Natural Language Systematicity from a Constraint on Excess Entropy

Richard Futrell
Department of Language Science
University of California, Irvine

Abstract

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 stochastic 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, which means that that parts of form correspond regularly to components of meaning.[1] 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**—roughly corresponding to words.

Natural language utterances, such as the examples given, typically consist of a concatenation of morphemes. When morphemes are combined by other means,[2] 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. I will call this property of natural language **locality**.

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 systematicity in human language can be explained by positing that human language operates under a constraint on excess entropy [22], a measure of complexity, which corresponds to a general constraint on control and information processing in incremental production and comprehension of language.

I consider a **language** to be any mapping $L : \mathcal{M} \to \Sigma^*$ from meanings $\mathcal{M}$ to forms (strings) drawn from a vocabulary 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 [33, 25].

The definition of systematicity is crucially relative to a chosen decomposition of the meanings and a chosen string combining function. There are 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

[1]From the perspective of semantics, this property is related to the more general concept of compositionality [9, 20, 12].

[2]For example, in Semitic nonconcatenative morphology, or Celtic consonant mutations.

Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

[34, 32, 28]. Likewise, the string combining function · 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 [14, 24, 26, 15]. 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 kernels on meanings and/or strings, or via implicit inductive biases built into learners [21, 1, 31, 8, 28].

In contrast, our 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 accounts for the particular properties of the meaning decomposition and the string combining function · in natural language. Regarding the latter, a good theory should predict that morphemes are usually combined by concatenation, and when not, something that maintains locality.

2 Excess Entropy

For a stationary stochastic process generating symbols $X_1, X_2, \ldots$, the **excess entropy E** [22, Def. 13] is defined as the mutual information between the past of the process (all the symbols up to an arbitrary time index, say t) and the future of the process (all the symbols at or after some time index):

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

Intuitively, it measures (a lower bound on) the amount of information about the past of a process that must be stored in order to reproduce the future of the process accurately.

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 $x = L(m)$, and then concatenating the strings x with a delimiter between them (a construction also used in [10]).

Calculation of Excess Entropy 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, $h = \lim_{t\to\infty} h_t$ [23]. Then the excess entropy can be read off of the curve of h_t for growing t [3, 4, 6, 5, 18]:[3]

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

Cognitive motivation I motivate the idea that excess entropy is constrained in natural language based on three facts 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 [16, 30, 7, 27], and (3) humans have limited incremental memory resources [19, 10, 11]. If the excess entropy of a language exceeds humans' memory capacities, then humans cannot produce and comprehend it accurately.

3 Examples

Here 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 give reasons for this.

[3]The (Relaxed) Hilberg Conjecture implies that $\mathbf{E}$ for natural language texts does not converge [13, 2, 6]. Our results are also consistent with a constraint that h_t decay quickly, even if Eq. 4 does not converge, or with a constraint on the predictive information bottleneck curve [29, 17, 10]. Furthermore, our results are for isolated utterances of language, not texts of unbounded length.

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: Two Huffman codes for the given source.

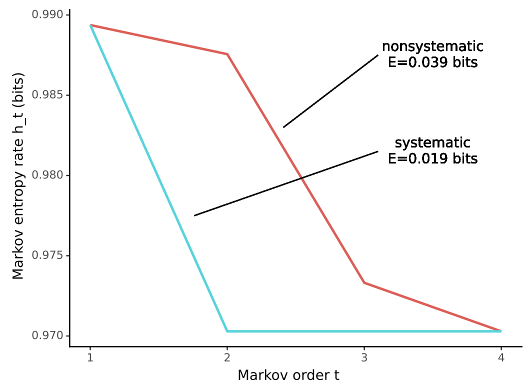

Figure 1: Excess entropies and Markov entropy rates h_t as a function of t.

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.

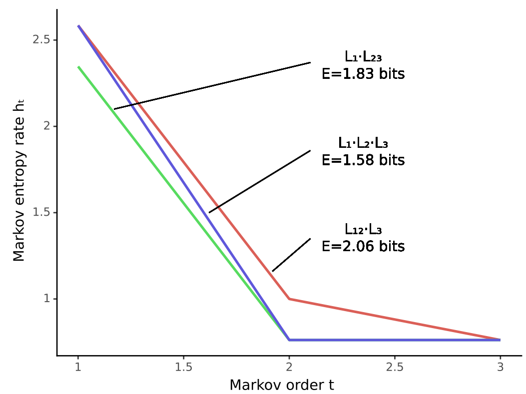

Figure 2: Excess entropies and Markov entropy rates for the three languages. The source induces mutual information between features 2 and 3.

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. Only the first Huffman code is systematic with respect to this decomposition—the first bit corresponds to the first meaning component, and the remaining bits to the second. Figure 1 shows entropy rate curves and excess entropies. The systematic code has lower excess entropy.

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

I consider languages expressing meanings with three binary features shown in Table 2. The first language, notated as $L_1 \cdot L_2 \cdot L_3$, is fully systematic in the three features: I 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 $\text{H}[M_1] = \text{H}[M_2] = \text{H}[M_3] = 1$ 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.

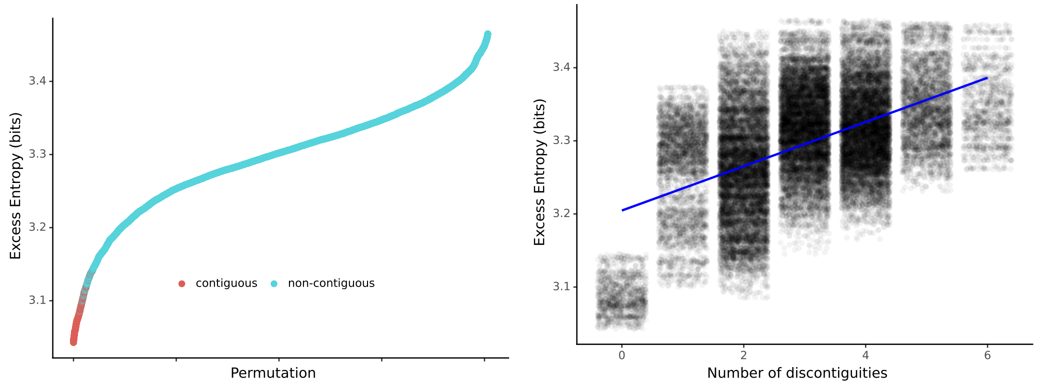

Figure 3: Excess entropy of permuted systematic languages. **Left**, ordered by excess entropy: permutations that maintain contiguity of morphemes in red. **Right**, by number of discontiguities (number of transitions from one morpheme to another within a string, minus one).

To see how systematicity of low-MI features lowers the excess entropy, I compare the fully systematic language $L_1 \cdot L_2 \cdot L_3$ against the partially-systematic $L_{12} \cdot L_3$. 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 the each character X_i encodes the meaning component 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. The conditional entropy of X_3 cannot be reduced without taking more context into account, increasing the 2nd-order Markov entropy rate and thus the excess entropy.

The finding that low-MI features tend to be expressed systematically gives us some traction on the question of how meanings may be decomposed into features in language. In particular, it means that languages constrained to have low excess entropy will appear to be systematic with respect to a set of features that are relatively statistically independent of each other.

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 $\{m^1, \dots, m^{10}\}$, 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 $L(m_k)$ is a mapping from a feature to a morpheme, a random string in $\{\texttt{0}, \texttt{1}\}^4$. 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 $L(m)$. Most of these languages interleave the two morphemes in various ways; a few leave the morphemes contiguous.

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, which is penalized by excess entropy.

4 Conclusion

I have demonstrated through some case studies that codes which have minimal excess entropy seem to have natural-language-like systematicity, in the sense that they consist of morphemes that regularly correspond to features of meaning which are concatenated together. Notably, our approach does not assume that string concatenation is a privileged operation, nor does it assume or require any pre-existing decomposition of meaning into features, nor indeed any structure to the meanings. It appears that languages constrained by excess entropy will tend to factorize meanings into features that are relatively statistically independent, which are then combined together systematically.

References

[1] Jeffrey A Barrett. The evolution of coding in signaling games. *Theory and Decision*, 67: 223–237, 2009.

[2] William Bialek, Ilya Nemenman, and Naftali Tishby. Complexity through nonextensivity. *Physica A: Statistical Mechanics and its Applications*, 302(1-4):89–99, 2001.

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

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

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

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

[7] Fernanda Ferreira and Benjamin Swets. 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, 2002.

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

[9] Gottlob Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892.

[10] Michael Hahn, Judith Degen, and Richard Futrell. Modeling word and morpheme order in natural language as an efficient tradeoff of memory and surprisal. *Psychological Review*, 128 (4):726–756, 2021.

[11] Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119, 2022.

[12] Irene Heim and Angelika Kratzer. *Semantics in Generative Grammar*. Wiley-Blackwell, Malden, MA, 1998.

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

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

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

[16] Willem J M Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA, 1989.

[17] Sarah E Marzen and James P Crutchfield. Predictive rate–distortion for infinite-order Markov procceses. *Journal of Statistical Physics*, 163:1312–1338, 2016.

[18] Sarah E Marzen and James P Crutchfield. Nearly maximally predictive features and their dimensions. *Physical Review E*, 95(5):051301, 2017.

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

[20] Richard Montague. English as a formal language. In Bruno Visentini, editor, *Linguaggi nella società e nella tecnica*, pages 189–223. 1970.

[21] Martin A Nowak and David C Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences*, 96:8028–8033, 1999.

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

[23] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656, 1948.

[24] Kenny Smith, Henry Brighton, and Simon Kirby. Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4):537–558, 2003.

[25] Kenny Smith, Simon Kirby, and Henry Brighton. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003.

[26] Kenny Smith, Monica Tamariz, and Simon Kirby. 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, 2013.

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

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

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

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

[31] Francesca Tria, Bruno Galantucci, and Vittorio Loreto. Naming a structured world: A cultural route to duality of patterning. *PLOS one*, 7(6):e37744, 2012.

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

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

[34] Wlodek Zadrozny. From compositional to systematic semantics. *Linguistics and Philosophy*, 17:329–342, 1994.