Natural-Language-Like Systematicity from a Constraint on Excess Entropy

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Human language is **systematic**: parts of form correspond regularly to components of meaning. For example, in the sentences *I saw the cat*, *a cat ate food*, etc., the part *cat* systematically refers to a particular aspect of meaning. Across languages, these parts are usually combined by concatenation. When they are not,[1] the resulting string still usually has subsequences that correspond to components of meaning, and these parts remain fairly contiguous. We call this property **locality**. Here we argue that local systematicity in natural language arises from minimization of excess entropy, a measure of the complexity of incremental information processing.

Formally, we consider a language to be any mapping $L : \mathcal{M} \to \Sigma^*$ from meanings $\mathcal{M}$ to forms (strings with characters $\in \Sigma$). If a meaning $m \in \mathcal{M}$ can be written as a product of two features as $m = m_1 \times m_2$, then we say a language is systematic if the form can be decomposed as $L(m_1 \times m_2) = L(m_1) \cdot L(m_2)$, with $\cdot$ a string combining function such as concatenation. We seek maximally general principles that explain (1) how meanings are decomposed and (2) why strings are combined locally in natural language.

Excess Entropy For a stationary stochastic process generating symbols $X_1, X_2, \ldots$, the **excess entropy** $\mathbf{E}$ is the mutual information between all the symbols up to an arbitrary time index and all the symbols at or after it (Shalizi and Crutchfield, 2001, §6). Intuitively, excess entropy measures the amount of information that an incremental predictor or generator must store about the past of the process in order to reproduce its future; it is the minimal amount of memory required to achieve the lowest possible average surprisal per character (Hahn et al., 2021). We calculate the excess entropy of a language L as the excess entropy of the stream of characters generated by repeatedly sampling meanings $m \in \mathcal{M}$ from a source $p_{\mathcal{M}}$ and translating them to strings as $s = L(m)$.

[1]For example, in Semitic nonconcatenative morphology, or Celtic consonant mutations, or when concatenation of underlying forms is obscured by phonological processes.

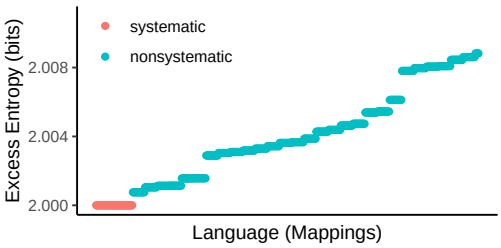

Figure 1: Excess entropy of languages (mappings from meanings to strings) for a source with three independent components, ordered by increasing $\mathbf{E}$.

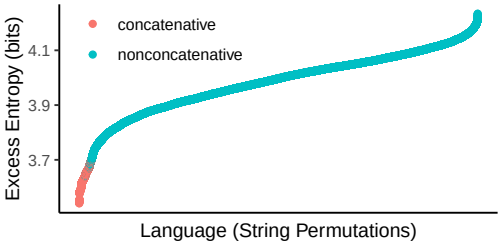

Figure 2: Excess entropy for languages $L_f(m) = f(L(m_1) \cdot L(m_2))$ for all permutations f. Languages that concatenate the two components have the lowest $\mathbf{E}$.

1 Simulations

These simulations show that languages minimize $\mathbf{E}$ when they concatenate substrings that systematically correspond to relatively independent components of the source.

Systematicity We consider all possible bijective languages $\mathcal{M} \to \{0,1\}^3$ for a source with three independent components, $p_{\mathcal{M}} \sim \text{Bernoulli}(.5) \times \text{Bernoulli}(.55) \times \text{Bernoulli}(.6)$. Figure 1 shows that the languages which minimize $\mathbf{E}$ are those which are systematic with respect to these components.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 336-337.
Irvine, California, June 27-29, 2024

	Real	Nonconcat.	Nonsys.	Nonsys. (L)
English	10.1	13.0 (<.01)	10.4 (.01)	10.3 (.01)
Czech	10.3	14.4 (<.01)	10.6 (.01)	10.6 (<.01)

Table 1: Character-level excess entropy (in bits) for adjective–noun pairs in UD corpora, compared with mean **E** (with standard deviation) from 1000 baseline samples. Baselines as in Figure 3.

Locality We consider languages $L_f(m) = f(L(m_1) \cdot L(m_2))$ for all permutations f, with $L(m_i)$ mapping to a random string in $\{0,1\}^4$, and a Zipfian source distribution $p(m) \propto m^{-1}$ over meanings $m \in \{00, 01, \ldots, 99\}$. The permutations f represent different possible string combination functions. Figure 2 shows that concatenation of the two strings yields the lowest excess entropy.

2 Crosslinguistic Corpus Studies

These studies show that systematicity and locality in actual language create minimal excess entropy.

Morphology Figure 3 shows empirical **E** for noun morphology in three languages with systematic (agglutinative) morphology, and one language (Latin) with non-systematic (fusional) morphology. We calculate excess entropy over all morphological forms of the noun, represented as a dummy stem plus affixes (for example, *Xoknak* for the dative plural in Hungarian) using frequencies of morphological features derived from Universal Dependencies (UD) corpora. We compare the empirical excess entropy against three baselines, representing non-systematic and non-concatenative morphology. For the agglutinative languages, real **E** is lower than the majority of baselines.

Syntax Table 1 shows empirical **E**, computed at the character level, for pairs of nouns and adjectives modifying them. These pairs are extracted from UD corpora and compared against baselines representing non-systematic and non-concatenative ways of combining adjective with the noun. The real languages, in which the meaning of an adjective–noun pair is systematic and the words are concatenated (with agreement in Czech), have lower **E** than all baseline samples.

3 Conclusion

We argue that natural languages are codes that minimize excess entropy. We showed that codes which minimize **E** consist of concatenated substrings corresponding to approximately independent components of the source, and that systematicity in natural languages coincides with minimization of **E**.

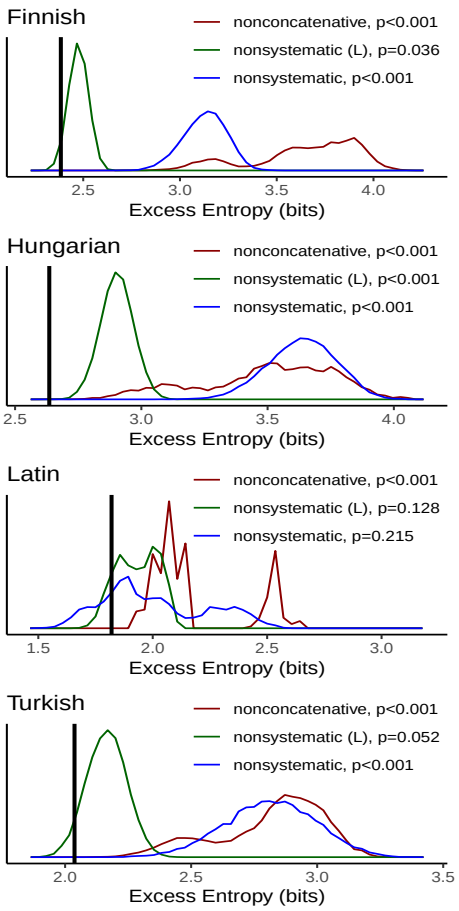

Figure 3: Empirical **E** (black vertical lines) of noun morphology, compared with three baselines: (1) a **nonconcatenative** baseline where the characters in each form are permuted; (2) a **nonsystematic** baseline where the assignments of forms to meanings (sets of features) is shuffled; and (3) a length-controlled **nonsystematic (L)** baseline which shuffles form–meaning assignments while preserving form length. p indicates the proportion of baseline samples with lower **E** than real forms.

References

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

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