nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02336-w

Linguistic structure from a bottleneck on sequential information processing

Received: 1 November 2024

Accepted: 6 October 2025

Published online: 24 November 2025

Check for updates

Richard Futrell [1] & Michael Hahn [2]

Human language has a distinct systematic structure, where utterances break into individually meaningful words that are combined to form phrases. Here we show that natural-language-like systematicity arises in codes that are constrained by a statistical measure of complexity called predictive information, also known as excess entropy. Predictive information is the mutual information between the past and future of a stochastic process. In simulations, we find that codes that minimize predictive information break messages into groups of approximately independent features that are expressed systematically and locally, corresponding to words and phrases. Next, drawing on cross-linguistic text corpora, we find that actual human languages are structured in a way that yields low predictive information compared with baselines at the levels of phonology, morphology, syntax and lexical semantics. Our results establish a link between the statistical and algebraic structure of language and reinforce the idea that these structures are shaped by communication under general cognitive constraints.

Human language is organized around a systematic, compositional correspondence between the structure of utterances and the structure of the meanings that they express[1]. For example, an English speaker will describe an image such as Fig. 1a with an utterance such as 'a cat with a dog', in which the parts of the the image correspond regularly with parts of the utterance such as 'cat'—what we call words. This way of relating form and meaning may seem natural, but it is not logically necessary. For example, Fig. 1b shows an utterance in a hypothetical counterfactual language where meaning is decomposed in a way that most people would find unnatural: here, we have a word 'gol', which refers to a cat head and a dog head together, and another word 'nar', which refers to a cat body and a dog body together. Similarly, Fig. 1c presents a hypothetical language that is systematic but with an unnatural way of decomposing the utterance: here, the utterance contains individually meaningful subsequences 'a cat', 'with' and 'a dog', but these are interleaved together, rather than concatenated as they are in English. We can even conceive of languages such as in Fig. 1d, where each meaning is expressed holistically as a single unanalysable form[2,3]—in fact, this lack of systematic structure is expected in optimal codes like Huffman codes[4,5]. Why is human language the way it is, and not like these counterfactuals?

We argue that the particular structure of human language can be derived from general constraints on sequential information processing. We start from three observations:

(1) Utterances consist, to a first approximation, of one-dimensional sequences of discrete symbols (for example, phonemes).
(2) The ease of production and comprehension of these utterances is influenced by the sequential predictability of these symbols down to the smallest timescales[6–11].
(3) Humans have limited cognitive resources for use in sequential prediction[12–16].

Thus, we posit that language is structured in a way that minimizes the complexity of sequential prediction, as measured using a quantity called predictive information: the amount of information about the past of a sequence that any predictor must use to predict its future[17,18], also called excess entropy[19,20]. Below, we find that codes that are constrained to have low predictive information within signals have systematic structure resembling natural language, and we provide massively cross-linguistic empirical evidence based on large text corpora showing that natural language has lower predictive information than would be expected if it had different kinds of structure.

[1]University of California, Irvine, Irvine, CA, USA. [2]Saarland University, Saarbrücken, Germany. e-mail: rfutrell@uci.edu

https://doi.org/10.1038/s41562-025-02336-w

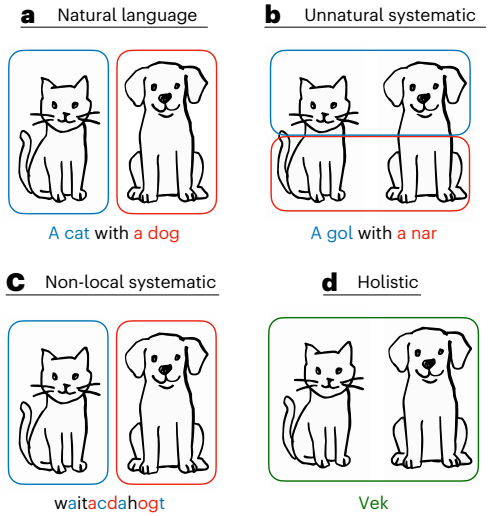

Fig. 1 | Example utterances describing an image in English and various hypothetical languages. a, An English utterance exhibiting natural local systematicity. **b**, An unnatural systematic language in which 'gol' means a cat head paired with a dog head and 'nar' means a cat body paired with a dog body. **c**, A non-local but systematic language in which an utterance is formed by interleaving the words for 'cat' and 'dog'. **d**, A holistic language in which the form 'vek' means 'a cat with a dog' with no correspondence between parts of form and parts of meaning.

Results

Explananda

First, we clarify what we want to explain. Taking a maximally general stance, we think of a language as a function mapping meanings to forms, where meanings are any objects in a set $\mathcal{M}$, and forms are strings drawn from a finite alphabet of letters Σ, typically standing for phonemes. We say a language is systematic when it is a homomorphism[21,22], as illustrated in Fig. 2. That is, if a meaning m can be decomposed into parts (say $m = m_1 \times m_2$), then the string for that meaning decomposes in the same way:

$$L(m_1 \times m_2) = L(m_1) \cdot L(m_2), \tag{1}$$

where '$\cdot$' is some means of combining two strings, such as concatenation. For example, an object ■ would be described in English as L(■) = blue square. The meaning ■ is decomposed into features for colour and shape, and these features are expressed systematically as the words 'blue' and 'square' concatenated together.

We wish to explain why human languages are systematic, why they decompose meanings in the way they do, and why they combine strings in the way they do. In particular, meanings are decomposed in a way that seems natural to humans (that is, like Fig. 1a and not Fig. 1b), a property we call 'naturalness'. Also, strings are usually combined by concatenation (that is, like Fig. 1a and not like Fig. 1c), or more generally by some process that keeps relevant parts of the string relatively close together. We call this property 'locality'.

Influential accounts have held that human language is systematic because language learners need to generalize to produce forms for never-before-seen meanings[23–26]. Such accounts successfully motivate systematicity in the abstract sense, but on their own they do not explain naturalness and locality. However, a theory of systematicity must have something to say about these properties, because if we are free to choose any arbitrary functions '$\times$' and '$\cdot$', then any function L can be considered systematic in the sense of equation (1), and the idea of systematicity becomes vacuous[27].

In existing work, naturalness and locality are explained via (implicit or explicit) inductive biases built into language learners[23,28–35] or stipulations about the mental representation or perception of meanings[36–40]. By contrast, we aim to explain natural local systematicity in language from maximally general principles, without any assumptions about the mental representation of meaning, and with extremely minimal assumptions about the structure of forms—only that they are ultimately expressed as one-dimensional sequences of discrete symbols.

Predictive Information

We measure the complexity of sequential prediction using predictive information, which is the amount of information that any predictor must use about the past of a stochastic process to predict its future (below, we assume familiarity with information-theoretic quantities of entropy and mutual information[41]). Given a stationary stochastic process generating a stream of symbols $\ldots, X_{t-1}, X_t, X_{t+1}, \ldots$, we split it into 'the past' X_{past}, representing all symbols up to time t, and 'the future' X_{future}, representing all symbols at time t or after. The predictive information or excess entropy[18,19] E is the mutual information between the past and the future:

$$E = \mathrm{I}[X_{\text{past}} : X_{\text{future}}]. \tag{2}$$

We calculate the predictive information of a language L as the predictive information of the stream of letters generated by repeatedly sampling meanings $m \in \mathcal{M}$ from a source distribution, translating them to strings as $s = L(m)$ and concatenating them with a delimiter in between.

Predictive information can be calculated in a simple way that gives intuition about its behaviour. Let h_n represent the n-gram entropy of a process, that is, the average entropy of a symbol given a window of $n-1$ previous symbols:

$$h_n = \mathrm{H}[X_t | X_{t-n+1}, \ldots, X_{t-1}]. \tag{3}$$

As the window size increases, the n-gram entropy decreases to an asymptotic value called the entropy rate h. The predictive information represents the convergence to the entropy rate,

$$E = \sum_{n=1}^{\infty} (h_n - h), \tag{4}$$

as illustrated in Fig. 3. This calculation reveals that predictive information is low when symbols can be predicted accurately on the basis of local contexts, that is, when h_n is close to h for small n.

Simulations

The following simulations show that, when languages minimize predictive information, they express approximately independent features

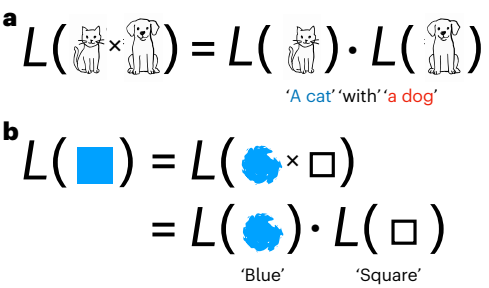

Fig. 2 | Two examples of linguistic systematicity as a homomorphism. $L(\cdot)$ stands for the English language, seen as a function from meanings to forms (strings). **a**, The meaning naturally decomposes into two features corresponding to the two animals. The form 'a cat with a dog' decomposes systematically into forms for the cat and the dog, concatenated together with the string 'with' between them. **b**, The meaning naturally decomposes into two features, corresponding to colour and shape. The form 'blue square' decomposes systematically into forms for the colour and the shape, concatenated together.

https://doi.org/10.1038/s41562-025-02336-w

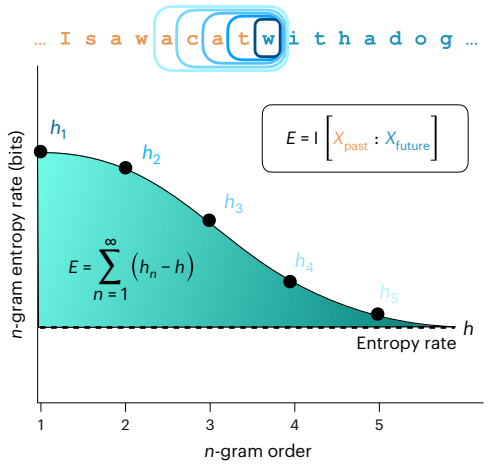

Fig. 3 | Schematic calculation of predictive information as the sum of n-gram entropies h_n minus the asymptotic entropy rate h.

systematically and locally in a way that corresponds to words and phrases in natural language.

Systematic expression of independent features. Consider a set of meanings consisting of the outcomes of three weighted coin flips. In a natural systematic language, we would expect each string to have contiguous 'words' corresponding to the outcome of each individual coin, whereas a holistic language would have no such structure, as shown in the examples in Fig. 4a. It turns out that, for these example languages, the natural systematic one has lower predictive information, as shown in Fig. 4b. In fact, among all possible unambiguous length-3 binary languages, predictive information is minimized exclusively in the systematic languages, as shown in Fig. 4c.

Intuitively, the reason systematic languages minimize predictive information here is that the features of meaning expressed in each individual letter are independent of each other, and so there is no statistical dependence among letters in the string. The general pattern is that an unambiguous language that minimizes predictive information will find features that have minimal mutual information and express them systematically. See Supplementary Section A for formal arguments to this effect.

Holistic expression of correlated components. What happens to predictive information when the source distribution cannot be expressed in terms of fully independent features? In that case, it is better to express the more correlated features holistically, without systematic structure. This holistic mapping is what we find in natural language for individual words (or, more precisely, morphemes), according to the principle of arbitrariness of the sign[42]. For example, the word 'cat' has no identifiable parts that systematically correspond to features of its meaning. Furthermore, as we will discuss below, morphemes in language typically encode categories whose semantic features are highly correlated with each other[43].

We demonstrate this effect in simulations by varying the coin-flip scenario above. Denote the three coin flips as M_1, M_2 and M_3. Imagine the second and third coins and are tied together, so that their outcomes M_2 and M_3 are correlated, as in the example in Fig. 4d. In the limit where M_2 and M_3 are fully correlated, these coin flips have effectively become one feature. Figure 4e shows predictive information for a number of possible languages in this setting, as a function of the mutual information between the tied coin flips M_2 and M_3. In the low-mutual-information regime—where M_2 and M_3 are nearly independent—the best language is still fully systematic. However, as mutual information increases, the best language is one that expresses the tied coin flips M_2 and M_3 together holistically, as a single 'word'. An unnatural language that expresses the uncorrelated coin flips M_1 and M_2 holistically is much worse, as is a non-local systematic language that breaks up the 'word' corresponding to the correlated coin flips M_2 and M_3.

Locality. Next, we show that minimization of predictive information yields languages where features of meaning correspond to localized parts of strings, corresponding to words. We consider a Zipfian distribution over 100 meanings, and a language L in which forms consist of two length-4 'words'. We then consider scrambled languages formed by applying permutations to the string output of L. For example, if the original language expresses a meaning with two words such as $L(m_1 \times m_2)$ = aaaa · bbbb, a possible scrambled language would have $L'(m_1 \times m_2)$ = baaabbab. These scrambled languages instantiate possible string combination functions other than concatenation.

Calculating predictive information for all possible scrambled languages, we find that the languages in which the 'words' remain contiguous have the lowest predictive information, as shown in Fig. 5a. This happens because the coding procedure above creates correlations among letters within a word. When these correlated letters are separated from each other—such as when letters from another word intervene—then predictive information increases. Interestingly, not every concatenative language is better than every non-concatenative one. This corresponds to the reality of natural language, in which limited non-concatenative and non-local morphophonological processes do exist, for example, in Semitic non-concatenative morphology[44].

Hierarchical structure. Natural language sentences typically have well-nested hierarchical syntactic structures, of the kind generated by a context-free grammar[45]: for example, the sentence '[[the big dog] chased [a small cat]]' has two noun phrases, indicated by brackets, which are contiguous and nested within the sentence. Minimization of predictive information creates these well-nested word orders, with phrases corresponding to groups of words that are more or less strongly correlated[46]. We demonstrate this effect using a source distribution defined over six random variables $M_1, \ldots, M_6$ with a covariance structure shown in the inset of Fig. 5b: each of the variable pairs (M_1, M_2) and (M_4, M_5) are highly internally correlated; these pairs are weakly correlated with M_3 and M_6, respectively; and both groups of variables are very weakly correlated with each other. As above, we consider all possible permutations of a systematic code for these source variables. The codes that minimize predictive information are those that are well nested with respect to the correlation structure of the source, keeping the letters corresponding to all groups of correlated features contiguous. Further simulation results involving context-free languages are found in Supplementary Section G. For a mathematical analysis of predictive information in local and random orders for structured sources, see Supplementary Section A.

Cross-linguistic empirical results

Here, we present cross-linguistic empirical evidence that the systematic structure of language has the effect of reducing predictive information at the levels of phonotactics, morphology, syntax and semantics, compared against systems that lack natural local systematicity.

Phonotactics. Languages have restrictions on what sequences of sounds may occur within words: for example, 'blick' seems like a possible English word, whereas 'bnick' does not, even though it is pronounceable in other languages[47]. These systems of restrictions are called phonotactics. Here, we show that actual phonotactic systems of human languages, which involve primarily local constraints on what sounds may co-occur, result in lower predictive information compared with counterfactual phonotactic systems. We compare phonemically transcribed wordforms in vocabulary lists of 61 languages against counterfactual alternatives generated by deterministically scrambling phonemes within a word while preserving manner of

https://doi.org/10.1038/s41562-025-02336-w

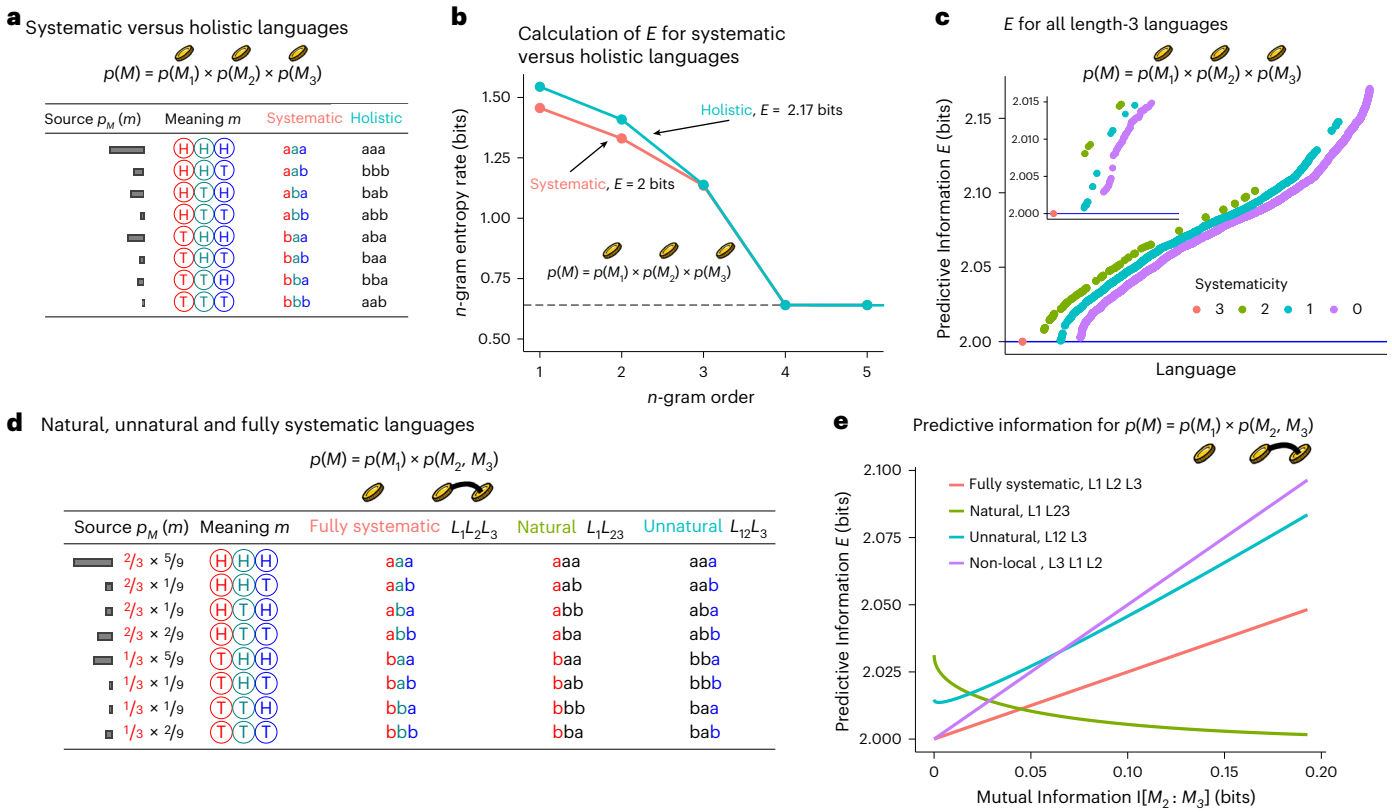

Fig. 4 | Simulations of languages for coin-flip distributions. a, Two unambiguous languages for meanings consisting of three weighted coin flips. In the systematic language, each letter corresponds to the outcome from one coin flip. In the holistic language, there is no natural systematic relationship between the form and the meaning. **b**, Calculation of predictive information for the source and two languages in **a**. The systematic language has lower predictive information. **c**, Predictive information of all bijective mappings from meanings to length-3 binary strings, for the meanings and source in **a**. Languages are ordered by predictive information and coloured by the number of coin flips expressed systematically: 3 for a fully systematic language and 0 for a fully holistic language. The inset box zooms in on the region of low predictive information. **d**, Languages used in **e** along with an example source, which has mutual information $I[M_2 : M_3] \approx 0.18$ bits. **e**, Predictive information of various languages for varying levels of mutual information between coin flips M_2 and M_3 (see text). Zero mutual information corresponds to **b** and **c**. The 'natural' language expresses M_2 and M_3 together holistically. The 'unnatural' language expresses M_1 and M_2 together holistically.

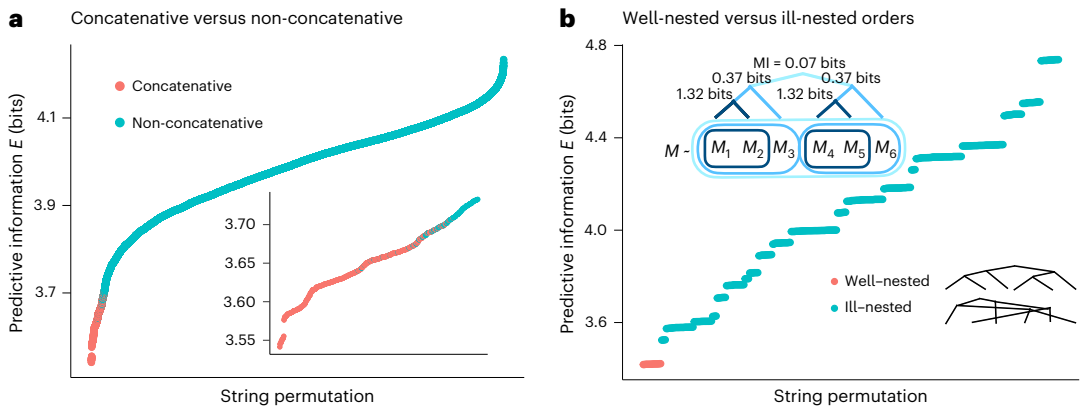

Fig. 5 | Simulations of codes with different orders of elements. a, Predictive information of all string permutations of a systematic language for a Zipfian source. Languages that combine components by concatenation, marked in red, achieve the lowest predictive information. The inset zooms in on the 2,000 permutations with the lowest predictive information. **b**, A hierarchically structured source distribution (see text) and predictive information of all permutations of a systematic language for this source. A language is well nested when all groups of letters corresponding to groupings in the inset tree figure are contiguous. The well-nested languages achieve the lowest predictive information.

articulation. This ensures that the resulting counterfactual forms are roughly possible to articulate. For example, an English word 'fasted' might be scrambled to form 'sefdat'. Calculating predictive information, we find that the real vocabulary lists have lower predictive information than the counterfactual variants in all languages tested. Results for six languages with diverse sound systems are shown in Fig. 6a. Results for the remaining 55 languages are presented in Supplementary Section C.

https://doi.org/10.1038/s41562-025-02336-w

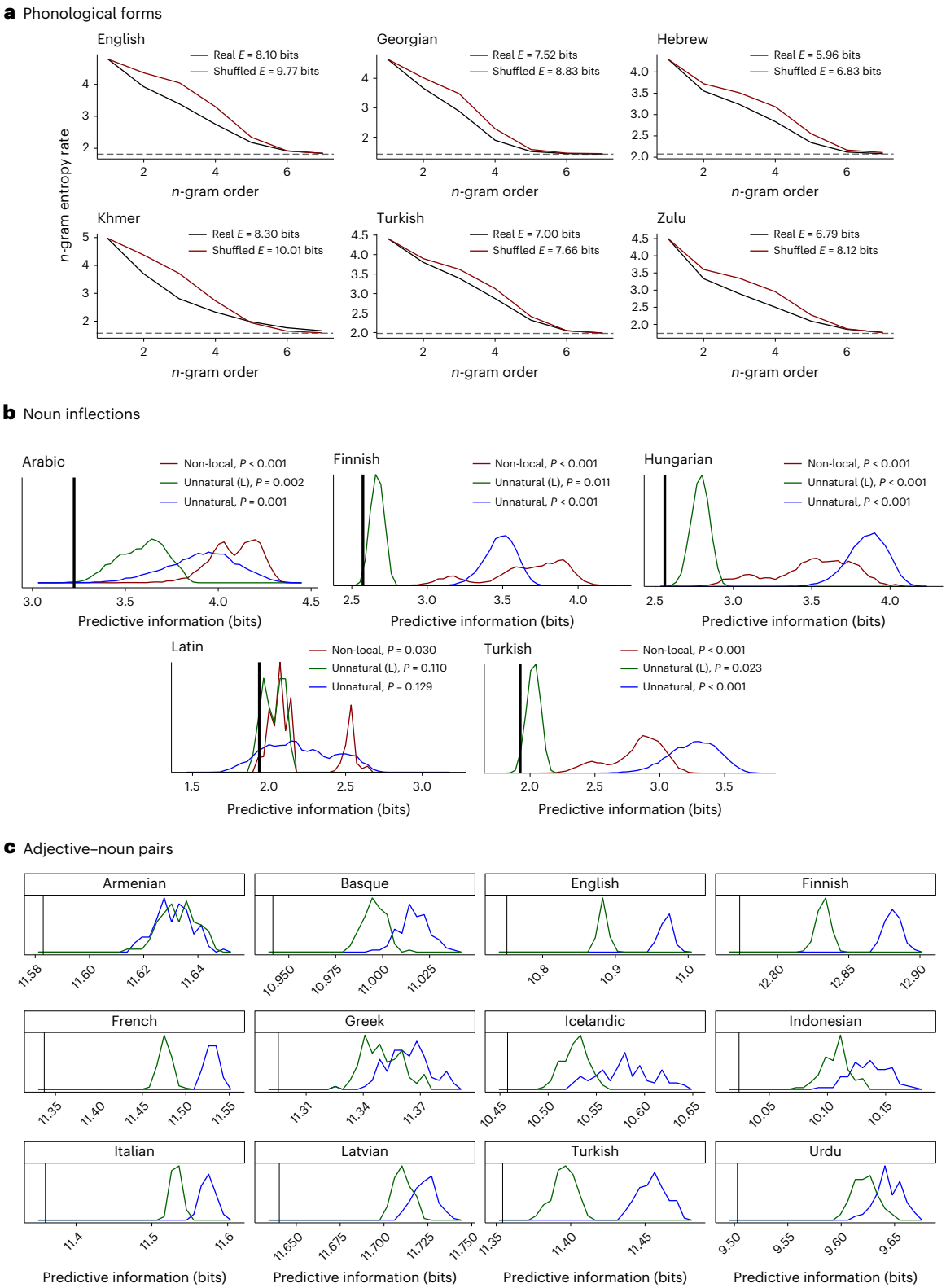

Fig. 6 | Evidence that natural languages are configured in a way that reduces predictive information, in phonotactics, morphology and syntax. **a**, Predictive information calculation for phonological forms in selected languages, comparing the attested forms against forms that have been deterministically shuffled while preserving manner of articulation. **b**, Letter-level predictive information of noun morphology (vertical black line), compared against predictive information values for four random baselines (densities of 10,000 samples; see text). *P* values indicate the proportion of baseline samples with lower predictive information than the attested forms. **c**, Letter-level predictive information of adjective–noun pairs from 12 languages, compared with baselines. Non-local baselines always generate much higher predictive information than the attested forms and are not shown.

https://doi.org/10.1038/s41562-025-02336-w

a Hungarian morphology

Frequency	Features	Hungarian form	Form (unnatural)	Form (non-local)	Meaning
	Nominative singular	ember	emberekhez	ember	person (subj.)
	Accusative singular	embert	embert	embert	person (obj.)
	Dative singular	embernek	embereknek	kembeern	to the person
	Allative singular	emberhez	embereket	zembeerh	toward the person
	Nominative plural	emberek	emberhez	embekre	people (subj.)
	Accusative plural	emberekét	emberek	eembektre	people (obj.)
	Dative plural	embereknek	embernek	knembekere	to the people
	Allative plural	emberekhez	ember	zhembekere	toward the people

b English syntax

Frequency	Meaning	English form	Form (unnatural)	Form (non-local)
		Blue square	Green square	squbluaree
		Green square	Blue square	squgreareen
		Blue circle	Blue circle	cirbluclee
		Green circle	Green circle	cirgrecleen

Fig. 7 | Examples of systematic morphology and syntax, and baselines used in experiments. a, Forms of the Hungarian noun 'ember' meaning 'person', along with examples of the unnatural and non-local baseline used in Fig. 6b. An additional 231 forms are not shown. The 'Frequency' column illustrates the joint frequency of grammatical features in the Hungarian Szeged UD corpus[100,106]. **b**, English forms for the given meanings, along with frequencies from the English Common Crawl web corpus[107]. Example unnatural and non-local baseline forms are shown.

Morphology. Words change form to express grammatical features in a way that is often systematic. For example, the forms of the Hungarian noun shown in Fig. 7a are locally systematic with respect to case and number features. In Fig. 6b, we show that the local systematic structure of affixes for case, number, possession and definiteness in five languages has the effect of reducing predictive information when comparing against baselines that disrupt this structure. We estimate predictive information of these morphological affixes across five languages, with source distributions proportional to empirical corpus counts of the joint frequencies of grammatical features. We compare the predictive information of the attested forms against three alternatives: (1) a non-local baseline generated by applying a deterministic permutation function to each form, (2) an unnatural baseline generated by permuting the assignment of forms to meanings (features) and (3) a more controlled unnatural baseline that permutes the form–meaning mapping while preserving form length. The unnatural baselines preserve the phonotactics of the original forms; only the form–meaning relationship is changed. We generate 10,000 samples (permutations) for each of the three baselines per language.

Across the languages, we find that the attested forms have lower predictive information than the majority of samples of the baselines. The weakest effect is in Latin, which also has the most fusional and least systematic morphology[48]. Note that Arabic nouns often show non-concatenative morphology in the form of so-called broken plurals: for example, the plural of the loanword 'film' meaning 'film' is *'aflām*. This pattern is represented in the forms used to generate Fig. 6b, and yet Arabic noun forms still have lower predictive information than the majority of baseline samples. This suggests that the limited form of non-concatenative morphology present in Arabic is still consistent with the idea that languages are configured in a way that keeps predictive information low.

Syntax. Phrases such as 'blue square' have natural local systematicity, as shown in Fig. 7b. We compare real adjective–noun combinations in corpora of 12 languages against unnatural and non-local baselines generated the same way as in the morphology study: permuting the letters within a form to disrupt locality, or permuting the assignment of forms to meanings to disrupt naturalness. We estimate the probability of a meaning as proportional to the frequency of the corresponding adjective–noun pair. Results are shown in Fig. 6c. The real adjective–noun pairs have lower predictive information than a large majority of baselines across all languages tested.

Word order. In an English noun phrase such as 'the three cute cats', the elements Determiner (D, 'the'), Numeral (N, 'three'), Adjective (A, 'cute') and Noun (n, 'cats') are combined in the order D–N–A–n. This order varies across languages—for example, Spanish has D–N–n–A ('los tres gatos lindos')—but certain orders are more common than others[49]. We aim to explain the cross-linguistic distribution of these orders through reduction of predictive information, which drives words that are statistically predictive of each other to be close to each other, an intuition shared with existing models of adjective order[40,46,50]. To do so, we estimate source probabilities for noun phrases (consisting of single head lemmas for a noun along with an optional adjective, numeral and determiner) based on corpus frequencies. We then calculate predictive information at the word level (treating words as single atomic symbols) for all possible permutations of D–N–A–n. Predictive information is symmetric with respect to time reversal, so we cannot distinguish orders such as D–N–A–n from n–A–N–D and so on. As shown in Fig. 8a, the orders with lower predictive information are also the orders that are more frequent cross-linguistically. A number of alternative source distributions also yield this downward correlation, as shown in Supplementary Section D.

Lexical semantics. Considering a word such as 'cats', all the semantic features of a cat (furriness, mammalianness and so on) are expressed holistically in the morpheme 'cat', while the feature of numerosity is separated into the plural marker '–s'. Plural marking like this is common across languages[51,52]. From reduction of predictive information, we expect relatively uncorrelated components of meaning to be expressed systematically, and relatively correlated components to be expressed together holistically. Thus, we hold that numerosity is selected to be expressed systematically in a separate morpheme because it is relatively independent of the other features of nouns, which are in turn highly correlated with each other. Our theory thus derives the intuition that natural categories arise from the correlational structure of experience[43].

We validate this prediction in a study of semantic features in English, using the Lancaster Sensorimotor Norms[53] to provide semantic features for English words and using the English Universal Dependencies

https://doi.org/10.1038/s41562-025-02336-w

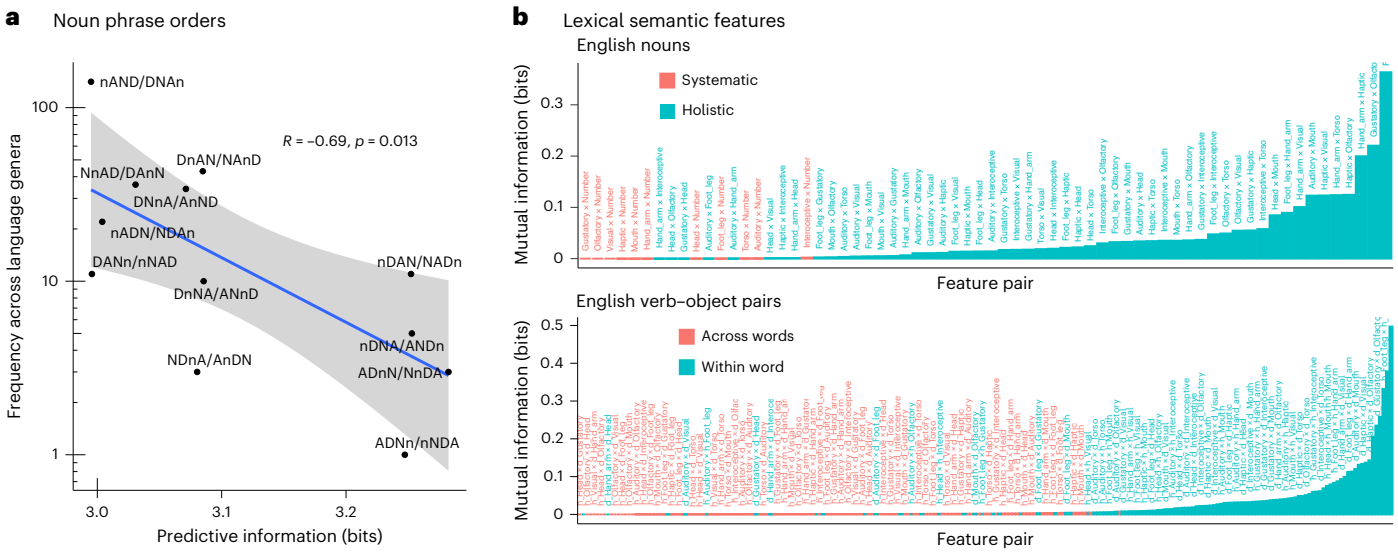

Fig. 8 | Evidence that word order and lexical semantics are configured in ways that reduce predictive information. a, Typological frequency of noun phrase orders (number of unrelated language genera showing the given order[49]) as a function of predictive information. More frequent orders have lower predictive information. The blue line shows a linear regression predicting log typological frequency from predictive information. Error bars indicate a 95% confidence interval of the slope of this regression. The negative correlation is significant with Pearson's $R = -0.69$ and $P = 0.013$. **b**, Top: pairwise mutual information of semantic features from the Lancaster Sensorimotor Norms[53] in addition to a number feature, as indicated by plural morphology. The number feature is expressed systematically; all others are holistic. Bottom: pairwise mutual information values for Lancaster Sensorimotor Norm features across and within words, for pairs of verbs and their objects.

(UD) corpus to provide a frequency distribution over words. The Lancaster Sensorimotor Norms provide human ratings for words based on sensorimotor dimensions, such as whether they involve the head or arms. As shown in Fig. 8b (top), we find that the semantic norm features are highly correlated with each other, and relatively uncorrelated with numerosity, as predicted by the theory.

For the same reason, the theory also predicts that semantic features should be more correlated within words than across words. In Fig. 8b (bottom), we show within-word and cross-word correlations of the semantic norm features for pairs of verbs and their objects taken from the English UD corpus. As predicted, the across-word correlations are weaker. Correlations based on features drawn from other semantic norms are presented in Supplementary Section E.

Discussion

Our results underscore the fundamental roles of prediction and memory in human cognition and provide a link between the algebraic structure of human language and information-theoretic concepts used in machine learning and neuroscience. Our work joins the growing body of information-theoretic models of human language based on resource-rational efficiency[54–59].

Language models

Large language models are based on neural networks trained to predict the next token of text given previous tokens. Our results suggest that language is structured in a way that makes this next-token prediction relatively easy, by minimizing the amount of information that needs to be extracted from the previous tokens to predict the following tokens. Although it has been claimed that large language models have little to tell us about the structure of human language—because their architectures do not reflect formal properties of grammars and because they can putatively learn unnatural languages as well as natural ones[60–62]—our results suggest that these models have succeeded so well precisely because natural language is structured in a way that makes their prediction task relatively simple. Indeed, neural sequence architectures struggle to learn languages that lack information locality[63,64].

Machine learning

Our results establish a connection between the structure of human language and ideas from machine learning. In particular, minimization of mutual information (a technique known as independent components analysis, ICA[65,66]) is widely deployed to create representations that are 'disentangled' or compositional[67], and to detect object boundaries in images, under the assumption that pixels belonging to the same object exhibit higher statistical dependence than pixels belonging to different objects[68]. (Although general nonlinear ICA with real-valued outputs does not yield unique solutions[69], we have found above that minimization of predictive information does find useful structure in our setting, with discrete string-valued outputs and a deterministic function mapping meaning to form.) We propose that human language follows a similar principle: it reduces predictive information, which amounts to performing a generalized sequential ICA on the source distribution on meanings, factoring it into groups of relatively independent components that are expressed systematically as words and phrases, with more statistical dependence within these units than across them. This provides an explanation for why ICA-like objectives yield representations that are intuitively disentangled, compositional, or interpretable: they yield the same kinds of concepts that we find encoded in natural language.

Neuroscience

Similarly, neural codes have been characterized as maximizing information throughput subject to information-theoretic and physiological constraints[70,71], including explicit constraints on predictive information[72,73]. These models predict that, in many cases, neural codes are decorrelated: distinct neural populations encode statistically independent components of sensory input[74]. Our results suggest that language operates on similar principles: it expresses meanings in a way that is temporally decorrelated. This view is compatible with neuroscientific evidence on language processing: minimization of predictive information (while holding overall predictability constant) equates to maximization of local predictability of the linguistic signal, a driver of the neural response to language[10,75].

Information theory and language

Previous work[76] derived locality in natural language from a related information-theoretic concept, the memory–surprisal trade-off or predictive information bottleneck curve, which describes the best achievable sequential predictability as a function of memory usage[77]. The current theory is a simplification that looks at only one part of the curve: predictive information is the minimal memory at which sequential predictability is maximized. A more complete information-theoretic view of language may have to consider the whole curve.

We join existing work attempting to explain linguistic structure on the basis of information-theoretic analysis of language as a stochastic process, for example, the study of lexical scaling laws as a function of redundancy and non-ergodicity in text[78]. Other work on predictive information in language has focused on the long-range scaling of the n-gram entropy in connected texts, with results seeming to imply that the predictive information diverges[79,80]. By contrast, we have focused on only single utterances, effectively considering only relatively short-range predictive information.

Cognitive status of predictive information

Predictive information is a fundamental measure of complexity, which may manifest explicitly or implicitly in various ways in the actual mechanisms of language production, comprehension and learning. For example, in a recent model of online language comprehension[81], comprehenders predict upcoming words on the basis of memory representations that are constrained to store only a small number of words. The fundamental limits of predictive information apply implicitly in this model because comprehenders' predictions cannot be more accurate than if they stored an equivalent amount of predictive information. As another example, a model of language production based on short stored chunks[46] would effectively produce language with low predictive information, because these chunks would be relatively independent of each other, while predictive relationships inside the stored chunks would be preserved. Predictive information has also been linked to difficulty of learning: processes containing more predictive information require more parameters and data to be learned[18], and any learner with limited ability to learn long-term dependencies will have an effective inductive bias towards languages with low predictive information. Predictive information is not meant as a complete model of the constraints on language, which would certainly involve factors beyond predictive information as well as separate, potentially competing pressures from comprehension and production[82].

Relatedly, while we have shown that natural language is configured in a way that keeps predictive information low, we have not speculated on how languages come to be configured in this way, in terms of language evolution and change. We believe there are multiple pathways for this to happen. For example, efficiency pressures in individual interactions could give rise to overall efficient conventions[83], or memory limits in learning[84,85] could cause learners to form low-predictive-information generalizations from their input. Identifying the causal mechanisms that control predictive information in language is a critical topic for future work.

Linguistics

Our theory of linguistic systematicity is independent of theoretical assumptions about mental representations of grammars, linguistic forms or the meanings expressed in language. Predictive information is a function only of the probability distribution on forms, seen as one-dimensional sequences of symbols unfolding in time. This independence from representational assumptions is an advantage, because there is as yet no consensus about the basic nature of the mental representations underlying human language[86,87].

Our results reflect and formalize a widespread intuition about human language, first formulated as Behaghel's Law[88]: 'that which is mentally closely related is also placed close together'. For example, words are contiguous units and the order of morphemes within them is determined by a principle of relevance[89,90], and important aspects of word order across languages have been explained in terms of dependency locality, the principle that syntactically linked words are close[91–94].

A constraint on predictive information predicts information locality: elements of a linguistic form should be close to each other when they predict each other[50]. We propose that information locality subsumes existing intuitive locality ideas. Thus, because words have a high level of statistical interpredictability among their parts[95], they are mostly contiguous, and as a residual effect of this binding force, related words are also close together. Furthermore, we have found that the same formal principle predicts the existence of linguistic systematicity and the way that languages divide the world into natural kinds[37,43].

Limitations

Much work is required to push our hypothesis to its limit. We have assumed throughout that languages are one-to-one mappings between form and meaning; the behaviour of ambiguous or non-deterministic codes, where ambiguity might trade off with predictive information, may yield additional insight. Furthermore, we have examined predictive information only within isolated utterances. It remains to be seen whether reduction of predictive information, applied at the level of many connected utterances, would be able to explain aspects of discourse structure such as the hierarchical organization of topics and topic–focus structure[96].

One known limitation of our theory is that predictive information is symmetric with respect to time reversal, so (at least when applied at the utterance level) it cannot explain time-asymmetric properties of language such as the pattern of 'accessible' (frequent, animate, definite and given) words appearing earlier within utterances than inaccessible ones[97,98]. There is also the fact that non-local and non-concatenative structures do exist in language, for example, long-term coreference relationships among discourse entities, and long-distance filler–gap dependencies, which would seem to contravene the idea that predictive information is constrained. An important area for future research will be to determine what effect these structures really have on predictive information, and what other constraints on language might explain them.

Methods

Constructing a stochastic process from a language

We define a language as a mapping from a set of meanings to a set of strings, $L : \mathcal{M} \to \Sigma^*$. To define predictive information of a language, we need a way to derive a stationary stochastic process generated by that language. We use the following mathematical construction that generates an infinite stream of symbols: (1) meanings $m \sim p_M$ are sampled i.i.d. from the source distribution p_M, (2) each meaning is translated into a string as $s = L(m)$, and (3) the strings s are concatenated end-to-end in both directions with a delimiter $\# \notin \Sigma$ between them. Finally, a string is chosen with probability reweighted by its length, and a time index t (relative to the closest delimiter to the left) is selected uniformly at random within this form.

This construction has the effect of zeroing out any mutual information between symbols with the delimiter between them. Thus, when we compute n-gram statistics, we can treat each form as having infinite padding symbols to the left and right. This is the standard method for collecting n-gram statistics in natural language processing[99].

Three-feature source simulation

For Fig. 4b,c, the source distribution is distributed as a product of three Bernoulli distributions:

$$M \sim \text{Bernoulli}\left(\frac{2}{3}\right) \times \text{Bernoulli}\left(\frac{2}{3}+\varepsilon\right) \times \text{Bernoulli}\left(\frac{2}{3}+2\varepsilon\right), \quad (5)$$

with $\varepsilon = 0.05$.

For Fig. 4e, we need to generate distributions of the form $p(M) = p(M_1) \times p(M_2, M_3)$ while varying the mutual information $I[M_2: M_3]$. We start with the source from equation (5) (whose components are here denoted p_{indep}) and mix it with a source that creates a correlation between M_2 and M_3:

$$p_\alpha(M = ijk) = p_{\text{indep}}(M_1 = i) \times \left[(1-\alpha)\left(p_{\text{indep}}(M_2 = j) \times p_{\text{indep}}(M_3 = k)\right) + \frac{\alpha}{2}\delta_{jk}\right], \quad (6)$$

with $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. The mixture weight α controls the level of mutual information, ranging from 0 at $\alpha = 0$ to at most 1 bit at $\alpha = 1$. A more comprehensive study of the relationship between feature correlation, systematicity and predictive information is given in Supplementary Section B, which examines systematic and holistic codes for a comprehensive grid of possible distributions on the simplex over four outcomes.

Locality simulation

For the simulation shown in Fig. 5a, we consider a source over 100 objects labelled $\{m^{00}, m^{01}, \ldots, m^{99}\}$, following a Zipfian distribution $p(M = m^i) \propto (i+1)^{-1}$. We consider a language based on a decomposition of the meanings based on the digits of their index, with for example m^{89} decomposing into features as $m_1^8 \times m_2^9$. Each utterance decomposes into two 'words' as $L(m_1 \times m_2) = L(m_1) \cdot L(m_2)$, where the word for each feature m^k is a random string in $\{0,1\}^4$, maintaining a one-to-one mapping between features m^k and words.

Hierarchy simulation

For the simulation shown in Fig. 5b, we consider a source M over $5^6 = 15{,}625$ meanings, which may be expressed in terms of six random variables $\langle M_1, M_2, M_3, M_4, M_5, M_6\rangle$ each over five outcomes, with a probability distribution as follows:

$$\begin{aligned} p(M) \;=\; & \alpha q(M_1, M_2, M_3, M_4, M_5, M_6) + (1-\alpha) \\ & \left(\left[\beta q(M_1, M_2, M_3) + (1-\beta)\left[\gamma q(M_1, M_2) + (1-\gamma)\, q(M_1) q(M_2)\right] q(M_3)\right]\right. \\ & \times \left[\beta q(M_4, M_5, M_6) + (1-\beta)\right. \\ & \left.\left.\left[\gamma\, q(M_4, M_5) + (1-\gamma)\, q(M_4) q(M_5)\right] q(M_6)\right]\right), \end{aligned} \quad (7)$$

where $\alpha = 0.01$, $\beta = 0.20$ and $\gamma = 0.99$ are coupling constants, and each $q(\cdot)$ is a Zipfian distribution as above. The coupling constants control the strengths of the correlations shown in Fig. 5b.

Phonotactics

We assume a uniform distribution over forms found in WOLEX. Supplementary Section F shows results for four languages using corpus-based word frequency estimates to form the source distribution, with similar results.

Morphology

We estimate the source distribution on grammatical features (number, case, possessor and definiteness) using the feature annotations from UD corpora, summing over all nouns, with add-1/2 smoothing. The dependency treebanks are drawn from UD v2.8[100]: for Arabic, NYUAD Arabic UD Treebank; for Finnish, Turku Dependency Treebank; for Turkish, Turkish Penn Treebank; for Latin, Index Thomisticus Treebank; for Hungarian, Szeged Dependency Treebank. Forms are represented with a dummy symbol 'X' standing for the stem, and then orthographic forms for suffixes, such as 'Xoknak' for the Hungarian dative plural. For Hungarian, Finnish and Turkish, we use the forms corresponding to back unrounded vowel harmony. For Latin, we use first-declension forms. For Arabic, we use regular masculine triptote forms with a broken plural; to do so, we represent the root using three dummy symbols, and the plural using a common 'broken' form[101], with, for example, 'XaYZun' for the nominative indefinite singular and ''aXYāZun' for the nominative indefinite plural. Results using an alternate broken plural form 'XiYāZun' are nearly identical.

Adjective–noun pairs

From UD corpora, we extract adjective–noun pairs, defined as a head wordform with part-of-speech 'NOUN' modified by an adjacent dependent wordform with relation 'amod' and part-of-speech 'ADJ'. The forms over which predictive information is computed consist of the pair of adjective and noun from the corpus, in their original order, in original orthographic form with a whitespace between them. The source distribution is directly proportional to the frequencies of the forms.

Noun phrase order

The source distribution on noun phrases is estimated from the empirical frequency of noun phrases in the German GSD UD corpus, which has the largest number of such noun phrases among the UD corpora. To estimate this source, we define a noun phrase as a head lemma of part-of-speech 'NOUN' along with the head lemmas for all dependents of type 'amod' (with part-of-speech 'ADJ'), 'nummod' (with part-of-speech 'NUM') and 'det' (with part-of-speech 'DET'). We extract these noun phrase forms from the corpus. When a noun phrase has multiple adjectives, one of the adjectives is chosen randomly and the others are discarded. The result is counts of noun phrases of the form below:

Determiner	Numeral	Adjective	Noun	Count
die	—	—	Hand	234
ein	—	alt	Kind	4
—	drei	—	Buch	2
ein	—	einzigartig	Parfümeur	1
...	...	...	...	...

The source distribution is directly proportional to these counts. We then compute predictive information at the word level over the attested noun phrases for all possible permutations of determiner, numeral, adjective and noun. Typological frequencies are as given by ref. 49.

Semantic features

We binarize the Lancaster Sensorimotor Norms[53] by recoding each norm as 1 if it exceeds the mean value for that feature across all words, and 0 otherwise. Word frequencies are calculated by maximum likelihood based on lemma frequencies in the concatenation of the English GUM[102], GUMReddit[103] and EWT[104] corpora from UD 2.8. The 'Number' feature is calculated based on the value of the 'Number' feature in the UD annotations. Verb–object pairs were identified as a head wordform with part-of-speech 'VERB' with a dependent wordform of relation 'obj' and part-of-speech 'NOUN'.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Unique data required to reproduce our results are available via GitHub at http://github.com/Futrell/infolocality. Corpus count data are drawn from Universal Dependencies v2.8, available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3683. The Lancaster Sensorimotor Norms are available at https://osf.io/7emr6/. Wordform data from the WOLEX database[105] are not publicly available, but a subset can be made available upon request to the authors.

Code availability

Code to reproduce our results is available via GitHub at http://github.com/Futrell/infolocality.

References

1. Frege, G. Gedankengefüge. *Beitr. Philos. Deutsch. Ideal.* **3**, 36–51 (1923).
2. Jespersen, O. *Language: Its Nature, Development, and Origin* (W. W. Norton and Company, 1922).
3. Wray, A. Protolanguage as a holistic system for social interaction. *Lang. Commun.* **18**, 47–67 (1998).
4. Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proc. IRE* **40**, 1098–1101 (1952).
5. Futrell, R. & Hahn, M. Information theory as a bridge between language function and language form. *Front. Commun.* **7**, 657725 (2022).
6. Goldman-Eisler, F. Speech production and language statistics. *Nature* **180**, 1497–1497 (1957).
7. Ferreira, F. & Swets, B. How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *J. Mem. Lang.* **46**, 57–84 (2002).
8. Bell, A., Brenier, J. M., Gregory, M., Girand, C. & Jurafsky, D. Predictability effects on durations of content and function words in conversational English. *J. Mem. Lang.* **60**, 92–111 (2009).
9. Smith, N. J. & Levy, R. P. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
10. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & De Lange, F. P. A hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
11. Ryskin, R. & Nieuwland, M. S. Prediction during language comprehension: what is next? *Trends Cogn. Sci.* **27**, 1032–1052 (2023).
12. Miller, G. A. & Chomsky, N. Finitary models of language users. *Handb. Math. Psychol.* **2**, 419–491 (1963).
13. Bratman, J., Shvartsman, M., Lewis, R. L. & Singh, S. A new approach to exploring language emergence as boundedly optimal control in the face of environmental and cognitive constraints. In *Proc. 10th International Conference on Cognitive Modeling* (eds Salvucci, D. D. & Gunzelmann, G.) 7–12 (Drexel University, 2010).
14. Christiansen, M. H. & Chater, N. The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* **39**, e62 (2016).
15. Futrell, R., Gibson, E. & Levy, R. P. Lossy-context surprisal: an information-theoretic model of memory effects in sentence processing. *Cogn. Sci.* **44**, e12814 (2020).
16. Ferdinand, V., Yu, A. & Marzen, S. Humans are resource-rational predictors in a sequence learning task. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.21.619537 (2024).
17. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **25**, 907–938 (1986).
18. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
19. Crutchfield, J. P. & Feldman, D. P. Regularities unseen, randomness observed: levels of entropy convergence. *Chaos* **13**, 25–54 (2003).
20. Dębowski, Ł. *Information Theory Meets Power Laws: Stochastic Processes and Language Models* (John Wiley & Sons, 2020).
21. Montague, R. Universal grammar. *Theoria* **36**, 373–398 (1970).
22. Janssen, T. M. V. & Partee, B. H. Compositionality. In *Handbook of Logic and Language* (eds van Benthem, J. & ter Meulen, A. G. B.) 417–473 (Elsevier, 1997).
23. Kirby, S. Syntax out of learning: the cultural evolution of structured communication in a population of induction algorithms. In *Advances in Artificial Life* (eds Floreano, D., Nicoud, J.-D. & Mondada, F.) 694–703 (Springer, 1999).
24. Smith, K., Brighton, H. & Kirby, S. Complex systems in language evolution: the cultural emergence of compositional structure. *Adv. Complex Syst.* **6**, 537–558 (2003).
25. Franke, M. Creative compositionality from reinforcement learning in signaling games. In *Evolution of Language: Proc. 10th International Conference (EVOLANG10)* (eds Cartmill, E. A. et al.) 82–89 (World Scientific, 2014).
26. Kirby, S., Tamariz, M., Cornish, H. & Smith, K. Compression and communication in the cultural evolution of linguistic structure. *Cognition* **141**, 87–102 (2015).
27. Zadrozny, W. From compositional to systematic semantics. *Ling. Philos.* **17**, 329–342 (1994).
28. Batali, J. Computational simulations of the emergence of grammar. In *Approaches to the Evolution of Language: Social and Cognitive Bases* (eds Hurford, J. R., Studdert-Kennedy, M. & Knight, C.) 405–426 (Cambridge Univ. Press, 1998).
29. Ke, J. & Holland, J. H. Language origin from an emergentist perspective. *Appl. Ling.* **27**, 691–716 (2006).
30. Tria, F., Galantucci, B. & Loreto, V. Naming a structured world: a cultural route to duality of patterning. *PLoS ONE* **7**, 1–8 (2012).
31. Lazaridou, A., Peysakhovich, A. & Baroni, M. Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations* (2017).
32. Mordatch, I. & Abbeel, P. Emergence of grounded compositional language in multi-agent populations. In *The Thirty-Second AAAI Conference on Artificial Intelligence* (eds Weinberger, K. Q. & McIlraith, S. A.) 1495–1502 (AAAI Press, 2018).
33. Steinert-Threlkeld, S. Toward the emergence of nontrivial compositionality. *Philos. Sci.* **87**, 897–909 (2020).
34. Kuciński, Ł., Korbak, T., Kołodziej, P. & Miłoś, P. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. *Adv. Neural Inf. Process. Syst.* **34**, 23075–23088 (2021).
35. Beguš, G., Lu, T. and Wang, Z. Basic syntax from speech: spontaneous concatenation in unsupervised deep neural networks. In *Proc. Annual Meeting of the Cognitive Science Society* Vol. 46 (2024); https://escholarship.org/uc/item/1ks8q4q9
36. Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. The evolution of syntactic communication. *Nature* **404**, 495–498 (2000).
37. Barrett, J. A. Dynamic partitioning and the conventionality of kinds. *Philos. Sci.* **74**, 527–546 (2007).
38. Franke, M. The evolution of compositionality in signaling games. *J. Logic Lang. Inf.* **25**, 355–377 (2016).
39. Barrett, J. A., Cochran, C. & Skyrms, B. On the evolution of compositional language. *Philos. Sci.* **87**, 910–920 (2020).
40. Culbertson, J., Schouwstra, M. & Kirby, S. From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language* **96**, 696–717 (2020).
41. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2006).
42. de Saussure, F. *Cours de linguistique générale* (Payot, 1916).
43. Rosch, E. Principles of categorization. In *Cognition and Categorization* (eds Rosch, E. & Lloyd, B. B.) 27–48 (Lawrence Elbaum Associates, 1978).
44. McCarthy, J. J. A prosodic theory of nonconcatenative morphology. *Ling. Inquiry* **12**, 373–418 (1981).
45. Chomsky, N. *Syntactic Structures* (Walter de Gruyter, 1957).
46. Mansfield, J. & Kemp, C. The emergence of grammatical structure from inter-predictability. In *A Festschrift for Jane Simpson* (eds O'Shannessy, C. & Gray, J.) 100–120 (ANU Press, 2025).
47. Chomsky, N. & Halle, M. *The Sound Pattern of English* (Harper and Row, 1968).
48. Rathi, N., Hahn, M. & Futrell, R. An information-theoretic characterization of morphological fusion. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 10115–10120 (Association for Computational Linguistics, 2021); https://doi.org/10.18653/v1/2021.emnlp-main.793

49. Dryer, M. S. On the order of demonstrative, numeral, adjective, and noun. *Language* **94**, 798–833 (2018).
50. Futrell, R. Information-theoretic locality properties of natural language. In *Proc. First Workshop on Quantitative Syntax* (eds Chen, X. & Ferrer-i-Cancho, R.) 2–15 (Association for Computational Linguistics, 2019); https://www.aclweb.org/anthology/W19-7902
51. Corbett, G. G. *Number* (Cambridge Univ. Press, 2000).
52. Garner, W. R. *The Processing of Information and Structure* (Lawrence Earlbaum Associates, 1978).
53. Lynott, D., Connell, L., Brysbaert, M., Brand, J. & Carney, J. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods* **52**, 1271–1291 (2020).
54. Ferrer-i-Cancho, R. & Solé, R. V. Least effort and the origins of scaling in human language. *Proc. Natl Acad. Sci. USA* **100**, 788 (2003).
55. Jaeger, T. F. & Tily, H. J. On language 'utility': processing complexity and communicative efficiency. *Wiley Interdisc. Rev. Cogn. Sci.* **2**, 323–335 (2011).
56. Kemp, C. & Regier, T. Kinship categories across languages reflect general communicative principles. *Science* **336**, 1049–1054 (2012).
57. Zaslavsky, N., Kemp, C., Regier, T. & Tishby, N. Efficient compression in color naming and its evolution. *Proc. Natl Acad. Sci. USA* **115**, 7937–7942 (2018).
58. Gibson, E. et al. How efficiency shapes human language. *Trends Cogn. Sci.* **23**, 389–407 (2019).
59. Levshina, N. *Communicative Efficiency* (Cambridge Univ. Press, 2022).
60. Mitchell, J. & Bowers, J. Priorless recurrent networks learn curiously. In *Proc. 28th International Conference on Computational Linguistics* (eds Scott, D., Bel, N. & Zong, C.) 5147–5158 (International Committee on Computational Linguistics, 2020); https://doi.org/10.18653/v1/2020.coling-main.451
61. Chomsky, N., Roberts, I. & Watumull, J. Noam Chomsky: the false promise of ChatGPT. *The New York Times* https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html (2023).
62. Moro, A., Greco, M. & Cappa, S. F. Large languages, impossible languages and human brains. *Cortex* **167**, 82–85 (2023).
63. Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K. & Potts, C. Mission: impossible language models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 14691–14714 (Association for Computational Linguistics, 2024); https://doi.org/10.18653/v1/2024.acl-long.787
64. Someya, T. et al. Information locality as an inductive bias for neural language models. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics* (eds Che, W. et al.) 27995–28013 (Association for Computational Linguistics, 2025); https://doi.org/10.18653/v1/2025.acl-long.1357
65. Ans, B., Hérault, J. & Jutten, C. Architectures neuromimétiques adaptatives: détection de primitives. *Cognitiva* **85**, 593–597 (1985).
66. Bell, A. J. & Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159 (1995).
67. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
68. Isola, P., Zoran, D., Krishnan, D. & Adelson, E. H. Crisp boundary detection using pointwise mutual information. In *Computer Vision–ECCV 2014: 13th European Conference, Proceedings, Part III 13* (eds Fleet, D. et al.) 799–814 (Springer, 2014).
69. Hyvärinen, A. & Pajunen, P. Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw.* **12**, 429–439 (1999).
70. Linsker, R. Self-organization in a perceptual network. *Computer* **21**, 105–117 (1988).
71. Stone, J. V. *Principles of Neural Information Theory: Computational Neuroscience and Metabolic Efficiency* (Sebtel Press, 2018).
72. Bialek, W., De Ruyter Van Steveninck, R. R. & Tishby, N. Efficient representation as a design principle for neural coding and computation. In *2006 IEEE International Symposium on Information Theory* 659–663 (IEEE, 2006).
73. Palmer, S. E., Marre, O., Berry, M. J. & Bialek, W. Predictive information in a sensory population. *Proc. Natl Acad. Sci. USA* **112**, 6908–6913 (2015).
74. Barlow, H. B. Unsupervised learning. *Neural Comput.* **1**, 295–311 (1989).
75. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
76. Hahn, M., Degen, J. & Futrell, R. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychol. Rev.* **128**, 726–756 (2021).
77. Still, S. Information bottleneck approach to predictive inference. *Entropy* **16**, 968–989 (2014).
78. Dębowski, Ł. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Inf. Theory* **57**, 4589–4599 (2011).
79. Dębowski, Ł. Excess entropy in natural language: present state and perspectives. *Chaos* **21**, 037105 (2011).
80. Dębowski, Ł. The relaxed Hilberg conjecture: a review and new experimental support. *J. Quant. Ling.* **22**, 311–337 (2015).
81. Hahn, M., Futrell, R., Levy, R. & Gibson, E. A resource-rational model of human processing of recursive linguistic structure. *Proc. Natl Acad. Sci. USA* **119**, e2122602119 (2022).
82. Dell, G. S. & Gordon, J. K. Neighbors in the lexicon: friends or foes? In *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (eds Schiller, N. O. & Meyer, A.) 9–38 (Mouton De Gruyter, 2003).
83. Hawkins, R. D. et al. From partners to populations: a hierarchical Bayesian account of coordination and convention. *Psychol. Rev.* **130**, 977 (2023).
84. Newport, E. L. Maturational constraints on language learning. *Cogn. Sci.* **14**, 11–28 (1990).
85. Cochran, B. P., McDonald, J. L. & Parault, S. J. Too smart for their own good: the disadvantage of a superior processing capacity for adult language learners. *J. Mem. Lang.* **41**, 30–58 (1999).
86. Jackendoff, R. Linguistics in cognitive science: the state of the art. *Ling. Rev.* **24**, 347–402 (2007).
87. Goldberg, A. E. Constructions work. *Cogn. Ling.* **20**, 201–224 (2009).
88. Behaghel, O. *Deutsche Syntax: Eine geschichtliche Darstellung. Band IV: Wortstellung* (Carl Winter, 1932).
89. Bybee, J. L. *Morphology: A Study of the Relation between Meaning and Form* (John Benjamins, 1985).
90. Givón, T. Isomorphism in the grammatical code: cognitive and biological considerations. *Stud. Lang.* **15**, 85–114 (1991).
91. Hawkins, J. A. *Efficiency and Complexity in Grammars* (Oxford Univ. Press, 2004).
92. Liu, H., Xu, C. & Liang, J. Dependency distance: a new perspective on syntactic patterns in natural languages. *Phys. Life Rev.* **21**, 171–193 (2017).
93. Temperley, D. & Gildea, D. Minimizing syntactic dependency lengths: typological/cognitive universal? *Annu. Rev. Ling.* **4**, 1–15 (2018).

nature human behaviour

Supplementary information

https://doi.org/10.1038/s41562-025-02336-w

Linguistic structure from a bottleneck on sequential information processing

In the format provided by the authors and unedited

Contents

A Basic formal results

In the Main Text, based on numerical simulations, we claimed that codes that minimize predictive information tend to (1) factorize their source distribution into approximately independent components, and (2) express these components systematically in local parts of strings. Here we provide some elementary theorems about codes that minimize predictive information, which illustrate these generalizations. Although a full mathematical analysis of such codes is beyond the scope of the current work, the results below serve to establish their general behavior.

A.1 Forms of predictive information

In the Main Text, we claimed that the predictive information of a stochastic process can be thought of in terms of successive approximations to the entropy rate based on n-gram models with successively larger context size n. This has previously been shown by Crutchfield and Feldman (2003, Prop. 8), among others. Here we provide the same result by means of a different and more direct proof.

Consider a stationary stochastic process generating symbols labelled $\ldots, X_{t-1}, X_t, X_{t+1}, \ldots$ extending into the infinite past and future. Predictive information is defined as the limit of the mutual information of large blocks of M symbols before and N symbols after an arbitrary time index t:

$$E = \lim_{N\to\infty} \lim_{M\to\infty} \mathrm{I}[X_{t-M:t} : X_{t:t+N}], \tag{1}$$

where $X_{a:b} = X_a, \ldots, X_{b-1}$ represents a block of symbols X indexed by an exclusive range. We write Eq. 1 in shorthand as the mutual information between the infinite past $X_{<t}$ and infinite future $X_{\geq t}$ of the process,

$$E = \mathrm{I}[X_{<t} : X_{\geq t}]. \tag{2}$$

Now we can state the theorem relating predictive information to entropy rates derived from n-gram models.

Theorem 1. *The predictive information E can be written as*

$$E = \lim_{N\to\infty} \sum_{n=1}^{N} (h_n - h)\,, \tag{3}$$

where h_n is the n-gram entropy rate

$$h_n = \mathrm{H}[X_t \mid X_{t-n+1:t}], \tag{4}$$

and h is the asymptotic entropy rate

$$h = \lim_{n\to\infty} h_n. \tag{5}$$

Proof. Invoking stationarity, we set $t = 1$ without loss of generality. Using the chain rule for mutual information, we rewrite the predictive information as a sum of conditional mutual informations:

$$E = \lim_{N\to\infty} \lim_{M\to\infty} \sum_{n=1}^{N} \mathrm{I}[X_{1-M:1} : X_n \mid X_{1:n}]. \tag{6}$$

Now we break each mutual information term into a difference of conditional entropies:

$$E = \lim_{N\to\infty} \lim_{M\to\infty} \sum_{n=1}^{N} \left(\mathrm{H}[X_n \mid X_{1:n}] - \mathrm{H}[X_n \mid X_{1-M:n}]\right). \tag{7}$$

Because conditioning reduces entropy, the terms $\mathrm{H}[X_n \mid X_{1-M:n}]$ (which are finite) converge monotonically downward in M, so we may swap the sum and the limit on M. Then, invoking stationarity again, we notice that the resulting two terms are the n-gram entropy rate and the asymptotic entropy rate:

$$E = \lim_{N\to\infty} \sum_{n=1}^{N} \left(\underbrace{\mathrm{H}[X_n \mid X_{1:n}]}_{n\text{-gram entropy rate}} - \underbrace{\lim_{M\to\infty} \mathrm{H}[X_n \mid X_{1-M:n}]}_{\text{asymptotic entropy rate}} \right) \tag{8}$$

$$= \lim_{N\to\infty} \sum_{n=1}^{N} (h_n - h)\,, \tag{9}$$

$$\tag{10}$$

as claimed.

□

A.2 Predictive information for a finite-state source

The following result shows that predictive information is bounded at a constant when a language puts symbols in an order that respects the correlational structure of the source distribution, when the source distribution has the form of a Hidden Markov Model. On the other hand, we will show in Section A.6 that random orders have average predictive information that grows linearly with the sequence length.

Theorem 2. *Let $(S_t)_{t\geq 0}$ be a Hidden Markov Model (HMM) with finite state space $\mathcal{S}$ and finite emission alphabet $\mathcal{A}$ generating a bi-infinite stationary stochastic process $\ldots, X_{-1}, X_0, X_1, \ldots$. Let $L \in \mathbb{N}$, and consider the length-L language given by $X_1 \ldots X_L$. The predictive information is bounded independently of the sequence length L:*

$$\frac{1}{L}\sum_{i=1}^{L} \mathrm{I}\left[X_{1\ldots i} : X_{i+1\ldots L}\right] = O(1) \tag{11}$$

where $O(1)$ contains constants depending on the HMM but not L.

Proof. Let $s_i \in \mathcal{S}$ be the state of the HMM after generating $\ldots X_{i-2}X_{i-1}X_i$. Note that s_i is a random variable with $\mathrm{H}[s_i] \leq \log|\mathcal{S}|$. Further, $\mathrm{I}[X_{1\ldots i} : X_{i+1\ldots L}|s_i] = 0$. Hence, by the Data Processing Inequality, $\mathrm{I}[X_{1\ldots i} : X_{i+1\ldots L}] \leq \mathrm{H}[s_i] \leq \log|\mathcal{S}| = O(1)$ independently of L. □

A.3 Length-2 languages

We now analyze the most basic case of a code that minimizes predictive information, one in which every meaning is expressed in a string of length 2. We find that a code which minimizes predictive information in this setting performs Independent Components Analysis on the source distribution, with the two characters of the output string representing the two maximally independent factors of the source.

Let $\mathcal{M}$ be a set of meanings with source distribution p_M, Σ_1 and Σ_2 be disjoint sets of symbols, and $\mathcal{L}$ be a set of languages defined as bijections $L : \mathcal{M} \to \Sigma_1 \times \Sigma_2$. The predictive information of a language $E(L)$ is the predictive information of the stream of symbols generated by repeatedly sampling meanings from p_M, translating them to strings as $s = L(m)$, and concatenating the resulting strings with a delimiter $\# \notin \Sigma_1, \notin \Sigma_2$ between them.

Theorem 3. *Any language $L^* \in \mathcal{L}$ that achieves $E(L^*) = \min_{L\in\mathcal{L}} E(L)$ has the form*

$$L^*(m) = \ell_1(m) \cdot \ell_2(m), \tag{12}$$

where ℓ_i denotes some mapping $\ell_i : \mathcal{M} \to \Sigma_i$ and where the outputs from ℓ_1 and ℓ_2 have minimal mutual information:

$$\ell_1, \ell_2 = \arg\min \mathrm{I}[\ell_1(M) : \ell_2(M)], \tag{13}$$

with the minimization performed over all mappings $\mathcal{M} \to \Sigma_i$.

Proof. Because the languages have strings of length 2, we calculate predictive information as

$$E = h_1 + h_2 + h_3 - 3h, \tag{14}$$

up to length 3, accounting for the delimiter # attached after the end of the string. The entropy rate $h = \frac{1}{3}\mathrm{H}[M]$ is constant across all languages because they are all bijections, so we ignore the entropy rate going forward. Furthermore, there is no decrease in n-gram entropy rate for $n > 3$, so we have $h_3 = h$. Dropping all irrelevant constants, E is thus

$$E \sim h_1 + h_2. \tag{15}$$

Calculation of h_1: The unigram entropy rate is the entropy of the distribution over symbols generated by first sampling a time index t relative to the most recent delimiter, and then looking at

the symbol at that position. For a code of length T (including the delimiter to the right), this is

$$\begin{aligned}
h_1 &= -\sum_{t=1}^{T} p(t) \sum_{x \in \Sigma_t} p(X_t = x) \log p(t) p(X_t = x) \\
&= -\frac{1}{T} \sum_{t=1}^{T} \sum_{x \in \Sigma_t} p(X_t = x) \log \frac{1}{T} p(X_t = x) \\
&= -\frac{1}{T} \sum_{t=1}^{T} \log \frac{1}{T} - \frac{1}{T} \sum_{t=1}^{T} \sum_{x \in \Sigma_t} p(X_t = x) \log p(X_t = x) \\
&= \log T + \frac{1}{T} \sum_{t=1}^{T} \mathrm{H}[X_t],
\end{aligned}$$

that is, a constant reflecting how much information is contained in each symbol about its position in the string, plus the average entropy of symbols found in each position. Ignoring constants not affected by the choice of language L, in our case with $T = 3$ this is

$$h_1 \sim \mathrm{H}[X_1] + \mathrm{H}[X_2] + \underbrace{\mathrm{H}[X_3]}_{=0}, \tag{16}$$

where $\mathrm{H}[X_3] = 0$ because we always have $X_3 = \#$.

Calculation of h_2: The bigram entropy rate h_2 can be calculated following the same logic, yielding

$$h_2 \sim \underbrace{\mathrm{H}[X_1 \mid X_0]}_{=\mathrm{H}[X_1]} + \mathrm{H}[X_2 \mid X_1] + \underbrace{\mathrm{H}[X_3 \mid X_2]}_{=0}, \tag{17}$$

where $H[X_1 \mid X_0] = H[X_1]$ because X_0 is the left delimiter, which is uninformative about the value of X_1.

Putting these together and ignoring irrelevant constants yields

$$\begin{aligned}
E &\sim h_1 + h_2 && (18)\\
&\sim \mathrm{H}[X_1] + \mathrm{H}[X_2] + \mathrm{H}[X_1] + \mathrm{H}[X_2 \mid X_1] && (19)\\
&= \mathrm{H}[X_1] + \mathrm{H}[X_2 \mid X_1] + \mathrm{I}[X_1 : X_2] + \mathrm{H}[X_1] + \mathrm{H}[X_2 \mid X_1] && (20)\\
&= 2\,\mathrm{H}[X_1] + 2\,\mathrm{H}[X_2 \mid X_1] + \mathrm{I}[X_1 : X_2] && (21)\\
&= 2\,\mathrm{H}[X_1, X_2] + \mathrm{I}[X_1 : X_2] && (22)\\
&= 2\,\mathrm{H}[M] + \mathrm{I}[X_1 : X_2]. && (23)
\end{aligned}$$

Thus, we are left with

$$E \sim \mathrm{I}[X_1 : X_2], \tag{24}$$

where all remaining constants do not depend on the choice of language L. Without loss of generality, we can write $X_1 = \ell_1(M)$ and $X_2 = \ell_2(M)$ for any language L with the appropriate choice of the ℓ_1, ℓ_2, and thus we have that minimal predictive information is achieved by finding functions ℓ_1, ℓ_2 to minimize mutual information:

$$\arg\min \mathrm{I}[\ell_1(M) : \ell_2(M)]. \tag{25}$$

□

Remark. As predictive information is symmetrical with respect to time reversal, the solutions here are symmetric with respect to swapping ℓ_1 and ℓ_2.

Remark. The argument reveals that there is a degenerate solution when $|\Sigma_i| \geq |\mathcal{M}|$: you could encode the source M entirely with ℓ_i, with the other $\ell_{j\neq i}$ a constant function. In that case it is always possible to achieve $\mathrm{I}[\ell_1(M) : \ell_2(M)] = 0$. This result mirrors the claim from Nowak et al. (2000) that combinatorial communication requires that the number of available signals is less than the number of available meanings.

A.4 Length-3 languages

We now consider codes consisting of strings of length 3. We find that, in this setting, the *order* of the characters in the string is determined by information locality: the non-adjacent characters should be maximally uncorrelated, while the adjacent characters may be more correlated.

Now consider bijective languages $L : \mathcal{M} \to \Sigma_1 \times \Sigma_2 \times \Sigma_3$ producing strings of length 3, with the alphabets Σ_i all disjoint. Now we no longer have invariance with respect to interchanging the features ℓ_1, ℓ_2, ℓ_3: the order in which features are expressed now matters. Below, we show that languages which minimize E order these features so as to minimize the mutual information of the nonlocal features ℓ_1 and ℓ_3.

Theorem 4. *Any length-3 language $L^* \in \mathcal{L}$ that achieves $E(L^*) = \min_{L\in\mathcal{L}} E(L)$ has the form*

$$L^*(m) = \ell_1(m) \cdot \ell_2(m) \cdot \ell_3(m) \tag{26}$$

where the functions $\{\ell_i\}$ are ordered so that $\mathrm{I}[\ell_1(M) : \ell_3(M)]$ is minimal.

Proof. Dropping irrelevant constants in the length-3 case yields

$$E \sim \mathrm{I}[X_1 : X_2] + \mathrm{I}[X_2 : X_3] + 2\,\mathrm{I}[X_1 : X_3 \mid X_2]. \tag{27}$$

This expression can be written out and then rearranged as so:

$$\begin{aligned}
E &\sim \mathbb{E}\left[\ln \frac{p(X_1, X_2)}{p(X_1)p(X_2)}\right] + \mathbb{E}\left[\ln \frac{p(X_2, X_3)}{p(X_2)p(X_3)}\right] + 2\,\mathbb{E}\left[\ln \frac{p(X_1, X_2, X_3)p(X_2)}{p(X_1, X_2)p(X_2, X_3)}\right] && (28)\\
&= \mathbb{E}\left[\ln \frac{p(X_1, X_2, X_3)p(X_1, X_2, X_3)p(X_2)p(X_2)}{p(X_1)p(X_2)p(X_3)p(X_1, X_2)p(X_2, X_3)}\right] && (29)\\
&= \mathbb{E}\left[\ln \frac{p(X_1, X_2, X_3)}{p(X_1)p(X_2)p(X_3)}\right] + \mathbb{E}\left[\ln \frac{p(X_1, X_3 \mid X_2)}{p(X_1 \mid X_2)p(X_3 \mid X_2)}\right] && (30)\\
&= \mathrm{TC}[X_1 : X_2 : X_3] + \mathrm{I}[X_1 : X_3 \mid X_2] && (31)\\
&= \underbrace{\mathrm{TC}[X_1 : X_2 : X_3] - \mathrm{I}[X_1 : X_2 : X_3]}_{\text{Order-independent}} + \underbrace{\mathrm{I}[X_1 : X_3]}_{\text{Order-dependent}}, && (32)
\end{aligned}$$

where $\mathrm{TC}[\cdot : \cdot : \cdot]$ is total correlation (Watanabe, 1960) and $\mathrm{I}[\cdot : \cdot : \cdot]$ is multivariate mutual information (McGill, 1955). Both the TC term and the multivariate mutual information term are invariant to permutations, so the ordering of X_1, X_2, X_3 does not matter for them. The only term that depends on the order of symbols is $\mathrm{I}[X_1 : X_3]$. Thus any candidate optimal language L may be improved by permuting the functions ℓ_1, ℓ_2, ℓ_3 to minimize $I[\ell_1(M) : \ell_3(M)]$. □

Remark. The multivariate information term $\mathrm{I}[X_1 : X_2 : X_3]$ may be positive or negative. If it is positive, the situation is called redundancy. If it is negative, the situation is called synergy. The result above shows that codes with synergy among the three symbols X_1, X_2, X_3 are dispreferred, and codes with redundancy are preferred.

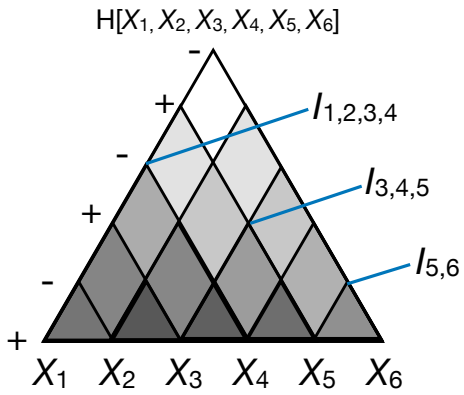

Figure 1: Schematic for coinformation in a set of 6 random variables, based on Bell (2003, Fig. 2). The joint entropy of $X_1, \ldots, X_6$ may be found by summing all the coinformations of all the strict subsets of these variables, weighted by the signs given to the left of the triangle. A few coinformations are highlighted. The true lattice of coinformations is a 6D Boolean hypercube; the figure shows a 2D reduction for visual clarity.

A.5 Length-T languages

Next, we consider the more general case of languages with utterance of a fixed length T, maintaining the setting where each position in the string has symbols from disjoint alphabets. We show that the predictive information for these languages may be expressed in terms of the coinformation lattice (Bell, 2003) among random variables corresponding to positions in the string. We find that predictive information is a function of the amount of coinformation in sets of variables and the *span size* of those sets, defined as the linear distance from the first character to the last character in the set. This gives a generalized form of information locality, where predictive information is low whenever any *set* of characters with high *synergy* are all close to each other.

Before stating the result, it is helpful to review the concept of coinformation. Consider a set of T random variables $X_1, \ldots, X_T$, and a set of indices such as, for example, $E = \{2, 3, 4\}$. Let X_E denote the random variables indexed by the set E, for example $X_E = \{X_2, X_3, X_4\}$. The **coinformation** among the random variables indexed by E is defined as

$$I_E = -\sum_{F \subseteq E} (-1)^{|F|} \, \mathrm{H}[X_F], \tag{33}$$

that is, the sum of entropies of all the subsets of X_E, weighted by 1 if the subset is of odd cardinality and -1 if the subset of is of even cardinality. For example, for $E = \{2, 3, 4\}$, the coinformation is

$$I_{2,3,4} = H[X_2] + H[X_3] + H[X_4] - H[X_2, X_3] - H[X_3, X_4] - H[X_2, X_4] + H[X_2, X_3, X_4]. \tag{34}$$

The coinformation generalizes entropy and mutual information. For a single variable, for example $E = \{1\}$, we recover the univariate entropy, $I_1 = \mathrm{H}[X_1]$. For two random variables, for example $E = \{1, 2\}$, we recover mutual information: $I_{1,2} = \mathrm{H}[X_1] + \mathrm{H}[X_2] - \mathrm{H}[X_1, X_2] = \mathrm{I}[X_1 : X_2]$. Coinformation for a set of variables is organized in a lattice structure, as illustrated in Figure A.5.

For $|E|$ odd (except for $|E| = 1$), coinformation can be negative, which corresponds to **synergy**, which happens for three variables when $I[X_1 : X_2 \mid X_3] > I[X_1 : X_2]$. Positive coinformation for an

odd number of variables corresponds to **redundancy**, which occurs when $I[X_1 : X_2 \mid X_3] < I[X_1 : X_2]$.

Coinformation may be interpreted as the amount of covariance among variables that cannot be detected from any strict subset of those variables. Synergy reflects a case when there is *more* such covariance, and redundancy reflects a case where there is less. Therefore, we can make the quantity of coinformation somewhat more interpretable by transforming it to **synergistic information** S, which makes synergy positive and redundancy negative:

$$S_E = (-1)^{|E|} I_E, \tag{35}$$

that is, to define synergistic information, we reverse the sign of coinformation for odd-numbered sets of variables. Synergistic information is positive when there is synergy among an odd-numbered set of variables, negative when there is redundancy, and positive when there is any coinformation among an even-numbered set of variables.

We can now state our result about predictive information in languages consisting of strings of length T.

Theorem 5. *For a language generating strings of fixed length T and disjoint alphabets for each string position, the predictive information E up to additive and multiplicative constants is*

$$E \sim \sum_{1 \le a < \cdots < z \le T} (z - a)\, S_{a,\ldots,z}, \tag{36}$$

where $S_{a,\ldots,z}$ is the synergistic information among the set of random variables $\{X_a, \ldots, X_z\}$ corresponding to characters at positions $a, \ldots, z$.

Proof. Up to additive and multiplicative constants, the predictive information in this language is

$$E \sim \sum_{t=1}^{T} \mathrm{I}[X_1, \ldots, X_t : X_{t+1}, \ldots, X_T] \tag{37}$$

$$= \sum_{t=1}^{T} \left(\mathrm{H}[X_1, \ldots, X_t] + \mathrm{H}[X_{t+1}, \ldots, X_T] - \mathrm{H}[X_1, \ldots, X_T]\right). \tag{38}$$

Next, we note that, inverting the definition of coinformation, we can write the entropy in a set of N variables as (Bell, 2003, p. 922)

$$\mathrm{H}[X_1, \ldots, X_N] = -\sum_{1 \le a < \cdots < z \le N} (-1)^{|a,\ldots,z|} I_{a,\ldots,z} \tag{39}$$

$$= -\sum_{1 \le a < \cdots < z \le N} S_{a,\ldots,z}, \tag{40}$$

where $a, \ldots, z$ is a set of indices defining a subset of the variables $X_1, \ldots, X_N$. We can use this to rewrite the predictive information in terms of synergistic information:

$$E \sim \sum_{t=1}^{T} \left(-\sum_{1 \le a < \cdots < z \le t} S_{a,\ldots,z} - \sum_{t+1 \le a < \cdots < z \le T} S_{a,\ldots,z} + \sum_{1 \le a < \cdots < z \le T} S_{a,\ldots,z} \right). \tag{41}$$

The question now is how many times we are adding in each synergistic information term $S_{a,\ldots,z}$ to get the total. We can imagine the whole expression as a sum over cut points t which split the string

into two parts, left and right. Within this sum, for each cut point, the last term adds in a synergistic information term $S_{a,\ldots,z}$ for each subset of indices $a, \ldots, z$, and the first two terms subtract all of the synergistic information terms whose indices are either entirely to the left of the cut point or to its right, leaving only those terms whose indices 'straddle' the cut, in the sense that at least one index is $\leq t$ and at least one index is $> t$. Thus, we can rewrite predictive information using an indicator variable for whether the set of indices $a, \ldots, z$ straddles the cut t. Then we count how often this indicator variable is equal to 1, yielding the result:

$$E \sim \sum_{t=1}^{T} \sum_{1 \leq a < \cdots < z \leq T} 1_{a \leq t < z} S_{a,\ldots,z} \tag{42}$$

$$= \sum_{1 \leq a < \cdots < z \leq T} \left(\sum_{t=1}^{T} 1_{a \leq t < z} \right) S_{a,\ldots,z} \tag{43}$$

$$= \sum_{1 \leq a < \cdots < z \leq T} (z - a)\, S_{a,\ldots,z}. \tag{44}$$

□

Remark. It can easily be checked that the formula recovers Eq. 24, which was used in the proof of Theorem 2, for $T = 2$.

Remark. This result goes some way toward linking the hierarchical and well-nested structure of human language with predictive information. In fixed-length languages that minimize predictive information, *groups* of words or letters will tend to be close to each other as a function of how much they covary, in a way that is nested according to the structure of the coinformation lattice. Ill-nested configurations, in which groups of variables with high synergistic information are placed in such a way that other variables intervene, would contribute more to the predictive information, since the synergistic information in groups of variables is weighted by the span of those variables.

A.6 Predictive information for a random permutation

The following result shows that random orders have average predictive information that grows linearly with the sequence length. This is in contrast to our results from Section A.2 showing that, for finite-state processes, the predictive information is bounded independently of L.

Theorem 6. *Let $\ldots, X_{-1}, X_0, X_1, \ldots$ be a bi-infinite stationary process. Let $L \in \mathbb{N}$, and consider the length-L language given by $X_1 \ldots X_L$. Assume the process contains predictive information beyond its ergodic components (Dębowski, 2009), in the sense that:*

$$\inf_{\Delta > 0} \mathrm{I}[X_w : X_{\ldots w-\Delta}] < \mathrm{I}[X_w : X_{\ldots,w-2,w-1}] \tag{45}$$

Consider the uniform distribution over bijections $\rho : [1, \ldots, L] \to [1, \ldots, L]$. Then

$$\mathbb{E}_\rho \left[\frac{1}{L} \sum_{i=1}^{L} \mathrm{I}\left[X_{\rho(1\ldots i)} : X_{\rho(i+1\ldots L)} \right] \right] = \Theta(L) \tag{46}$$

where the expectation describes an average over all bijections ρ, and constants in $\Theta(L)$ depend on the HMM but not L.

The intuition is that for any process with local statistical structure, beyond its ergodic components, permutations of the positions will tend to disrupt this local structure and create long-range dependencies.

Proof. The expectation is evidently $O(L)$; we need to show it is $\Omega(L)$. Define $A = \rho(1 \ldots i)$, $B = \rho(i+1 \ldots L)$. The proof idea is to focus attention on positions w where $w \in A$ but a contiguous sequence of positions to its left is in B. Such situations create opportunity for X_B to provide predictive information about X_A. Formally, for any $\Delta > 0$:

$$
\begin{aligned}
&\mathbb{E}[\mathrm{I}[X_A : X_B]] \\
&= \mathbb{E}\left[\sum_{w \in A} \mathrm{I}[X_w : X_{j \in B} | X_{j<w, j \in A}]\right] \\
&= \sum_{w=1}^{L} \mathbb{E}\left[1_{w \in A} \mathrm{I}[X_w : X_{j \in B} | X_{j<w, j \in A}]\right] \\
&\geq \sum_{w=1}^{L} \mathbb{E}\left[1_{w \in A} 1_{[w-\Delta, w-1] \cap A = \emptyset} \mathrm{I}[X_w : X_{j \in B} | X_{j<w, j \in A}]\right] \\
&\geq \sum_{w=1}^{L} \mathbb{E}\left[1_{w \in A} 1_{[w-\Delta, w-1] \cap A = \emptyset} \mathrm{I}[X_w : X_{[w-\Delta, w-1]} | X_{j<w, j \in A}]\right] \\
&= \sum_{w=1}^{L} \mathbb{E}\left[1_{w \in A} 1_{[w-\Delta, w-1] \cap A = \emptyset} \mathrm{I}[X_w : X_{[w-\Delta, w-1]} | X_{j<w-\Delta, j \in A}]\right] \\
&= \sum_{w=1}^{L} p_\rho(w \in A; [w-\Delta, w-1] \cap A = \emptyset) \mathbb{E}\left[\mathrm{I}[X_w : X_{[w-\Delta, w-1]} | X_{j<w-\Delta, j \in A}] | w \in A, 1_{[w-\Delta, w-1] \cap A = \emptyset}\right]
\end{aligned}
$$

We now need to show that, for large Δ,

$$
\mathrm{I}[X_i : X_{[w-\Delta, w-1]} | X_{j<w-\Delta, j \in A}] \tag{47}
$$

is bounded away from 0 uniformly over A. Consider[1]

$$
\begin{aligned}
&\mathrm{I}[X_w : X_{[w-\Delta, w-1]} \mid X_{j<w-\Delta, j \in A}] \\
&= \mathrm{I}[X_w : X_{[w-\Delta, w-1]}] - \mathrm{I}[X_w : X_{j<w-\Delta, j \in A}] + \mathrm{I}[X_w : X_{j<w-\Delta, j \in A} \mid X_{[w-\Delta, w-1]}] \\
&\geq \mathrm{I}[X_w : X_{[w-\Delta, w-1]}] - \mathrm{I}[X_w : X_{j<w-\Delta, j \in A}] \\
&\geq \mathrm{I}[X_w : X_{[w-\Delta, w-1]}] - \mathrm{I}[X_w : X_{j<w-\Delta}]
\end{aligned}
$$

When $\Delta \to \infty$, the first term converges to $\mathrm{I}[X_w | X_{\ldots w-2, w-1}]$. By assumption, the difference between this and the second term is strictly greater than zero. Overall, this shows (47) is bounded strictly

[1] Reflecting the general identity

$$
\begin{aligned}
\mathrm{I}[A : B \mid C] &= H[A \mid C] - H[A \mid C, B] \\
&= H[A] - H[A \mid B] - H[A] + H[A \mid C] + H[A \mid B] - H[A \mid B, C] \\
&= \mathrm{I}[A : B] - \mathrm{I}[A : C] + \mathrm{I}[A : C \mid B]
\end{aligned}
$$

away from zero independently of A, for some sufficiently large Δ which we henceforth fix for the given HMM, independently of L. Let $C > 0$ be this lower bound for (47).

It remains to understand why, assuming $|A|$ and $|B|$ are sufficiently large, $\mathbb{E}[\mathrm{I}[X_A : X_B]]$ is $\Omega(L)$. Given the Δ we have fixed,

$$p_\rho(w \in A; [w-\Delta, w-1] \cap A = \emptyset) \geq D > 0 \tag{48}$$

for a constant D independent of w, for L sufficiently large, when $0.1L < |A| < 0.9L$. For, in this case, we have

$$\begin{aligned}
&p_\rho(w \in A; [w-\Delta, w-1] \cap A = \emptyset) \\
&= p_\rho(w \in A) \cdot \prod_{j=1}^{\Delta} p_\rho(w-j \in B | w \in A, w-1 \in B, \ldots, w-j+1 \in B) \\
&= \underbrace{p(\rho(w) \leq i)}_{=\frac{i}{L}} \cdot \prod_{j=1}^{\Delta} \underbrace{p\left(\rho(w-j) > i | \rho(w) \leq i, \rho(w-1) > i, \ldots, \rho(w-j+1) > i\right)}_{=\frac{L-i-j+1}{L-j}} \\
&\geq \frac{i}{L} \cdot \left(\frac{L-i-\Delta}{L}\right)^{\Delta} \\
&\geq \frac{1}{10} \left(\frac{0.1L-\Delta}{L}\right)^{\Delta} \\
&\geq \frac{1}{10} \cdot \frac{1}{20^{\Delta}} =: D
\end{aligned}$$

where the last step holds when $L > 20\Delta$. Taken together,

$$\mathbb{E}[\mathrm{I}[X_A : X_B]] \geq L \cdot D \cdot C = \Omega(L) \tag{49}$$

when $0.1L < |A| < 0.9L$. The claim follows. □

We note that one can strengthen the proof to provide a high-probability bound, showing that *most* permutations ρ satisfy such linear scaling. The reason is that a random permutation, when $|A|$ and $|B|$ are both large, is very likely to satisfy the event described in (48) on a constant fraction of positions w.

B Sources over Two Features

Simulation results in the main text are based on distributions of the form

$$p(M) = p(M_1) \times p(M_2, M_3) \tag{50}$$

for varying levels of correlation between the binary random variables M_2 and M_3. The main result is that when M_2 and M_3 have lower mutual information, a systematic code for these features minimizes predictive information, but as mutual information increases, a holistic code is more preferred. Here we complement these results with a more in-depth study of a source distribution over two features of the form $p(M) = p(M_1, M_2)$ for binary random variables M_1 and M_2, looking at a grid of possible distributions over 4 outcomes. This comprehensive approach allows us to examine the effects of the marginal probabilities for M_1 and M_2, as well as the effects of different *kinds* of correlations between features on the relative preference for systematic vs. holistic codes.

Outcome	Corr. Source	Anticorr. Source	Systematic	cnot(1,2)	cnot(2,1)
00	3/8	0	ac	ac	ac
01	1/8	1/4	ad	ad	bd
10	1/8	1/4	bc	bd	bc
11	3/8	1/2	ad	bc	ad

Table 1: Some possible sources and codes for the two binary random variables M_1, M_2. The correlated source has Pearson's $r = 1/2$. The anticorrelated source has $r = -1/3$.

The main result is shown in Figure 2, which shows predictive information for all possible mappings from the four outcomes of M to strings in $\{\texttt{a}, \texttt{b}\} \times \{\texttt{c}, \texttt{d}\}$. The rows indicate different marginal probabilities for $p(M_1 = 1)$, the columns indicate different marginal probabilities for $p(M_2 = 1)$, and the x axis indicates the Pearson correlation between M_1 and M_2. The Pearson correlation is necessary to make sense of the pattern here, because two kinds of correlation can induce mutual information between M_1 and M_2: a *positive* correlation between the most probable outcomes and a *negative* correlation, as shown in Table 1. In the positive correlation case, the features M_1 and M_2 are effectively 'fused'—at maximal correlation, there is actually only one feature here, as we always have $M_1 = M_2$. In the negative correlation case, it is as if one of the four outcomes has been effectively removed from the probability distribution.

There are two conclusions to be drawn from Figure 2 beyond the conclusions in the main text. First, the level of preference for systematicity in the low-correlation case depends on the marginal distributions being imbalanced: at $p(M_1) = p(M_2) = \frac{1}{2}$, even when there is zero correlation between the features, the holistic code is just as good as the systematic code. This makes sense because for a uniform distribution over 4 outcomes, there is no reason to favor any one factorization over another. However, as the marginals become more imbalanced (moving downward or to the right in the figure), the systematic code becomes better in the low-correlation range. For these imbalanced marginals, there are generally two red lines to be seen in the figure, corresponding to the two possible classes of non-systematic codes for the source: cnot$(1, 2)$ and cnot$(2, 1)$, which differ in which feature is used as the control bit to flip the other one.

The second conclusion to be drawn from Figure 2 is that there is different behavior for positive and negative feature correlations when the marginals for M_1 and M_2 are both imbalanced. In particular, in the lower right corner, the systematic code is sometimes better than the holistic code when there is a negative correlation. This happens because, in the negatively correlated source, the systematic code allows the appearance of individual symbols to be correlated with the overall probability of the string: for example, in the systematic code for the negatively correlated source in Table 1, high-frequency strings always have `d`, and `a` only appears in low-frequency strings. The result is that the unigram entropy is minimized by the systematic code for such a source.

Figure 3 shows predictive information for codes as a function of mutual information between random variables M_1 and M_2, with the negatively-correlated sources separated out and indicated with a dotted line. We see that the preference for holistic codes as a function of mutual information is weaker for the negatively correlated sources, and also that these sources cannot achieve mutual information as high as the positively correlated ones.

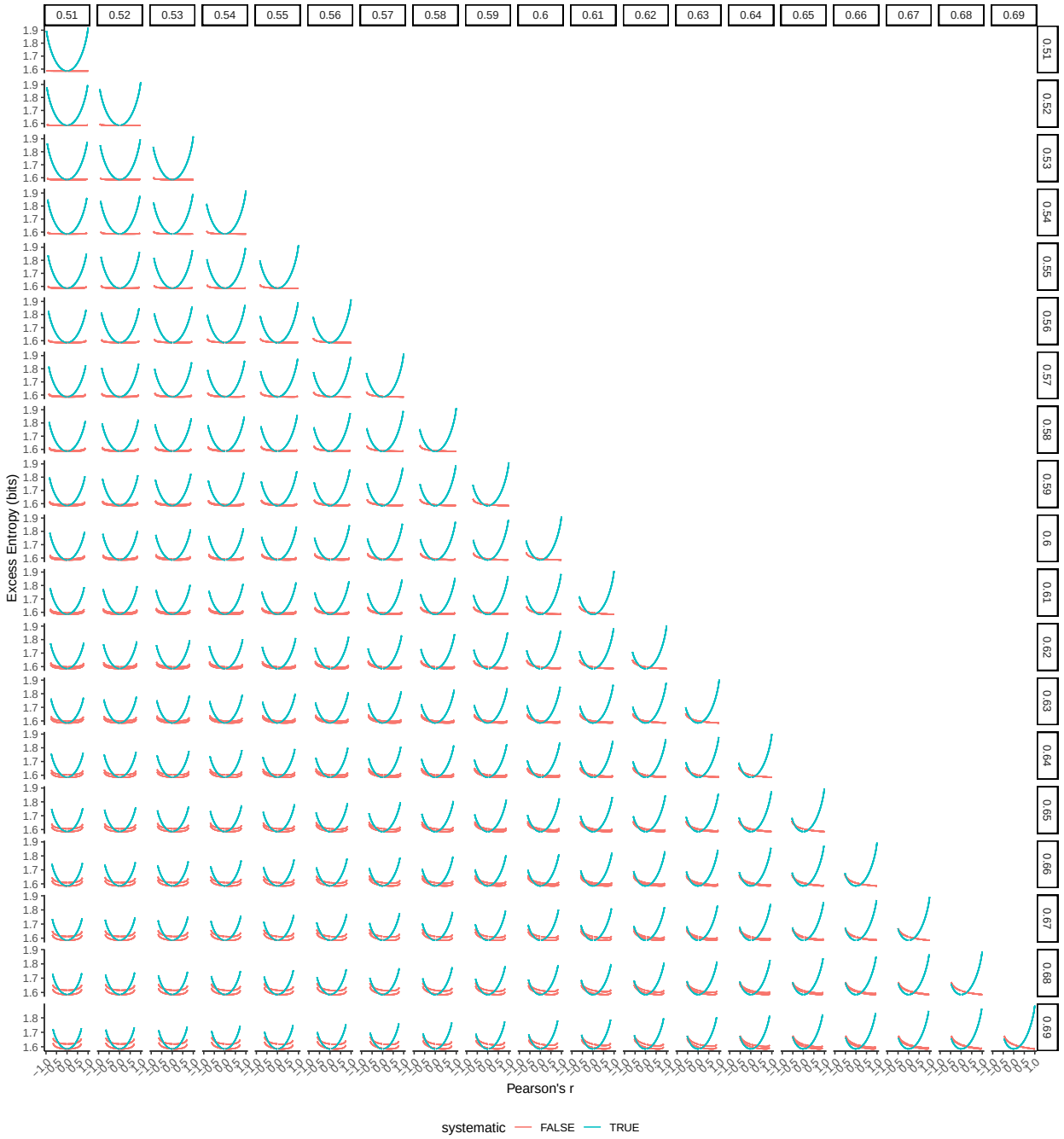

Figure 2: Predictive information (labelled as excess entropy) of length-2 codes for a grid over the simplex of possible sources over two binary random variables, $p(M) = p(M_1, M_2)$. Rows show the marginal probability $p(M_1 = 1)$. Columns show the marginal probability $p(M_2 = 1)$. The x axis shows the Pearson correlation between M_1 and M_2.

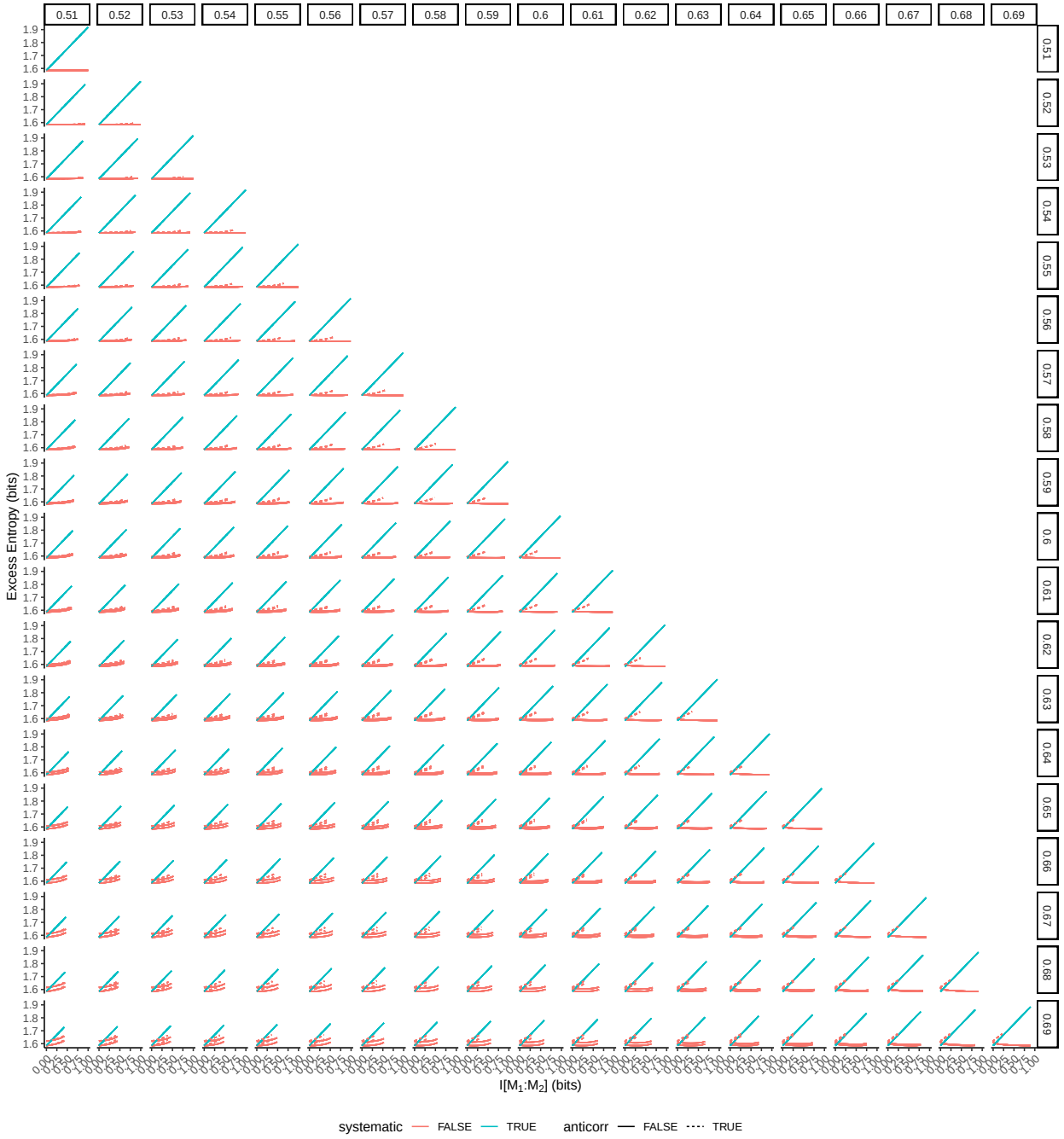

Figure 3: Predictive information (labelled as excess entropy) of codes for a grid over the simplex of possible sources over two binary random variables as in Figure 2, but now by mutual information instead of Pearson correlation. Dotted lines indicate codes for sources whose Pearson correlation is negative.

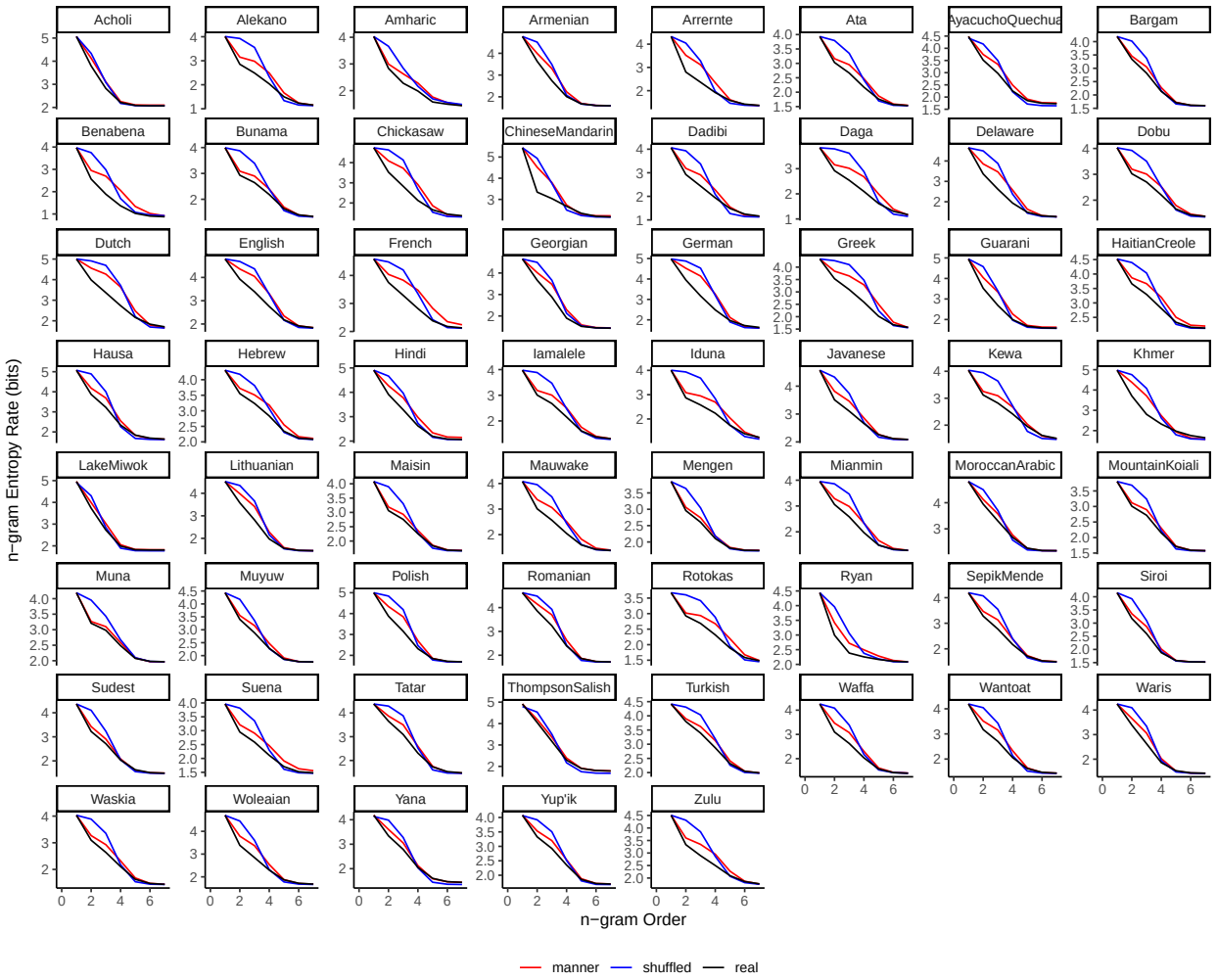

Figure 4: Calculation of predictive information for all 61 languages in the WOLEX database, for the attested forms (black), a deterministic shuffle that preserves manner of articulation (red), and a general deterministic shuffle (blue).

C Phonological Locality in 61 languages

Figure 4 shows the calculation of predictive information for all 61 languages in the WOLEX database, for the real languages compared against two baselines generated by applying deterministic shuffling functions to the attested forms. Table 2 shows the calculated predictive information values.

D NP Orders with Other Source Distributions

The noun phrase ordering results in the main text were derived using a source distribution over NPs estimated from the German Universal Dependencies corpus. Here we show results using other naturalistic source distributions, from corpora of Spanish (Figure 5), English (Figure 6), Czech (Figure 7), Icelandic (Figure 8), and Latin (Figure 9). We also show results using the artificial source developed by Mansfield and Kemp (2023) to study NP order in Figure 10.

Language	Real	Manner	Shuffled
Acholi	5.64	6.17	6.39
Alekano	7.42	8.79	9.60
Amharic	5.62	6.67	7.42
Armenian	6.91	8.04	8.60
Arrernte	6.34	8.17	8.37
Ata	5.98	6.76	7.59
Ayacucho Quechua	6.49	7.23	7.82
Bargam	6.26	6.74	7.45
Benabena	6.57	8.92	9.33
Bunama	6.94	7.59	8.69
Chickasaw	8.60	10.80	11.30
Dadibi	7.43	8.43	9.15
Daga	7.99	9.70	10.60
Delaware	7.88	9.97	10.60
Dobu	6.99	7.86	8.91
Dutch	9.38	11.90	12.40
English	6.91	8.34	8.65
French	6.35	7.78	8.50
Georgian	7.52	8.83	9.38
German	8.83	11.20	11.60
Greek	8.10	10.20	10.90
Guarani	7.00	8.21	8.66
Haitian Creole	6.05	6.82	7.63
Hausa	8.35	9.21	9.83
Hebrew	5.96	6.83	7.33
Hindi	6.64	7.56	8.12
Iamalele	7.21	8.10	9.09
Iduna	8.02	9.38	10.60
Javanese	5.72	6.57	7.08
Kewa	7.22	7.92	8.82
Khmer	8.30	10.00	10.50
Lake Miwok	6.20	6.75	6.87
Lithuanian	7.24	8.54	9.00
Maisin	5.72	6.07	7.03
Mandarin Chinese	6.06	7.65	8.05
Mauwake	6.53	8.05	8.79
Mengen	4.66	4.95	5.85
Mianmin	6.80	8.01	8.83
Moroccan Arabic	6.20	6.71	6.99
Mountain Koiali	5.55	5.99	6.72
Muna	5.13	5.46	6.53
Muyuw	6.12	6.79	7.36
Polish	7.85	9.35	9.90
Romanian	7.44	8.37	8.67
Rotokas	6.49	7.49	8.41
Ryan	3.87	5.02	5.63
Sepik Mende	6.77	7.48	8.47
Siroi	5.79	6.37	7.06
Sudest	6.59	6.96	7.90
Suena	6.08	6.89	7.74
Tatar	7.96	8.82	9.39
Thompson Salish	7.49	7.87	8.38
Turkish	7.00	7.66	8.37
Waffa	6.64	7.74	8.49
Wantoat	6.63	7.69	8.17
Waris	6.55	7.39	7.98
Waskia	6.40	7.09	7.85
Woleaian	6.83	7.92	8.57
Yana	6.82	7.38	8.13
Yup'ik	6.09	6.75	7.43
Zulu	6.79	8.12	9.00

Table 2: Predictive information values (in bits) for 61 languages of the WOLEX sample, visualized in Figure 4. 'Real' is the predictive information of the attested wordforms. 'Manner' is for wordforms shuffled while preserving manner. 'Shuffled' is for wordforms shuffled without regard for manner.

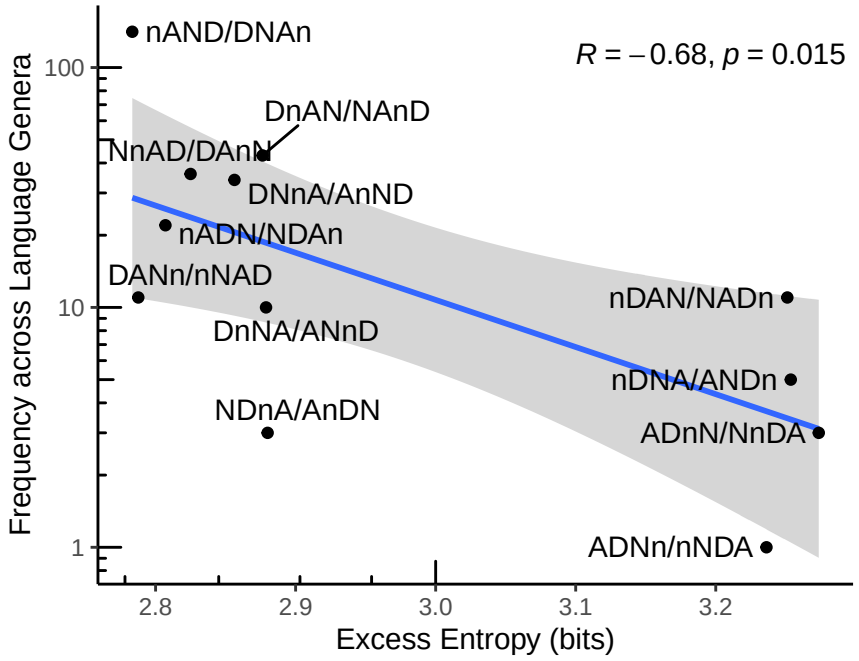

Figure 5: Typology frequencies of NP orders by predictive information estimated using the **Spanish UD source** (Mariona Taulé and Recasens, 2008). Lines and statistics as in the figure in the main text.

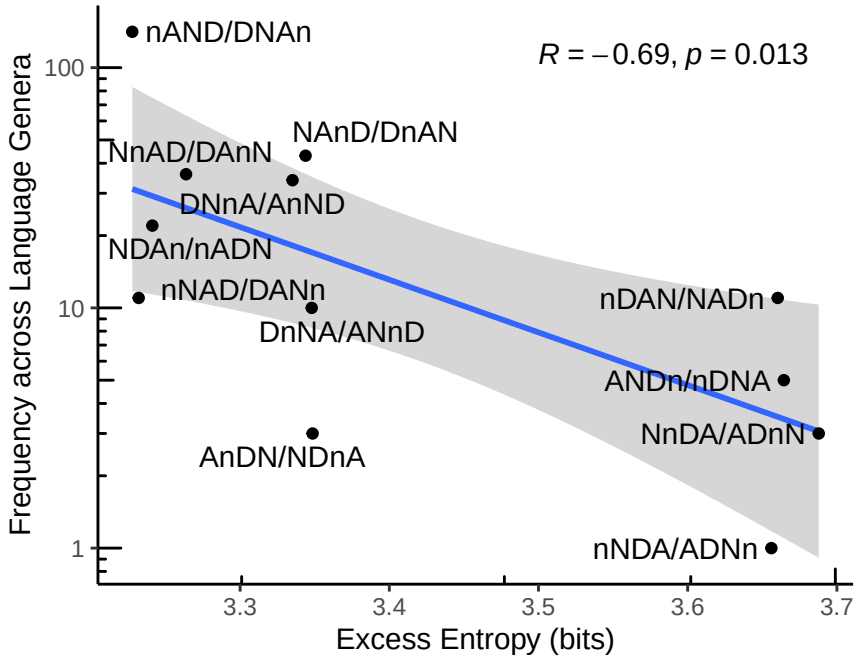

Figure 6: Typology frequencies of NP orders by predictive information estimated using the **English UD source** (Zeldes, 2017).

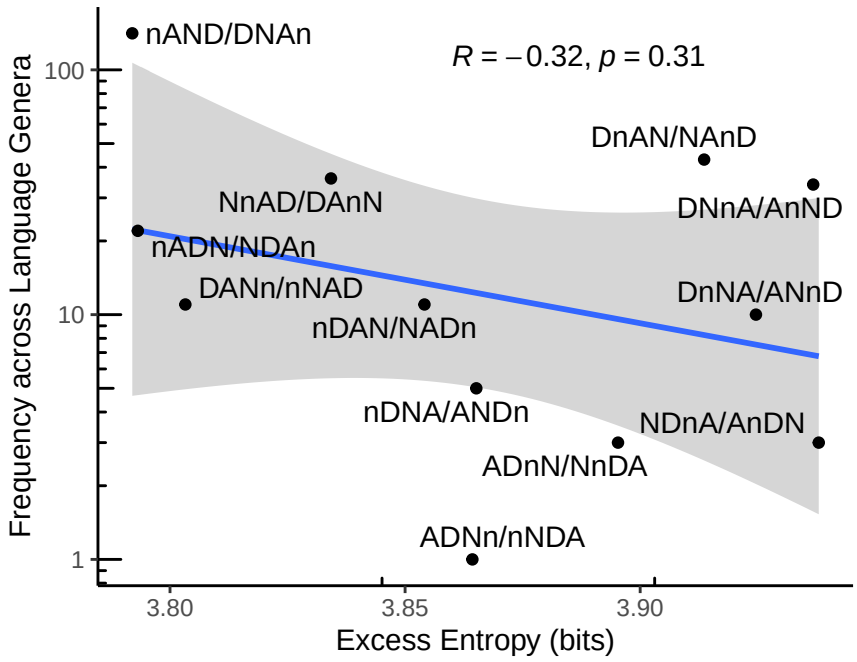

Figure 7: Typology frequencies of NP orders by predictive information estimated using the **Czech UD source** (Hladká et al., 2008). We believe the weaker correlation here is due to the rarity of determiners in the Czech corpus.

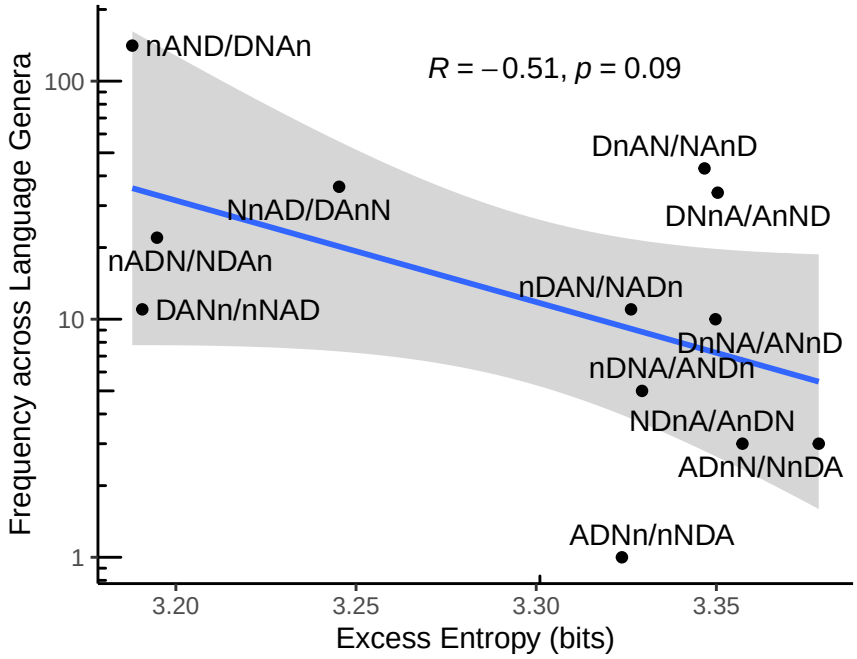

Figure 8: Typology frequencies of NP orders by predictive information estimated using the **Icelandic UD source** (Arnardóttir et al., 2020).

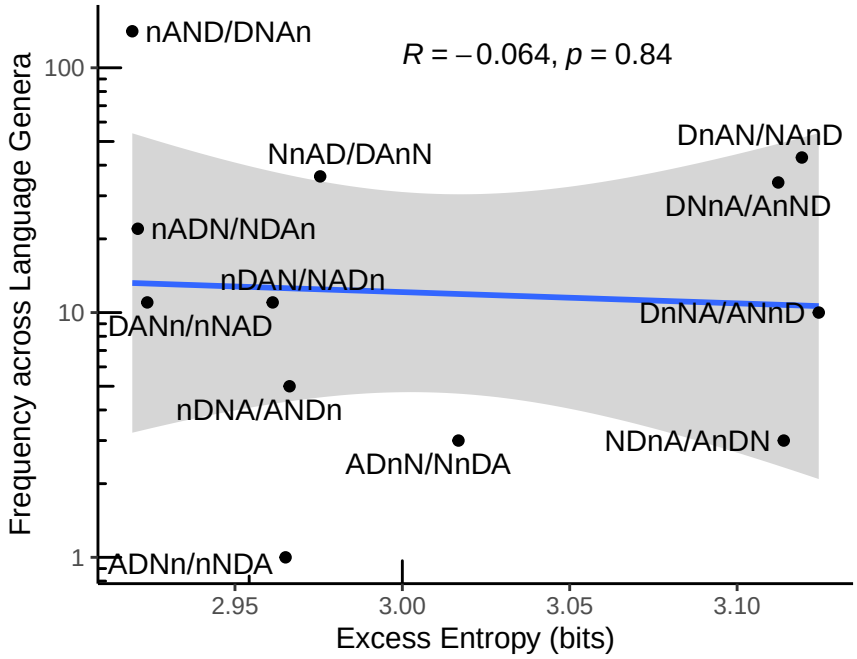

Figure 9: Typology frequencies of NP orders by predictive information estimated using the **Latin UD source** (based on all Latin UD corpora). As the text genre for this corpus is highly unusual (consisting of over 1000 years' worth of text, much of it poetry or written by non-native speakers), we believe that the distribution of NPs in this corpus is not representative of the 'true' source distribution over NP meanings.

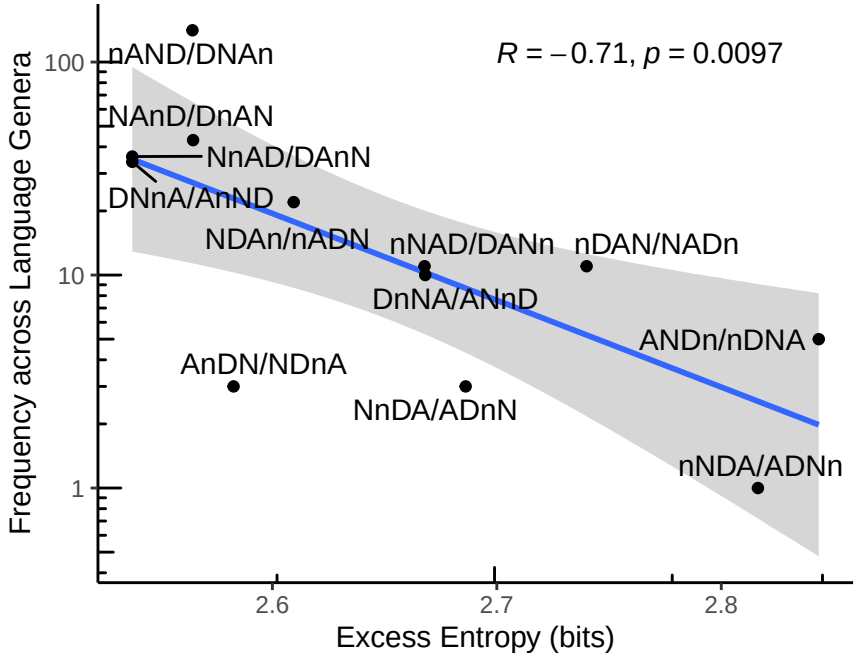

Figure 10: Typology frequencies of NP orders by predictive information estimated using the artificial **MK23 source** (Mansfield and Kemp, 2023).

E Correlation of Semantic Features

In Figure 11 we present results of the study on correlation of semantic features, but using the semantic feature norms from the Glasgow Word Norms (Scott et al., 2017), which rate words for features such as dominance, valence, and arousal. Features are binarized and their frequencies and pairwise MIs are calculated as in the main text. Results are similar to the main text: the across-morpheme and across-word features largely have lower mutual information than within-morpheme and within-word features.

F Phonotactic Results with Corpus Frequencies

Recall that our analysis of phonotactics assumed a uniform distribution over forms. This is because the phonological forms, as listed in the WOLEX database, cannot straightforwardly be matched to corpus data. However, for four languages (Dutch, English, French, and German), WOLEX provides orthographic forms. Using these, we derived corpus frequencies from the full Wikipedia texts in these languages. We applied simple Laplace smoothing at $\alpha = 1$. Results as shown in Figure 12 closely agree with those derived under a uniform distribution.

G Hierarchically-Structured Sources

G.1 Varying Coupling Parameters in Tree Structures

We created further sources by keeping the tree structure from Main Paper, Figure 2F, but varying the parameters $\alpha, \beta, \gamma \in [0, 1]$ randomly subject to the constraint $4\alpha < 2\beta < \gamma$. We created 70 random samples. Results, shown in Figure 13, reproduce the pattern from Main Paper, Figure 2F.

G.2 Sources Defined by PCFGs

We constructed probabilistic context-free grammars (PCFGs) defined by 5 terminals and 5 nonterminals. For each nonterminal a, we considered the 100 possible binary productions $a \to bc$ where b, c are terminals or nonterminals. For each nonterminal, we defined a distribution over these 100 possible productions $a \to bc$ by defining

$$p(a \to bc) \propto \exp(T p_{a \to bc}), \tag{51}$$

where $T > 0$ is an inverse temperature parameter and each $p_{a \to bc} \in [0, 1]$ is a random number (cf. DeGiuli, 2019). The probabilities are normalized to sum up to one for each left-hand side a. The inverse temperature parameter controls the variability in the probabilities of different productions; higher values result in a sparser source.

We then enumerated all 5^6 strings of length 6 over the given nonterminals, and used the CKY algorithm to compute the probabilities of all of these strings under the given PCFG. This defines a source over all strings of length 6.

At inverse temperatures $T = 1, 2, 3, 4, 10, 20$ we sampled 10 PCFGs each, and compared the predictive information of the language given by the PCFG (systematic and local), deterministic permutations of the 6 positions (systematic and nonlocal), and 360 randomly chosen shuffles of the mapping between forms and probabilities (neither local nor systematic).

Results (Figure 14) show that local orderings usually achieve lower predictive information. Nonsystematic codes have much higher predictive information, very closely concentrated around values clearly separated from the systematic codes.

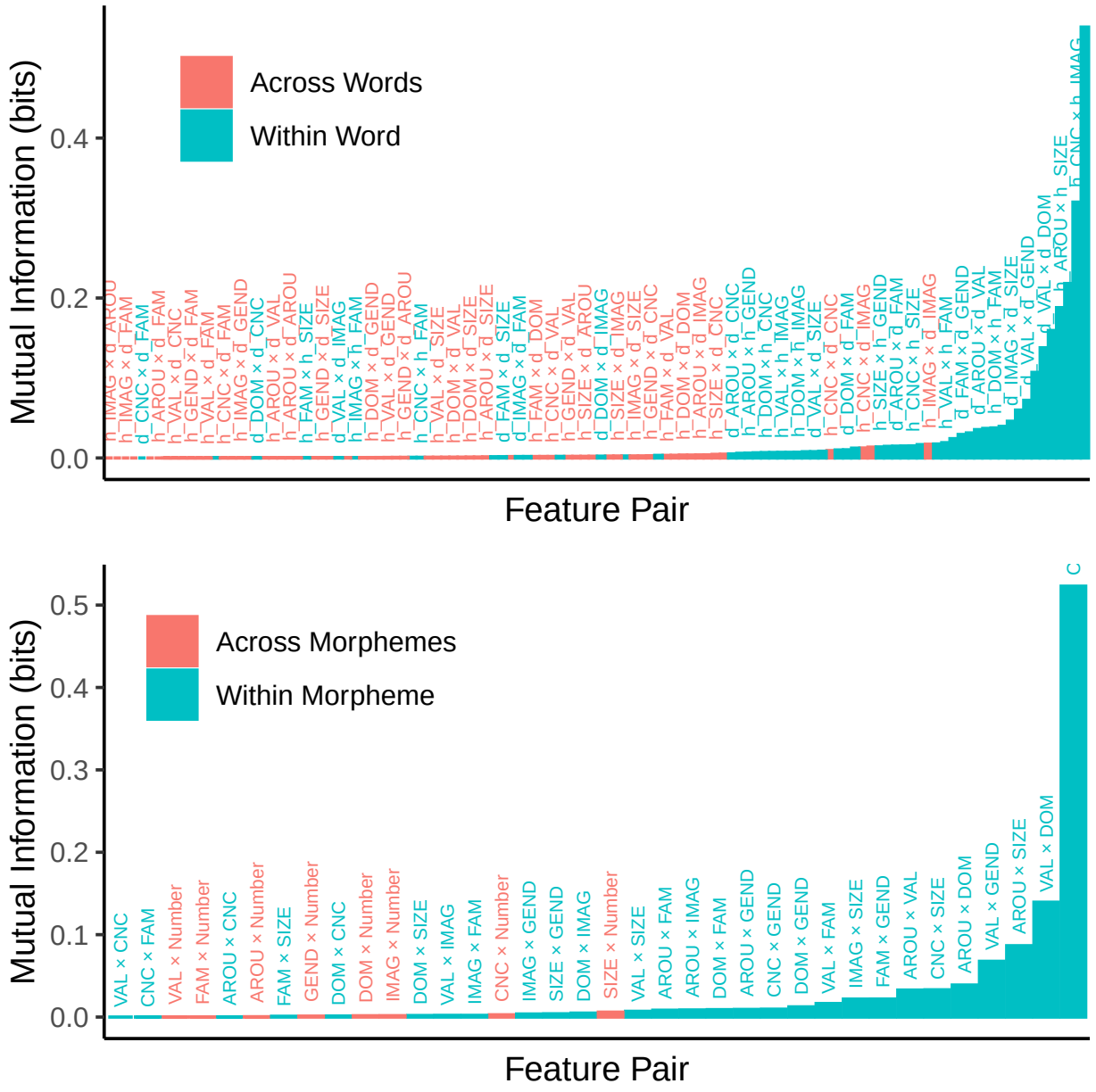

Figure 11: (A) Pairwise mutual information of semantic features from the Glasgow Word Norms (Scott et al., 2017) across words and within words, for pairs of verbs and objects in Universal Dependencies English corpora. (B) Pairwise mutual information of the Glasgow Word Norms along with a number feature indicated by plural morphology. The across-word and across-morpheme features have generally lower MI than the within-word and within-morpheme features.

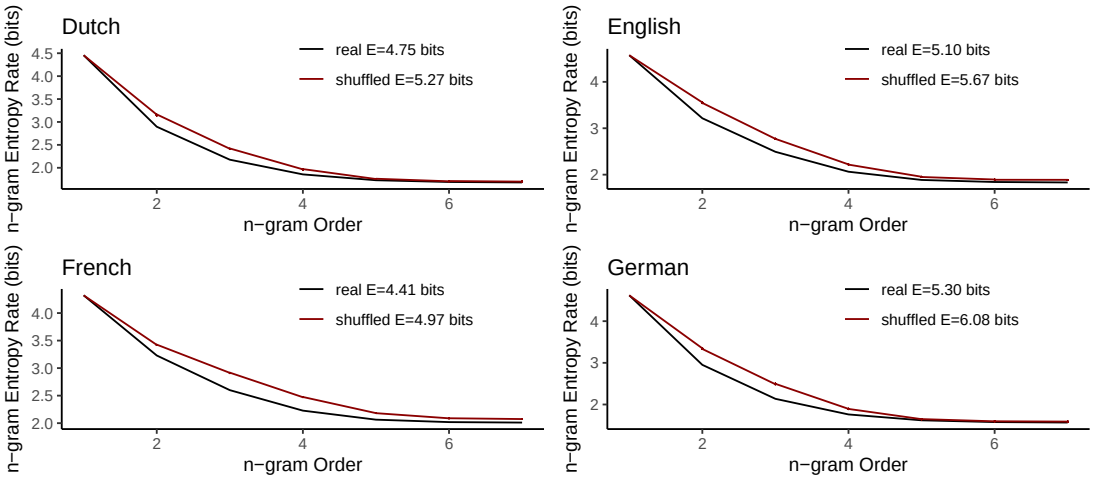

Figure 12: Calculation of predictive information using corpus frequencies, for the 4 languages in the WOLEX database for which orthographic forms are available in WOLEX. We show the attested forms (black) and a deterministic shuffle that preserves manner of articulation (red). Results match those found with uniform distributions.

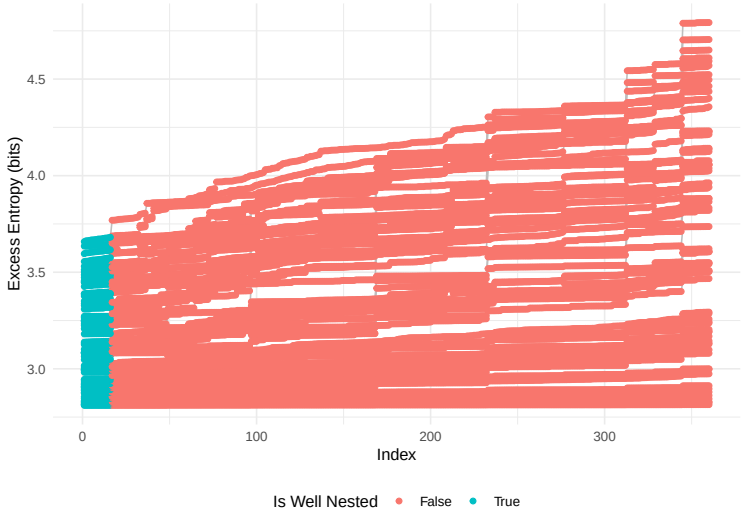

Figure 13: Results for 70 sampled combinations of coupling parameters for the tree structure in Main Paper, Figure 2F. Across samples, well-nested orderings achieve lower predictive information than non-well-nested orderings.

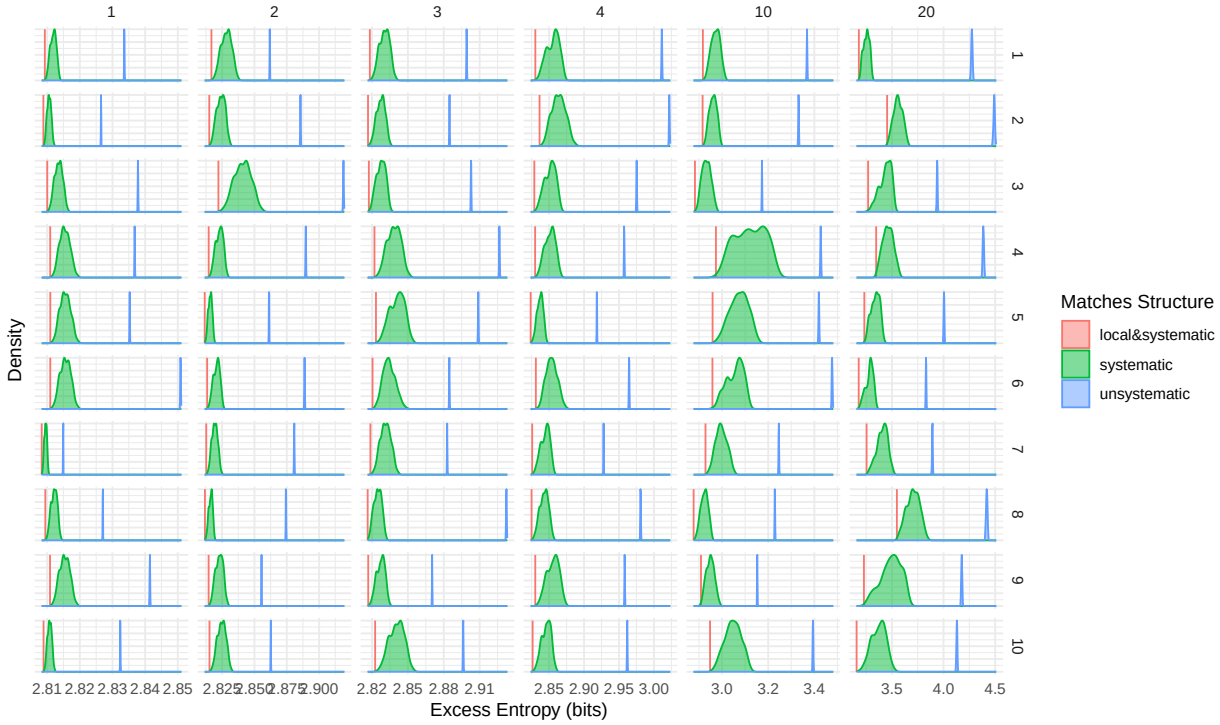

Figure 14: Distribution of predictive information, for 10 randomly constructed PCFG sources for length-6 strings over 5 symbols, at six different inverse temperature parameters (T in (51)). We compare the local and systematic code given the PCFG (red) with the systematic codes given by the deterministic shuffles of the six positions (green), and an equal number (360, up to reversal) of unsystematic codes given by shuffles of the mapping between forms and probabilities (blue). Local and systematic codes tend to achieve lower predictive information than other systematic codes. Unsystematic codes strongly concentrate at substantially higher predictive information.

References

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., and Steingrímsson, S. (2020). A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online). Association for Computational Linguistics.

Bell, A. J. (2003). The co-information lattice. In *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 921–926.

Crutchfield, J. P. and Feldman, D. P. (2003). Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54.

Dębowski, Ł. (2009). A general definition of conditional information and its application to ergodic decomposition. *Statistics & Probability Letters*, 79(9):1260–1268.

DeGiuli, E. (2019). Random language model. *Physical Review Letters*, 122(12):128301.

Hladká, B., Hajic, J., Hana, J., Hlavácová, J., Mírovskỳ, J., and Raab, J. (2008). The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41.

Mansfield, J. and Kemp, C. (2023). The emergence of grammatical structure from inter-predictability. *PsyArXiv*.

Mariona Taulé, M. A. M. and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

McGill, W. J. (1955). Multivariate information transmission. *IEEE Transactions on Information Theory*, 4(4):93–111.

Nowak, M. A., Plotkin, J. B., and Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, 404(6777):495–498.

Scott, G. G., Keitel, A., Becirspahic, M., O'Donnell, P. J., and Sereno, S. C. (2017). The Glasgow Norms: Ratings of 5,500 words on 9 scales.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82.

Zeldes, A. (2017). The GUM Corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

94. Futrell, R., Levy, R. P. & Gibson, E. Dependency locality as an explanatory principle for word order. *Language* **96**, 371–413 (2020).
95. Mansfield, J. The word as a unit of internal predictability. *Linguistics* **59**, 1427–1472 (2021).
96. Chafe, W. L. Givenness, contrastiveness, definiteness, subjects, topics and points of view. In *Subject and Topic* (ed. Li, C. N.) 27–55 (Academic Press, 1976).
97. Bock, J. K. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychol. Rev.* **89**, 1–47 (1982).
98. Bresnan, J., Cueni, A., Nikitina, T. & Baayen, H. Predicting the dative alternation. In *Cognitive Foundations of Interpretation* (eds Bouma, G., Krämer, I. & Zwarts, J.) 69–94 (Royal Netherlands Academy of Science, 2007).
99. Chen, S. F. & Goodman, J. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **13**, 359–393 (1999).
100. Nivre, J. et al. *Universal Dependencies 1.0* (Universal Dependencies Consortium, 2015); http://hdl.handle.net/11234/1-1464
101. Thackston, W. M. *An Introduction to Koranic and Classical Arabic: An Elementary Grammar of the Language* (IBEX Publishers, 1994).
102. Zeldes, A. The GUM Corpus: creating multilayer resources in the classroom. *Lang. Resour. Eval.* **51**, 581–612 (2017).
103. Behzad, S. & Zeldes, A. A cross-genre ensemble approach to robust Reddit part of speech tagging. In *Proc. 12th Web as Corpus Workshop*, (eds Barbaresi, A. et al.) 50–56 (European Language Resources Association, 2020); https://aclanthology.org/2020.wac-1.7
104. Silveira, N. et al. A gold standard dependency corpus for English. In *Proc. Ninth International Conference on Language Resources and Evaluation* (eds Calzolari, N. et al.) 2897–2904 (European Language Resources Association, 2014).
105. Graff, P. *Communicative Efficiency in the Lexicon*. PhD thesis, Massachusetts Institute of Technology (2012).
106. Vincze, V. et al. Hungarian dependency treebank. In *Proc. Seventh International Conference on Language Resources and Evaluation* (eds Calzolari, N. et al.) (European Language Resources Association, 2010); http://www.lrec-conf.org/proceedings/lrec2010/pdf/465_Paper.pdf
107. Buck, C., Heafield, K. & van Ooyen, B. *N*-gram counts and language models from the Common Crawl. In *Proc. Ninth International Conference on Language Resources and Evaluation* (eds Calzolari, N. et al.) 3579–3584 (European Language Resources Association, 2014); http://www.lrec-conf.org/proceedings/lrec2014/pdf/1097_Paper.pdf

Acknowledgements

We thank S. Piantadosi, N. Rathi, G. Scontras, K. Mahowald, N. Zaslavsky, T. Pimentel, R. Hawkins, N. Imel, R. Sun, Z. Pizlo, B. Skyrms, J. Barrett, J. Andreas, M. Marcolli, J. P. Vigneaux Ariztia, Ł. Dębowski, A. Nini and audiences at NeurIPS InfoCog 2023, the UCI Center for Theoretical Behavioral Sciences, EvoLang 2024, TedLab, the Society for Computation in Linguistics 2024, the Quantitative Cognitive Linguistics Network and the CalTech Seminar on Information and Geometry for discussion. We received no specific funding for this work.

Author contributions

R.F. designed and ran studies in the main text. R.F. and M.H. performed mathematical analyses, designed and ran studies in the Supplementary Information, and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02336-w.

Correspondence and requests for materials should be addressed to Richard Futrell.

Peer review information *Nature Human Behaviour* thanks Łukasz Dębowski, Byung-Doh Oh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Richard Futrell

Last updated by author(s): Jul 5, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No new data was collected for this paper.
Data analysis	All code used to analyze data is available online as described in the Code Availability statement.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Unique data required to reproduce our results is available at http://github.com/Futrell/infolocality. Corpus count data is drawn from Universal Dependencies v2.8, available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3683. The Lancaster Sensorimotor Norms are available at https://osf.io/7emr6/. Wordform data from the WOLEX database is not publicly available, but a subset can be made available upon request to the authors.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No data involving human subjects is reported.
Reporting on race, ethnicity, or other socially relevant groupings	No data involving human subjects is reported.
Population characteristics	No data involving human subjects is reported.
Recruitment	No data involving human subjects is reported.
Ethics oversight	No data involving human subjects is reported.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Numbers of baseline samples are reported in the main text and methods.
Data exclusions	There were no data exclusions.
Replication	A number of replications involving slightly different data sources and baselines are reported in the SI. Code for reproducing all results is made publicly available.
Randomization	Generation of random baselines is described in the main text and methods.
Blinding	Blinding is not possible as the results consist entirely of computational simulations based on corpus data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*