and which is the object when these are confusable [77, 78]. However, there is evidence that SOV order is favored by more general, still not fully understood cognitive biases: it is the preferred order in emergent sign languages and ad-hoc gestural communication [79, 80], although these languages and communication systems are typically not strictly SOV. The noisy-channel theory suggests that languages with SOV order must use some other strategy to robustly communicate subject and object: in fact these languages usually have morphological case-marking distinguishing subject from object [81], and in ad-hoc gestural communication the preferred order switches to SVO when subject and object are confusable [77].

Dependency locality. Dependency locality is a theory that the linear distance between words linked in dependencies (dependency length) should be as short as possible [82, 9, 83, 84, 85, 86, 87, 88, 89, 90]. Minimizing dependency length increases processing efficiency because it minimizes the difficulty associated with working memory retrievals during incremental production and comprehension. From the perspective of communicative efficiency, dependency length is a component of the effort involved in sending and receiving utterances, beyond utterance length. For example, in Figure 4A, sentence 4 is notably awkward-sounding, in part because of the long dependency between “threw” and “out”. Such long dependencies cause measurable online comprehension difficulty [91, 92].

Dependency locality provides a unified explanation for several word order phenomena. First, it explains the **harmonic word order** correlations, a set of correlated word order patterns across languages (for example, languages where numerals precede nouns usually also have adjectives preceding nouns) [3, 93, 94, 9], in addition to certain exceptions to these correlations [87]. Second, it explains the general preference to place short constituents before long constituents in certain languages such as English [95, 84], and the reverse preference in other languages such as Japanese [96]. Third, it possibly explains why dependency trees rarely have **crossing links** [97], which correspond to deviations from context-free syntax [98].

Numerous corpus studies provide evidence for dependency locality as a pressure on word order [84, 86, 99, 87, 100]. For example, [88] show that dependency length in hand-parsed corpora of 37 languages is minimized compared to several random baselines instantiating other independently-motivated constraints on word order (see Figure 4B). A dependency locality bias also emerges in artificial language learning [101], suggesting it may influence languages causally via learning biases (see also [102], who find a learning bias in favor of harmonic word orders).

A tight integration of dependency locality with information theory is possible via recent developments in models of online comprehension difficulty [103, 104, 105, 106]. Under the assumption that comprehenders’ memory for linguistic context is noisy, a principle of **information locality** can be derived, such that all words with high **mutual information** should be close for processing efficiency, where mutual information is a measure of how strongly the distribution of two words is constrained [107]. Under this hypothesis, linguistic heads and dependents are word pairs that have especially high mutual information, meaning that dependency locality becomes a special case of information locality. Information locality extends the theoretical reach of dependency locality from strictly syntactic dependencies to all statistical dependencies. Information locality might thus potentially explain word order phenomena that go beyond traditional dependency locality, such as preferences in the relative order of adjectives [108, 109].

Morphology

Morphology refers to the ways in which word forms change in order to express gradations in meaning: for example, English marks plural number on nouns by adding an “-s”. Across languages, morphological marking indicates dimensions of meaning such as number, gender, the person of the subject and object of verbs, and more generally the syntactic relationships among words.

Tradeoff of morphology and fixed word order. Languages adopt two basic strategies to indicate syntactic dependency structure: it can be encoded in fixed word order—as in English—or in morphology, in which words are marked with morphemes which encode their dependency relations to other words. Such marking can take the form of case marking or agreement [110, 111]. Fixed word order and morphology are two means by which a latent tree structure can be encoded into a sequence of words.

A basic efficiency prediction is that when a language has fixed word order, it does not need morphology, because dependency-marking morphemes would be redundant. Furthermore, if a language has morphology, then it does not need fixed word order with respect to dependency structure: word order can instead be repurposed to convey other aspects of meaning, such as information structure [112]. Thus morphological marking should trade off with fixed word order.

Such a trade-off has long been described qualitatively in the linguistic literature [113, 114, 115] and has recently been quantified in corpus studies: an information-theoretic measure of word-order freedom correlates negatively with the presence of morphological marking in dependency-parsed corpora for certain dependency structures [116], and more generally there is a tradeoff of word-order and word-internal information content [117]. Such a tradeoff also emerges in artificial language learning [118].

One possible factor influencing why some languages opt for more rigid word order as opposed to morphology is population structure: languages spoken by populations with more second language speakers tend to disprefer morphology [119, 120], possibly because morphological rules are harder for second language speakers to learn [120].

Redundancy in morphology. One puzzling property of languages from the perspective of efficiency is that they often go out of their way to redundantly mark apparently irrelevant properties of words, such as gender marking for inanimate nouns [121]. In many cases, the gender marking assigned to nouns is nearly arbitrary. More generally, noun classification systems divide nouns into potentially dozens of arbitrary categories which must be marked. How does such an arbitrary and difficult-to-learn aspect of language fit in with the idea that languages are shaped by efficiency?

A recent line of research has proposed that redundant marking of features such as gender may ease language processing effort. Among word classes and across languages, the part of speech associated with most difficulty is nouns [122]. Since language processing effort is associated with predictability [103, 104], an element that makes nouns more predictable in context would make them easier to produce and comprehend. Grammatical gender marking serves exactly this purpose: knowing ahead of time whether a noun is “masculine” or “feminine” reduces uncertainty about lexical identity [123, 124]. In languages without gender marking, this same role may be filled by redundant prenominal adjectives [125]. More generally, such redundant marking may play a role similar to parity bits in information-theoretic codes: they provide redundancy that makes the linguistic signal robust to noise.

Intriguingly, from the communicative perspective, the arbitrariness of gender systems could be a feature, not a bug. It has been argued that if gender markers were not arbitrary—if they were totally predictable—they would not provide unique information about nouns [124]. From the purely communicative perspective, the most useful gender system might thus be an entirely random mapping of nouns to genders—but an entirely arbitrary mapping of nouns to genders would be complex to learn and deploy, suggesting that a complexity–efficiency tradeoff may exist in gender systems. Pending further investigation into this and related hypotheses, the possibility that communicative efficiency considerations play a causal role in shaping the mapping from noun to gender remains speculative.

How Learnability interacts with Efficiency

Traditional theories of language, especially in the generative tradition, have emphasized the problem of learnability: the idea that certain languages may be impossible to learn by children without strong innate principles [126]. Such theories hypothesized that universals arise due to these innate principles, [13], but had difficulty explaining and predicting the empirical range of variation found in languages [12, 127]. More recent arguments based on language learnability have focused on the idea that linguistic systems can be differentially complex, with the more complex systems being harder to learn (but not impossible) [128, 129]. Then universals can be explained under the assumption that simpler systems are preferred [130, 46].

The efficiency perspective on learnability is somewhat different from the traditional perspective in formal linguistics in terms of Universal Grammar (UG) [131]. In typical UG theories, certain languages are learnable because they are compatible with UG, and languages incompatible with UG are unlearnable. More recent work has taken a graded approach to learnability: within the space of languages that can be learned at all, we can ask whether certain languages might be harder or easier to learn. Essentially, a language that is more complex is harder to learn than one which is simpler, a general property of algorithmic and statistical theories of learning [132, 133, 129].

The emergence of compositionality in language transmission

An even more basic question than the explanation of lexical and syntactic universals is the origin of linguistic structure itself. For example, in the section on syntactic efficiency, we discussed how languages use word order and morphology to indicate syntactic dependency trees, but a deeper question is why sentences consist of words that can be arranged in dependency trees in the first place.

A potential solution to this problem lies in the classic learning problem for language: how learners can acquire facility with an infinitely expressive system mapping forms to meanings from limited exposure? For the most frequently intended meanings, learners encounter sufficient learning instances to support a holistic, memorized relationship between utterance form and meaning, such as the English greeting "Hello". But in light of the complete repertoire of utterance meanings that a speaker may need to convey or understand, the totality of any native speaker's linguistic experience is extraordinarily sparse. For our species, the solution to this problem is that language is compositional: smaller meaningful forms can be put together into a novel, larger form whose meaning is a predictable function of its parts [5]. By mastering the basic units and composition functions of a compositional system through simplicity principles [128, 133], a learner can acquire an infinitely expressive set of form–meaning mappings. Humans might have a strong inductive bias constraining language learning to some set of compositional systems, which may have emerged as a side effect of multiple evolutionary events in our lineage [134] or which may have been selected for directly [135]. An alternative, usage-based possibility is that languages might have become compositional through diachronic selection from the transmission bottleneck in their cultural evolution [136, 137, 138]. Even if a non-compositional system were perfectly human-learnable from sufficient evidence, it could not survive over generations, due to input sparsity for any individual learner.

Learnability beyond compositionality: Iconicity & Systematicity

Since the advent of Saussurean structural linguistics [139], the relationship between a word's form and its meaning is generally seen to be arbitrary [5]. The world's languages offer many examples of arbitrariness: for instance, English speakers use the sound sequence "shoe" to label footwear, while French speakers use the same sequence of sounds to label cabbage. This **arbitrariness of the sign** frees speakers to easily coin new words to suit communicative needs. But if wordforms themselves give no clues as to meaning, it poses an acquisition challenge: any given caregiver utterance of "shoe", for example, is compatible with a vast range of plausible intended meanings from among which an infant learner would have to guess [140].

However, accumulating evidence [141, 142], both experimental [143, 144, 145] and computational [146, 147, 148, 149, 150, 151], suggests less arbitrariness in the lexicons of both spoken and signed languages [152] than previously appreciated. For example, it has been shown that the lexicons of natural languages show significantly more phonological clustering over and above phonotactic and morphological regularities [149] (Figure 5). That is, given the constraints imposed by their phonotactic rules, the lexicons of natural languages use a smaller portion of the phonological space available to them. The clustered nature of the lexicon may result from a non-arbitrary relationship between semantics and phonology. Non-arbitrariness can involve **iconicity**, whereby certain acoustic or visual characteristics of sounds or signs are intrinsically good matches to meanings for the human mind (e.g., certain vowel contrasts corresponding to contrasts in magnitude). Non-arbitrariness can also involve language-specific systematicity, such as correlations between sound and grammatical category (in English, nouns' stressed syllables tend to contain back vowels like "school", whereas verbs' stressed syllables tend to contain front vowels like "meet" [153]), or between sound and meaning (in English, words beginning with the phonaestheme "gl-", like "glimmer" and "glow" often relate to light; words beginning with "sn-", like "sneeze" or "sniff", often relate to the nose [144]).

Advances in computing power, natural language processing methods, and statistical analysis have made possible large-scale quantitative analysis of the lexicon to reveal the degree and nature of clustering between form and meaning below the level of the morpheme. For example, positive correlations have been found between word-pair phonological distances and vector space-representation semantic distances significantly above what would be expected under random wordform assignment, in a variety of languages [149]. Other work using nonparametric regression techniques has recovered the best-known English phonaesthemes like "gl-" and "sn-" [148]. These form–meaning regularities exist in a wide range of typologically unrelated languages [147, 150] over and above what would be expected by chance [146] suggesting a fundamental drive for phonological and semantic clustering in the lexicon.

Non-arbitrariness may help learners bootstrap their way into language [154]. Consistent with that hypothesis, corpus analyses have demonstrated that the first words acquired by children display more non-arbitrariness than later acquired words [155, 146, 156], suggesting that non-arbitrariness plays a crucial role in early language acquisition. Experimental studies have also contributed to the evidence that non-arbitrariness benefits learning: Non-arbitrariness helps to bootstrap the acquisition of individual words at the onset of vocabulary development [157] and, as the vocabulary grows, aids learning broad categorical distinctions such as nouns vs. verbs, and supports categorical generalization to novel words [158, 159, 157].

If non-arbitrariness in the lexicon facilitates learning, then learning might in turn shape the structure of the lexicon. This hypothesis is supported by experiments on language evolution in the laboratory [e.g., 137], where arbitrary signals can become systematic after repeated generations of language transmission [160, 138, 161]. As language learners learn the meanings and the functions of words, arbitrary and non-arbitrary mappings each will bring their own selective advantages and disadvantages [146]. Over generations of learners, such advantages and disadvantages will shape vocabulary structure, influencing the presence and the distribution of (non-)arbitrary form-meaning mappings within and across languages.

Challenges for the Efficiency Approach

There are many challenges for the efficiency-based approach to explaining language structure, both methodological and empirical. These challenges primarily have to do with sharpening the theory to extend its empirical reach while maintaining consistent notions of efficiency and constraints on efficiency.

Specification and measurement of complexity. While communicative efficiency can be quantified using information-theoretic concepts, there is less agreement on how to characterize the complexity/learnability of a language. In particular, the complexity of a language may be measured as the length of its description in some meta-language, e.g. as in [47], but this measure depends on the particular description meta-language used. While complexity estimates of this form are to some extent independent of the meta-language [129], they may still vary by large magnitudes. The precise specification of the proper complexity metric will be necessary to further formalize and make precise the relationship between learnability, compositionality, systematicity, and iconicity.

Specification of null hypotheses. To show that an aspect of language is efficient in terms of communication and learnability, it is often necessary to contrast the observed language data with some counterfactual baseline indicating what a language might look like without a pressure for efficiency. For example, in an analysis of kinship systems, natural language kinship systems are compared with an enumeration of logically possible kinship systems [47], and in an analysis of syntactic dependencies, dependency lengths in real sentences are compared to dependency lengths in randomly reordered sentences [88]. Specification of these counterfactual baselines is often the most challenging part of making an efficiency claim. Consider for example if we wanted to claim that a particular utterance is the most efficient way to express a certain message in some context: to make this claim most rigorously, we would have to compare the utterance against all the other things a speaker might have said, given all the degrees of freedom available to the speaker, including pragmatics. The generation of such baseline utterances would be enormously complex. Currently there is no method of generating baseline utterances and languages which is general and satisfactory across all cases.

Measurement of information-theoretic quantities. To quantify communicative efficiency, the information content / entropy, etc., of words, sentences, and messages needs to be estimated. These values are estimated by fitting probability distributions from corpora. However, it is difficult to get reliable estimates of these information-theoretic quantities without very large datasets [162, 163]. For example, [33] use web-scale n-gram corpora provided by Google. The dependence on large datasets makes empirical studies of many languages difficult and confines many studies to modern English, for which the most data is available.

Specification of communicative utility. The information-theoretic notion of efficiency has to do with how well a language communicates arbitrary messages. However, in real natural language communication, there are other factors influencing the utility of an utterance. For example, languages are often embedded in societies that dictate taboo words: words which must be avoided lest the speaker incur social repercussions. The avoidance of such words must figure into the calculation of the utility of an utterance: an utterance containing a taboo word will be avoided even if it is otherwise efficient. It is currently unclear whether such factors can be expressed as part of general communicative efficiency as described above, or if they require additional terms to be added to the equations describing communicative utility. Nor is it clear whether such phenomena have an explanation based in efficiency.

Differences across populations and cultures. Languages around the world vary in sound inventories, grammatical rules, and the organization of the lexicon. The populations in which these languages are spoken vary in many other ways, too: from social organization to technology to culture-specific values. These latter differences can change the distribution of communicative needs: in some populations it will more often be important to convey a given type of meaning than in others. These differences in communicative needs rapidly affect the lexicon—new technologies immediately lead to new words to name and describe them, for example—and in some cases may be reflected in specialized grammaticalized subsystems within the language, such as the grammar of honorific expressions in Japanese [164]. Differences in factors such as population structure [119] and prevalence of non-native speakers [120] may affect the amount of common knowledge shared among speakers, which might in turn affect the linguistic encodings of intended speaker meanings required to ensure accurate communication. But the extent to which these considerations can lead to explanatory theories of cross-linguistic differences remains unclear.

Pathways of historical change. The variation among languages across the world suggests a diversity of ways that a language can be organized that offers a good solution to the language user and learner's problem of efficiency–complexity tradeoff. Major typological features of languages can change rapidly, but not all changes are equally likely. For example, a language's preferred ordering of subject, object, and verb can change completely in the span of a few hundred years, as happened in early English [165], but not all changes among the six logically possible orderings are equally likely [166, 167]. Models of historical language change are crucial to an explanatory theory of the variations seen in linguistic structure [168, 169, 170], but our understanding of the constraints on these models remains limited.

Concluding remarks

The rich collection of studies reviewed here reveals numerous linguistic phenomena which are driven by considerations of efficiency, balancing efficient use and complexity. These results highlight that theories of the evolutionary and ontogenetic origins of language systems are likely to be most productively constrained by the social and communicative function of language.

Interestingly, many of the results have applied tools from information theory—which assumes an idealized version of communication—in order to understand language. One of the most productive directions for ongoing work will be continuing to connect information theory to empirically testable cognitive processes: after all, information is quantified via probabilistic expectations, and so the psychological capacity for and limitations to computing these expectations will tightly constrain our theories. These constraints can only be discovered by further work to understand the mechanisms of language processing. Thus, a compelling future for language research lies in connecting linguistics, information theory, and cognitive psychology in order to provide formalized accounts of how language communicates information, in comprehension and learning alike.

Outstanding Questions

When communicative efficiency and learnability come into conflict, what generalizations can be made about how this conflict is resolved? What features and structures in language reflect such conflict, and how?

To what extent can principles of efficient communication and learnability help explain finer-grained grammatical phenomena investigated in generative linguistics, such as syntactic island?

Can the cognitive constraints shaping communicative efficiency and learnability be connected to and grounded in neural principles of human brain organization?

To what extent does variability within human populations, including individual differences in language processing capabilities and social dynamics, also shape language structure?

Can principles of efficiency in usage and learnability also help us understand symbolic systems used by humans other than natural language, such as gesture, programming languages, and mathematics?

In usage-based theories, the distribution of messages that people want to communicate plays a crucial causal role in influencing language structure. But to what extent does a language's structure in turn influence the distribution of messages that its speakers want to communicate?

Box 1: Methods to investigate cross-linguistic generalizations

The development and testing of efficiency-based hypothesis makes heavy use of analysis of corpora [33, 150, 146, 88, 72]—collection of linguistic data, typically collected from naturalistic sources, such as books, newspapers, websites, radio or television broadcasts, or spontaneous conversations recorded with the consent of the participants. Naturalistic sources play a key role because efficiency in communication and learning must be evaluated with respect to the distribution of messages that speakers intend to convey to one another, and with respect to the distribution of linguistic forms to which a comprehender or learner is exposed. Corpora must often be manually or automatically annotated for additional linguistic structure, such as word boundaries in languages that are not tokenized, morphological structure within words, and syntactic trees within sentences. Advances in natural language processing have played a crucial enabling role in corpus-based work. In work on the partitioning of semantic spaces in the lexicon, electronic dictionaries and similar databases amenable to computational analysis have played a parallel role [61, 47, 58, 17].

Achieving a complete understanding of the range and character of linguistic structure also involves deep analysis of single languages [13, 65, 171, 40, 64]. Many languages have structures which are highly creative, but also extremely infrequent, and therefore difficult to observe in natural corpora. For example, the sentence “Onto the table jumped the cat” involves locative inversion, in which the positions of the subject “the cat” and the locative phrase “onto the table” are swapped relative to ordinary English word order. This construction is quite rare in everyday speech and writing, but it is perfectly understandable by fluent English speakers, and native speakers command a sophisticated tacit understanding of its conditions of felicitous use. This construction appears in many languages, in variable yet constrained forms [171]. Detailed study of such structures is a central focus of generative linguistics. Just as rare neurological conditions may reveal unexpected properties of how the brain is organized, these infrequent linguistic phenomena can give considerable insight into the computational properties of human language.

Discovering cross-linguistic empirical generalizations and testing hypotheses about universality or relative prevalances of various aspects of language structure often makes use of typological databases [3, 68]. Some degree of typological information is available for thousands of languages [172], but for the vast majority of these languages the available information remains highly limited. A key future methodological direction for language science is to make corpus resources and deep linguistic analyses available for a wider range of languages. Such resources and analysis can play a transformative enabling role in testing and developing theories of language structure based on principles of efficiency [88], learnability, and formal parsimony.

Frequently, efficiency-based theories make predictions that can be tested experimentally using the behavioral methods of psychological science [34, 91]. These methods allow dense data to be collected to help understand structures that are rare in naturalistic use, to deconfound by experimental design factors that are correlated in naturalistic data, and to perform interventions to clarify causal relationships. Experimental studies can also involve artificial languages [118, 102] with researcher-designed grammatical properties to test learnability, the role of the communicative channel, and consequences of iterated transmission [137, 138, 160].

Computational modeling plays an increasingly central role in formalizing and testing efficiency-based theories. Computational models help verify the formal soundness of informally stated theoretical proposals [103, 74], estimate quantities (such as conditional word probabilities or vector-based word meaning representations) that play key roles in these theories from corpora [33, 148, 107, 150], clarify the structural tendencies of human languages by comparing them against counterfactual simulated languages lacking efficiency-based properties [87, 146, 88, 150], and gain insight into theoretical dynamics of language change through simulated intergenerational transmission [136, 138].

Box 2: Fundamental Concepts from Information Theory

Information theory is the mathematical theory linking the notions of probability, information, and efficient communication [16, 173].

The fundamental insight of information theory is that the information content of a discrete event x is given by its log inverse probability, or **surprisal**:

$$h(x) = \log_k \frac{1}{p(x)}. \tag{1}$$

When the logarithm is taken in base $k = 2$, then information content is measured in **bits**; base $k = 10$ corresponds to units of **bans**, and the natural logarithm corresponds to units of **nats**. Information content gives the length of the shortest uniquely decipherable code that can be written for the event x given an alphabet of k distinct letters. Going forward we assume $k = 2$.

Given a random variable X, the **entropy** of X is the average information content of samples from X:

$$H[X] = \sum_x p(x) \log \frac{1}{p(x)}. \tag{2}$$

Entropy can be interpreted as the degree of uncertainty about the value X. Entropy is non-negative.

Given two random variables X and C in a joint distribution, we can ask how much uncertainty remains about X after an observer learns the value of C. This quantity is **conditional entropy**:

$$H[X|C] = \sum_{x} \sum_{c} p(x, c) \log \frac{1}{p(x|c)}. \tag{3}$$

Conditional entropy has a value between zero and unconditional entropy: $0 \leq H[X|C] \leq H[X]$ for all X and C [173]. This fact is key for the proof of the utility of ambiguity in [14].

When two random variables X and C predict each other, then intuitively we say that they share information content. The amount of shared information content is given by the **mutual information**:

$$I[X:C] = H[X] - H[X|C], \tag{4}$$

which is the difference between the unconditional entropy of X and the conditional entropy of X given C (or vice versa). Mutual information measures how many bits of information you get about X on average when you learn the value of C. For all X and C, mutual information is non-negative and symmetric in its two arguments:

$$\begin{aligned} I[X:C] &= I[C:X] \\ &\geq 0. \end{aligned} \tag{5}$$

Communicative success in information theory is defined as the event where the information source's intended message is reconstructed accurately at the destination (see Figure 1). The success criterion may be that the received message is exactly equal to the intended message, or it may be that the reconstructed message is only approximately equal. The study of communicative efficiency under the criterion of approximate equality is called **rate–distortion theory**, in which rate quantifies the amount of information transferred through the channel, and distortion is a measure of the divergence of the reconstructed message from the intended message.

Glossary

Arbitrariness of the sign is a central property of natural language that was emphasized in early European linguistics, especially by Ferdinand de Saussure.

Channel coding: The task of choosing a code to enable communication in the presence of noise, according to the characteristics of the channel.

CIELAB color space: is a color space defined by the International Commission on Illumination (CIE) in 1976.

The **communicativity color score** of a color c, is a score reflecting how easy it is to communicate c given a distribution of words w for c in the language.

Compositional: Two words combine / compose together to make a larger phrase, with a meaning that is composed of the meanings of the parts.

Crossing links: Most dependency links in human languages do not cross other dependency links in the same sentence.

Dependency syntax / trees: Two words are in a dependency relation if they combine together in a phrase to make a larger meaning, such as “the” and “dog”, or “dog” and “sleeps”, in the sentence “the dog sleeps”.

Dependency length: the length in words between two words that are dependent on one another for meaning.

Dependency locality: Words in sentences of all human languages tend to be closer together than any reasonable baseline.

Efficiency: A code is efficient if successful communication can be achieved with minimal effort on average by the sender and receiver, usually by minimizing the message length.

Harmonic word orders: A set of correlated word order patterns across languages. For example, languages with prepositions will tend to have verbs before their objects; languages with postpositions will tend to have verbs following their objects [3].

Iconic: An iconic sign is one where its form somehow resembles its meaning.

Minimal pair: two words that differ on only one sound / phoneme, such as “dog” vs. “log”; and “dog” vs. “dot”.

Information theory: the mathematical theory linking the notions of probability, information, and efficient communication.

Information locality: The hypothesis that all words with high mutual information should be close, for processing efficiency [107].

Language utility: The overall usefulness of a language, including both communicative efficiency and complexity as factors.

Pareto efficiency: a state of allocation of resources from which it is impossible to reallocate so as to make any one individual or preference criterion better off without making at least one individual or preference criterion worse off.

Pareto frontier: the set of all Pareto efficient allocations.

Source coding: The task of choosing a code that makes signals short on average, according to the distribution over messages coming out of the information source.

Surprisal: the negative log probability of a discrete event.

The **total information** of a set of color terms the weighted average of the communicativity of each color in the space.

Uniform information density: The hypothesis that the information content of an utterance per unit time should be roughly constant [70, 71, 72]

Universal Grammar (UG) is the name given to the Chomksyan hypothesis that certain aspects of language are innate to humans [131]. There is no consensus on the details of what these aspects might be.

The **World Color Survey** is a set of color labels from languages from 110 non-industrialized cultures around the world [61, 62]. It includes labels from 20-30 participants for each of 330 colors.

References

[1] Ray Jackendoff. *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford University Press, 2003.

[2] Alexandra Y. Aikhenvald and R. M. W. Dixon. *The Cambridge Handbook of Linguistic Typology.* Cambridge University Press, 2017.

[3] Joseph H Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113, 1963.

[4] George Kingsley Zipf. Human behaviour and the principle of least-effort. cambridge ma edn. *Reading: Addison-Wesley*, 1949.

[5] Charles F Hockett and Charles D Hockett. The origin of speech. *Scientific American*, 203(3):88–97, 1960.

[6] Dan I. Slobin. Cognitive prerequisites for the development of grammar. In Dan I. Slobin and C. A. Ferguson, editors, *Studies of Child Language Development.* Holf, Rinehart & Winston, New York, 1973.

[7] Bernard Comrie. *Language Universals and Linguistic Typology.* University of Chicago Press, Chicago, 1st edition, 1981.

[8] Talmy Givón. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Stud Lang*, 15:335–370, 1991.

[9] John A Hawkins. *A performance theory of order and constituency*, volume 73. Cambridge University Press, 1994.

[10] William A. Croft. Functional approaches to grammar. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 6323–6330. Elsevier Sciences, Oxford, 2001.

[11] Joan Bybee. From usage to grammar: The mind's response to repetition. *Language*, pages 711–733, 2006.

[12] Martin Haspelmath. Parametric versus functional explanations of syntactic universals. In T. Biberauer, editor, *The limits of syntactic variation*, pages 75–107. Benjamins, 2008.

[13] Noam Chomsky. Reflections on language. *New York: Pantheon*, 212, 1975.

[14] Steven T Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012.

[15] T Florian Jaeger and Harry Tily. On language ‘utility’: Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335, 2011.

[16] Claude E Shannon. A mathematical theory of communications. *Bell Systems Technical Journal*, 27(4):623–656, 1948.

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

[18] Claude E Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.

[19] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127, 2010.

[20] Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879, 2001.

[21] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[22] Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer, 2012.

[23] Tim Genewein, Felix Leibfried, Jordi Grau-Moya, and Daniel Alexander Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27, 2015.

[24] Fernando Pereira. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253, 2000.

[25] Thomas Wasow, Amy Perfors, and David Beaver. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282, 2005.

[26] George Armitage Miller. *Language and communication.* McGraw-Hill, 1951.

[27] Fibla L. Fievet A. C. Dautriche, I. and A Christophe. Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, 104:83–105, 2018.

[28] Herbert H Clark. *Using language.* Cambridge University Press, 1996.

[29] Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, 7:1116, 2016.

[30] George K. Zipf. *The Psycho-Biology of Language: An Introdution to Dynamic Philology.* MIT Press, 1935.

[31] Ian H Witten, Radford M Neal, and John G Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

[32] William Marslen-Wilson. Sentence perception as an interactive parallel process. *Science*, 189(4198):226–228, 1975.

[33] Steven T Piantadosi, Harry Tily, and Edward Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

[34] Kyle Mahowald, Evelina Fedorenko, Steven T Piantadosi, and Edward Gibson. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318, 2013.

[35] George Peter Murdock et al. Social structure. Technical report, Free Press New York, 1949.

[36] Sara Nerlove and A Kimball Romney. Sibling terminology and cross-sex behavior 1. *American Anthropologist*, 69(2):179–187, 1967.

[37] Brent Berlin and Paul Kay. *Basic color terms: Their university and evolution.* University of California Press, 1969.

[38] Paul Kay and Chad K McDaniel. The linguistic significance of the meanings of basic color terms. *Language*, pages 610–646, 1978.

[39] Melissa Bowerman et al. Learning how to structure space for language: A crosslinguistic perspective. *Language and Space*, pages 385–436, 1996.

[40] Leonard Talmy. How language structures space. In H. Pick and L. Acredelo, editors, *Spatial orientation: Theory, research, and application*, pages 225–282. Plenum Press, 1983.

[41] Stephen C Levinson and Sérgio Meira. ‘Natural concepts’ in the spatial topological domain—adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language*, 79(3):485–516, 2003.

[42] Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel BM Haun, and Stephen C Levinson. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3):108–114, 2004.

[43] Joseph H. Greenberg. Generalizations about numeral systems. In Joseph Harold Greenberg, Charles Albert Ferguson, and Edith A. Moravcsik, editors, *Universals of human language*, volume 3, pages 249–295. Stanford University Press Stanford, 1978.

[44] Bernard Comrie. Numeral bases. In *The world atlas of language structures*, pages 530–533. Oxford Univ. Press, 2005.

[45] Terry Regier, Charles Kemp, and Paul Kay. Word meanings across languages support efficient communication. *The handbook of language emergence*, 87:237, 2015.

[46] Charles Kemp, Yang Xu, , and Terry Regier. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4:109–128, 2018.

[47] Charles Kemp and Terry Regier. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054, 2012.

[48] John A. Lucy. The linguistics of “color”. In C.L. Hardin and L. Maffi, editors, *Color categories in thought and language*, pages 320–346. Cambridge University Press, 1997.

[49] Kimberly Jameson and Roy G D’Andrade. It’s not really red, green, yellow, blue: an inquiry into perceptual color space. pages 295–319. Cambridge University Press, 1997.

[50] Delwin T Lindsey and Angela M Brown. Universality of color names. *Proceedings of the National Academy of Sciences*, 103(44):16608–16613, 2006.

[51] Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441, 2007.

[52] Luc Steels, Tony Belpaeme, et al. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–488, 2005.

[53] Andrea Baronchelli, Tao Gong, Andrea Puglisi, and Vittorio Loreto. Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107(6):2403–2407, 2010.

[54] David H Brainard. Color appearance and color difference specification. *The Science of Color*, 2:191–216, 2003.

[55] Asifa Majid and Nicole Kruspe. Hunter-gatherer olfaction is special. *Current Biology*, 28(3):409–413, 2018.

[56] Asifa Majid, Sean Roberts, Ludy Cilissen, Karen Emmorey, Brenda Nicodemus, Lucinda O'Grady, Bencie Woll, Barbara LeLan, Hilário de Sousa, Brian L. Cansler, Shakila Shayan, Connie de Vos, Gunter Senft, N. J. Enfield, Rogayah A. Razak, Sebastian Fedden, Sylvia Tufvesson, Mark Dingemanse, Özge Öztürk, Penelope Brown, Clair Hill, Olivier Le Guen, Vincent Hirtzel, Rik van Gijn, Mark A. Sicoli, , and Stephen C. Levinson. Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, 115(45):11369–11376, 2018.

[57] Delwin T Lindsey and Angela M Brown. World color survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences*, 106(47):19785–19790, 2009.

[58] Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi, and Bevil R Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017.

[59] Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Color naming reflects both perceptual structure and communicative need. *arXiv preprint arXiv:1805.06165*, 2018.

[60] Delwin T Lindsey, Angela M Brown, David H Brainard, and Coren L Apicella. Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, 25(18):2441–2446, 2015.

[61] Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. *The world color survey*. CSLI Publications, 2009.

[62] Paul Kay and Luisa Maffi. Color appearance and the emergence and evolution of basic color lexicons. *American Anthropologist*, 101(4):743–760, 1999.

[63] Richard A Hudson. *English word grammar*, volume 108. Basil Blackwell Oxford, 1990.

[64] Carl Pollard and Ivan Sag. *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press and Stanford: CSLI Publications., 1994.

[65] Ronald M. Kaplan and Joan Bresnan. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge, MA, 1982. Reprinted in Mary Dalrymple, Ronald M. Kaplan, John Maxwell, and Annie Zaenen, eds., *Formal Issues in Lexical-Functional Grammar*, 29–130. Stanford: Center for the Study of Language and Information. 1995.

[66] Noam Chomsky. *The Minimalist Program (Current Studies in Linguistics 28).* MIT Press, 1995.

[67] Greville G. Corbett, Norman M. Fraser, and Scott McGlashan, editors. *Heads in Grammatical Theory.* Cambridge University Press, 1993.

[68] Matthew S Dryer. SVO languages and the OV:VO typology. *Journal of Linguistics*, 27(2):443–482, 1991.

[69] Ramon Ferrer i Cancho. The placement of the head that minimizes online memory. *Language Dynamics and Change*, 5(1):114–137, 2015.

[70] August Fenk and Gertraud Fenk. Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27:400–414, 1980.

[71] Dmitriy Genzel and Eugene Charniak. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 199–206. Association for Computational Linguistics, 2002.

[72] Roger Levy and T. Florian Jaeger. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, pages 849–856, 2007.

[73] Luke Maurits, Dan Navarro, and Amy Perfors. Why are some word orders more common than others? a Uniform Information Density account. In *Advances in neural information processing systems*, pages 1585–1593, 2010.

[74] Roger Levy. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Waikiki, Honolulu, 2008.

[75] Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50):21086–21090, 2009.

[76] Edward Gibson, Leon Bergen, and Steven T Piantadosi. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, page 201216438, 2013.

[77] Edward Gibson, Steven T Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7):1079–1088, 2013.

[78] Richard Futrell, Tina Hickey, Aldrin Lee, Eunice Lim, Elena Luchkina, and Edward Gibson. Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition*, 136:215–221, 2015.

[79] Susan Goldin-Meadow, Wing Chee So, Aslı Özyürek, and Carolyn Mylander. The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, 105(27):9163–9168, 2008.

[80] Marieke Schouwstra and Henriëtte de Swart. The semantic origins of word order. *Cognition*, 131(3):431–436, 2014.

[81] Matthew S Dryer. Case distinctions, rich verb agreement, and word order type (comments on Hawkins' paper). *Theoretical Linguistics*, 28(2):151–158, 2002.

[82] Jan Rijkhoff. Explaining word order in the noun phrase. *Linguistics*, 28(1):5–42, 1990.

[83] Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

[84] Thomas Wasow. *Postverbal behavior.* Number 145. CSLI Publications, 2002.

[85] Ramon Ferrer i Cancho. Euclidean distance between syntactically linked words. *Physical Review E*, 70(5):056135, 2004.

[86] Haitao Liu. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191, 2008.

[87] Daniel Gildea and David Temperley. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310, 2010.

[88] Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, 2015.

[89] Haitao Liu, Chunshan Xu, and Junying Liang. Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193, 2017.

[90] David Temperley and Daniel Gildea. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80, 2018.

[91] Daniel Grodner and Edward Gibson. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29(2):261–290, 2005.

[92] Brian Bartek, Richard L Lewis, Shravan Vasishth, and Mason R Smith. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178, 2011.

[93] Theo Vennemann. Theoretical word order studies: results and problems. *Papiere zur Linguistik*, 7(1974):5–25, 1974.

[94] Matthew S Dryer. The Greenbergian word order correlations. *Language*, pages 81–138, 1992.

[95] Lynne M Stallings, Maryellen C MacDonald, and Padraig G O'Seaghdha. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417, 1998.

[96] Hiroko Yamashita and Franklin Chang. "Long before short" preference in the production of a head-final language. *Cognition*, 81(2):B45–B55, 2001.

[97] Ramon Ferrer i Cancho. Why do syntactic links not cross? *EPL (Europhysics Letters)*, 76(6):1228–1234, 2006.

[98] Marco Kuhlmann. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387, 2013.

[99] Y. Albert Park and Roger Levy. Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the 10th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) conference*, pages 335–343, Boulder, Colorado, USA, 2009.

[100] Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, 155:204–232, 2016.

[101] Maryia Fedzechkina, Becky Chu, and T Florian Jaeger. Human information processing shapes language change. *Psychological Science*, 29(1):72–82, 2018.

[102] Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Learning biases predict a word order universal. *Cognition*, 122(3):306–329, 2012.

[103] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[104] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.

[105] Roger Levy. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, 2011.

[106] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.

[107] Richard Futrell and Roger Levy. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 688–698, 2017.

[108] Gregory Scontras, Judith Degen, and Noah D Goodman. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66, 2017.

[109] Michael Hahn, Judith Degen, Noah Goodman, Dan Jurafsky, and Richard Futrell. An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*, 2018.

[110] Johanna Nichols. Head-marking and dependent-marking grammar. *Language*, pages 56–119, 1986.

[111] Greville G Corbett. *Agreement*, volume 109. Cambridge University Press, 2006.

[112] Jennifer E Arnold, Elsi Kaiser, Jason M Kahn, and Lucy K Kim. Information structure: linguistic, cognitive, and processing approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(4):403–413, 2013.

[113] Edward Sapir. *Language: An introduction to the study of speech.* Harcourt, Brace and Co., 1921.

[114] Paul Kiparsky. The rise of positional licensing. In Ans van Kemenade and Nigel VIncent, editors, *Parameters of morphosyntactic change*, pages 460–494. Cambridge University Press, 1997.

[115] Thomas McFadden. On morphological case and word-order freedom. In *Annual Meeting of the Berkeley Linguistics Society*, volume 29, pages 295–306, 2003.

[116] Richard Futrell, Kyle Mahowald, and Edward Gibson. Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, 2015.

[117] Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Mueller-Spitzer. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS One*, 12(3):e0173614, 2017.

[118] Maryia Fedzechkina, T Florian Jaeger, and Elissa L Newport. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, page 201215776, 2012.

[119] Gary Lupyan and Rick Dale. Language structure is partly determined by social structure. *PloS One*, 5(1):e8559, 2010.

[120] Christian Bentz and Aleksandrs Berdicevskis. Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 222–232, 2016.

[121] Greville G. Corbett. Gender, 1991.

[122] Frank Seifart, Jan Strunk, Swintha Danielsen, Iren Hartmann, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich, Nivja H de Jong, and Balthasar Bickel. Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences*, 115(22):5720–5725, 2018.

[123] Richard Futrell. German noun class as a nominal protection device. *Unpublished honors thesis, Stanford University*, 2010.

[124] Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. A functional theory of gender paradigms. *Morphological Paradigms and Functions. Leiden: Brill*, 2016.

[125] Melody Dye, Petar Milin, Richard Futrell, and Michael Ramscar. Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*, 10(1):209–224, 2018.

[126] E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

[127] Nicholas Evans and Stephen C Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448, 2009.

[128] Ray J Solomonoff. A formal theory of inductive inference. part I. *Information and control*, 7(1):1–22, 1964.

[129] M. Li and P.M.B. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, 2008.

[130] Kenny Smith, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda, and Elizabeth Wonnacott. Language learning, language use and the evolution of linguistic variation. *Phil. Trans. R. Soc. B*, 372(1711):20160051, 2017.

[131] Noam Chomsky. Approaching UG from below. In *Interfaces + recursion = language?: Chomsky's Minimalism and the view from syntax-semantics*. Mouton de Gruyter, 2007.

[132] Nick Chater and Paul Vitányi. Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1):19–22, 2003.

[133] Nick Chater and Paul Vitányi. 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology*, 51(3):135–163, 2007.

[134] Robert Berwick, Gabriel Beckers, Kazuo Okanoya, and Johan Bolhuis. A bird's eye view of human language evolution. *Frontiers in Evolutionary Neuroscience*, 4:1–25, 2012.

[135] Steven Pinker and Paul Bloom. Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–727, 1990.

[136] Simon Kirby. Syntax without natural selection. In M. Studdert-Kennedy C. Knight and J.R. Hurford, editors, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pages 303–323. Cambridge Univ Press, Cambridge, UK, 2000.

[137] Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008.

[138] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

[139] Ferdinand de Saussure. Course in general linguistics. *Trans. Roy Harris. London: Duckworth*, 1916.

[140] Willard Van Orman Quine. *Word and object.* MIT Press, 1960.

[141] Pamela Perniss, Robin Thompson, and Gabriella Vigliocco. Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227, 2010.

[142] Mark Dingemanse, Damián E Blasi, Gary Lupyan, Morten H Christiansen, and Padraic Monaghan. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603–615, 2015.

[143] Janis B Nuckolls. The case for sound symbolism. *Annual Review of Anthropology*, 28(1):225–252, 1999.

[144] Benjamin K Bergen. The psychological reality of phonaesthemes. *Language*, 80(2):290–311, 2004.

[145] Jamie Reilly, Jinyi Hung, and Chris Westbury. Non-arbitrariness in mapping word form to meaning: Cross-linguistic formal markers of word concreteness. *Cognitive Science*, 41(4):1071–1089, 2017.

[146] Padraic Monaghan, Richard C Shillcock, Morten H Christiansen, and Simon Kirby. How arbitrary is language? *Phil. Trans. R. Soc. B*, 369(1651):20130299, 2014.

[147] Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823, 2016.

[148] E Dario Gutiérrez, Roger Levy, and Benjamin Bergen. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2379–2388, 2016.

[149] Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T Piantadosi. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145, 2017.

[150] Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T Piantadosi. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8):2149–2169, 2017.

[151] Nelson F. Liu, Gina-Anne Levow, and Noah A. Smith. Discovering phonesthemes with sparse regularization. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 49–54, 2018.

[152] Brent Strickland, Carlo Geraci, Emmanuel Chemla, Philippe Schlenker, Meltem Kelepir, and Roland Pfau. Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, pages 5968–5973, 2015.

[153] Michael H Kelly. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99(2):349–364, 1992.

[154] Mutsumi Imai and Sotaro Kita. The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Phil. Trans. R. Soc. B*, 369:20130298, 2014.

[155] Catherine E Laing. A phonological analysis of onomatopoeia in early word production. *First Language*, 34(5):387–405, 2014.

[156] Lynn K Perry, Marcus Perlman, and Gary Lupyan. Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PloS One*, 10(9):e0137147, 2015.

[157] James Brand, Padraic Monaghan, and Peter Walker. The changing role of sound-symbolism for small versus large vocabularies. *Cognitive Science*, 42:578–590, 2018.

[158] Mutsumi Imai, Sotaro Kita, Miho Nagumo, and Hiroyuki Okada. Sound symbolism facilitates early verb learning. *Cognition*, 109(1):54–65, 2008.

[159] Stanka A Fitneva, Morten H Christiansen, and Padraic Monaghan. From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, 36(5):967–997, 2009.

[160] Catriona Silvey, Simon Kirby, and Kenny Smith. Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39(1):212–226, 2015.

[161] Tessa Verhoef, Seán G Roberts, and Mark Dingemanse. Emergence of systematic iconicity: Transmission, interaction and analogy. In *37th Annual Meeting of the*

Cognitive Science Society (CogSci 2015), pages 2481–2486. Cognitive Science Society, 2015.

[162] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[163] Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer i Cancho. The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19:275–307, 2017.

[164] Sizuo Mizutani. Taiguu hyougen no sikumi (structure of honorific expressions). In *Unyou (The Pragmatics)*. Asakura, 1983.

[165] Brady Z Clark. *A stochastic optimality theory approach to syntactic change*. PhD thesis, Stanford University Ph.D. dissertation, 2004.

[166] Murray Gell-Mann and Merritt Ruhlen. The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, 108(42):17290–17295, 2011.

[167] Luke Maurits and Thomas L Griffiths. Tracing the roots of syntax with bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 111(37):13576–13581, 2014.

[168] Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive science*, 31(3):441–480, 2007.

[169] Daniel J Hruschka, Morten H Christiansen, Richard A Blythe, William Croft, Paul Heggarty, Salikoko S Mufwene, Janet B Pierrehumbert, and Shana Poplack. Building social cognitive models of language change. *Trends in cognitive sciences*, 13(11):464–469, 2009.

[170] Mitchell G Newberry, Christopher A Ahern, Robin Clark, and Joshua B Plotkin. Detecting evolutionary forces in language change. *Nature*, 551(7679):223, 2017.

[171] Joan Bresnan and Jonni M Kanerva. Locative inversion in Chicheŵa: a case study of factorization in grammar. *Linguistic Inquiry*, pages 1–50, 1989.

[172] Matthew S. Dryer and Martin Haspelmath, editors. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013.

[173] Thomas M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ, 2006.

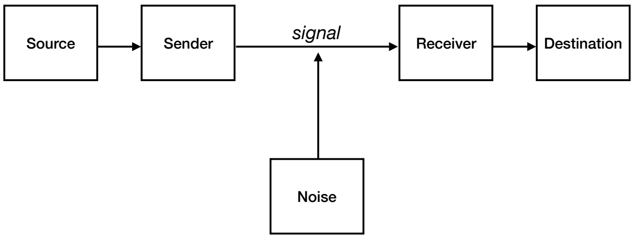


Figure 1. **Information-theoretic picture of communication.**

First, the information source selects a message to be transmitted. Next, the message is encoded into a signal, and that signal is sent to a receiver through some medium called a channel. The receiver then decodes the signal to recover the intended message. Successful communication means that the message recovered at the destination is approximately or exactly equal to the message selected at the source.

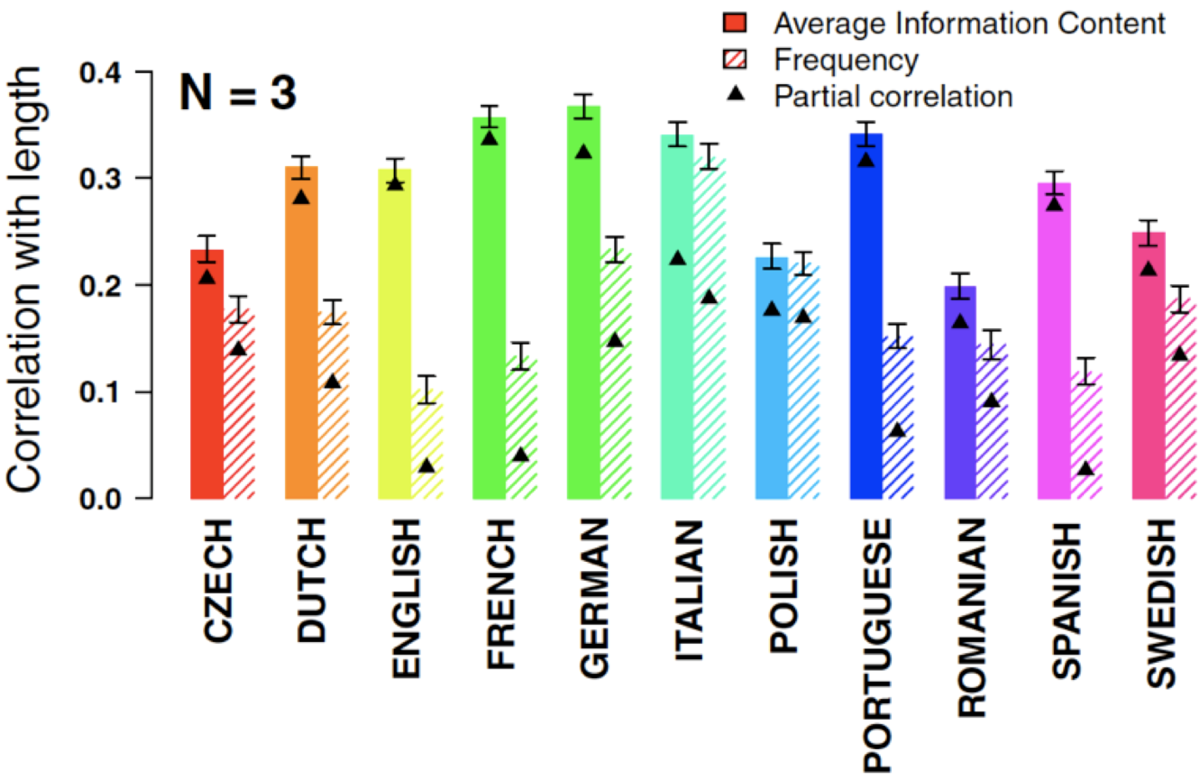


Figure 2. **Words' 3-gram predictability correlates with length, cross-linguistically.**

Solid bars show correlations between word length and a word's predictability according to the two previous words (3-grams). Dashed bars show correlations between word length and word frequency. The predictability correlations are higher than the frequency correlations across languages [33].

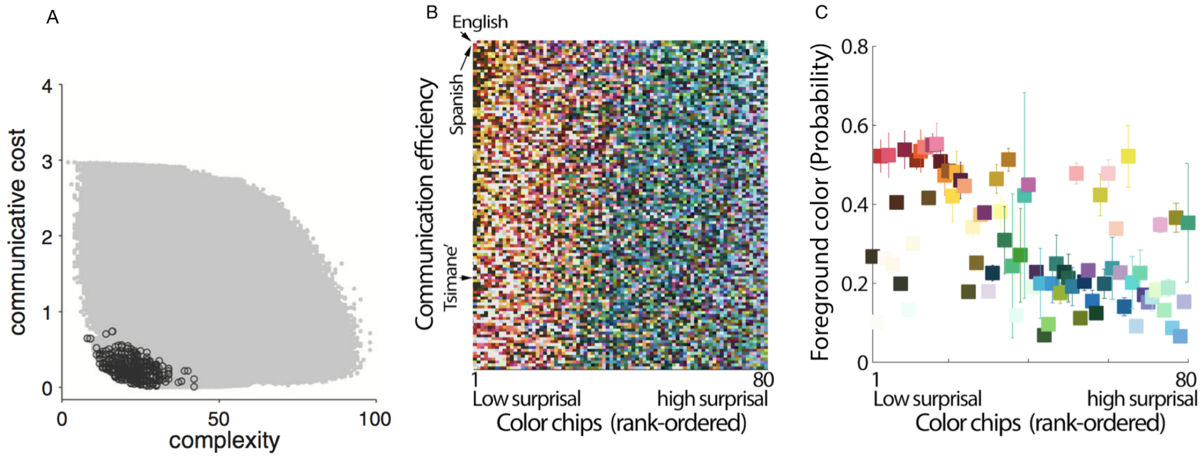


Figure 3. **Efficient word meanings in kinship and color labels cross-linguistically**

A: An analysis by Kemp and Regier [47] suggests that human languages lie on or near an optimal frontier of efficient trade-off between informativeness and complexity in the domain of kinship labels. B: Color chips rank-ordered by their average surprisal for all languages in the world color survey, and English, Bolivian-Spanish and Tsimane' (from Gibson et al. [58]). Each row shows data for a given language, and the languages are ordered according to their overall communication efficiency. C: The color statistics of objects predicts the average surprisal of colors. Probability of a pixel being in a foreground object of the Microsoft Research Asia (MRSA) database of 20,000 natural images compared with the rank-order of average surprisal of a color, in the Tsimane' language. There is a similar relationship for other languages.

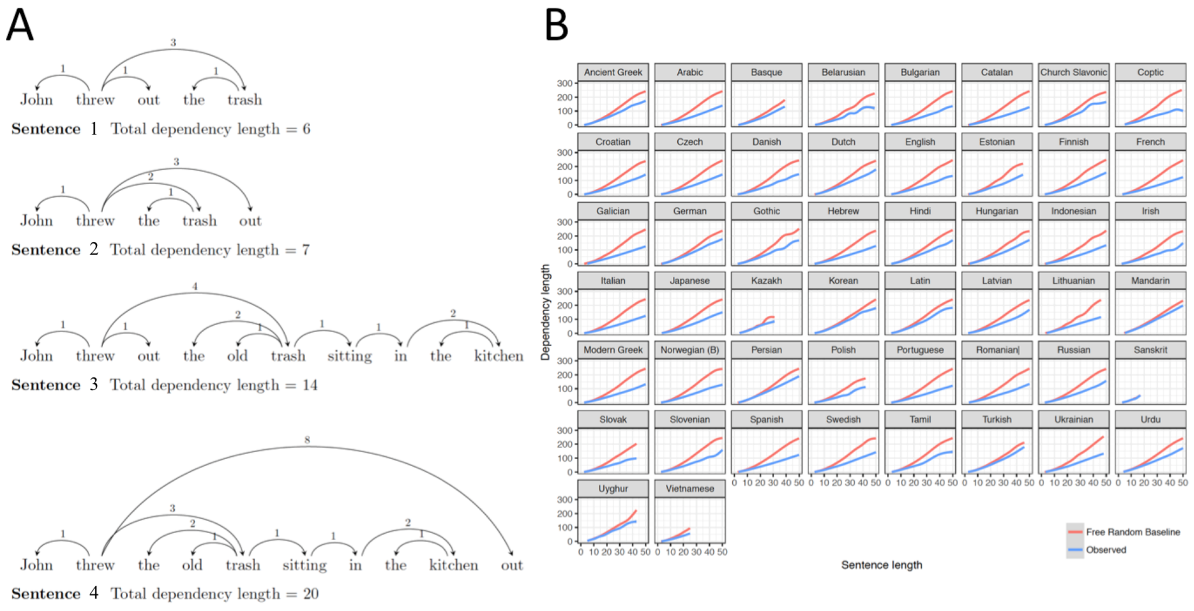


Figure 4. **Languages minimize syntactic dependency lengths cross-linguistically.**

A: Syntactic dependency structures for four English sentences. Sentence 4, with a long-distance dependency between the verb “threw” and the particle “out” is more awkward than sentence 3, with a local dependency between these words, thus illustrating dependency locality. B: For each language analyzed, the average dependency length (blue line) is less than a random baseline (red line) as the sentences get longer [88]. We show results from a replication of [88] on 50 languages.

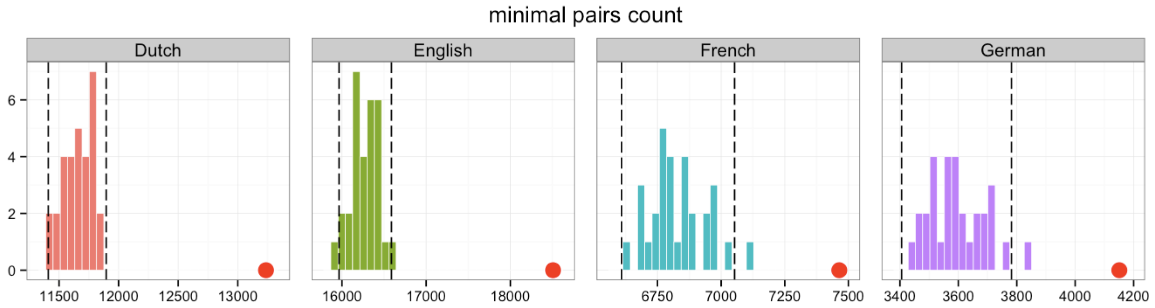


Figure 5. **The lexicon has more minimal pairs than expected**

Dautriche et al. [149] compared the amount of phonological clustering present in the lexicons of 4 languages to a set of simulated lexicons, that acted as a baseline, and that were probabilistically generated based solely on phonotactic and articulatory information. The measure presented here is the total number of **minimal pairs** (pairs of words that differ on only one phoneme) for each language (red dot) together with the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. For all four languages, the lexicon has significantly more minimal pairs than predicted by a phonotactic baseline. Similar results were found across a variety of phonological clustering measures.