Crosslinguistic patterns of phrasal word order support efficient prediction

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

Abstract

Across languages, word order within phrases shows robust statistical regularities: dependents are typically ordered consistently by type, word order tends to be harmonic with the head near an edge, and more strongly associated dependents tend to occur closer to the head. Existing accounts often derive these properties from separate mechanisms or stipulations. I propose a unified efficiency-based explanation grounded in information theory: the idea is that languages are organized so that it is possible to make good predictions using only small amounts of memory. Using simulations of nouns modified by multiple adjectives differing in association strength to the noun, I compare linguistically motivated ordering policies that vary in consistency, harmony, and locality. Across sampled source distributions, I find that memory requirements for prediction are minimal in orders that are simultaneously harmonic, consistent, and local. I analyze why these effects arise: consistency makes upcoming category structure predictable from small amounts of context, and harmony makes phrase boundaries easy to anticipate. The results suggest that diverse crosslinguistic patterns of phrasal order may reflect a general pressure to make incremental prediction simple.

Keywords: language efficiency; information theory; word order; language and memory

Introduction

Across languages, the order of elements within phrases shows constrained variation, often of a statistical nature where some orders are overwhelmingly more common than others (Greenberg, 1963; Verkerk et al., 2026). By elements in a phrase, I am referring to a word (called a head) and all the words that directly syntactically modify it (called dependents). For example, in the English noun phrase *those three cute cats* the head noun is *cats* and the other words are its dependents. The words must go in this kind of order in English; alternatives like **three those cats cute* are ungrammatical.

Below I summarize three key general properties of phrasal order across languages. These general properties will be my explananda for the paper.

- **Consistency.** The order of dependents depends primarily only on their type. For example, in English, noun phrases have the order Determiner–Numeral–Adjective–Noun, regardless of exactly which determiners, numerals, adjectives, or nouns are involved. One rarely (although not never) finds inconsistencies, for example, languages where a determiner must go first in the noun phrase if it is *those* but third if it is *these*.

- **Harmony.** Across types, dependents tend to be primarily to the right of the head or primarily to the left of the head, regardless of dependent or head type (Greenberg, 1963; Vennemann, 1974; Dryer, 1992). The result is that the head is typically *peripheral* (initial or final) within its phrase. For example, in unrelated languages like Japanese, Quechua, and Tamil, nouns appear at the end of noun phrases and verbs appear at the end of verb phrases; these languages are close to strictly head-final (Kuno, 1973; Cole, 1982; Lehman, 1993). Exceptions are common: for example, Spanish basic noun phrase order is Determiner–Numeral–Noun–Adjective (*esos tres gatos lindos*) with the numeral and determiner before the noun but (most) adjectives after it.

- **Locality.** The order of dependents often seems to be determined by some notion of 'distance' to the head. The nature of this distance varies in different theories. Dependency locality, as a theory, holds that word order is organized in a way that minimizes linear distance from heads to dependents (Hawkins, 2004; Liu et al., 2017; Temperley and Gildea, 2018; Futrell et al., 2020b). Other theories have held that various measures of semantic or statistical association can predict the relative order of dependents, with dependents more associated with the noun appearing closer (Givón, 1991; Futrell, 2019; Culbertson et al., 2020).

The locality property, in the form of association, is well exemplified by the order of adjectives when multiple adjectives modify a noun. For example, English speakers find orders such as *small cute cat* more acceptable than *cute small cat*, which either seems wrong or like it expresses some marked meaning. The ordering of adjectives, which shows universal tendencies across languages (Dixon, 1982; Sproat and Shih, 1991; Scontras, 2023), seems to be predictable to some extent from statistical association with the noun in the form of pointwise mutual information (pmi), with adjectives more associated with the noun (such as *cute*) going closer to it (Futrell et al., 2020a; Dyer et al., 2023). Culbertson et al. (2020) find that locality based on pmi can explain the relative ordering of color and shape adjectives as well as numerals and determiners in the noun phrase, suggesting that this locality idea is on the way to providing a unified view of the relative order of dependents in a phrase.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

My main result is to show that, in simple simulations, a single notion of processing efficiency (the amount of memory required to make good predictions) robustly favors all three of consistency, harmony, and locality. I quantify the memory required for prediction using the information-theoretic concept of predictive information. I suggest that this single concept may provide an explanation for why languages are structured this way.

Related Work

Existing accounts of phrasal order have derived consistency, harmony, and locality from separate stipulations.

For example, dependency locality on its own predicts that the head should be *central* in its phrase, with its dependents flanking it to the left and right, violating harmony, and if the dependents themselves are phrases of different lengths, then their order should be determined by those lengths, violating consistency. A reasonable match to attested word orders (including capturing harmony to some extent) is only achieved through an additional stipulation of consistency (Gildea and Temperley, 2007, 2010; Hahn et al., 2020; Hahn and Xu, 2022).

Harmony, on the other hand, has often been explained through an innate head-direction parameter in learners which applies globally across syntactic categories (Travis, 1984; Chomsky, 1995; Lasnik and Lohndal, 2010). This parameter on its own does not capture consistency or locality, although it is typically embedded within a framework that implicitly stipulates consistency.

Another class of explanations involves a simplicity bias in grammars: the idea that consistency and harmony are properties of grammars with a relatively small number of (meta-)rules, which are thus easier to learn (Culbertson and Kirby, 2016, 2022). However, to explain locality, this kind of approach requires many stipulations about the set and order of syntactic categories within the grammar—for example, what I have been calling noun phrase order might be explained in terms of a structure where a determiner phrase contains a numeral phrase, which contains an adjective phrase, etc. (Cinque, 2005). Extending this approach to adjective order requires a very large number of syntactic categories for different semantic classes of adjectives, along with a fixed ordering over them (Cinque, 1994; Cinque and Rizzi, 2008; Leung et al., 2020), for example a 'size phrase' containing a 'nationality phrase' containing a 'color phrase', etc.[1]

There have also been attempts to explain similar phenomena in terms of simple dynamics of diachronic language change and chunking in language production. Mansfield and Kemp (2025) aim to explain locality in this way, and Mansfield and Krapp (2025) aim for harmony, finding that within their model of language change, syntactic categories are driven to the edge of a phrase when they are not optional. Consistency is also related to the idea of category clustering in this literature (Mansfield et al., 2020, 2022).

These accounts are complementary to the one I am advancing. While they provide particular chunking algorithms which might be good descriptions of language producers and learners, I will provide a complexity metric (predictive information) that circumscribes the performance of *any possible agent* that predicts or produces language using any possible algorithm. In particular, if language is structured in a way that yields low predictive information, then it will also be easy to produce using a production algorithm based on short chunks as found in Mansfield and Kemp (2025), as well as any number of other similar algorithms.

Predictive Information

Imagine you are faced with a linguistic context, such as *those three* . . . , and you have not yet encountered its continuation, such as . . . *cute cats*. If you are a comprehender, then you are already forming expectations about what might follow (Clark, 2013; Ryskin and Nieuwland, 2023), and the way you process the continuation will depend on how predictable it is (Smith and Levy, 2013). If you are a learner, then the amount of information the continuation gives you about the language will depend on how well you can predict it from the context (Futrell, 2025). If you are a producer and you just produced the context, then your task is to produce a continuation that expresses your thought while not deviating too much from the grammatical and statistical expectations set by the context (Levy and Jaeger, 2007; Futrell, 2023).

In all cases, your task involves extracting and/or storing information about the context in order to predict (or generate) the continuation. Thus, I propose that a natural complexity for metric for language is *how much information* must be extracted from contexts in order to predict continuations: a language is more complex if good predictions require more memory for context. An agent with limited memory capacity would be unable to accurately comprehend, learn, or produce a language that imposes too high a memory requirement.

I operationalize this idea primarily using **predictive information (PI)**, which is the mutual information between contexts and continuations (Bialek et al., 2001; Shalizi and Crutchfield, 2001). Predictive information is the minimal average amount of information about contexts which must be used by any agent that predicts continuations with the best possible accuracy.

More technically, predictive information is defined in terms of stochastic processes. A stochastic process, such as a stream of observed forms in a language, can be thought of as a probability distribution on sequences of symbols (such as words) labeled with time indices as $\ldots, X_{-1}, X_0, X_1, \ldots$. Let the process be stationary, meaning that the distribution on each symbol does not depend on the numerical value of the time index.

[1] These accounts might seem to become less stipulative if one posits that the categories and their order are induced by learners based on the statistics of their input (Culbertson et al., 2020). But then one is then left with the question of why their input is structured in such a way that gives rise to certain categories in certain orders rather than others, and/or why learners prefer to process their input into certain categories and orders rather than others.

Predictive information is the mutual information[2] between an unboundedly long block of symbols before an arbitrary time index 0 and everything that follows:

$$\mathrm{PI} \triangleq \mathrm{I}[X_{<0} : X_{\geq 0}]. \tag{1}$$

Remarkably, the computation of the actual value of predictive information is simple in a number of cases. Consider a language containing a finite number of strings, strings are drawn from a vocabulary of symbols Σ. We can consider the stochastic process generated by randomly sampling strings from the language according to some source distribution reflecting communicative need (Kemp et al., 2018), and concatenating the strings end-to-end with a delimiter. In that scenario, PI can be calculated using ***n*-gram entropy** values h_n, which represent the uncertainty about the n'th symbol given the block of $n-1$ previous symbols $X_1, \ldots, X_{n-1}$:

$$\begin{aligned} h_n &\triangleq \mathrm{H}[X_n \mid X_1, \ldots, X_{n-1}] \\ &= -\sum_{x_1,\ldots,x_n \in \Sigma^n} p(x_1, \ldots, x_n) \log p(x_n \mid x_1, \ldots, x_{n-1}). \end{aligned} \tag{2}$$

The n-gram entropies h_n get smaller with larger n. In a language whose longest string has length N, the lowest n-gram entropy is given by h_N. Then the PI can be calculated in terms of how quickly the h_n curve approaches h_N:

$$\mathrm{PI} = \sum_{n=1}^{N} (h_n - h_N), \tag{3}$$

as illustrated in Figure 1 (Crutchfield and Feldman, 2003). For a language with a known probability distribution, a small maximum length N, and a small vocabulary Σ, it is not hard to compute Eq. 3 directly.

The formula Eq. 3 lets one exactly compute PI for simulated languages, and it also reveals something about what PI means, and what it would mean for a language to be configured in a way that minimizes PI. Eq. 3 will be small when h_n is close to h_N for small n: that is, when it is possible to predict symbols as well as possible on the basis of only small windows of context. Thus, PI will be relatively low in languages where statistical dependencies are *mostly* local—hinting at the locality property described in the Introduction.

An important limitation is that PI (in fact, the whole h_n curve: Shannon, 1951, p. 58) is invariant to time reversal. This means that the PI for any word order will be exactly the same as the PI for the mirror image order. This symmetry means that we cannot distinguish between mirror image orders (for example, ABCD vs. DCBA) in a theory based on minimizing PI.

Predictive Information and Language

Predictive information has been proposed as a general complexity measure for language by Futrell and Hahn (2025), who

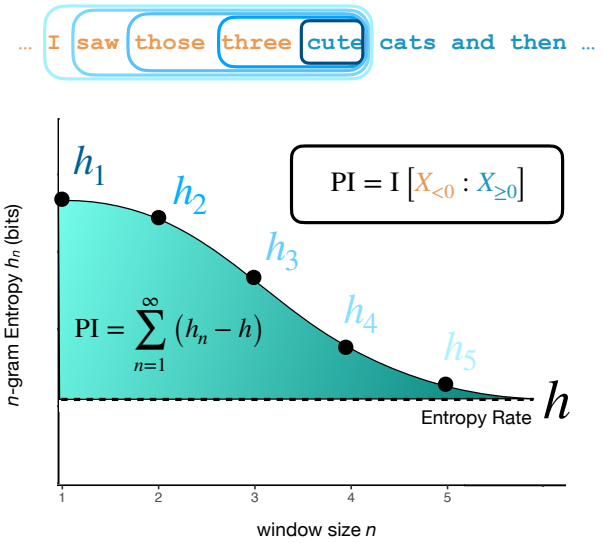

Figure 1: Schematic for calculation of predictive information. Each n-gram entropy h_n is the entropy for the last in a window of n words. As window sizes increases, the n-gram entropy h_n decreases down to an asymptotic value called entropy rate h. Predictive information, which is the mutual information between past and future blocks of words, can be calculated from the h_n curve as shown.

show that a number of basic properties of human language emerge in codes that are structured in a way that keeps PI low. They also report some results on modeling the crosslinguistic distribution of noun phrase orders. Rathi et al. (2025) find that, in simulations, PI can predict speakers' soft preferences for a number of syntactic alternations. PI, under the name 'excess entropy', has also been applied to language in the quantitative linguistics and statistical physics literatures (Ebeling and Pöschel, 1994; Debowski, 2011). Regarding the relation between PI and learning, Someya et al. (2025) find that languages are more learnable by autoregressive language models when they have low n-gram entropy h_n.

Predictive information is closely related to the memory–surprisal tradeoff (Hahn et al., 2021a,b; Rathi et al., 2022, 2026), which also expresses the intuition that languages are complex when good predictions can only be made using a large amount of memory. The memory–surprisal tradeoff is also based on the h_n curve,[3] and we will see that both PI and the memory–surprisal tradeoff can model phrasal orders equally well in simulations. The difference in interpretation between the two measures is subtle: the memory–surprisal tradeoff describes the extent to which the future can be predicted *approximately* on the basis of small amounts of information about the past, whereas PI describes the minimal amount of information about the past needed for the best possible pre-

[2] I assume basic familiarity with information-theoretic quantities of entropy and mutual information. For an introduction/reference, see Cover and Thomas (2006, Ch. 2).

[3] Complexity according to the memory–surprisal tradeoff is the area under the curve defined by $x = \sum_{n=1}^{N} (h_n - h_N)$, $y = h_{N+1} - h$ for $N = 1, 2, \ldots$.

diction of the future. I consider these two measures to be complementary quantifications of the same intuitive idea: that languages are simple when little memory is needed to make good predictions. The results I show here do not distinguish between them.

I also compare against another information-theoretic notion of linguistic complexity, Uniform Information Density (UID: Fenk and Fenk, 1980; Levy and Jaeger, 2007; Clark et al., 2023), which holds that language is processed most efficiently when the incremental surprisal of each word is low variance. Unlike the memory–surprisal tradeoff, surprisal variance has no known connection to the amount of memory needed for prediction, nor to the h_n curve. I will find that UID does not have predictive power for the word order patterns of interest.

Simulations

My specific aim is to simulate word order in phrases consisting of a head and a variable number of single-word dependents, and to calculate predictive information for the different orders. To make things more concrete, when describing the simulations below, I will refer to the head as the *noun* and the dependents as *adjectives*, but the simulation is really meant to model any phrases with a similar distribution—in particular, it applies most directly to heads modified by adjuncts. In the discussion, I will address how my results carry over to other linguistic dependencies.

Below, I describe a source distribution over nouns and sets of adjectives modifying them. I then describe a number of ordering policies for the nouns and adjectives, and evaluate the predictive information of noun phrases under those policies.

Source distribution I create an artificial source distribution meant to capture the intuition that a noun may be modified by multiple adjectives, organized roughly into semantic types that have different degrees of association with it.

Concretely I simulate a source distribution over nouns modified by up to 5 adjectives, drawn from a set of C adjective classes $\{1,\dots,C\}$ (for example, if $C=3$ one can think of these as the classes of shape, color, and size adjectives). There is a vocabulary of nouns of size V_N and, within each adjective class, a distinct vocabulary of adjectives of size V_A. The probability to generate a noun n modified by adjectives $a_1,\dots,a_K$ is

$$p(n,a_1,\dots,a_K) \propto p_\mathrm{noun}(n)\prod_{i=1}^{K}(1-p_\mathrm{halt})\,p(a_k \mid n), \quad (4)$$

where the conditional probability for an adjective a belonging to class c given a noun n is

$$p(a \mid n) \propto p_\mathrm{adj}(a)\exp\left(\frac{\beta}{c}E(a,n)\right), \quad (5)$$

where $E(a,n)$ is a random number drawn from a standard Normal distribution for each adjective a and noun n. This probability distribution induces probabilistic coupling between the adjectives and nouns, with weaker coupling for higher adjective class index c. The underlying distributions $p_\mathrm{noun}(\cdot)$ and $p_\mathrm{adj}(\cdot)$ provide baseline frequencies for nouns and adjectives. These are Zipfian over the vocabulary of nouns or adjectives in the appropriate class: for nouns n_i numbered $i=1,\dots,V_\mathrm{N}$ I set $p_\mathrm{noun}(n_i)\propto i^{-1}$, and for adjectives a_j numbered $j=1,\dots,V_\mathrm{A}$ within class c I set $p_\mathrm{adj}(a_j)\propto \frac{1}{C}j^{-1}$. I run simulations with $\beta=2$, $C=3$, $V_\mathrm{N}=V_\mathrm{A}=5$, and $p_\mathrm{halt}=1/2$, and average the results over 100 randomly-sampled source distributions. For the first, second, and third adjective classes, their average mutual information with the noun is 0.68, 0.29, and 0.15 bits respectively.

Statistically, this distribution creates noun phrases where adjectives appearing together are conditionally independent of each other given the noun. However, the adjectives are *unconditionally* dependent on each other, so adjective–adjective predictability is an important factor for ordering. This conditional independence pattern is what is expected under a probabilistic context-free grammar for such phrases.

Ordering policies Of the many logically possible ordering policies for adjectives and nouns, I consider a few which are of linguistic interest. The orders are given along with a mnemonic representation, where N is the noun, $ABC\dots$ are adjective classes by descending MI with nouns, and $abc\dots$ are specific adjectives by descending pmi with the specific noun.

1. Consistent Local Harmonic (*NABC*): The noun appears initially, and adjectives are ordered consistently by class in order of decreasing MI with nouns. This is (the mirror image of) the kind of adjective order found in English.

2. Consistent Nonlocal Harmonic (*NCBA*): The same as the above, except adjective classes are ordered by increasing MI with nouns.

3. Consistent Local Medial (*ANBC*): Order consists of any class-1 adjectives, followed by the noun, followed by any remaining adjectives consistently by class in order of decreasing MI with nouns. This is the kind of adjective order found in Romance languages like French: there is a certain class of adjectives that may appear before the noun, otherwise they appear after.

4. Consistent Local Nonharmonic (*BNAC*): Starting from the noun in the middle, adjectives are ordered outward alternating on the left and right, in order of increasing class MI. This order minimizes the average distance from nouns to associated adjectives; it is inspired an algorithm for minimizing dependency length (Gildea and Temperley, 2007).

5. Consistent Nonlocal Nonharmonic (*ACNB*): Like the above, but adjectives are placed in order of *decreasing* class MI with nouns, outward from the noun on alternating sides.

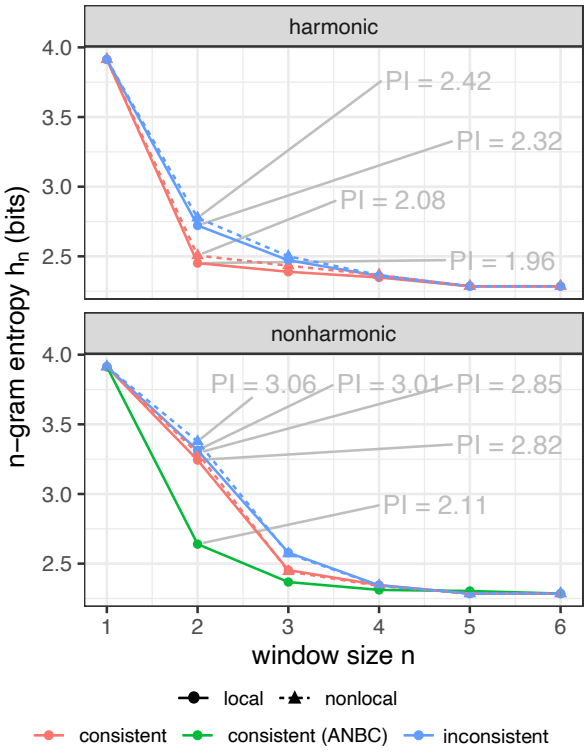

Figure 2: Curves of n-gram entropy h_n for the simulated phrases under different ordering policies (average of 100 sampled distributions). PI (in bits) of each ordering policy is indicated in gray. Harmonic, consistent, and local orders give the lowest PI.

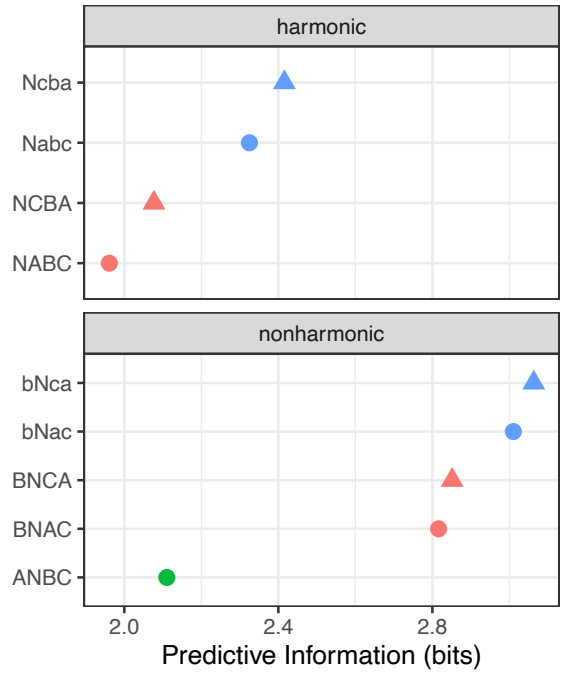

Figure 3: Predictive information for different ordering policies (average of 100 sampled source distributions). Shapes and colors as in Figure 2.

6. Inconsistent Local Harmonic (*Nabc*): The noun appears initially, and adjectives are ordered by decreasing pmi with the specific noun, regardless of adjective class.

7. Inconsistent Nonlocal Harmonic (*Ncba*): the same as the above, but adjectives are ordered by increasing pmi with the spcific head noun.

8. Inconsistent Local Nonharmonic (*bNac*): Adjectives are ordered by decreasing pmi with the specific noun, alternating left and right.

9. Inconsistent Nonlocal Nonharmonic (*Nabc*): The same as the above, but adjective are ordered by increasing pmi with the specific noun.

Results

As seen in Figure 2, Table 1, and Figure 3, I find that the consistent, harmonic, and local orders generate robustly lower PI than other orders. The memory–surprisal tradeoff yields the same result. UID yields no typologically interesting pattern.

Among these ordering policies, the most important factor for lowering PI is harmony, followed by consistency and locality.An interesting exception to the importance of harmony is the Romance-type *ANBC* order, in which any class-1 adjective goes before the noun, and all other adjectives go after it. This order is different from the other nonharmonic orders because the exceptional adjective class, which goes before the noun, is always consistent (always class 1), whereas in the other nonharmonic orders, the exceptional adjective class is determined by the relative mutual information with the noun.

Why does it work?

Why are consistency and harmony so advantageous from the perspective of predictive information? The answer is that, although these properties may move associated adjectives farther from the noun, they result in the noun phrase *as a whole* being more predictable from small amounts of contexts.

Consistency. Consider the case when we have three adjectives modifying a noun. If these are arranged by type, as *NABC*, then after encountering an *B*-type adjective, we know that what follows must be either another *B*-type adjective, a *C*-type adjective, or the end of the noun phrase. It is thus possible to make a good prediction about the next symbol while only considering a small amount of information about the context (that the most recent word was an *B*-type adjective). On the other hand, if adjectives are arranged by decreasing pmi with the noun, as *Nabc*, then after encountering adjective *b*, we know that what follows is either the end of the phrase or any adjective whose pmi with the noun is lower than *b*. These predictions are dependent on knowing *both* the specific adjective *b* and the noun *N*, and thus require more predictive

Ordering Policy	Harmonic?	Consistent?	Local?	Predictive Info.	Memory–Surprisal AUC	Surprisal Variance
NABC	✓	✓	✓	1.96	5.83	5.03
NCBA	✓	✓	✗	2.08	6.11	4.86
ANBC	∼	✓	✓	2.11	6.21	5.02
Nabc	✓	✗	✓	2.32	6.76	4.98
Ncba	✓	✗	✗	2.42	6.99	4.90
BNAC	✗	✓	✓	2.82	8.24	5.67
BNCA	✗	✓	✗	2.85	8.36	5.56
bNac	✗	✗	✓	3.01	8.78	5.77
bNca	✗	✗	✗	3.06	8.97	5.77

Table 1: Predictive information, area under the memory–surprisal tradeoff curve, and surprisal variance, all in units of bits, for the simulated phrases for different ordering policies (average of 100 sampled source distributions). Predictive information and the memory–surprisal tradeoff favor ordering policies that are consistent, harmonic, and local. Surprisal variance shows no typologically interesting pattern. The PI results are visualized in Figure 3.

information.

Harmony. Harmony (that is, head peripherality) lowers PI because it makes the edge of the phrase easy to predict. Once one encounters the noun *N*, one knows that the phrase boundary is adjacent. The noun is an effective and simple marker of the phrase boundary in these simulations because (1) there must be at least one noun per phrase, (2) there may be no more than one noun per phrase, and (3) the nouns are lexically distinct from the adjectives. This motivation for harmony mirrors intuitions from Bever (1970) and Mansfield and Krapp (2025).

Linguistic generality

The phrases in this simulation most naturally capture heads modified by *adjuncts*, which are optional and can be iterated (for example, many adjectives can modify a noun). More generally, heads can also have *arguments*, which are singular and (often) mandatory (for example, a transitive verb takes exactly one direct object). In this connection, Futrell and Hahn (2025) already report some positive results for modeling the order of noun–determiner and noun–numeral dependencies using PI, based on realistic source distributions. More generally, I believe the results here will largely carry over to argument dependencies. In terms of predictive information, harmony is still beneficial for argument dependencies as long as the head is lexically distinct from its arguments (and otherwise I expect less of a pressure for harmony). As for consistency and locality, the factors that make them beneficial for predictive information in my simulations will also hold for argument dependencies.

Limitations

The simulation setting involves only small vocabularies and dependents consisting of single words. It models adjunct dependencies more naturally than argument dependencies. Most seriously, it uses artificial source distributions which are meant to idealize real distributions, but which are likely different from them in important ways. It will be important to see how well the theory fares using realistic data and probability distributions. This will require overcoming nontrivial challenges in estimation of PI from sparse data.

Taking the case of adjective order in particular, it is not yet clear to what extent PI and related ideas can provide a complete theory of the crosslinguistic phenomena. Some challenges include, for example, the fact that ordering constraints are weaker or non-existent when a conjunction marker or linker intervenes between adjectives (Rosales Jr. and Scontras, 2019; Samonte and Scontras, 2019), and any left–right asymmetries which may exist, which remain unaccounted for.

Conclusion

I have given simple simulation-based evidence that a theory of word order based on predictive complexity—the idea that languages are complex for comprehenders, producers, and learners when good predictions can only be made using a large amount of memory—can capture general patterns of phrasal order in a simpler way than existing accounts. I see that indices of predictive complexity are naturally lower in languages where dependents are ordered *consistently by type* in order of degree of association with the head, and that word order should be harmonic: the relative order of head and dependent should be consistent across different classes, such that the head is consistently final or initial. Alternative accounts can only account for these phenomena through multiple different assumptions.

Acknowledgments

I thank Greg Scontras, Connor Mayer, David Inman, and the UCI QuantLang Collective for comments. Code to reproduce these results is available at `https://github.com/Futrell/infolocality`.

References

Bever, T. G. (1970). The cognitive basis for linguistic structures. In Hayes, J. R., editor, *Cognition and the Development of Language*. Wiley, New York.

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

Chomsky, N. (1995). *The Minimalist Program*. Cambridge University Press.

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In *Paths towards Universal Grammar*, pages 85–110. Georgetown University Press.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, 36(3):315–332.

Cinque, G. and Rizzi, L. (2008). The cartography of syntactic structures. *Studies in Linguistics*, 2:42–58.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36:181–204.

Clark, T. H., Meister, C., Pimentel, T., Hahn, M., Cotterell, R., Futrell, R., and Levy, R. (2023). A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Cole, P. (1982). *Imbabura Quechua*, volume 5 of *Lingua Descriptive Studies*. North-Holland, Amsterdam.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

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.

Culbertson, J. and Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, 6:1964.

Culbertson, J. and Kirby, S. (2022). Syntactic harmony arises from a domain-general learning bias. In Culbertson, J., Rabagliati, H., Ramenzoni, V. C., and Perfors, A., editors, *Proceedings of the 44th Annual Meeting of the Cognitive Science Society, CogSci 2022*, Toronto.

Culbertson, J., Schouwstra, M., and Kirby, S. (2020). From the world to word order: Deriving biases in noun phrase order from statistical properties of the world. *Language*, 96(3):696–717.

Dixon, R. M. W. (1982). *Where Have All the Adjectives Gone? And Other Essays in Semantics and Syntax*. Mouton, Berlin, Germany.

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

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68(1):81–138.

Dyer, W., Torres, C., Scontras, G., and Futrell, R. (2023). Evaluating a century of progress on the cognitive science of adjective ordering. *Transactions of the Association for Computational Linguistics*, 11:1185–1200.

Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters*, 26(4):241.

Fenk, A. and Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie*, 27(3):400–414.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In Chen, X. and Ferrer-i-Cancho, R., editors, *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

Futrell, R. (2023). Information-theoretic principles in incremental language production. *Proceedings of the National Academy of Sciences*, 120(39):e2220593120.

Futrell, R. (2025). Language learning as codebreaking: The key roles of redundancy and locality. In Anderson, C. J., Mailhot, F., and Prasad, G., editors, *Proceedings of the Society for Computation in Linguistics 2025*, pages 54–63, Eugene, Oregon. Association for Computational Linguistics.

Futrell, R., Dyer, W., and Scontras, G. (2020a). What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2003–2012, Online. Association for Computational Linguistics.

Futrell, R. and Hahn, M. (2025). Linguistic structure from a bottleneck on sequential information processing. *Nature Human Behaviour*.

Futrell, R., Levy, R. P., and Gibson, E. (2020b). Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–413.

Gildea, D. and Temperley, D. (2007). Optimizing grammars for minimum dependency length. In Zaenen, A. and van den Bosch, A., editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic. Association for Computational Linguistics.

Gildea, D. and Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.

Givón, T. (1991). Isomorphism in the grammatical code: Cognitive and biological considerations. *Studies in Language*, 15(1):85–114.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of Language*, pages 73–113. MIT Press, Cambridge, MA.

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., Jurafsky, D., and Futrell, R. (2020). Universals of word order reflect optimization of grammars for efficient

communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353.

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

Hahn, M. and Xu, Y. (2022). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences*, 119(24):e2122604119.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press, Oxford.

Kemp, C., Xu, Y., and Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Kuno, S. (1973). *The Structure of the Japanese Language*, volume 3 of *Current Studies in Linguistics*. MIT Press, Cambridge, MA.

Lasnik, H. and Lohndal, T. (2010). Government–binding/principles and parameters theory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):40–50.

Lehman, T. (1993). *A Grammar of Modern Tamil*. Institute of Language and Culture, Pondicherry.

Leung, J. Y., Emerson, G., and Cotterell, R. (2020). Investigating cross-linguistic adjective ordering tendencies with a latent-variable model. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4016–4028, Online. Association for Computational Linguistics.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19:849–856.

Liu, H., Xu, C., and Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Mansfield, J. and Kemp, C. (2025). The emergence of grammatical structure from inter-predictability. In *A Festschrift for Jane Simpson*, pages 100–120. ANU Press.

Mansfield, J. and Krapp, L. S. (2025). A simple explanation for harmonic word order. *Cognitive Science*, 49(4):e70056.

Mansfield, J., Saldana, C., Hurst, P., Nordlinger, R., Stoll, S., Bickel, B., and Perfors, A. (2022). Category clustering and morphological learning. *Cognitive Science*, 46(2):e13107.

Mansfield, J., Stoll, S., and Bickel, B. (2020). Category clustering: A probabilistic bias in the morphology of verbal agreement marking. *Language*, 96(2):255–293.

Rathi, N., Futrell, R., and Jurafsky, D. (2025). Soft production preferences emerge from a bottleneck on memory. In Ruggeri, A., Barner, D., Walker, C., and Bramley, N., editors, *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, pages 65–72.

Rathi, N., Hahn, M., and Futrell, R. (2022). Explaining patterns of fusion in morphological paradigms using the memory–surprisal tradeoff. In Culbertson, J., Rabagliati, H., Ramenzoni, V. C., and Perfors, A., editors, *Proceedings of the 44th Annual Meeting of the Cognitive Science Society, CogSci 2022*, Toronto.

Rathi, N., Hahn, M., and Futrell, R. (2026). Towards an information-theoretic model of morphological fusion based on an efficient tradeoff of memory and surprisal. *Language*.

Rosales Jr., C. M. and Scontras, G. (2019). On the role of conjunction in adjective ordering preferences. *Proceedings of the Linguistic Society of America*, 4:32–1.

Ryskin, R. and Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences*, 27(11):1032–1052.

Samonte, S. and Scontras, G. (2019). Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. *Proceedings of the Linguistic Society of America*, 4:33–1.

Scontras, G. (2023). Adjective ordering across languages. *Annual Review of Linguistics*, 9:357–376.

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. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

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

Someya, T., Svete, A., DuSell, B., O'Donnell, T. J., Giulianelli, M., and Cotterell, R. (2025). Information locality as an inductive bias for neural language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27995–28013, Vienna, Austria. Association for Computational Linguistics.

Sproat, R. and Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In Georgopoulos, C. and Ishihara, R., editors, *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, pages 565–593. Kluwer Academic, Dordrecht, Netherlands.

Temperley, D. and Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:1–15.

Travis, L. (1984). *Parameters and effects of word order variation*. PhD thesis, Massachusetts Institute of Technology.

Vennemann, T. (1974). Theoretical word order studies: Results and problems. *Papiere zur Linguistik*, 7:5–25.

Verkerk, A., Shcherbakova, O., Haynie, H. J., Skirgård, H., Rzymski, C., Atkinson, Q. D., Greenhill, S. J., and Gray, R. D. (2026). Enduring constraints on grammar revealed by Bayesian spatiophylogenetic analyses. *Nature Human Behaviour*, 10(1):126–136.