Information-theoretic locality properties of natural language

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

Abstract

I present theoretical arguments and new empirical evidence for an information-theoretic principle of word order: information locality, the idea that words that strongly predict each other should be close to each other in linear order. I show that information locality can be derived under the assumption that natural language is a code that enables efficient communication while minimizing information-processing costs involved in online language comprehension, using recent psycholinguistic theories to characterize those processing costs information-theoretically. I argue that information locality subsumes and extends the previously-proposed principle of dependency length minimization (DLM), which has shown great explanatory power for predicting word order in many languages. Finally, I show corpus evidence that information locality has improved explanatory power over DLM in two domains: in predicting which dependencies will have shorter and longer lengths across 50 languages, and in predicting the preferred order of adjectives in English.

1 Introduction

The field of functional linguistics has long argued that the distinctive properties of natural language are best explained in terms of what makes for an efficient communication system under the cognitive constraints particular to human beings. The idea is that the properties of language are determined by the pressure to enable efficient communication while minimizing the information-processing effort required for language production and comprehension by humans.

Within that field, a particularly promising concept is the principle of **dependency length minimization** (DLM): the idea that words linked in syntactic dependencies are under a pressure to be close in linear order. DLM provides a single unified explanation for many of the word order properties of natural language: Greenberg's harmonic word order universals (Greenberg, 1963; Dryer, 1992; Hawkins, 1994, 2004, 2014) and exceptions to them (Temperley, 2007), the rarity of crossing dependencies (Ferrer-i-Cancho, 2006) (corresponding to deviations from context-free grammar: Kuhlmann, 2013), ordering preferences based on constituent length such as Heavy NP shift (Wasow, 2002; Gildea and Temperley, 2010), and the statistical distribution of orders in treebank corpora (Liu, 2008; Futrell et al., 2015). See Liu et al. (2017) and Temperley and Gildea (2018) for recent reviews. The theoretical motivation for DLM is based on efficiency of language processing: the idea is that long dependencies tax the working memory capacities of speakers and listeners (Gibson, 1998, 2000); in line with this view, there is observable processing cost in terms of reading time for long dependencies (Grodner and Gibson, 2005; Bartek et al., 2011).

At the same time, there have been attempts to derive the properties of human language formally from information-theoretic models of efficiency (Ferrer-i-Cancho and Solé, 2003; Ferrer-i-Cancho and Díaz-Guilera, 2007). But it is not yet clear how a principle such as DLM, which appears to be necessary for explaining the syntactic properties of natural language, would fit into these theories, or more generally into the information-theoretic view of language as an efficient code. The motivation for DLM is based on heuristic arguments about memory usage and on empirical results from studies of online processing, and it is not clear how to translate this motivation into the language of information theory.

Here I bridge this gap by providing theoretical arguments and empirical evidence for a new, information-theoretic principle of word order, grounded in empirical findings from the psycholinguistic literature and in the theory of communication in a noisy channel. I assume that linguistic speakers and listeners are processing language incrementally using lossy memory representations of linguistic context. Under these circumstances, we can derive a principle of **information locality**, which states that an efficient language will minimize the linear distance between elements with high **mutual information**, an information-theoretic measure of how strongly two words predict each other. Furthermore, assuming a particular probabilistic interpretation of dependency grammar (Eisner, 1996; Klein and Manning, 2004), I show that DLM falls out as an approximation to information locality. Finally, I

present two new pieces of empirical evidence that information locality provides improved explanatory power over DLM in predicting word orders in corpora.

The remainder of the paper is structured as follows. Section 2 reviews relevant psycholinguistic results and information-theoretic models of online processing difficulty, concluding that they are inadequate for predicting word order patterns. Section 3 shows how to derive the principle of information locality from a modified model of online processing difficulty, and how DLM can be seen as a special case of information locality. In Section 4, I give corpus evidence that information locality makes correct predictions in two cases where DLM makes no predictions: in predicting the distance between words in dependencies in general across 50 languages, and in predicting the relative order of adjectives in English.

2 Background: Efficient communication under information processing constraints

I am interested in the question: What would a maximally efficient communication system look like, subject to human information processing constraints? To answer this question, we need a model of those information processing constraints. Here I review a leading theory of information processing constraints operative during on-line language comprehension, called **surprisal theory** (Hale, 2001; Levy, 2008; Hale, 2016), which is mathematically grounded in information theory, and discuss the relevance of surprisal theory for word order patterns in languages. Perhaps surprisingly, it turns out surprisal theory has very little to say about word order, which will necessitate an update to the theory described in Section 3.

Surprisal theory holds that the incremental processing difficulty for a word w given preceding context c (comprising the previous words as well as extra-linguistic context) is proportional to the **surprisal** of the word given the context:

$$\text{Difficulty}(w|c) \propto -\log p(w|c), \tag{1}$$

where the surprisal is measured in bits when the logarithm is taken to base 2. This quantity is also interpretable as the **information content** of the word in context. It indicates the extent to which a word is unpredictable in context. Under surprisal theory, the average processing difficulty per word in language is proportional to the **entropy rate** of the language: the average surprisal of each word given an unbounded amount of context information.

There are multiple convergent theoretical motivations for surprisal theory (Levy, 2013), and it is in line with recent theories of information processing difficulty from robotics, artificial intelligence, and neuroscience in that it proposes a certain amount of cost per bit of information processed (Friston, 2010; Tishby and Polani, 2011; Genewein et al., 2015).

Surprisal theory also has excellent empirical coverage of psycholinguistic data: for example, taking word-by-word reading times as a measure of processing difficulty, Smith and Levy (2013) find that empirically observed reading times in naturalistic text are a robustly linear function of surprisal over 8 orders of magnitude. Levy (2008) shows that surprisal theory can explain many diverse phenomena studied in the previous psycholinguistic literature.

The fact that processing time is a linear function of surprisal will be important for deriving predictions about word order: it tightly constrains theories about the interactions of word order and processing difficulty. In fact, surprisal theory in the form of Eq. 1 leads to the prediction that the average processing difficulty per word is not at all a function of the word order rules of a language, provided that different word order rules do not affect the entropy rate of the language. To see this, consider a sentence of n words $w_1, ..., w_n$ in some language L. The total information-processing difficulty for comprehending this sentence ends up being equal to the quantity of information content of the sentence in the language:

$$\begin{aligned}
\text{Difficulty}(w_1, ..., w_n) &= \sum_{i=1}^{n} \text{Difficulty}(w_i|w_1, ..., w_{i-1}) && (2)\\
&\propto \sum_{i=1}^{n} -\log p_L(w_i|w_1, ..., w_{i-1})\\
&= -\log \prod_{i=1}^{n} p_L(w_i|w_1, ..., w_{i-1})\\
&= -\log p_L(w_1, ..., w_n). && (3)
\end{aligned}$$

Now let us consider how this sentence might look in another language L' with other rules for ordering words. As long as the total probability of the sentence in L' is the same as the equivalent sentence in L—regardless of the order of words—the predicted processing difficulty for the sentence is the same. For example, maybe L is English and L' is reversed-English: a language which is identical to English except that all sentences are reversed in order. Then the English sentence $w_1, ..., w_n$ would come out as the reversed-English sentence $w_n, w_{n-1}, ..., w_1$, with the same total probability and thus exactly the same predicted processing difficulty under surprisal theory.

The general expressed by Eq. 3 is that, under surprisal theory, the word order patterns of a language do not affect the overall processing difficulty of the language unless they increase or decrease the average total surprisal of sentences of the language, or in other words the entropy over sentences in a language. The predicted processing difficulty is not affected by word order rules except inasmuch as they decrease the entropy over sentences (by introducing ambiguities) or increase the entropy over sentences (by removing ambiguities) (Levy, 2005, §2.8.3) (Futrell, 2017, §5.1). Essentially, all that surprisal theory has to say about word order is that less frequent orders within a language are more costly.

This invariance to order is problematic for theories that have attempted to explain word order patterns in terms of maximizing the predictability of words (Gildea and Jaeger, 2015; Ferrer-i-Cancho, 2017): such theories have derived predictions about word order by introducing auxiliary assumptions. For example, Gildea and Jaeger (2015) show that word order rules in languages minimize surprisal as calculated from a trigram model, rather than a full probability model; this ends up being a special case of the theory we advocate below in Section 3. Ferrer-i-Cancho (2017) implicitly assumes that the predictability of the verb is more impactful for processing difficulty than the predictability of other words, such that orders that minimize the surprisal of the verb are favorable.

There are at least two general ways to modify surprisal theory to break its order-invariance. The first would be to posit that processing difficulty is some non-linear function of surprisal. This route is not attractive, because the current state of empirical knowledge is that processing time is determined linearly by surprisal (Smith and Levy, 2013). The second way of modifying surprisal theory would be to posit that the relevant probability distribution of words given contexts does not take into account full information from the context, or is distorted in some way relative to the true distribution of words given contexts. As we will see below, this solution allows us to derive information locality.

3 Lossy-context surprisal and information locality

I propose to modify surprisal theory in the manner described in Futrell and Levy (2017). The contents of this section are a simplified exposition of the derivations presented in that paper.

In the modified suprisal theory, the predicted processing difficulty per word w is a function of the word's expected log probability given a *lossy* or *noisy* **memory representation** m of the context c. That is:

$$\text{Difficulty}(w|c) \propto \mathop{\mathbb{E}}_{m|c} \left[-\log p(w|m)\right], \tag{4}$$

where $m|c$ indicates the conditional distribution of lossy memory representations given contexts, called the **memory encoding function**. I call this model **lossy-context surprisal**, because the predicted processing difficulty depends on a lossy memory m, rather than the objective context c. In general, due to the Data Processing Inequality (Cover and Thomas, 2006), m can be seen as a representation of c to which noise has been added. Taking c to be the sequence of word tokens leading up to a given token w_i, we can write Eq. 4 as:

$$\text{Difficulty}(w_i|w_{1:i-1}) \propto \mathop{\mathbb{E}}_{m|w_{1:i-1}} \left[-\log p(w_i|m)\right], \tag{5}$$

where $w_{1:i-1}$ denotes the sequence of words from index 1 to index $i-1$ inclusive.

Unlike plain surprisal theory, lossy-context surprisal predicts that some systems of word order rules will result in more processing efficiency than others. In particular, it predicts locality effects (Gibson, 1998, 2000) in the form of **information locality**: there will be difficulty when elements that have high mutual information are distant from each other in linear order. The basic intuition is that, when two elements that predict each other *in principle* are separated in time, they will not be able to predict each other *in practice* because by the time the processor gets to the second element, the first one has been partially forgotten. The result is that the second element is less predictable than it could have been, causing excess processing cost.

3.1 Derivation of information locality

Assume that the memory encoding function $m|c$ is structured such that some proportion of the information available in a word is lost depending on how long the word has been in memory. For a word which has been in memory for one timestep, the proportion of information which is lost is a constant e_1; for a word which has been in memory for two timesteps, the proportion of information lost is e_2; in general for a word which has been in memory for t timesteps, the proportion of information lost is e_t. Assume further that e_t is monotonically increasing in t: i.e. $t < \tau$ implies $e_t \leq e_\tau$. This process reflects the fact that information in a memory representation can only become degraded over time, in the spirit of the Data Processing Inequality (Cover and Thomas, 2006).

This memory model is equivalent to assuming that the context is subject to **erasure noise**, a commonly used noise model in information theory (Cover and Thomas, 2006). In erasure noise, a symbol x is stochastically *erased*

(replaced with a special erasure symbol E) with some probability e. The noise model here further assumes that the erasure rate increases with time: I call this noise model **progressive erasure noise**.

I will now show that subjective surprisal, under the assumption of progressive erasure noise, gives rise to information locality.

Under progressive erasure noise, the context $w_{1:i-1}$ can be represented as a sequence of symbols $m_{1:i-1}$. Each symbol m_j, called a **memory symbol**, is equal either to the context word w_j or to the erasure symbol E. The surprisal of a word w_i given the memory representation $m_{1:i-1}$ can be written in two terms:

$$-\log p(w_i|m_{1:i-1}) = -\log p(w_i) - \text{pmi}(w_i; m_{1:i-1}),$$

where $\text{pmi}(w_i; m_{1:i-1}) = \log \frac{p(w_i|m_{1:i-1})}{p(w_i)}$ is the **pointwise mutual information** (Fano, 1961; Church and Hanks, 1990) of the word and the memory representation, giving the extent to which the particular memory representation predicts the particular word. We can now use the chain rule to break the pointwise mutual information into separate terms, one for each symbol in the memory representation:

$$\begin{aligned}
\text{pmi}(w_i; m_{1:i-1}) &= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j|m_{1:j-1}) \\
&= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) - \sum_{j=1} \text{pmi}(w_i; m_j; m_{1:j-1}) \\
&= \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) - R, \qquad (6)
\end{aligned}$$

where $\text{pmi}(x; y; z)$ is the three-way pointwise **interaction information** of three variables (Bell, 2003), indicating the extent to which the conditional $\text{pmi}(w_i; m_j|m_{1:j-1})$ differs from the unconditional $\text{pmi}(w_i; m_j)$. These higher-order interaction terms are then grouped together in a term called R.

Now substituting Eq. 6 into Eq. 5, we get an expression for processing difficulty in terms of the pmi of each memory symbol with the current word:

$$\begin{aligned}
\text{Difficulty}(w_i|w_{1:i-1}) &\propto \mathop{\mathbb{E}}_{m|w_{1:i-1}} [-\log p(w_i|m)] \qquad (5) \\
&= \mathop{\mathbb{E}}_{m|w_{1:i-1}} \left[-\log p(w_i) - \sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) + R \right] \\
&= -\log p(w_i) - \mathop{\mathbb{E}}_{m|w_{1:i-1}} \left[\sum_{j=1}^{i-1} \text{pmi}(w_i; m_j) + R \right] \\
&= -\log p(w_i) - \sum_{j=1}^{i-1} \mathop{\mathbb{E}}_{m_j|w_j} [\text{pmi}(w_i; m_j)] + \mathop{\mathbb{E}}_{m|w_{1:i-1}} [R]\,. \qquad (7)
\end{aligned}$$

It remains to calculate the expected pmi of the current word and a memory symbol given the distribution of possible memory symbols. Recall that each m_j is either equal to the erasure symbol E (with probability e_{i-j}) or to the word w_j (with probability $1 - e_{i-j}$). If $m_j = \text{E}$, then $\text{pmi}(w_i; m_j) = 0$; otherwise $\text{pmi}(w_i; m_j) = \text{pmi}(w_i; w_j)$. Therefore the expected pmi between a word w_i and a memory symbol m_j is $(1 - e_{i-j})\text{pmi}(w_i; w_j)$. The effect of erasure noise on the higher-order terms collected in R is more complicated, but in general will have the effect of reducing their value, because a higher-order interaction information term will have a value of 0 if any single variable in it is erased. Therefore we can write the expected processing difficulty per word as:

$$\text{Difficulty}(w_i|w_{1:i-1}) \propto -\log p(w_i) - \sum_{j=1}^{i-1} (1 - e_{i-j})\text{pmi}(w_i; w_j) + o(R), \qquad (8)$$

where $o(R)$ indicates a value that is is upper-bounded by R. Assuming the higher-order terms $o(R)$ are negligible, then the expected processing difficulty as a function of word order is purely determined by the expression

$$-\sum_{j=1}^{i-1} (1 - e_{i-j})\text{pmi}(w_i; w_j). \qquad (9)$$

As words w_i and w_j become more distant from each other, the value of the survival probability $(1 - e_{i-j})$ must decrease, so the value of (9) must increase, such that the theory predicts increased processing difficulty in proportion to the pairwise pmi between w_i and w_j.

In general, Eq. 8 holds that processing difficulty as a function of word order increases monotonically as elements with high pointwise mutual information are separated in linear order.[1] It will be minimized when elements with the highest pointwise mutual information are closest to each other. If word orders are shaped by a pressure for processing efficiency, then information locality comes out to a kind of attraction between words with high pmi.[2]

3.2 DLM as an approximation to information locality

The principle of information locality holds that groups of words with high mutual information will tend to be close to each other, in order to maximize online processing efficiency. I wish to argue that this result subsumes the principle of dependency length minimization (DLM), which holds that all words in syntactic dependencies will tend to be close to each other. This connection requires a linking hypothesis: that syntactic dependencies correspond to the word pairs with high mutual information within a sentence. I call this hypothesis the **Head–Dependent Mutual Information (HDMI) hypothesis**.

There are good theoretical and empirical reasons to believe the HDMI hypothesis is true. The empirically-measured mutual information of words pairs in head–dependent relationships has been found to be greater than various baselines in Futrell and Levy (2017) across languages. Theoretically, it makes sense for word pairs in dependency relationships to have the highest mutual information because mutual information is a measure of the strength of covariance between two variables, and words in dependencies are by definition those word pairs whose covariance is directly constrained by grammatical rules. More formally, in distributions over dependency trees generated by head-outward generative models such as those presented in Eisner (1996, "Model C"), heads and dependents will have the highest mutual information of any word pairs. The basic idea that dependencies correspond to high-mutual-information word pairs has a long history in computational linguistics (Resnik, 1996; de Paiva Alves, 1996; Yuret, 1998).

If we assume the strongest form of the HDMI hypothesis—that mutual information between words *not* linked in a dependency is completely negligible—then Eq. 8 implies that the expected processing cost for a sentence as a function of word order is a monotonically increasing function of dependency length, which is exactly the claim underlying DLM. This strong form of the HDMI hypothesis is surely false, but it shows how DLM can serve as an approximation to the predictions of information locality.

The notion of information locality is also linked to more general notions of complexity, such as the theory of statistical complexity (Crutchfield and Young, 1989), which apply to any stochastic process. The **statistical complexity** of a process is the entropy of the maximally compressed representation of the past of a process required to predict the future of the process with optimal accuracy. Among processes with the same entropy rate, processes with poor information locality properties (where elements with high mutual information are far from each other) will have higher statistical complexity, because each bit of predictive information will need to be retained in memory over more timesteps. If we view DLM as a special case of information locality, then that means that minimizing dependency length has the effect of lowering statistical complexity. Thus it may be the case that the word order properties of human language are a very general consequence of minimization of statistical complexity.

4 Information locality beyond DLM

Here I give new empirical evidence that natural langauges exhibit information locality in a way that goes beyond the predictions of DLM.

4.1 Strength of locality effect for different dependencies

DLM predicts that all words in dependencies will be under a pressure to be close to each other, but it does not make any predictions about *which* dependencies will be under especially strong pressure. However, empirically, DLM

[1] If we include the effects of the higher-order terms collected in R, then Eq. 7 also implies that processing difficulty will increase when groups of elements with high interaction information are separated from each other in time. Here "high interaction information" refers to a large positive value in the case of even-cardinality groups of elements, and a large negative value in the case of odd-cardinality groups of elements. See Bell (2003) for the relevant technical details on interaction information.

[2] If words are under a pressure to be close as a function of their pmi, then this raises the question of what is to be expected for nonce and novel words, for which no corpus co-ocurrence statistics are available. This issue was raised as an objection to information locality by Dyer (2017). While the probabilities that go into practically calculating pmi come from corpora, the probabilities that are truly important from the perspective of processing difficulty are the listener's subjective probabilities, which are only approximated by corpus-derived probabilities (Smith and Levy, 2011). A listener encountering a nonce word will have some hypothesis about its syntax and meaning, which means that the listener will have expectations about what words the nonce word will co-occur with, and thus the nonce word will have a nonzero (subjective) pmi value with other words for the listener. In an experimental study, the pmi values for nonce words could be measured using techniques such as the Cloze task (Taylor, 1953), which measures these subjective probabilities.

Language	β_{pmi}	p value	Language	β_{pmi}	p value
Ancient Greek	-0.18	< .001	Japanese	-0.32	<.001
Arabic	-0.26	<.001	*Kazakh*	*-1.18*	*0.01*
Basque	-0.22	<.001	Korean	-0.14	<.001
Belarusian	-0.20	<.001	Latin	-0.18	<.001
Bulgarian	-0.29	<.001	Latvian	-0.32	<.001
Catalan	-0.29	<.001	Lithuanian	-0.41	<.001
Church Slavonic	-0.23	<.001	Mandarin	-0.19	<.001
Coptic	-0.35	<.001	Modern Greek	-0.25	<.001
Croatian	-0.32	<.001	Norwegian	-0.37	<.001
Czech	-0.27	<.001	Persian	-0.19	<.001
Danish	-0.38	<.001	Polish	-0.23	<.001
Dutch	-0.10	<.001	Portuguese	-0.23	<.001
English	-0.38	<.001	Romanian	-0.36	<.001
Estonian	-0.32	<.001	Russian	-0.18	<.001
Finnish	-0.29	<.001	*Sanskrit*	*0.10*	*0.28*
French	-0.33	<.001	Slovak	-0.30	<.001
Galician	-0.35	<.001	Slovenian	-0.38	<.001
German	-0.25	<.001	Spanish	-0.37	<.001
Gothic	-0.19	<.001	Swedish	-0.35	<.001
Hebrew	-0.21	<.001	Tamil	-0.18	<.001
Hindi	-0.26	<.001	Turkish	-0.22	<.001
Hungarian	-0.11	<.001	Ukrainian	-0.29	<.001
Indonesian	-0.22	<.001	Urdu	-0.22	<.001
Irish	-0.37	<.001	*Uyghur*	*-0.04*	*0.79*
Italian	-0.35	<.001	Vietnamese	-0.27	<.001

Table 1: Regression coefficients predicting dependency length as a function of pmi between head and dependent. A negative sign indicates that words with higher pmi are closer to each other. Languages where the effect is not significant at $p < .001$ are in italics.

effects in word order preferences and also in online processing difficulty show asymmetries based on the details about particular dependencies (Stallings et al., 1998; Demberg and Keller, 2008).

Here I propose that the strength of attraction between two words linked in a dependency is modulated by the pointwise mutual information of the two words, as predicted by information locality.

I tested this hypothesis in 50 languages of the Universal Dependencies 2.1 treebanks (Nivre et al., 2017). I excluded all punctuation and root dependencies, and collapsed all strings of words linked by "flat", "fixed", "compound"q lo dependencies (which indicate multiword expressions) into single tokens. For each word pair $r = (h, d)$ in a head–dependent relationship, I fit a linear regression model to predict the distance between the two words y_r from their pmi:

$$y_r = \beta_0 + \beta_{\text{pmi}}\text{pmi}(h; d) + S_i + S_{i,\text{pmi}}\text{pmi}(h; d) + \epsilon_r, \tag{10}$$

where S_i and $S_{i,\text{pmi}}$ are by-sentence random intercepts and slopes subject to L_2 regularization, making this a mixed-effects regression model (Gelman and Hill, 2007; Baayen et al., 2008; Barr et al., 2013). These extra terms account for any per-sentence idiosyncratic behavior of dependencies (for example, effects of sentence length). The key coefficient is β_{pmi} which, if significantly negative, indicates that words with high pmi are especially attracted to each other. The pmi values were calculated between part-of-speech tags rather than wordforms in order to avoid data sparsity issues in the estimation of mutual information. For computational tractability, I include only at most 10,000 sentences per language and exclude sentences of length greater than 20 words.

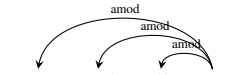
Table 1 shows the values of β_{pmi} and their significance for 50 languages. In all except 3 languages, I find the significant negative effect at $p < .001$, indicating information locality effects beyond DLM. The effect size is relatively stable across languages. In particular, the average value of β_{pmi} across languages is around -0.3, with standard error 0.02, indicating that for every bit of pmi between parts-of-speech, words in dependencies are about 0.3 words closer together on average, robustly across languages.

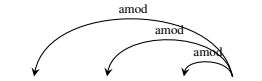
The code for this analysis is available online as part of `http://github.com/langprocgroup/cliqs`.

4.2 Adjective order

Speakers of many languages show robust, stable patterns of preferences in terms of how they order attributive adjectives that simultaneously modify a noun (Dixon, 1982; Scontras et al., 2017, 2019). For example, English

speakers generally prefer the order in Example 1 over 2, or they perceive Example 2 as expressing a different meaning which is marked relative to the first. As the dependency structures show, the classical theory of DLM would not make any predictions for the relative ordering of these adjectives, as all are in equivalent dependency relationships with the head noun. Classical syntactic theories of adjective order have assumed that adjectives can be sorted into semantic classes (e.g., value, color, nationality) and that there is a universal order of semantic classes in terms of which adjectives go closer to the noun (e.g., adjectives are placed close to the noun in the order nationality $>$ color $>$ value) (Cinque and Rizzi, 2008).

(1)


(2)


Here I suggest that the preferred order of adjectives is determined by information locality: that is, adjectives with higher mutual information with a noun go closer to that noun.

Previous work has shown that the best empirical predictor of adjective order is the rating of **subjectivity** given to adjectives by experimental participants, with more subjective adjectives going farther out from the head noun (Scontras et al., 2017, 2019), but this work did not compare predictions with mutual information. Simultaneously, Kirby et al. (2018) compared size adjectives and color adjectives—where color adjectives are preferred to be farther out from the noun in English—and found that color adjectives have lower pmi with the noun than size adjectives.

Here I compare subjectivity and mutual information as predictors of adjective order in large corpora of English. For subjectivity ratings, I use the data from Scontras et al. (2017). For co-occurrence and order data, I use the Google Syntactic n-grams corpus (Goldberg and Orwant, 2013). From this corpus, I collect all cases of two adjectives modifying a single following noun with relation type *amod*, where an adjective counts for inclusion if its part of speech is `JJ`, `JJR`, or `JJRS` and it is listed as an adjective in the CELEX database (Baayen et al., 1995), and a noun counts for inclusion if its part of speech is `NN` or `NNS` and it is listed as a noun in CELEX. The result is a dataset of adjective–adjective–noun (AAN) triples, containing 1,604,285 tokens and 16,681 types.

For the calculation of pointwise mutual information, we need the conditional probability of adjectives given nouns. I find this probability by maximum likelihood estimation, collecting all instances of single adjectives modifying a following noun, by the same criteria as above.

I tested the relationship between pmi, subjectivity, and order statistically using logistic regression. Given two adjectives preceding a noun, the question is which of the two is closer to the noun, as a function of the *difference* in pmi or subjectivity between the two. Given an observed pair of adjectives (A_1, A_2), ordered alphabetically, I fit a logistic regression to predict whether the alphabetically-second adjective A_2 is ordered closer to the noun N in the corpus, as a function of the pmi and subjectivity difference between A_1 and A_2. The regression equation is:

$$\log \frac{p(A_2 \text{ closer to } N)}{p(A_1 \text{ closer to } N)} = \beta_0 + \beta_S(S(A_1) - S(A_2)) + \beta_{\text{pmi}}(\text{pmi}(A_1; N) - \text{pmi}(A_2; N)) + \epsilon,$$

where $S(A)$ is the subjectivity rating of adjective A. This regression setup was used to predict order data in Morgan and Levy (2016). A positive coefficient β_S indicates that the adjective with greater subjectivity is likely to be farther from the noun. A negative coefficient β_{pmi} indicates that an adjective with greater pmi with the noun is likely to be closer to the noun.

To evaluate the accuracy of subjectivity and information locality as theories of adjective order, I separated out 10% of the AAN types as a test set (1,668 types), with the remaining types forming a training set (15,013 types). I fit the logistic regression to the token-level data for the AAN types in the training set, a total of 1,473,269 tokens. I find $\beta_0 = -0.7$, $\beta_S = 14.1$, and $\beta_{\text{pmi}} = -0.6$, with all coefficients significant at $p < .001$. The results mean that for each bit of pmi between an adjective A and a noun, beyond the pmi of the other adjective with the noun and controlling for subjectivity, the log-odds that A is closer to the noun increase by .6.[3]

I used the held-out AAN triples to test how well subjectivity and pmi would generalize when predicting adjective order in unseen data (a total of 131,016 tokens). Table 2 shows test-set accuracy of logistic regressions predicting adjective order using subjectivity, pmi, or both as predictors. Subjectivity and pmi have roughly equal accuracy on

[3]The values of subjectivity tend to be smaller than the values of pmi (average subjectivity of adjectives in the corpus is 0.5; average pmi is 2.3), so the larger coefficient β_S should not necessarily be interpeted as meaning that the effect of subjectivity is larger.

Subjectivity	PMI	Both
68.4%	66.9%	72.9%

Table 2: Accuracy of subjectivity and pmi as predictors of adjective order in logistic regressions, for held-out types of adjective–adjective–noun triples.

the held-out types, with pmi slightly lower. The highest accuracy is achieved when both subjectivity and pmi are included as predictors. This result shows that mutual information has predictive value for adjective order beyond what is accounted for by subjectivity.

The regression shows that both subjectivity and mutual information are good predictors of adjective order, so the question arises of whether the two predictors make overlapping or divergent predictions. For 53% of the test-set tokens, subjectivity and pmi make the same (correct) prediction. I qualitatively examined cases where pmi got the order right and subjectivity did not, and found that these usually consist of cases where two adjectives are in the same semantic class, and yet strong ordering preferences exist in the corpus, such as *big long beard* (preferred) vs. *long big beard* (dispreferred).

The results here do not adjudicate between subjectivity and mutual information as better predictors of adjective order. The two may be independent factors predicting adjective order (a hypothesis explored by Hahn et al., 2018, with a theoretical justification related to information locality), or they may be related. I posit that subjectivity and mutual information are conceptually related. The reasoning is: if an adjective is more subjective, then its applicability to any given noun is determined by some external factor outside than the noun itself—the speaker's subjective state. In contrast, the applicability of a less subjective adjective is more strongly determined by the noun itself due to the inherent properties of the noun. Mutual information is calculated from co-ocurrence statistics, where the speaker's subjective state is unknown and therefore appears as a noise variable affecting the distribution of adjectives. So from the perspective of co-ocurrence statistics, the distribution of more subjective adjectives is noisier, and therefore has less mutual information with the head noun. The relationship between subjectivity, mutual information, and adjective order may be the following: subjectivity determines the joint distribution of adjectives and nouns, which in turn dictates the mutual information, which then determines the preferred order via the principle of information locality.

In support of this idea, I found that the subjectivity score for an adjective is moderately anticorrelated with its average pmi with nouns at $r = -.32$, Spearman's $\rho = -.35$; the relationship between the two is shown in Figure 1. Note that the estimates of mutual information obtained from corpora are noisy: it is notoriously difficult to estimate quantities such as mutual information from count data (Paninski, 2003). Better estimates of mutual information, obtained through more data or more sophisticated estimation techniques, may show stronger correlations with subjectivity and with adjective order.

The code for this analysis is available online at `http://github.com/langprocgroup/adjorder`.

5 Conclusion

I presented a theoretical argument that, if languages are organized for efficient communication subject to human information processing constraints, then they will have the property of information locality: words that predict each other will appear close to each other in time. I presented two pieces of novel evidence in favor of information locality over previous theories of word order. I believe the principle of information locality will enrich the growing link between theories of syntax and notions of processing efficiency. By deriving the principle of dependency length minimization in an information-theoretic setting and demonstrating improved predictive power over simple dependency length minimization, it opens the way for unified information-theoretic models of human language.

Acknowledgments

I thank Roger Levy, Greg Scontras, and Ted Gibson for many conversations on these topics and the anonymous reviewers for helpful comments on the paper.

References

Baayen, R. H., Davidson, D., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

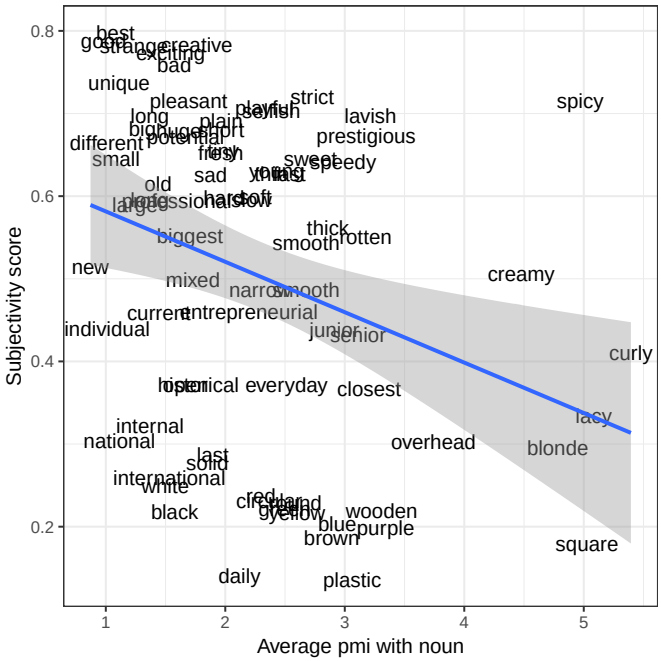


Figure 1: Relationship between adjective subjectivity score and average pmi with nouns in Google Syntactic n-Grams corpus.

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178–1198.

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

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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

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

Crutchfield, J. P. and Young, K. (1989). Inferring statistical complexity. *Physical Review Letters*, 63(2):105.

de Paiva Alves, E. (1996). The selection of the most probable dependency structure in Japanese using mutual information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 372–374.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Dixon, R. M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Mouton, Berlin, Germany.

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

Dyer, W. E. (2017). *Minimizing integration cost: A general theory of constituent order*. PhD thesis, University of California, Davis, Davis, CA.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communication*. MIT Press, Cambridge, MA.

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters*, 76(6):1228.

Ferrer-i-Cancho, R. (2017). The placement of the head that maximizes predictability: An information theoretic approach. *Glottometrics*, 39:38–71.

Ferrer-i-Cancho, R. and Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06009.

Ferrer-i-Cancho, R. and Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127.

Futrell, R. (2017). *Memory and locality in natural language*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.

Genewein, T., Leibfried, F., Grau-Moya, J., and Braun, D. A. (2015). Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2:27.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O'Neil, W., editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126.

Gildea, D. and Jaeger, T. F. (2015). Human languages order information efficiently. *arXiv*, 1510.02823.

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

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics*, volume 1, pages 241–247.

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.

Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261–290.

Hahn, M., Degen, J., Goodman, N., Jurafsky, D., and Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci)*.

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, pages 1–8.

Hale, J. T. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press, Cambridge.

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

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press, Oxford.

Kirby, S., Culbertson, J., and Schouwstra, M. (2018). The origins of word order universals: Evidence from corpus statistics and silent gesture. In *The Evolution of Language: Proceedings of the 12th International Conference (Evolangxii)*. NCU Press.

Klein, D. and Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 478.

Kuhlmann, M. (2013). Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.

Levy, R. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity*. PhD thesis, Stanford University, Stanford, CA.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In van Gompel, R. P. G., editor, *Sentence Processing*, page 78âĂŞ114. Hove: Psychology Press.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

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

Morgan, E. and Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, 157:382–402.

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cinková, S., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Laippala, V., Lambertino, L., Lando, T., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Östling, R., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Rama, T., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Romanenko, M., Rosa, R., Rovati, D., Sagot, B., Saleh, S., Samardžić, T., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Washington, J. N., Wirén, M., Wong, T.-s., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Scontras, G., Degen, J., and Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1(1):53–65.

Scontras, G., Degen, J., and Goodman, N. D. (2019). On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.

Smith, N. and Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

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

Stallings, L. M., MacDonald, M. C., and O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39(3):392–417.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

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

Tishby, N. and Polani, D. (2011). Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer.

Wasow, T. (2002). *Postverbal Behavior*. CSLI Publications, Stanford, CA.

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.