Natural-language-like systematicity from a constraint on excess entropy

Richard Futrell

Department of Language Science
University of California, Irvine
rfutrell@uci.edu

Natural language is systematic: utterances are composed of individually meaningful parts which are typically concatenated together. I argue that natural-language-like systematicity arises in codes when they are constrained by excess entropy, the mutual information between the past and the future of a process. In three examples, I show that codes with natural-language-like systematicity have lower excess entropy than matched alternatives.

1. Introduction

A key property of human language is that it is systematic: parts of form correspond regularly to components of meaning. For example, in the English sentences *I saw the cat*, *the cat meowed*, *a cat ate food*, etc., the substring *cat* systematically refers to a particular aspect of meaning: that these sentences all have to do with domestic felines. These substrings which make a regular contribution to meaning are called **morphemes**. This property is related to the more general concept of compositionality (Frege, 1892; Montague, 1970; Heim and Kratzer, 1998).

Natural language utterances, such as the examples given, typically consist of a concatenation of morphemes. When morphemes are combined by other means—for example, in Semitic nonconcatenative morphology, or Celtic consonant mutations, or when concatenation of 'underlying forms' is obscured by phonological processes—the resulting string still usually has subsequences that regularly correspond to aspects of meaning, and these parts remain fairly contiguous or close to each other within the string. I will call this property of natural language **locality**. Furthermore, these parts have a high level of statistical inter-predictability (Mansfield, 2021; Mansfield and Kemp, 2023).

Systematicity is not a property of efficient codes as studied in coding theory, which raises the question of why human language has it. I propose that the particular form of systematicity that we see in human language can be explained by positing that human language operates under a constraint on excess entropy (Shalizi and Crutchfield, 2001), a measure of the complexity of incremental prediction.

I consider a **language** to be any mapping $L : \mathcal{M} \to \Sigma^*$ from meanings $\mathcal{M}$ to forms (strings) drawn from an alphabet Σ. Suppose that a meaning can be

represented in terms of **features**: that is, a meaning $m \in \mathcal{M}$ can be written as a product of two features as $m = m_1 \times m_2$. Then I say a language is **systematic** if the form associated with that meaning can be decomposed in the same way:

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

for a string combining function $\cdot$, such as concatenation. A language is **holistic** otherwise (Wray, 1998; Smith et al., 2003b).

The definition of systematicity is crucially relative to a chosen decomposition of the meanings and a chosen string combining function. There are infinitely many ways meanings can be decomposed into features. If we are free to choose any such decomposition, then any function L can be made systematic (Zadrozny, 1994; Westerståhl, 1998; Andreas, 2019). Likewise, the string combining function $\cdot$ needs to be constrained, or else systematicity can be achieved trivially.

In existing accounts, the emergence of systematicity in language is often motivated by learners' need to generalize in order to produce forms for never-before-seen meanings (Kirby, 2002; Smith et al., 2003a, 2013; Kirby et al., 2015). Such accounts successfully motivate systematicity in the abstract sense of Eq. 1, but they (explicitly or implicitly) require independent specification of the meaning decomposition and string combination function, via stipulations about the structure of the space of meanings and/or strings, or via inductive biases built into learners (Batali, 1998; Nowak and Krakauer, 1999; Barrett, 2009; Tria et al., 2012; Franke, 2016; Steinert-Threlkeld, 2020).

In contrast, my goal is not only to explain why natural language has systematicity in the abstract sense, but also to give a theory based on maximally general principles that can explain the meaning decomposition ($\times$) and the string combining function ($\cdot$) that we find in natural language, based on minimal assumptions about the structure of meanings, strings, or learners.

2. Excess Entropy

For a stationary stochastic process generating symbols $X_1, X_2, \ldots$, the **excess entropy** $\mathbf{E}$ measures the complexity of incremental prediction. It is (a lower bound on) the amount of information that needs to be stored about the past of the process in order to accurately reproduce its future (Shalizi and Crutchfield, 2001, §6). Formally, it is defined as the mutual information between all the symbols up to an arbitrary time index (say t) and all the symbols at or after that time index:[1]

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

[1] I assume familiarity with information-theoretic quantities. Briefly, the **entropy** $\mathrm{H}[X] = -\sum_x p(x) \log p(x)$ of a random variable X is its average information content. The **conditional entropy** $\mathrm{H}[X \mid Y] = -\sum_{x,y} p(x, y) \log p(x \mid y)$ of X given another random variable Y is its average information content given knowledge of Y. The **mutual information (MI)** between X and Y is the amount of information contained in X about Y: $\mathrm{I}[X : Y] = \mathrm{H}[X] - \mathrm{H}[X \mid Y] \geq 0$.

In order to apply this concept to languages as defined in Section 1, it is necessary to construct an appropriate stochastic process from the outputs of a language L. This can be done by sampling meanings $m \in \mathcal{M}$ iid from a source M, translating them to strings $s = L(m)$, and then concatenating the strings s with a delimiter between them. This construction was introduced by Hahn et al. (2021a).

Calculation Let h_t represent the t**'th-order Markov entropy rate** of a process, that is, the conditional entropy of symbols given $t-1$ previous symbols:

$$h_t = \mathrm{H}[X_t \mid X_1, \ldots, X_{t-1}]. \tag{3}$$

For a stationary process, the **entropy rate** h is the limit h_t as t goes to infinity, $h = \lim_{t\to\infty} h_t$ (Shannon, 1948). Then the excess entropy can be read off of the curve of h_t for growing t (Bialek et al., 2001b; Crutchfield and Feldman, 2003; Dębowski, 2011):[2]

$$\mathbf{E} = \sum_{t=1}^{\infty} (h_t - h)\,. \tag{4}$$

Cognitive motivation I motivate the idea that excess entropy is constrained in natural language based on three observations about how humans produce and comprehend language: (1) natural language utterances consist, to a first approximation, of one-dimensional sequences of symbols (phonemes), (2) (spoken) production and comprehension are highly incremental (Levelt, 1989; Tanenhaus et al., 1995; Ferreira and Swets, 2002; Smith and Levy, 2013), and (3) humans have limited incremental memory resources for use in language processing (Miller, 1956; Hahn et al., 2021a, 2022). Thus if the excess entropy of a language exceeds humans' memory capacities, then humans cannot produce and comprehend it accurately. Because excess entropy is a highly generic measure of complexity, other cognitive motivations are possible, including some based on learnability.

3. Examples

I consider a number of languages which are unambiguous and have the same entropy rate, but which differ in their systematicity and locality. I show that the systematic and local languages have lower excess entropy and explain why.

[2]The (Relaxed) Hilberg Conjecture implies that $\mathbf{E}$ for natural language texts does not converge (Hilberg, 1990; Bialek et al., 2001a; Dębowski, 2015). I note that the current results are for isolated utterances of language, not texts of unbounded length. Furthermore, the results below are also consistent with a constraint that h_t decay quickly, even if Eq. 4 does not converge, or with a constraint on conditional entropy as a function of the amount of information stored in memory (Still, 2014; Marzen and Crutchfield, 2016; Hahn and Futrell, 2019; Hahn et al., 2021a,b; Rathi et al., 2022).

3.1. *Systematic vs. nonsystematic Huffman codes*

The first example shows that minimizing code length does not produce systematicity. I consider two Huffman (minimal-length) codes for the source in Table 1, with a decomposition of the meanings into two features which are statistically independent. Only the first Huffman code is systematic with respect to this decomposition—the first bit corresponds to the first feature, and the remaining bits to the second. Figure 1 shows entropy rates and excess entropies for the two codes. The systematic code has lower excess entropy; the reason for this will be explained below.

Probability	Features	Form (Syst.)	Form (Nonsyst.)
$2/3 \times 1/2$	00	00	00
$2/3 \times 1/4$	01	010	110
$2/3 \times 1/8$	02	0110	0100
$2/3 \times 1/8$	03	0111	0101
$1/3 \times 1/2$	10	10	10
$1/3 \times 1/4$	11	110	111
$1/3 \times 1/8$	12	1110	0111
$1/3 \times 1/8$	13	1111	0110

Table 1.: A source and two Huffman codes for it.

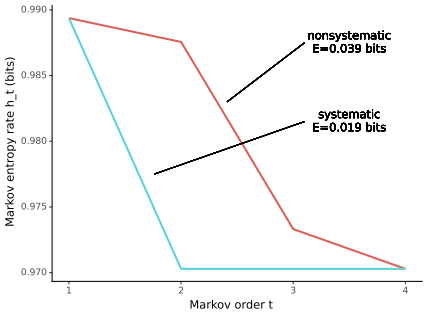

Figure 1.: Excess entropies and Markov entropy rates h_t as a function of t.

3.2. *Systematicity for low-MI features, holistic expression for high-MI features*

I consider languages expressing meanings that can be decomposed into three binary features shown in Table 2. The first language, notated as $L_1 \cdot L_2 \cdot L_3$, is fully systematic with respect to the three features: we have

$$L(m_1) = \begin{cases} \texttt{a} & m_1 = 0 \\ \texttt{b} & m_1 = 1 \end{cases}, \quad L(m_2) = \begin{cases} \texttt{c} & m_2 = 0 \\ \texttt{d} & m_2 = 1 \end{cases}, \quad L(m_3) = \begin{cases} \texttt{e} & m_3 = 0 \\ \texttt{f} & m_3 = 1 \end{cases}. \tag{5}$$

The second language $L_1 \cdot L_{23}$ expresses features 2 and 3 holistically, and the third language $L_{12} \cdot L_3$ expresses features 1 and 2 holistically. I calculate excess entropy for these languages using a source that yields an MI of 0.5 bits between features 2 and 3 and 0 bits between all other features, with the unconditional entropy of each feature equal to 1 bit.

Results are shown in Figure 2. The lowest-excess-entropy language is the one that expresses high-MI features holistically, followed by the fully systematic language, followed by the language that expresses low-MI features holistically.

Features	$L_1 \cdot L_2 \cdot L_3$	$L_1 \cdot L_{23}$	$L_{12} \cdot L_3$
000	ace	ace	ace
001	acf	acf	acf
010	ade	adf	ade
011	adf	ade	adf
100	bce	bce	bde
101	bcf	bcf	bdf
110	bde	bdf	bce
111	bdf	bde	bcf

Table 2.: Three languages for expressing meanings that are decomposed into three binary features.

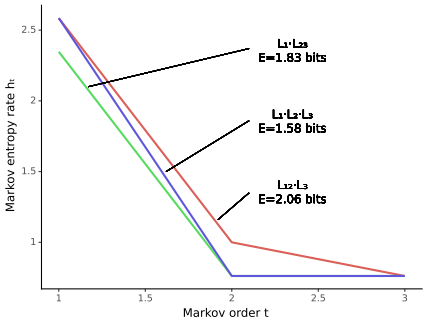

Figure 2.: Excess entropies and Markov entropy rates for the three languages. The source induces mutual information between features 2 and 3.

Explanation It is detrimental to express low-MI features holistically because this creates unnecessary long-range correlations between symbols. To see why, compare the fully systematic language $L_1 \cdot L_2 \cdot L_3$ against the partially-systematic $L_{12} \cdot L_3$, and consider the conditional entropy of the third character X_3. In the systematic code, this is $\mathrm{H}[X_3 \mid X_2] = \mathrm{H}[M_3 \mid M_2]$, because each character X_i encodes the feature M_i. But in the nonsystematic code, we have $\mathrm{H}[X_3 \mid X_2] = \mathrm{H}[M_3] \geq \mathrm{H}[M_3 \mid M_2]$, because the character X_2 is not informative on its own about the value of M_2. Thus the conditional entropy of X_3 cannot be reduced without taking long-range context into account, increasing excess entropy.

The finding that minimization of excess entropy drives low-MI features to be expressed systematically in these examples gives some traction on the question of how meanings may be decomposed into features in language. Languages constrained to minimize excess entropy will appear to be systematic with respect to features that are relatively statistically independent of each other. In contrast, more correlated features will tend to be expressed holistically.

3.3. *Locality*

Here I show that, when languages are systematic, minimization of excess entropy pushes them to maintain locality. I consider a language for a meaning source M over 10 objects $\{m^1, \ldots, m^{10}\}$, following a Zipfian distribution $p_M(m^i) \propto i^{-1}$. Each of these meanings is decomposed into two parts as $m = m_1 \times m_2$, with each utterance decomposing into two morphemes as $L(m_1 \times m_2) = L(m_1) \cdot L(m_2)$, where the morpheme $L(m_k)$ for feature m_k is a random string in $\{\mathtt{0}, \mathtt{1}\}^4$. Now I consider the excess entropy of every possible language $L_f(m) = f(L(m))$, where f is a deterministic permutation function applied to the characters of the string

of $L(m)$. The permuted languages L_f generally represent string combination functions other than concatenation. Most of these languages L_f interleave the two morphemes in various ways; a few leave the morphemes contiguous.

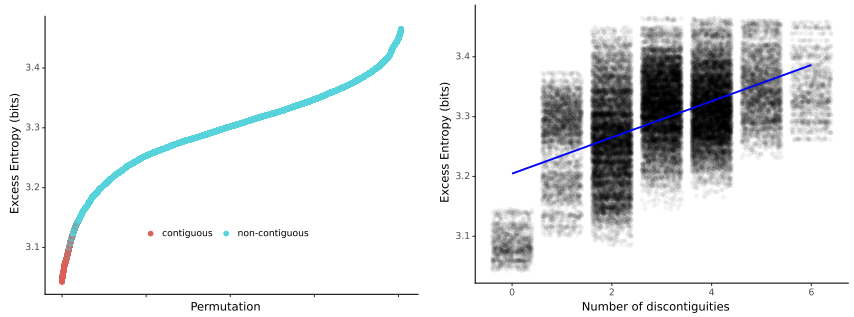

Figure 3.: Excess entropy of permuted systematic languages. **Left**, ordered by excess entropy: permutations that maintain contiguity of morphemes in red. **Right**, by number of discontiguities (number of transitions from one morpheme to another within a string, minus one).

Figure 3 shows the excess entropy for all permutations. The languages with the lowest excess entropy are the contiguous ones. This happens because the coding procedure above creates redundancy among characters within a morpheme. When these redundant characters are separated from each other by a large distance—such as when characters from another morpheme intervene—then the language has long-range mutual information, increasing excess entropy.

4. Conclusion

In the case studies given, the codes which have minimal excess entropy seem to have natural-language-like systematicity: they consist of morphemes which regularly correspond to features of meaning, which are concatenated together or at least kept minimally contiguous. Notably, the current approach does not assume any inductive biases of learners nor pre-existing structure to meanings or forms, except that they are strings. It appears that languages constrained by excess entropy tend to factorize meanings into features that are relatively statistically independent of each other and then combine strings for these features locally.

References

Andreas, J. (2019). Measuring compositionality in representation learning. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA.

Barrett, J. A. (2009). The evolution of coding in signaling games. *Theory and Decision*, 67:223–237.

Batali, J. (1998). Computational simulations of the emergence of grammar. *Approaches to the Evolution of Language–Social and Cognitive Bases*.

Bialek, W., Nemenman, I., and Tishby, N. (2001a). Complexity through nonextensivity. *Physica A: Statistical Mechanics and its Applications*, 302(1-4):89–99.

Bialek, W., Nemenman, I., and Tishby, N. (2001b). Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463.

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. (2015). The relaxed Hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4):311–337.

Dębowski, Ł. (2011). Excess entropy in natural language: Present state and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):037105.

Ferreira, F. and Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, 46(1):57–84.

Franke, M. (2016). The evolution of compositionality in signaling games. *Journal of Logic, Language and Information*, 25(3-4):355–377.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Hahn, M., Degen, J., and Futrell, R. (2021a). Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 128(4):726–756.

Hahn, M. and Futrell, R. (2019). Estimating predictive rate–distortion curves using neural variational inference. *Entropy*, 21:640.

Hahn, M., Futrell, R., Levy, R., and Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Hahn, M., Mathew, R., and Degen, J. (2021b). Morpheme ordering across languages reflects optimization for processing efficiency. *Open Mind*, 5:208–232.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Wiley-Blackwell, Malden, MA.

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten—eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44(9–10):243–248.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, E., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, pages 173–203. Cambridge University Press, Cambridge.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Mansfield, J. (2021). The word as a unit of internal predictability. *Linguistics*, 59(6):1427–1472.

Mansfield, J. and Kemp, C. (2023). The emergence of grammatical structure from interpredictability. In *A Festschrift for Jane Simpson*.

Marzen, S. E. and Crutchfield, J. P. (2016). Predictive rate–distortion for infinite-order Markov procceses. *Journal of Statistical Physics*, 163:1312–1338.

Miller, G. (1956). Human memory and the storage of information. *IRE Transactions on Information Theory*, 2(3):129–137.

Montague, R. (1970). English as a formal language. In Visentini, B., editor, *Linguaggi nella società e nella tecnica*, pages 189–223.

Nowak, M. A. and Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, 96:8028–8033.

Rathi, N., Hahn, M., and Futrell, R. (2022). Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Shalizi, C. R. and Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3-4):817–879.

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

Smith, K., Brighton, H., and Kirby, S. (2003a). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4):537–558.

Smith, K., Kirby, S., and Brighton, H. (2003b). Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386.

Smith, K., Tamariz, M., and Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *35th Annual Conference of the Cognitive Science Society*, pages 1348–1353. Cognitive Science Society.

Smith, N. J. and Levy, R. P. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Steinert-Threlkeld, S. (2020). Toward the emergence of nontrivial compositionality. *Philosophy of Science*, 87(5):897–909.

Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tria, F., Galantucci, B., and Loreto, V. (2012). Naming a structured world: A cultural route to duality of patterning. *PLOS one*, 7(6):e37744.

Westerståhl, D. (1998). On mathematical proofs of the vacuity of compositionality. *Linguistics and Philosophy*, 21(6):635–643.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language & Communication*, 18(1):47–67.

Zadrozny, W. (1994). From compositional to systematic semantics. *Linguistics and Philosophy*, 17:329–342.