

# Psychological Review

## Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-Off of Memory and Surprisal

Michael Hahn, Judith Degen, and Richard Futrell

Online First Publication, April 1, 2021. <http://dx.doi.org/10.1037/rev0000269>

### CITATION

Hahn, M., Degen, J., & Futrell, R. (2021, April 1). Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-Off of Memory and Surprisal. *Psychological Review*. Advance online publication. <http://dx.doi.org/10.1037/rev0000269>

# Modeling Word and Morpheme Order in Natural Language as an Efficient Trade-Off of Memory and Surprisal

Michael Hahn<sup>1</sup>, Judith Degen<sup>1</sup>, and Richard Futrell<sup>2</sup>

<sup>1</sup> Department of Linguistics, Stanford University

<sup>2</sup> Department of Language Science, University of California

Memory limitations are known to constrain language comprehension and production, and have been argued to account for crosslinguistic word order regularities. However, a systematic assessment of the role of memory limitations in language structure has proven elusive, in part because it is hard to extract precise large-scale quantitative generalizations about language from existing mechanistic models of memory use in sentence processing. We provide an architecture-independent information-theoretic formalization of memory limitations which enables a simple calculation of the memory efficiency of languages. Our notion of memory efficiency is based on the idea of a *memory-surprisal trade-off*: A certain level of average surprisal per word can only be achieved at the cost of storing some amount of information about the past context. Based on this notion of memory usage, we advance the *Efficient Trade-off Hypothesis*: The order of elements in natural language is under pressure to enable favorable memory-surprisal trade-offs. We derive that languages enable more efficient trade-offs when they exhibit *information locality*: When predictive information about an element is concentrated in its recent past. We provide empirical evidence from three test domains in support of the Efficient Trade-off Hypothesis: A reanalysis of a miniature artificial language learning experiment, a large-scale study of word order in corpora of 54 languages, and an analysis of morpheme order in two agglutinative languages. These results suggest that principles of order in natural language can be explained via highly generic cognitively motivated principles and lend support to efficiency-based models of the structure of human language.

**Keywords:** language universals, sentence processing, information theory

**Supplemental materials:** <https://doi.org/10.1037/rev0000269.supp>

Natural language is a powerful tool that allows humans to communicate, albeit under inherent cognitive resource limitations. Here, we investigate whether human languages are grammatically structured in a way that reduces the cognitive resource requirements for comprehension, compared to counterfactual languages that differ in grammatical structure.


The suggestion that the structure of human language reflects a need for efficient processing under resource limitations has been present in the linguistics and cognitive science literature for decades (Berwick & Weinberg, 1984; Chomsky, 2005; Gibson et al., 2019; Hahn et al., 2020; Hawkins, 1994; Jaeger & Tily, 2011; Yngve, 1960). The idea has been summed up in Hawkins's (2004) Performance-Grammar Correspondence Hypothesis (PGCH), which holds that grammars are structured so that the typical utterance is easy to produce and comprehend under performance constraints.

One major source of resource limitation in language processing is incremental memory use. When producing and comprehending language in real time, a language user must keep track of what they have already produced or heard in some kind of incremental memory store, which is subject to resource constraints. These memory constraints have been argued to underlie various *locality principles* which linguists have used to predict the orders of words within sentences and morphemes within words (e.g., Behaghel, 1932; Bybee, 1985; Givón, 1985; Hawkins, 1994, 2004, 2014; Rijkhoff, 1990; Temperley & Gildea, 2018). The idea is that language should be structured to reduce long-term dependencies of various kinds, by placing elements that depend on each other close to each other in linear order. That is, elements of utterances which are more "relevant" or "mentally connected" to each other are closer to each other.

Our contribution is to present a new, highly general formalization of the relationship between sequential order and incremental memory in language processing, from which we can derive a precise and empirically testable version of the idea that utterance elements which depend on each other should be close to each other. Our formalization allows us to predict the order of words within sentences, and morphemes within words directly by the minimization of memory usage.

We formalize the notion of memory constraints in terms of what we call the *memory-surprisal trade-off*: The idea that the ease of comprehension depends on the amount of computational resources invested into remembering previous linguistic elements, for

Michael Hahn  <https://orcid.org/0000-0003-4828-4834>

Judith Degen  <https://orcid.org/0000-0003-2513-0234>

An earlier version of this work was presented at the 32nd Annual CUNY Conference on Human Sentence Processing, 2019. All code and data are freely available at <https://github.com/m-hahn/memory-surprisal>

Correspondence concerning this article should be addressed to Michael Hahn, Department of Linguistics, Stanford University, Stanford, CA 94305-2150, United States. Email: [mhahn2@stanford.edu](mailto:mhahn2@stanford.edu)

example, words. Therefore, there exists a trade-off between the quantity of memory resources invested, and the ease of language processing. The shape of this trade-off depends on the grammar of a language, and in particular, the way that it structures information in time. We characterize memory resources using the theory of lossy data compression (Berger, 2003; Cover & Thomas, 2006).

Within our framework, we prove a theorem showing that lower memory requirements result when utterance elements that depend on each other statistically are placed close to each other. This theorem does not require any assumptions about the architecture or functioning of memory, except that it has a bounded capacity. Using this concept, we introduce the *Efficient Trade-off Hypothesis*: Order in natural language is structured so as to provide efficient memory–surprisal trade-off curves. We provide evidence for this hypothesis in three studies. We demonstrate that word orders with short dependencies do indeed engender lower working memory resource requirements in toy languages studied in the previous literature, and we show that real word orders in corpora of 54 languages have lower memory requirements than would be expected under artificial baseline comparison grammars. Finally, we show that we can predict the order of morphemes within words in two languages using our principle of the minimization of memory usage.

Our work not only formalizes and tests an old idea in functional linguistics and psycholinguistics but also it opens up connections between those fields and the statistical analysis of natural language (Bentz et al., 2017; Debowski, 2011; Lin & Tegmark, 2017), and more broadly, between linguistics and fields that have studied information-processing costs and resource requirements in brains (e.g., Friston, 2010) and general physical systems (e.g., Still et al., 2012).

## Background

A wide range of work has argued that information in natural language utterances is ordered in ways that reduce memory effort, by placing elements close together when they depend on each other in some way. Here, we review these arguments from linguistic and cognitive perspectives.

### Dependency Locality and Memory Constraints in Psycholinguistics

When producing and comprehending language in real time, a language user must keep track of what she has already produced or heard in some kind of incremental memory store, which is subject to resource constraints. An early example of this idea is the study by Miller and Chomsky (1963) who attributed the unacceptability of multiple center embeddings in English to limitations of human working memory. Concurrent and subsequent work studied how different grammars induce different memory requirements in terms of the number of symbols that must be stored at each point to produce or parse a sentence (Abney & Johnson, 1991; Gibson, 1991; Resnik, 1992; Yngve, 1960). In psycholinguistic studies, memory constraints typically manifest in the form of processing difficulty associated with long-term dependencies. For example, at the level of word-by-word online language comprehension, there is observable processing difficulty at moments when it seems that information about a word must be retrieved from working memory. This difficulty increases when there is a great deal of time or intervening

material between the point when a word is first encountered and the point when it must be retrieved from memory (Balling & Kizach, 2017; Bartek et al., 2011; Gibson, 1998, 2000; Gibson & Thomas, 1999; Lewis & Vasisht, 2005; McElree, 2000; Nicenboim et al., 2015). That is, language comprehension is harder for humans when words which depend on each other for their meaning are separated by many intervening words. This idea is most prominently associated with the *Dependency Locality Theory* (DLT) of human sentence processing (Gibson, 2000).

For example, Grodner and Gibson (2005) studied word-by-word reading times in a series of sentences such as (1) below.

- (1) a. The *administrator* who the nurse *supervised* ...
- b. The *administrator* who the nurse from the clinic *supervised* ...
- c. The *administrator* who the nurse who was from the clinic *supervised* ...

In these sentences, the distance between the noun *administrator* and the verb *supervised* is successively increased. Grodner and Gibson (2005) found that as this distance increases, there is a concomitant increase in reading time at the verb *supervised* and following words.

The hypothesized reason for this reading time pattern is based on memory constraints. The idea goes as follows: At the word *supervised*, a comprehender who is trying to compute the meaning of the sentence must integrate a representation of the verb *supervised* with a representation of the noun *administrator*, which is a direct object of the verb. This integration requires retrieving the representation of *administrator* from working memory. If this representation has been in working memory for a long time—for example, as in Sentence 1c as opposed to 1a—then the retrieval is difficult or inaccurate, in a way that manifests as increased reading time. Essentially, there exists a dependency between the words *administrator* and *supervised*, and more excess processing difficulty is incurred the more the two words are separated; this excess difficulty is called a *dependency locality effect*.

The existences of dependency locality effects in human language processing, and their connection with working memory, are well-established (Fedorenko et al., 2013). These locality effects in online processing mirror locality effects in word order, described below.

### Locality and Crosslinguistic Universals of Order

Dependency locality in word order means that there is a pressure for words which depend on each other syntactically to be close to each other in linear order. There is ample evidence from corpus statistics indicating that dependency locality is a real property of word order across many languages (Ferrer-i-Cancho, 2004, p. 5; Futrell et al., 2015a; Gildea & Temperley, 2007, 2010; Liu, 2008; Liu et al., 2017; Temperley & Gildea, 2018). Hawkins (1994, 2003) formulates dependency locality as the Principle of Domain Minimization, and has shown that this principle can explain crosslinguistic universals of word order that have been documented by linguistic typologists for decades (Greenberg, 1963). Such a pressure can be motivated in terms of the documented online processing difficulty associated with long-term dependencies among words: Dependency locality in word order means that online processing is easier.

An example is order alternation in postverbal constituents in English. Although noun phrase (NP) objects ordinarily precede prepositional phrases (PPs) (2a, example from the study by Staub et al., 2006), this order is less preferred when the NP is very long (2c) in which case the inverse order becomes more felicitous (2d). The pattern in (2d) is known as Heavy NP Shift (Arnold et al., 2000; Ross, 1967; Stallings & MacDonald, 2011). Compared to (2c), it reduces the distance between the verb “ate” and the PP, while only modestly increasing the distance between the verb and object NP.

- (2) a. Lucy ate [the broccoli] with a fork.
- b. ? Lucy ate with a fork [the broccoli].
- c. Lucy ate [the extremely delicious, bright green broccoli] with a fork.
- d. Lucy ate with a fork [the extremely delicious, bright green broccoli].

Locality principles have also appeared in a more general form in the functional linguistics literature, in the form of the idea that elements which are more “relevant” to each other will appear closer to each other in linear order in utterances (Behaghel, 1932; Bybee, 1985; Givón, 1985, 1991; Newmeyer, 1992). Here, “elements” can refer to words or morphemes, and the definition of “relevance” varies. For example, Givón (1985)’s *Proximity Principle* states that elements are placed closer together in a sentence if they are closer conceptually. Applying a similar principle, Bybee (1985) studied the order or morphemes within words across languages, and argued that (for example) morphemes that indicate the valence of a verb (whether it takes zero, one, or two objects) are placed closer to the verb root than morphemes that indicate the plurality of the subject of the verb because the valence morphemes are more “relevant” to the verb root. Although these theories are widespread in the linguistics literature, there exists to date no quantifiable definition of “relevance” or “being closer conceptually.” One of our contributions is to derive such a notion of “relevance” from the minimization of memory usage during language processing.

## Architectural Assumptions

The connection between memory resources and locality principles relies on the idea that limitations in working memory will give rise to difficulty when elements that depend on each other are separated at a large distance in time. In previous work, this idea has been motivated in terms of specific assumptions about the architecture of memory. For example, models of memory in sentence processing differ in whether they assume limitations in storage capacity (e.g., “memory cost” in the model of Gibson, 1998) or the precision with which specific elements can be retrieved from memory (e.g., Lewis & Vasishth, 2005). Furthermore, to derive the connection between memory usage in such models and locality in word order, it has been necessary to stipulate that memory representations or activations decay over time in some way to explain why longer dependencies are harder to process. The question remains of whether these assumptions about memory architecture are necessary, or whether word orders across languages are optimized for memory independently of the implementation and architecture of human language processing.

In this work, we adopt an information-theoretic perspective on memory use in language processing, which abstracts away from the details of memory architecture. Within our framework, we will

establish the connection between memory resources and locality principles by providing general information-theoretic lower bounds on memory use. We quantify memory resources in terms of their information-theoretic capacity measured in bits, following models proposed for working memory in other domains (Brady et al., 2008, 2009; Sims et al., 2012). Our result immediately entails a link between locality and boundedness of memory, following only from the stipulation that memory is finite in capacity. In particular, our model does not require any assumption that memory representations or activations decay over time (as was required in previous work: Futrell, Gibson, & Levy, 2020; Gibson, 1998; Lewis & Vasishth, 2005). We will then show empirical evidence that the orders of words and morphemes in natural language are structured in a way that reduces our measure of memory use compared to the orders of counterfactual baseline languages.

The remainder of the article is structured as follows. We first introduce the memory–surprisal trade-off and introduce our Efficient Trade-off Hypothesis. Then, we test the Efficient Trade-off Hypothesis in three studies. In Study 1, we qualitatively test the Hypothesis in a reanalysis of word orders emerging in a miniature artificial language study (Fedzechkina et al., 2017). In Study 2, we quantitatively test the Hypothesis in a large-scale study of the word order of 54 languages. In Study 3, we test the Hypothesis on morpheme order in Japanese and Sesotho. Finally, we discuss the implications and limitations of the reported results.

## Memory–Surprisal Trade-Off

In this section, we introduce the main concept and hypothesis of the article. We provide a technical definition of the memory–surprisal trade-off curve, and we prove a theorem showing that more efficient memory–surprisal trade-offs are possible in languages exhibiting information locality, that is, in languages where words that depend on each other are close to each other. This theorem establishes the formal link between memory efficiency in online processing and locality in word order.

## An Information-Theoretic Model of Online Language Comprehension

We begin developing our model by considering the process of language comprehension, where a listener is processing a stream of words uttered by an interlocutor. Experimental research has established three properties of online language comprehension: (a) listeners maintain some information about the words received so far in incremental memory, (b) listeners form probabilistic expectations about the upcoming words (e.g., Altmann & Kamide, 1999; Kuperberg & Jaeger, 2016; Staub & Clifton, 2006), and (c) words are easy to process to the extent that they are predictable based on a listener’s memory of words received so far (Futrell, Gibson, & Levy, 2020; Hale, 2001; Levy, 2008). See the General Discussion for discussion of how our model is related to theories that do not explicitly make these assumptions.

We formalize these three observations into postulates intended to provide a simplified picture of what is known about online language comprehension. Consider a listener comprehending a sequence of words  $w_1, \dots, w_t, \dots, w_n$ , at an arbitrary time  $t$ .

1. Comprehension Postulate 1 (Incremental memory). At time  $t$ , the listener has an incremental *memory state*  $m_t$  that contains her stored information about previous words. The memory state is characterized by a memory encoding function  $M$  such that  $m_t = M(w_{t-1}, m_{t-1})$ .
2. Comprehension Postulate 2 (Incremental prediction). The listener has a subjective probability distribution at time  $t$  over the next word  $w_t$  as a function of the memory state  $m_t$ . This probability distribution is denoted  $P(w_t|m_t)$ .
3. Comprehension Postulate 3 (Linking hypothesis). Processing a word  $w_t$  incurs difficulty proportional to the **surprisal** of  $w_t$  given the memory state  $m_t$ .<sup>1</sup>

$$\text{Difficulty} \propto -\log_2 P(w_t|m_t). \quad (1)$$

The claim that processing difficulty should be directly proportional to surprisal comes from *surprisal theory* (Hale, 2001; Levy, 2008), an established psycholinguistic theory that can capture reading time effects related to garden-path disambiguation, antilocality effects, and effects of syntactic construction frequency. Surprisal is a robust linear predictor of reading times in large-scale eye-tracking studies based on naturalistic text (Aurnhammer & Frank, 2019; Frank & Hoeks, 2019; Goodkind & Bicknell, 2018; Smith & Levy, 2013; Wilcox et al., 2020), and effects of surprisal have been observed for units as small as phonemes (Gwilliams et al., 2020). There are several converging theoretical arguments for surprisal as a measure of processing cost (Levy, 2008; Smith & Levy, 2013). Surprisal theory is compatible with different views on the mechanisms underlying prediction, and can reflect different mechanisms such as preactivation and integration (Kuperberg & Jaeger, 2016). We do not assume that listeners explicitly compute a full-fledged distribution  $P(w_t|m_t)$ ; we view  $P(w_t|m_t)$  as a formalization of the probabilistic expectations that listeners form during comprehension.

Our expression (1) differs from the usual formulation of surprisal theory in that we consider predictability based on a (potentially lossy or noisy) memory representation  $m_t$ , rather than predictability based on the true complete context  $w_t, \dots, m_{t-1}$ . The generalization to lossy memory representations is necessary to capture empirically observed effects of memory limitations on language processing, such as dependency locality and structural forgetting (Futrell, Gibson, & Levy, 2020).

In this work, we are interested in using theories of processing difficulty to derive predictions about languages as a whole, not about individual words or sentences. Therefore, we need a measure of the processing difficulty associated with a language as a whole. For this, we consider the *average surprisal* per word in the language. We call this quantity the *average surprisal* of a language given a memory encoding function  $M$ , denoted  $S_M$ .

Crucially, the listener’s ability to predict upcoming words accurately depends on how much she remembers about previous words. As the precision of her memory increases, the accuracy of her predictions also increases, and the average surprisal  $S_M$  for each incoming word decreases. Taking an information-theoretic perspective, we can think about the amount of information (measured in bits) about previous words stored in the listener’s memory state. This quantity of information is given by the *entropy* of the memory

state, which we denote  $H_M$ . As the listener stores more and more bits of information about the previous words in her memory state, she can achieve lower and lower surprisal values for the upcoming words. This trade-off between memory and surprisal is the main object of study in this article.

The *memory–surprisal trade-off curve* answers the question as follows: For a given amount of information about previous words  $H_M$  stored in the listener’s memory state, what is the lowest achievable average surprisal  $S_M$ ? Two example trade-off curves are shown in Figure 1. In general, as the listener stores more information about previous words in her memory state, her lowest achievable average surprisal can only decrease. So the curve is always monotonically decreasing. However, the precise shape of the trade-off curve depends on the structure of the language being predicted. For example, Figure 1 shows how two hypothetical languages might engender different trade-off curves, with Language A allowing more favorable trade-offs than Language B. That is, for Language A, it is possible to achieve lower processing difficulty while investing less memory resources than in Language B.

## Main Hypothesis

Having conceptually introduced the memory–surprisal trade-off, we can state the main hypothesis of this work, the Efficient Trade-off Hypothesis.

### Efficient Trade-off Hypothesis:

The order of elements in natural language is characterized by a distinctively steeper memory–surprisal trade-off curve, compared to other possible orders.

A steep trade-off curve corresponds to memory efficiency, in the sense that it is possible to achieve a low level of processing difficulty (average surprisal  $S_M$ ) while storing a relatively small amount of information about previous words (entropy of memory  $H_M$ ). We hypothesize that this property is reflected in grammatical structure and usage preferences across languages.

## Formal Definition of the Memory–Surprisal Trade-Off

Here, we provide the technical definition of the memory–surprisal trade-off curve. Let  $W$  be a stochastic process generating a stream of symbols extending indefinitely into the past and future:  $\dots, w_{-2}, w_{-1}, w_0, w_1, w_2, \dots$ . These symbols can represent words, morphemes, or other units for decomposing sentences into a sequence of symbols. We model this process as stationary (Doob, 1953), that is, the joint probability distributions of symbols at different time points depend only on their relative positions in time, not their absolute positions (see SI Section 1.1.1 for more on this modeling assumption).

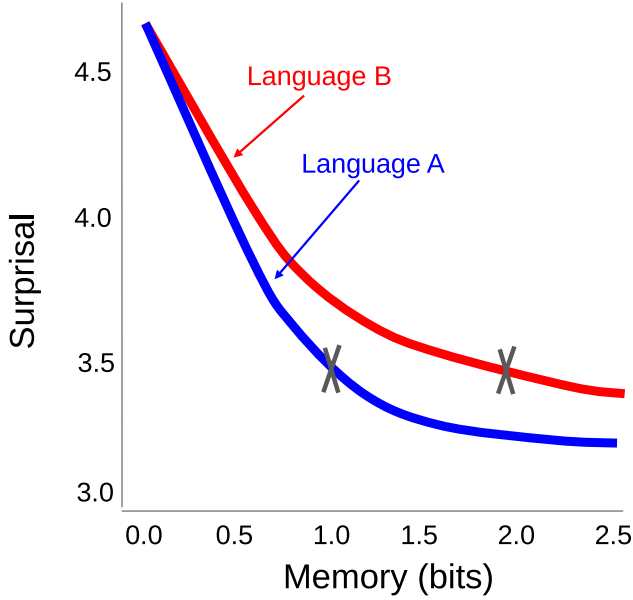
Let  $M$  be a memory encoding function. We consider memory and surprisal costs at an arbitrary time point  $t$ . Recall that the surprisal for a specific word  $w_t$  after a past word sequence  $\dots, w_{t-2}, w_{t-1}$  encoded into a memory state  $m_t$  is as follows:

<sup>1</sup> In this article, all logarithms are taken to base 2. As choosing another basis (e.g.,  $e$ ) would only result in multiplication with a proportionality constant, this assumption does not impact the generality of this linking hypothesis.



**Figure 1**

*Example Memory–Surprisal Trade-Off Curves for Two Languages, A and B. Achieving an Average Surprisal of 3.5 Bits Requires Storing at Least 1.0 Bits in Language A, While It Requires Storing 2.0 Bits in Language B. Language A Has a Steeper Memory–Surprisal Trade-Off Than Language B, and Requires Less Memory Resources to Achieve the Same Level of Processing Difficulty*



Note. See the online article for the color version of this figure.

$$-\log_2 P(w_t | m_t).$$

The *average surprisal* of the process  $W$  under the memory encoding function  $M$  is obtained by averaging over all possible past sequences  $\dots, w_{t-2}, w_{t-1}$  with associated memory states  $m_t$ , and possible next words  $w_t$ :

$$S_M \equiv - \sum_{w_t, m_t} P(m_t) P(w_t | m_t) \log_2 P(w_t | m_t).$$

where  $w_t$  ranges over possible symbols, and  $m_t$  ranges over possible outputs of the memory encoding function  $M$ . This quantity is known as the conditional entropy of  $w_t$  given  $m_t$  (Cover & Thomas, 2006, p. 17):

$$S_M = H[w_t | m_t].$$

Because the process  $W$  is stationary, the average surprisal  $S_M$  is independent of the choice of  $t$  (see SI Section 1.1.2). The lowest possible average surprisal for  $W$  is attained when  $m_t$  perfectly encodes all previous observed words. This quantity is called the *entropy rate* of  $W$  (Cover & Thomas, 2006, pp. 74–75):

$$S_\infty \equiv H[w_t | \dots, w_{t-2}, w_{t-1}],$$

which again is independent of  $t$  because  $W$  is stationary. We use the notation  $S_\infty$  to suggest this idea of unlimited resources. The entropy rate of a stochastic process is the irreducible unpredictability of the process: The extent to which a stream of symbols remains unpredictable even for a predictor with unlimited resources.

Because the memory state  $m_t$  is a function of the previous words  $\dots, w_{t-2}, w_{t-1}$ , we can prove by the Data Processing Inequality (Cover & Thomas, 2006, pp. 34–35) that the entropy rate must be less than or equal to the average surprisal for any memory encoding function  $M$ :

$$S_\infty \leq S_M.$$

If the memory state  $m_t$  stores all information about the previous words  $\dots, w_{t-2}, w_{t-1}$ , then we have  $S_M = S_\infty$ .

Having defined average surprisal, we now turn to the question of how to define memory capacity. The average amount of information stored in the memory states  $m_t$  is the average number of bits required to encode  $m_t$ . This is given by the entropy of the stationary distribution over memory states,  $H_M$ :

$$H_M \equiv H[m_t]$$

where

$$H[m_t] = - \sum_m P(m_t = m) \log_2 P(m_t = m)$$

where  $m$  runs over all possible states of the memory encoding  $m_t$ . Again, because  $W$  is stationary, this quantity does not depend on the choice of  $t$  (see SI Section 1.1.2).

We will be imposing bounds on  $H_M$  and studying the resulting values of  $S_M$ .

**Definition 1.** The *memory–surprisal trade-off curve* for a process  $W$  is the lowest achievable average surprisal  $S_M$  for each value of  $H_M$ . Let  $R$  denote an upper bound on the memory entropy  $H_M$ ; then, the memory–surprisal trade-off curve as a function of  $R$  is given by

$$D(R) \equiv \min_{M: H_M \leq R} S_M, \quad (2)$$

where the minimization is over all memory encoding functions  $M$  whose entropy  $H_M$  is less than or equal to  $R$ .

The memory state  $m_t$  is generally a lossy representation of the true context of words  $w_1, \dots, w_{t-1}$ , meaning that  $m_t$  does not contain all the possible information about  $w_1, \dots, w_{t-1}$ . The mathematical theory of lossy representations is *rate–distortion theory* (for an overview and key results, see Cover & Thomas, 2006, pp. 301–347); this theory has seen recent successful application in cognitive science and linguistics as a model of rational action under resource constraints (Brady et al., 2009; Gershman, 2020; Schach et al., 2018; Sims, 2018; Sims et al., 2012; Zaslavsky et al., 2018; Zénon et al., 2019). Rate–distortion theory studies curves of the form of Eq. 2, which quantify trade-offs between negative utility (“distortion”) and information (“rate”).

## Information Locality

The shape of the memory–surprisal trade-off is determined in part by the grammatical structure of a language. Some hypothetical languages enable more efficient trade-offs than others by allowing

a listener to store fewer bits in memory to achieve the same level of average surprisal.

Here, we will demonstrate that the memory–surprisal trade-off is optimized by languages with word orders exhibiting a property called *information locality*. Information locality means that words that depend on each other statistically are located close to each other in time. We will argue that information locality generalizes the well-known word order principle of dependency locality.

We will make our argument by defining a lower bound on the memory–surprisal trade-off curve (Eq. 2). This lower bound represents an unavoidable cost associated with a certain level of memory usage  $H_M$ ; the true average surprisal  $S_M$  might be higher than this bound.

Our argument will make use of a quantity called *mutual information*. Mutual information is the most general measure of statistical association between two random variables. The mutual information between two random variables  $X$  and  $Y$ , conditional on a third random variable  $Z$ , is defined as:

$$I[X:Y|Z] \equiv \sum_{x,y,z} P(x,y,z) \log_2 \frac{P(x,y|z)}{P(x|z)P(y|z)}. \quad (3)$$

Mutual information is always nonnegative. It is zero when  $X$  and  $Y$  are conditionally independent given  $Z$ , and positive whenever  $X$  gives any information that makes the value of  $Y$  more predictable, or vice versa.

We will study the mutual information structure of natural language sentences, and in particular, the mutual information between words at certain distances in linear order. We define the notation  $I_t$  to mean the mutual information between words at distance  $t$  from each other, conditional on the intervening words:

$$I_t \equiv I[w_t : w_0 | w_1, \dots, w_{t-1}].$$

This quantity, visualized in Figure 2(a), measures how much predictive information is provided about the current word by the word  $t$  steps in the past. It is a statistical property of the language, and can be estimated from large-scale text data.

Equipped with this notion of mutual information at a distance, we can now state our theorem:

**Theorem 1. (Information locality bound)** For any positive integer  $T$ , let  $M$  be a memory encoding function such that

$$H_M \leq \sum_{t=1}^T t I_t. \quad (4)$$

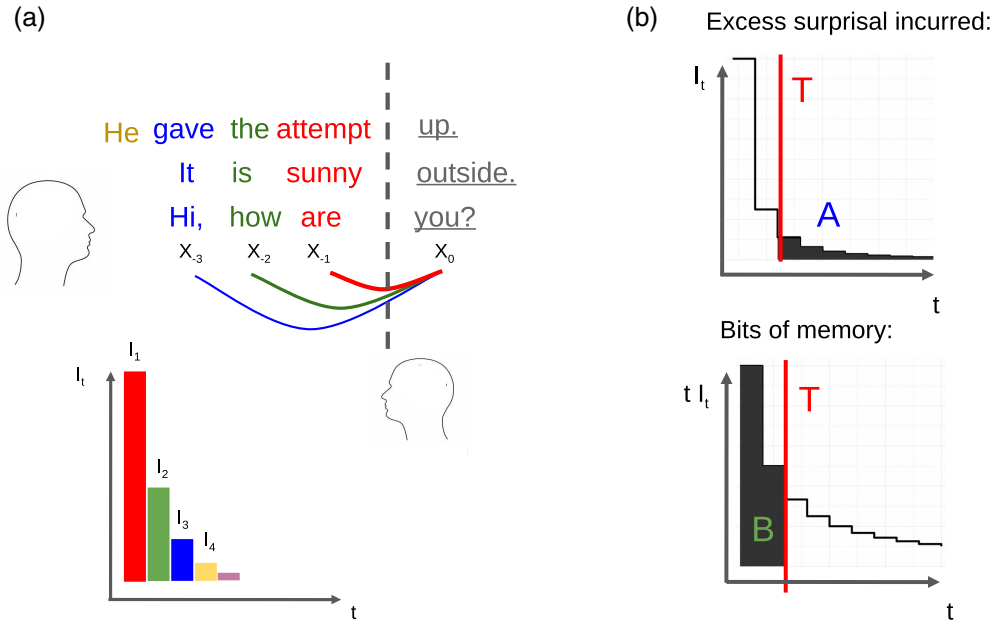
Then, we have a lower bound on the average surprisal under the memory encoding function  $M$ :

$$S_M \geq S_\infty + \sum_{t=T+1}^{\infty} I_t. \quad (5)$$

A formal proof based on the Comprehension Postulates 1–3 is given in SI Section 1.2. An intuitive argument, forming the basis of

**Figure 2**

(a) Conditional Mutual Information  $I_t$  Captures how Much Predictive Information About the Next Word Is Provided, on Average, by the Word  $t$  Steps in the Past. (b) Here, We Illustrate Our Theoretical Result. We Plot  $I_t$  (Top) and  $t I_t$  (Bottom) as Functions of  $t$ . for Any Choice of  $T$ , a Listener Using  $B$  Bits Memory (Bottom) to Represent Prior Observations Will Incur at Least  $A$  Bits of Extra Surprisal Beyond the Entropy Rate



Note. See the online article for the color version of this figure.

the proof, is the following. Suppose that a comprehender predicting the  $t$ 'th word  $w_t$  uses an average of  $I_t$  bits of information coming from a previous word  $w_0$ . Then these bits must have been carried over  $t$  timesteps and thus have occupied memory for  $t$  timesteps. Because this happens for every word in a sequence, there are, at any given point in time,  $t$  such packets of information, each with an average size of  $I_t$  bits, that have to be maintained, summing up to  $tI_t$ . In the specific setting where  $M$  encodes information from a contiguous span of the past  $T$  words, the total amount of encoded information thus sums up to  $\sum_{i=1}^T tI_t$ , while information from longer contexts is lost, increasing surprisal by  $\sum_{i=T+1}^{\infty} I_i$ . Although this informal argument specifically considers a memory encoding function that utilizes information from a contiguous span of the past  $T$  words, the formal proof extends this to all memory encoding functions  $M$  satisfying the Comprehension Postulates.

### Interpretation

The theorem means that a predictor with limited memory capacity will always be affected by surprisal cost arising from long-term statistical dependencies of length greater than  $T$ , for some finite  $T$ . This is why we call the result “information locality”: Processes are easier to predict when most statistical dependencies are short term (shorter than some  $T$ ). Below, we explain in more detail how this interpretation matches the mathematics of the theorem.

The quantities in the theorem are shown visually in Figure 2. Eq. 4 describes a memory encoding function which has enough capacity to remember the relevant information from at most  $T$  words in the immediate past. The minimal amount of memory capacity which would be required to retain this information is the sum  $\sum_{i=1}^T tI_t$ , reflecting the cost of holding  $I_t$  bits in memory for  $t$  timesteps up to  $t = T$ .

The information locality bound theorem says that the surprisal cost for this memory encoding function is at least  $S_{\infty} + \sum_{i=T+1}^{\infty} I_i$  (Eq. 5). The first term  $S_{\infty}$  is the entropy rate of the process, representing the bits of information in the process which could not have been predicted given any amount of memory. The second term  $\sum_{i=T+1}^{\infty} I_i$  is the sum of all the relevant information contained in words *more* than  $T$  timesteps in the past [see Figure 2(b)]. These correspond to bits of information in the process which *could have* been predicted given infinite memory resources, but which were not, due to the limit on memory usage.

The theorem gives a lower bound on the memory–surprisal trade-off curve, meaning that there is no memory encoding function  $M$  with capacity  $H_M$  which achieves lower average surprisal than Eq. 5. In terms of psycholinguistics, if memory usage is bounded by Eq. 4, then processing cost of at least Eq. 5 is inevitable. Importantly, the bound holds for *any* memory encoding function  $M$ , including functions that do not specifically keep track of a window of the past  $T$  words. The information locality bound theorem demonstrates in a highly general way that language comprehension requires less memory resources when statistical dependencies are mostly short term.

Because processing long-term dependencies requires higher memory usage, the theorem also implies that a language can be easier to process when most of the predictive information about a word is concentrated close to that word in time—that is, when  $I_t$  falls off rapidly as  $t \rightarrow \infty$ . When memory capacity is limited, then there

must be some timescale  $T$  such that a listener appears to be affected by excess surprisal arising from statistical dependencies of length greater than  $T$ . A language avoids such cost to the extent that it avoids dependencies with a time span larger than  $T$ .

We illustrate the theorem in Figure 3. We consider two hypothetical languages, LessEfficient and MoreEfficient, where  $I_t := 5t^{-1.5}$  for LessEfficient and  $I_t := 3.5t^{-2.5}$  for MoreEfficient.<sup>2</sup> The curves of  $I_t$ , as a function of the distance  $t$ , are shown in Figure 3 (left). In both cases,  $I_t$  converges to zero as  $t$  grows to infinity. However,  $I_t$  decays more quickly for language MoreEfficient. This means that predictive information about an observation is concentrated more strongly in the recent past. In Figure 3 (right), we show  $t \cdot I_t$  as a function of  $t$ . Note that the area under the curve (AUC) is equal to (4). This area is smaller for the MoreEfficient language, as  $I_t$  decays more quickly there. In Figure 4, we show the resulting bounds on memory–surprisal trade-offs of the two languages. As  $I_t$  decays faster for language MoreEfficient, it has a more efficient memory–surprisal trade-off, allowing a listener to achieve strictly lower surprisal across a range of memory values.

### Other Kinds of Memory Bottlenecks

We derived the memory–surprisal trade-off and the Information Locality Lower Bound by imposing a capacity limit on memory using the entropy  $H_M$ . The entropy  $H_M$  represents the average amount of information that can be stored in memory at any time. However, in some psycholinguistic theories, memory-related difficulty arises not because of a bound on memory capacity, but rather because of difficulties involved in retrieving information from memory (Lewis & Vasishth, 2005; McElree, 2000; Nicenboim & Vasishth, 2018; Vasishth et al., 2019).

It turns out that it is possible to derive results closely analogous to ours by imposing a capacity limit on the retrieval of information from memory, rather than the storage of information. Essentially, the constraint on the memory state in our Theorem 1 can be reinterpreted as a constraint on the capacity of a communication channel linking short-term memory to working memory. This result constrains average surprisal for memory models based on cue-based retrieval such as the Adaptive Control of Thought-Rational (ACT-R) model of Lewis and Vasishth (2005). In fact, the theorem based on retrieval capacity gives a tighter bound than the theorem based on storage capacity. For the full model and derivation, see SI Section 1.3.

We believe that concepts analogous to the memory–surprisal trade-off and the Information Locality Lower Bound are likely to be valid across a broad range of models of incremental processing and memory.

### Study 1: Memory and Dependency Length

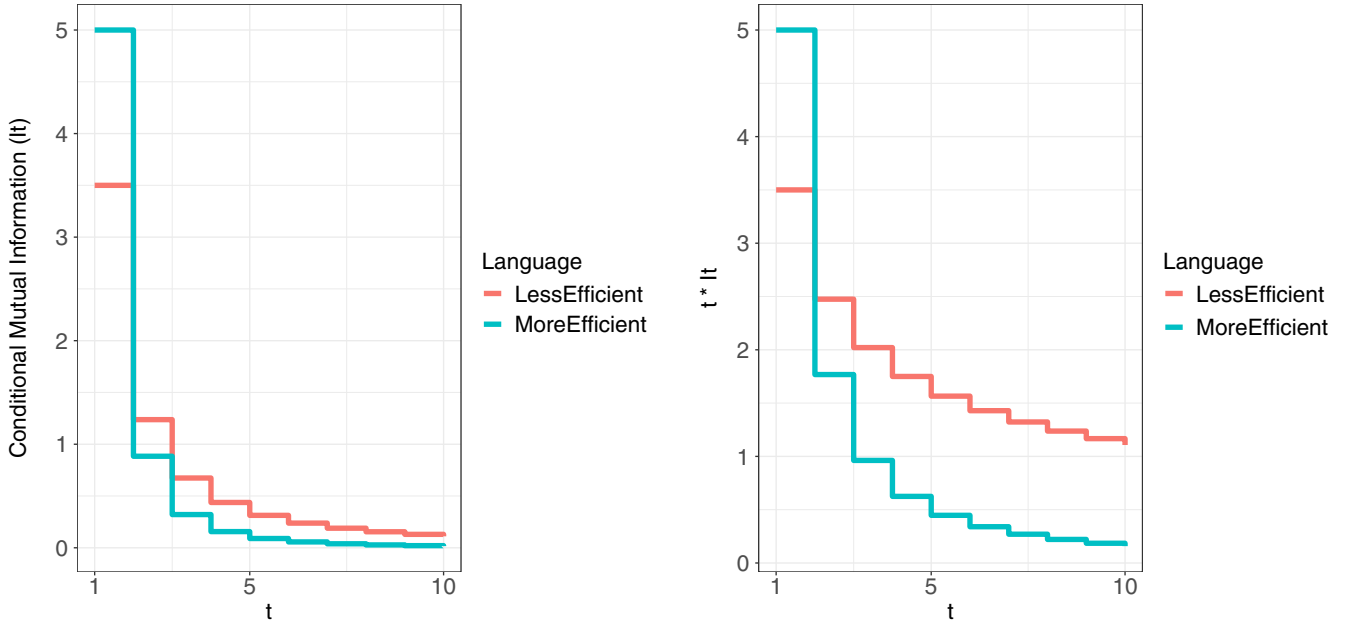
So far, we have proven that there exists a trade-off between memory invested and surprisal incurred during language processing, and that this trade-off is optimized when languages have relatively short-term dependencies. In this section, we qualitatively test the Efficient Trade-off Hypothesis by reanalyzing the data from the study by Fedzechkina et al., 2017. This is a miniature artificial

<sup>2</sup> Although these are purely mathematical examples, the  $I_t$  curve for natural languages does seem empirically to fall off as a power law as in these examples (Debowski, 2015).



**Figure 3**

Left:  $I_t$  as a Function of  $t$ , for Two Different Hypothetical Languages.  $I_t$  Decays Faster for the MoreEfficient Language: Predictive Information About the Present Observation Is Concentrated More Strongly in the Recent Past. Right:  $t \cdot I_t$  as a Function of  $t$  for the Same Languages



Note. See the online article for the color version of this figure.

language study that showed a bias for dependency locality in production in artificial language learning. We will show that, as predicted, the languages which were favored in the artificial language learning experiment are those which optimize the memory–surprisal trade-off.

### Background: Fedzechkina et al. (2017)

Fedzechkina et al. (2017) conducted a miniature artificial language learning experiment in which participants were exposed to videos describing simple events, paired with sentences in an artificial language of the form Subject–Object–Verb or Object–Subject–Verb, in equal proportion, with free variation between these two word orders. The subject and the object were either simple nouns, or complex noun phrases with modifiers. Participants were trained to produce sentences in response to videos.

Crucially, Fedzechkina et al. (2017) set up the experiment such that in all training trials, either the subject and the object were both simple or they were both complex. Then, after participants were sufficiently skilled in the use of the artificial language, they were asked to produce sentences describing videos with mixed complexity of noun phrases. The possible word orders that could be produced in this mixed-complexity setting are shown in Figure 5; the orders marked *A* would create short dependencies, and the orders marked *B* would create long dependencies.

Fedzechkina et al. (2017) found that participants favored the *A* orders over the *B* orders, despite the fact that there was no pattern in the participants' training input which would have favored *A* over *B*. That is, when exposed to input which was ambiguous with respect to language *A* or *B*, participants favored language *A*. Fedzechkina et al.

(2017) explained this result in terms of dependency locality: Because the *A* orders create short dependencies between the verb and its arguments and the *B* orders create long dependencies, participants preferred the *A* orders.

### Calculating the Memory–Surprisal Trade-Off for the Artificial Languages

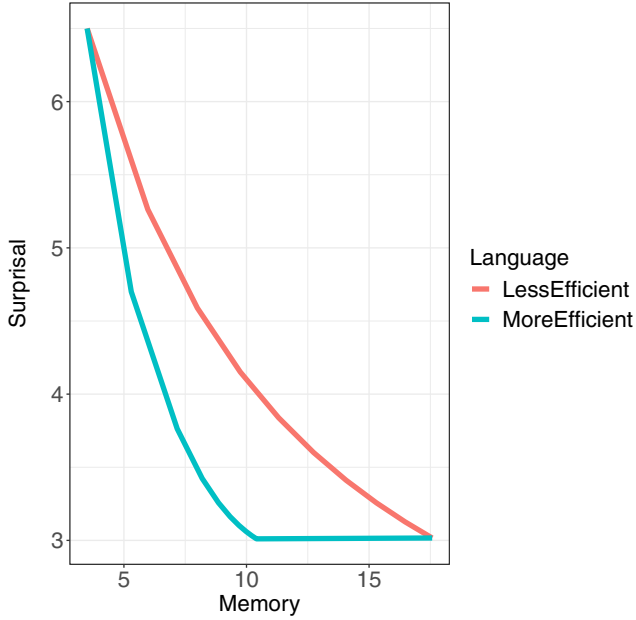
The Efficient Trade-off Hypothesis predicts that the favored language *A* has a steeper memory–surprisal trade-off curve than the disfavored language *B*. Because of the controlled nature of this artificial language, we are able to test this hypothesis by exactly computing the bound on memory as given in Theorem 1. In fact, for this toy process, we can prove that the bound provided by the theorem is achievable, meaning that our computations reflect the true memory–surprisal trade-off curve, and not only a lower bound on it.

We only consider the head-final version of Fedzechkina et al. (2017)'s artificial language. This is because our bound on the memory–surprisal trade-off curve is invariant under reversal of a language. That is, if we take a language and reverse the order of all the words in all its sentences, we would measure the same lower bound on the memory–surprisal trade-off curve (for a proof, see SI Section 1.5). Therefore, strictly head-final and strictly head-initial languages are equivalent under our bound.

We constructed a stochastic process representing the language consisting of sentences with the *A* orders from Figure 5, and one language consisting of the *B* orders. Following the experimental setup of Fedzechkina et al., (2017), we assigned equal probability to the two possible configurations per language, and used a separate set

**Figure 4**

*Listener's Memory–Surprisal Trade-Off for the Two Hypothetical Languages in Figure 3. Recall That the MoreEfficient Language Has a Faster Decay of Conditional Mutual Information  $I_t$ . Correspondingly, This Figure Shows That a Listener Can Achieve Lower Average Surprisal at the Same Level of Memory Load*



Note. See the online article for the color version of this figure.

of nouns (inanimate nouns) for the embedded noun in the long phrase.

We interpreted each of the two languages as a stationary processes, extending infinitely in both directions, by concatenating independent samples drawn from the language, and separating them with a special symbol indicating sentence boundaries. We computed the bounds on memory and surprisal (4–5) from Theorem 1 from a chain of 1,000 independently sampled sentences, for each of the two versions of the toy language.

**Figure 5**

*Production Targets in the Miniature Artificial Language From the Study by Fedzechkina et al. 2017. The Language Has Head-Final Order, With Free Variation Between SO and OS Orders. When One of the Arguments Is Much Longer Than the Other, Placing the Longer One First (A Orders) Shortens Syntactic Dependencies, Compared to B Orders*

#### A Orders: Short Dependencies

OSV: [[Adjective Noun Postposition] Noun-CASE] Noun Verb

SOV: [[Adjective Noun Postposition] Noun] Noun-CASE Verb

#### B Orders: Long Dependencies

SOV: Noun [[Adjective Noun Postposition] Noun-CASE] Verb

OSV: Noun-CASE [[Adjective Noun Postposition] Noun] Verb

## Results

Figure 6 (left) shows the curve of the conditional mutual information  $I_t$  as a function of the distance  $t$ , for the two languages  $A$  and  $B$ . The curves differ at  $t = 2$  and  $t = 5$ : About 0.105 bits of predictive information that are at distance  $t = 2$  in the  $A$  orders are moved to  $t = 5$  in the  $B$  orders.

The source of the difference lies in predicting the presence and absence of a case marker on the second argument. Conceptually, a comprehender may be in a state of uncertainty as to whether a subject or object might follow. Because surprisal is determined entirely by the statistical properties of distributions over wordforms, this uncertainty manifests as uncertainty about whether to expect an accusative case marker.<sup>3</sup> In the  $A$  orders, considering the last two words is sufficient to make this decision. In the  $B$  orders, it is necessary to consider the word before the long second constituent, which is five words in the past.

The total amount of predictive information—corresponding to the area under the  $I_t$  curve—is the same for both languages, indicating that both languages are equally predictable. However, the memory demands differ between the two languages. Figure 6 (right) shows the minimal memory requirements for remembering predictive information at a distance  $t \cdot (t \cdot I_t)$  as a function of  $t$ . As  $I_t$  decays faster in  $A$  orders, the total AUC differs between  $A$  and  $B$ , and is larger in  $B$ . Thus, achieving the same predictive accuracy in language  $B$  requires more memory resources than in language  $A$ .

Figure 7 shows the resulting memory–surprisal trade-off curve for the two versions of the artificial language from the study by Fedzechkina et al., 2017, obtained by tracing out all values of  $T = 1, 2, \dots$  in the theorem, and connecting the points linearly.<sup>4</sup> The curve shows that, at any desired level of surprisal, language  $A$  requires at most as much memory as language  $B$ . To reach optimal surprisal, the empirically favored language  $A$  requires strictly less memory.

## Discussion

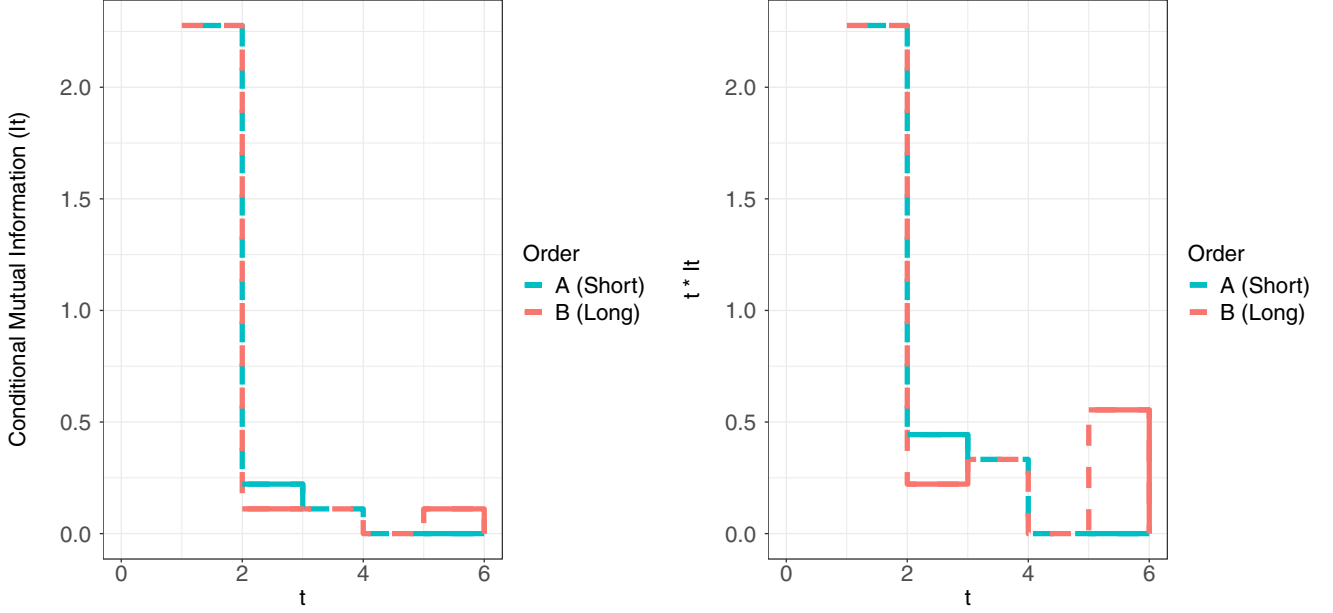
In a reinterpretation of previous experimental findings, we showed that the languages which are favored in an artificial language learning experiment are those which optimize the memory–surprisal trade-off. This is evidence that learners and/or speakers have a bias toward word orders that optimize the trade-off. Furthermore, this result solidifies the link between the memory–surprisal trade-off and more traditional notions from linguistics, such as dependency locality. We found that the word orders which are optimal from the perspective of dependency locality are also those orders which are optimal from the perspective of the memory–surprisal trade-off, in the setting of a small controlled artificial language. In Study 2, we scale this approach up to larger corpora of real text.

<sup>3</sup> In other languages lacking case markers, similar uncertainty may manifest as uncertainty about wordform because subjects and objects often have very different distributions over wordforms.

<sup>4</sup> Linear interpolation is justified because rate-distortion curves such as the memory–surprisal trade-off curve are convex (Berger, 2003).

**Figure 6**

Left: Decay of Conditional Mutual Information  $I_t$ , as a Function of the Distance  $t$ , for the Two Versions in the Artificial Language. The Areas Under the Two Curves Are Identical, Corresponding to the Fact That Both Orders Are Equally Predictable. However, Mutual Information Decays Faster in Language A. Right: the Minimal Memory Requirement  $tI_t$ , to Store  $I_t$  Bits of Information for Timespan  $t$ , as a Function of  $t$ . The Area Under the B Curve Is Larger, Corresponding to Larger Memory Demand for This Order



Note. See the online article for the color version of this figure.

## Study 2: Large-Scale Evidence That Word Orders Optimize Memory–Surprisal Trade-Off

To test whether word orders as found in natural language reflect optimization for the memory–surprisal trade-off more generally, we compare the memory–surprisal trade-offs of 54 actual languages to those of counterfactual baseline languages. These baseline languages differ from the actual languages only in their word order rules. This method of comparison against counterfactual baseline languages was introduced by Gildea and Temperley (2007, 2010) and has since been fruitfully applied to study optimization-based models of word order universals (Futrell et al., 2015a; Gildea & Jaeger, 2015; Hahn et al., 2020).

Here, we describe how we measure the memory–surprisal trade-off in corpora, and how we generate counterfactual baseline languages. We then compare the trade-off in real corpora against the trade-off in the counterfactual baselines. For the majority of languages, we find that the real languages have more favorable memory–surprisal trade-offs than the baselines, in line with the Efficient Trade-off Hypothesis.

### Measuring the Memory–Surprisal Trade-Off in Corpora

The key to evaluating the memory–surprisal trade-off from corpus data is the set of quantities  $I_t$ , the mutual information between words at distance  $t$  conditional on the intervening words. These quantities can be plugged in to Theorem 1 to give a lower bound on the memory–surprisal trade-off.

The quantities  $I_t$  can be estimated as the difference between the average surprisal of Markov models have access to windows of size  $t$  and  $t + 1$ . That is, if we have a  $t$ 'th-order Markov model with average surprisal

$$S_t = H[w_t | w_1, \dots, w_{t-1}] \quad (6)$$

and a  $(t + 1)$ 'th-order Markov model with average surprisal

$$S_{t+1} = H[w_{t+1} | w_0, \dots, w_t],$$

then, we can calculate  $I_t$  straightforwardly in the following way:

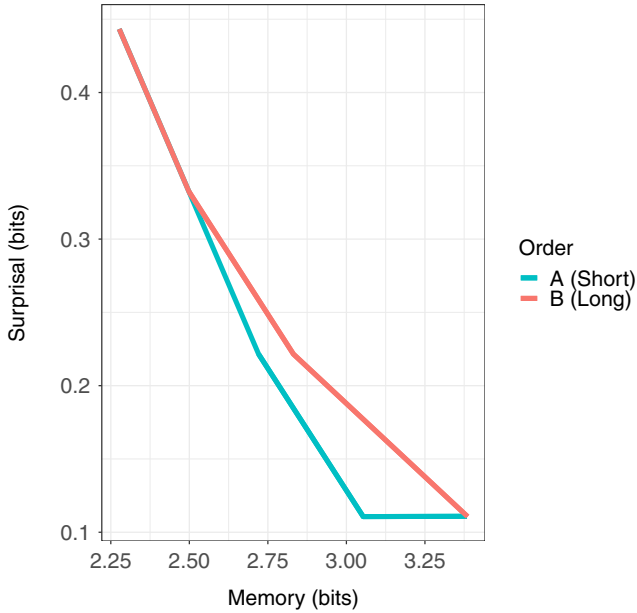
$$\begin{aligned} I_t &= I[w_t : w_0 | w_1, \dots, w_{t-1}] \\ &= S_t - S_{t+1}. \end{aligned}$$

Therefore, to evaluate  $I_t$ , all we need is a way of fitting Markov models of order  $t$  and  $t + 1$  and computing their average surprisals.

To fit Markov models to the data, we use neural language models. In particular, we use Recurrent Neural Networks with Long Short-Term Memory architectures (Hochreiter & Schmidhuber, 1997). Neural network models are the basis of the state-of-the-art in statistical modeling of language. Surprisal estimates derived from such models have been shown to best predict reading times, compared to other models, for example,  $n$ -gram models (Frank & Bod, 2011; Goodkind & Bicknell, 2018). See SI Section 3.2 for details on how these models were fit to data, and see SI Sections 3.4 and 3.5 for control studies using other methods of estimating  $I_t$  (based on  $n$ -gram models and PCFG chart parsers). These control

**Figure 7**

Trade-Off Between Listener Memory and Surprisal for the Two Versions of the Artificial Language From the Study by Fedzechkina et al. 2017. Language A Requires Less Memory at the Same Level of Surprisal



Note. See the online article for the color version of this figure.

studies yield the same qualitative results as the neural network-based studies presented here.

For each language, we run the neural network estimator multiple times with different random seeds, to control for variation in the random initialization of model parameters (see SI Section 3.2.3 for details).

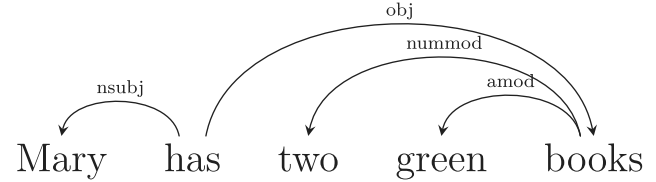
To evaluate the average surprisal values  $S_t$ , we computed the empirical word-by-word surprisal values under the  $t$ 'th-order Markov model for held-out data, different from the data that was used to train the model. By evaluating on held-out data, we avoid underestimating the value of  $S_t$  due to overfitting. We chose held-out data based on existing splits of corpora, see the section "Data" below.

## Data

We draw on syntactically annotated corpora, compiled by the Universal Dependencies project for several dozen languages (Nivre et al., 2017). These are annotated in the format of Dependency Grammar (Corbett et al., 1993; Hays, 1964; Hudson, 1984; Melcuk, 1988; Tesnière & Kahane, 2015). In such dependency corpora, sentences are annotated with *dependency trees* (Figure 8). These are directed trees describing the grammatical relations among words. For example, the arcs labeled "obj" represent that the noun in question is the *direct object* of the verb, rather than, for example, the subject or an indirect object. A dependency arc is drawn from a *head* (e.g., the verb "has") to a *dependent* (e.g., its object "book"). Dependency trees can be defined in terms of many different syntactic theories (Corbett et al., 1993). Although there are some differences in how different formalisms would draw trees for certain sentences, there is broad enough agreement about dependency trees

**Figure 8**

An English Sentence With Dependency Annotations, According to the Universal Dependencies 2.4 Standard Nivre-Universal-2017. We Visualize Grammatical Relations as Arcs Drawn From Heads (e.g., the Verb "Has") to Dependents (e.g., its Object "Book")



that it has been possible to develop large-scale dependency-annotated corpora of text from dozens of languages (Nivre et al., 2017).

We computed memory–surprisal trade-offs for all languages for which there are Universal Dependencies 2.4 treebanks with a total of at least 500 sentences of training data. We excluded data from historical languages, as these corpora often include poetry, translated text, or texts spanning several centuries.<sup>5</sup> This resulted in 54 languages. We also excluded corpora that primarily contain code-switched text<sup>6</sup> or text created by nonnative speakers.<sup>7</sup>

For each of these languages, we pooled all available corpora into one data set. Most Universal Dependencies corpora have a predefined split into *training*, *held-out* (also known as *development*), and *test* partitions. In most cases, we used the predefined data split, separately pooling data from the different partitions. For some languages with little data, there is no predefined training partition, or the training partition is smaller than the other partitions. In these cases, we redefined the split to obtain more training data. For these languages, we pooled all the available partitions, used 100 randomly selected sentences as held-out data, and used the remainder as training data.<sup>8</sup> We did not make use of the *test* partitions here. We provide the sizes of the resulting data sets in SI Section 3.1. The data sets ranged in size from 564 sentences (Armenian) to 114,304 sentences (Czech), with a median of 5,255 sentences per language. For each language, we obtain a stationary process by concatenating the sentences from the corpus in random order, separated with an end-of-sentence symbol.

## Defining Baselines

Testing the Efficient Trade-off Hypothesis requires comparing the memory–surprisal trade-offs of real grammars to those of baseline grammars. The baseline grammars we construct are counterfactual ordering grammars that define consistent ordering rules similar to those found in actual languages (Figure 9). For instance, these grammars specify which dependents precede or follow their heads (e.g., whether objects follow or precede verbs, whether adjectives follow or precede nouns), and the relative order of different dependents on the same side of the head (e.g., whether noun phrases have order adjective–numeral–noun or

<sup>5</sup> Historical languages excluded, Ancient Greek, Classical Chinese, Coptic, Gothic, Latin, Old Church Slavonic, and Old French.

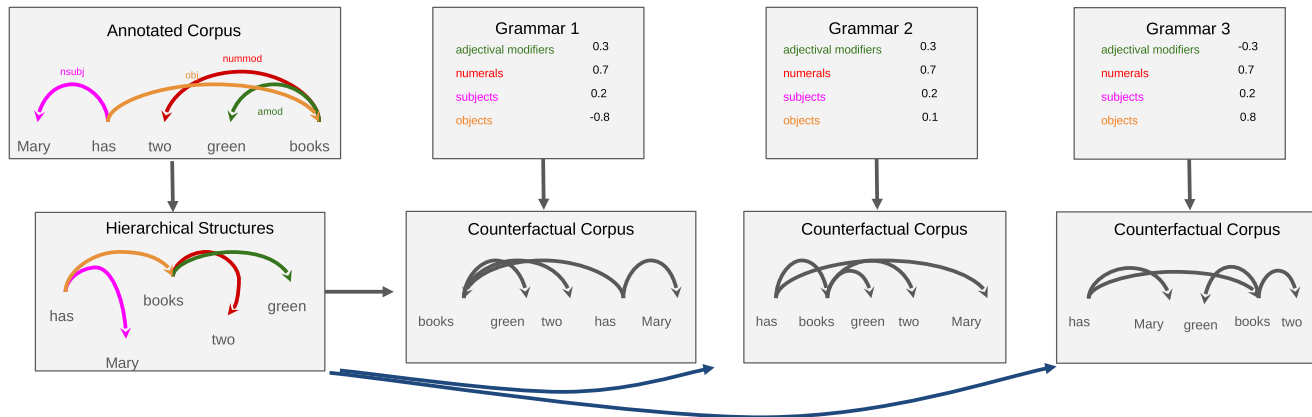
<sup>6</sup> Hindi English corpus.

<sup>7</sup> ESL for English, CFL for Chinese.

<sup>8</sup> This affects Amharic, Armenian, Breton, Buryat, Cantonese, Faroese, Kazakh, Kurmanji, Nijja, Thai, and Uyghur.

**Figure 9**

*Estimating Chance by Constructing Counterfactual Grammars and Languages: We Start From an Annotated Dependency Corpus of Sentences Annotated With Syntactic Dependencies (Top Left). We Then Extract the Raw Dependency Structures, Stripping Away Word Order Information (Bottom Left). We Construct Baseline Ordering Grammars That Provide Rules for Ordering the Words in Such Dependency Structures (Grammars 1–3). Applying Any Such Grammar to the Dependency Structures Yields a Counterfactual Corpus of a Hypothetical Language That Has the Same Dependency Structures as the Actual Language, but Different Word Order Rules*



*Note.* See the online article for the color version of this figure.

numeral–adjective–noun). Our formalism of ordering grammars was introduced in the study by Hahn et al. (2020), adapting the method of Gildea and Temperley (2007, 2010) to the setting of dependency corpora.

Universal Dependencies 2.4 defines 37 universal syntactic relations that are used to label dependency arcs across all corpora. These relations encode crosslinguistically meaningful relations such as subjects (*nsubj*, see Figure 8), objects (*obj*), and adjectival modifiers (*amod*). We define ordering grammars by assigning a parameter  $a_\tau \in [-1, 1]$  to every one of these 37 universal syntactic relations. Relations sometimes have language-specific subtypes; we do not distinguish these subtypes. Following Gildea and colleagues, this parameter defines how dependents are ordered relative to their head: Given a head and a set of dependents, we order each dependent by the parameter  $a_\tau$  assigned to the syntactic relation linking it to the head. Dependents with negative weights are placed to the left of the head; dependents with positive weights are placed to the right. Ordering grammars describe languages that have consistent word order. For instance, the subject is consistently ordered before or after the verb, depending on whether the parameter  $a_{nsubj}$  for the verb–subject dependency is positive or negative.

We constructed baseline grammars by randomly sampling the parameters  $a_\tau$ . Such baseline grammars define languages that have word order rules which are consistent but do not exhibit systematic preferences for patterns such as short dependencies.

We first constructed at least 10 baseline grammars for each of the 54 real languages. We then continued to construct baseline grammars until a precision-based stopping criterion was reached. This criterion was designed to ensure that enough grammars were sampled to reliably compare the trade-off curves of real and baseline grammars, without biasing results toward or against our hypothesis (see SI Section 3.2.3). The stopping criterion compared what fraction of baseline grammars had strictly more (or strictly less) efficient trade-off curves than the real ordering, and required a bootstrapped 95% confidence interval for that ratio to have

width  $\leq 0.15$ . The resulting number of baseline grammars ranged from 10 (Italian and Romanian) to 347 (Estonian).<sup>9</sup>

Due to the way ordering grammars are specified, certain kinds of rules cannot be modeled by our word order grammars. This includes rules sensitive to the category of the dependent, such as the difference between postverbal nominal objects and preverbal pronominal objects in Romance languages. It also includes rules sensitive to larger context, for example, the alternation between verb-final order in embedded clauses and verb-initial/verb-medial order in main clauses in German and Dutch. Furthermore, the model does not allow rules specifying interactions between different constituents, for instance, verb-second order, where exactly one dependent precedes the verb, and all others follow it. Finally, the model does not account for word order freedom, as all ordering choices are deterministic. In this sense, ordering grammars only represent an approximation to the kinds of ordering rules found in natural language. Other models described in the literature (Futrell & Gibson, 2015; Wang & Eisner, 2016) mostly share these limitations.

To ensure that results are not due to the representational restrictions of the word order grammar formalism, we also compared the baselines to the result of ordering the corpora according to grammars that approximate the real orders to the extent possible in the grammar formalism. These grammars have exactly the same representational constraints as the baseline grammars while approximating the real orderings. We expect these grammars to have better memory–surprisal trade-offs than comparable random baseline grammars across all languages. We created these ordering grammars by fitting them to the actual orderings of each language using the method of Hahn et al., 2020. They match the order of the actual language in those cases where order of a relation is fully consistent; for relations where order is variable, they approximate this by modeling the most frequent order. In representing word order rules,

<sup>9</sup> Due to a scripting error, 846 grammars were generated for Erzya even though this was not required by the stopping criterion.



they have the same limitations as baseline grammars have, for instance, they cannot specify rules sensitive to the category of the dependent or to larger context.

## Results

To test the Efficient Trade-off Hypothesis, we compare the trade-off curves for the real orders with those for random baseline grammars. In Figure 10, we show the estimated values of  $I_t$  for real and fitted orders and the median of  $I_t$  across different baseline grammars. In most languages,  $I_t$  is distinctly larger for the actual and fitted orderings compared to the baseline orderings. This means that real orderings tend to concentrate more predictive information at the immediately preceding word than baseline grammars.

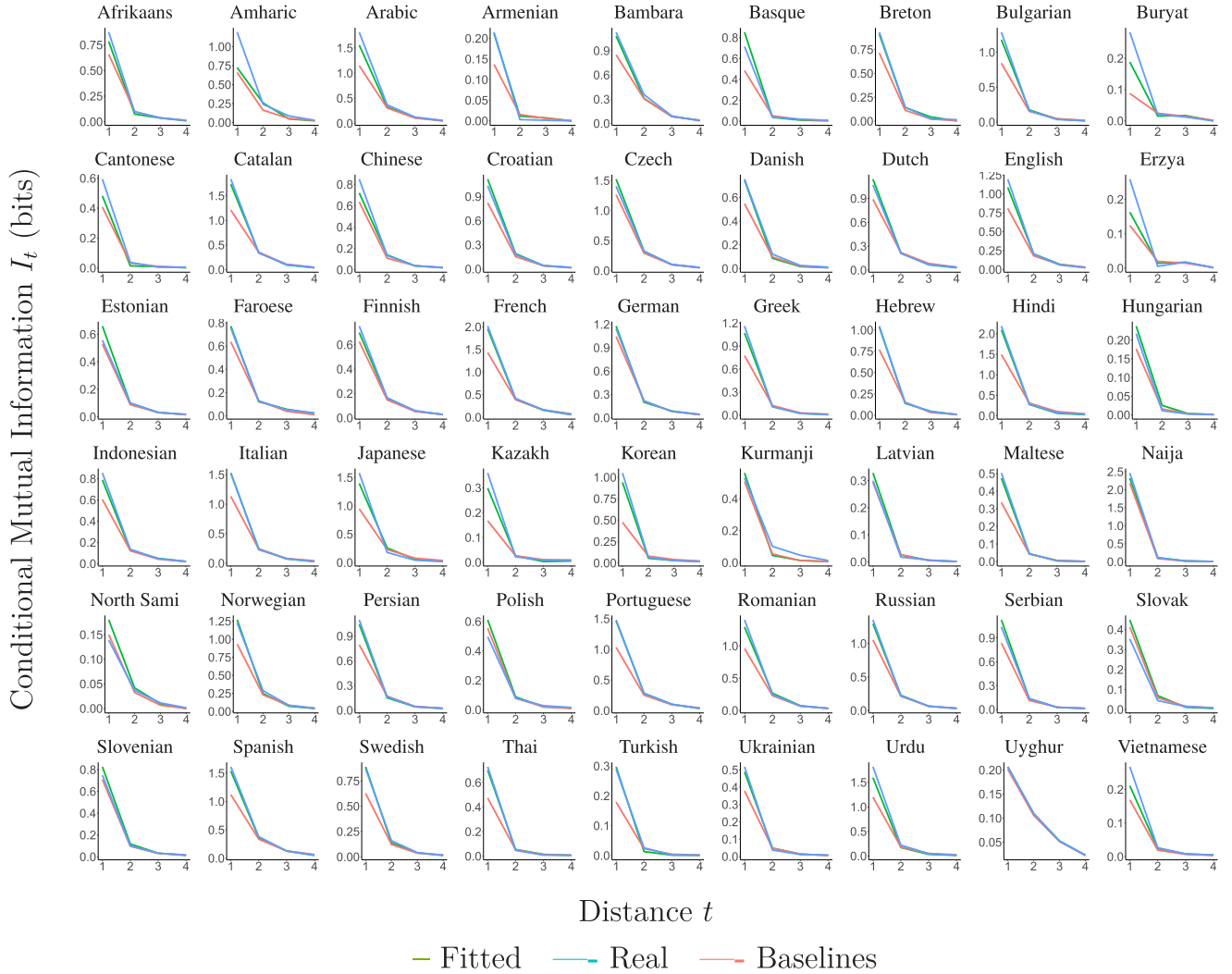
In Figure 11, we show the resulting bounds on the memory–surprisal trade-off curves, showing surprisals at given levels of

memory, for real and baseline languages. We compute surprisal at 40 evenly spaced points of memory (selected individually for each language, between 0 and the maximal memory capacity  $H_M$  obtained using Theorem 1), over real orders and baseline grammars. At each point, we then compute the median surprisal over all model runs for the real language, and over all baselines grammars. For each point, we compute an nonasymptotic and nonparametric 95% confidence interval for this median surprisal using the binomial test.

Numerically, the real language provides a better trade-off than the median of the baselines across all languages, with four exceptions (Latvian, North Sami, Polish, and Slovak). To quantify the degree of optimality of real orders, we further computed the area under the memory–surprisal trade-off curve (AUC) for real and baseline orderings. AUC is a general quantity evaluating the efficiency of a trade-off curve (Bradley, 1997). A *smaller* area indicates a *more efficient* memory–surprisal trade-off. In Figure 12, we plot the AUC

**Figure 10**

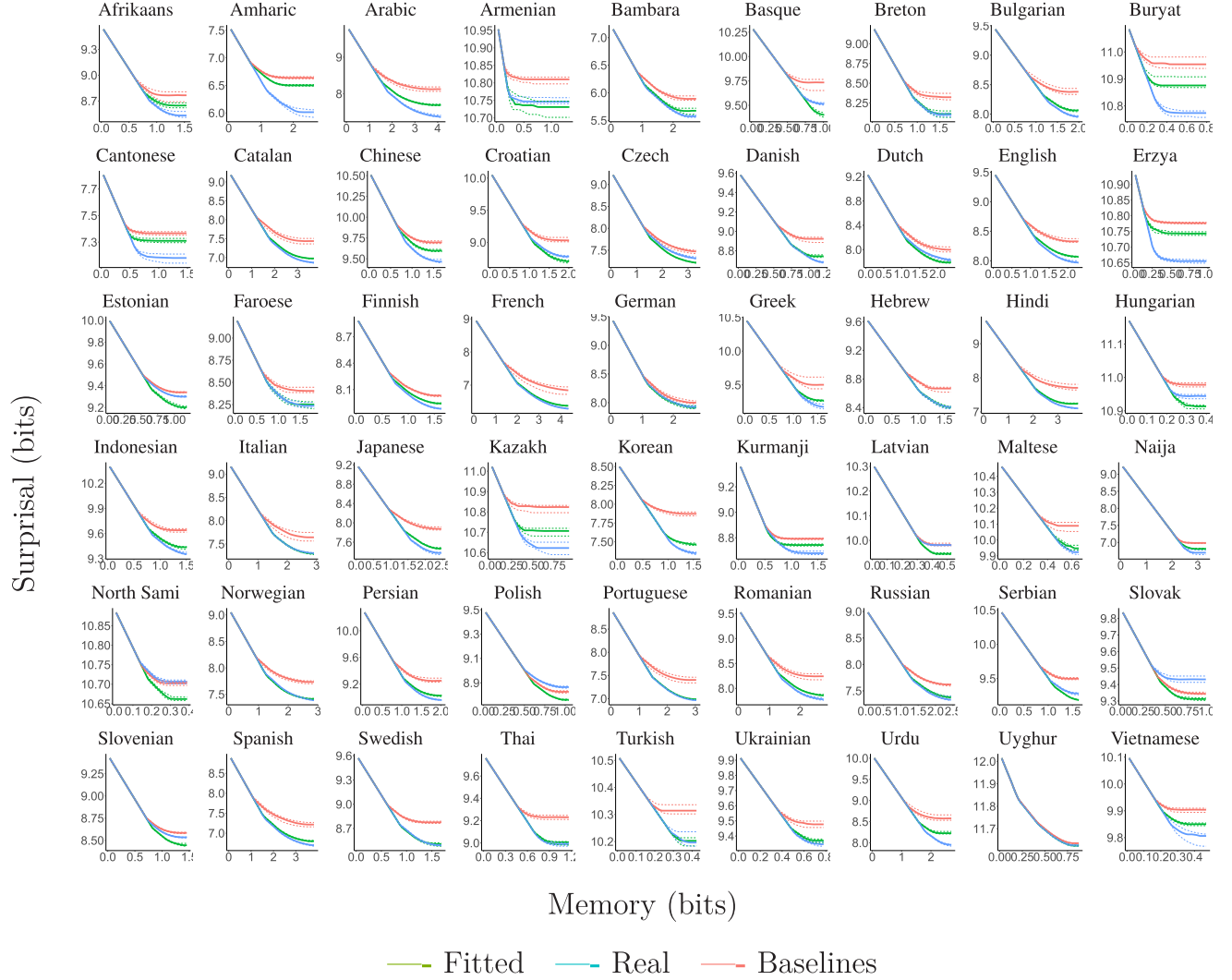
Conditional Mutual Information  $I_t$  (y-Axis) as a Function of  $t$  (x-Axis), for Real, Fitted and Baseline Orders. We Plot the Median Over All Sampled Baseline Grammars



*Note.* See the online article for the color version of this figure.

**Figure 11**

*Surprisal (y-Axis) at Given Memory Level (x-Axis), for Real, Fitted, and Baseline Orders. for the Real and Fitted Orders, We Provide the Median Across Multiple Random Seeds of the Neural Network Estimator for  $I_t$  (see SI Section 3.2.2), and 95% Confidence Bands. for the Baseline Grammars, We Provide the Median Across Both the Sampled Baseline Grammars and Multiple Random Seeds of the Estimator, and 95% Confidence Bands for This Median*



*Note.* See the online article for the color version of this figure.

for the real orderings, together with the distribution of AUCs for baseline grammars. We quantify the degree of optimality by the fraction of baseline grammars for which the AUC is higher than for the real orders: The real ordering is highly efficient if it results in a lower AUC than almost all baseline grammars. Numerically, the AUC is smaller in the real orderings than in at least 50% of baseline grammars in all but three languages (Polish, Slovak, and North Sami). We evaluated significance using a two-sided binomial test. In these three languages, the AUC is higher in the real orderings than in significantly less than 50% ( $p < .01$  in each language). In all other languages except for Latvian, the fraction of more efficient baseline grammars was significantly less than 50%, at  $p = .01$ , where we applied Hochberg's step-up procedure (Hochberg, 1988) to control

for multiple comparisons. In 42 of the 54 languages, the real language was more efficient than all of the sampled baseline grammars.

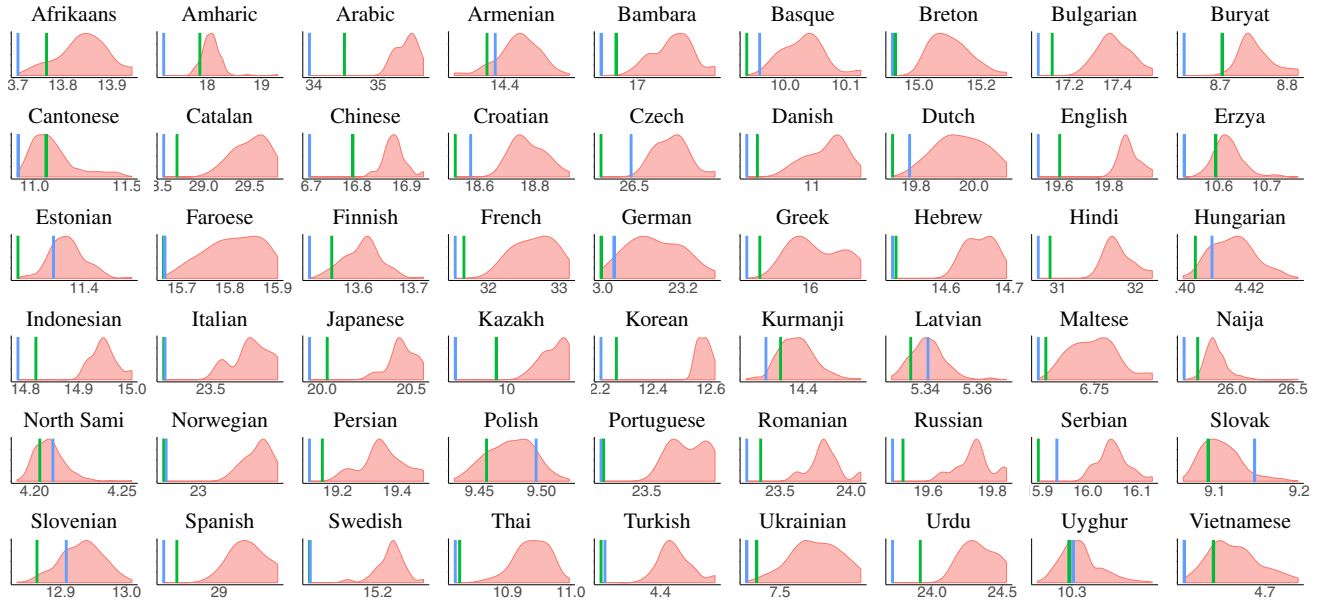
The AUC for the fitted grammars is lower than more than 50% of random baseline grammars in all 54 languages ( $p < .01$ , using two-sided Binomial test and Hochberg's step-up procedure). Thus, we replicate the result that ordering regularities of real languages provide more efficient trade-offs than most possible order grammars even when comparing within the same word order grammar formalism.

## Discussion

We have found that 50 out of 54 languages provide better memory-surprisal trade-offs than random baselines with consistent

**Figure 12**

*Histograms for the Area Under the Curve (AUC) for the Memory–Surprisal Trade-Offs for Real, Fitted, and Random Orders. We Provide a Kernel Density Smoothing Estimate of the Distribution of Random Baseline Orders. A Smaller AUC Value Indicates a More Efficient Trade-Off. In Most Cases, the Real and Fitted Orders Provide More Efficient Tradeoffs Than Most or All Baseline Grammars*



*Note.* See the online article for the color version of this figure.

but counterfactual word order rules. Numerically, we observed differences in memory and surprisal between real and baseline orders in the range of up to a few bits, often less than a bit (Figure 11). Although one bit of memory seems like a small difference, this is a difference in cost at *every word*, which accumulates over a sentence. In a sentence with 20 words, the overall number of bits that have to be encoded over time (though not simultaneously) additionally might add up to 20 bits.

Four languages provide exceptions; these are Latvian (Baltic), North Sami (Uralic), Polish, and Slovak (both Slavic). These four languages did not have significantly lower AUC values than half of the random baselines. One feature that unites these four languages is that they have strong word order freedom, as we will see below in Figure 13. Word order freedom plausibly makes sentences less predictable, as the same syntactic structure can receive different surface realizations. We thus hypothesized that word order freedom impacts the memory–surprisal trade-off, and that languages with more strongly fixed word order should display more optimal memory–surprisal trade-offs.

To test this hypothesis, we examined the correlation between word order freedom and the surprisal difference between real and baseline orderings. To quantify word order freedom, we used a corpus-based estimate, the *branching direction entropy* (Futrell et al., 2015b). This is the entropy of the ordering (head-first or dependent-first) of dependencies conditioned on the dependency label and the part-of-speech label of head and dependent. These two quantities are plotted in Figure 13. We found that branching direction entropy was strongly correlated with the surprisal difference between real and baseline orderings (Spearman correlation  $-.58, p < .0001$ ).

This result might mean that optimization of word orders for memory–surprisal trade-offs is indeed stronger in languages with more fixed word order, and that word order freedom leads to less efficient memory–surprisal trade-offs. A second possibility is that languages with seemingly free word order encode other information in word order, in particular, information about information structure (e.g., Firbas, 1966, 1974; Givón, 1988; Myhill, 1985). Next, we test the latter hypothesis by examining whether the degree of optimization changes when taking into account information structure.

### Controlling for Information Structure

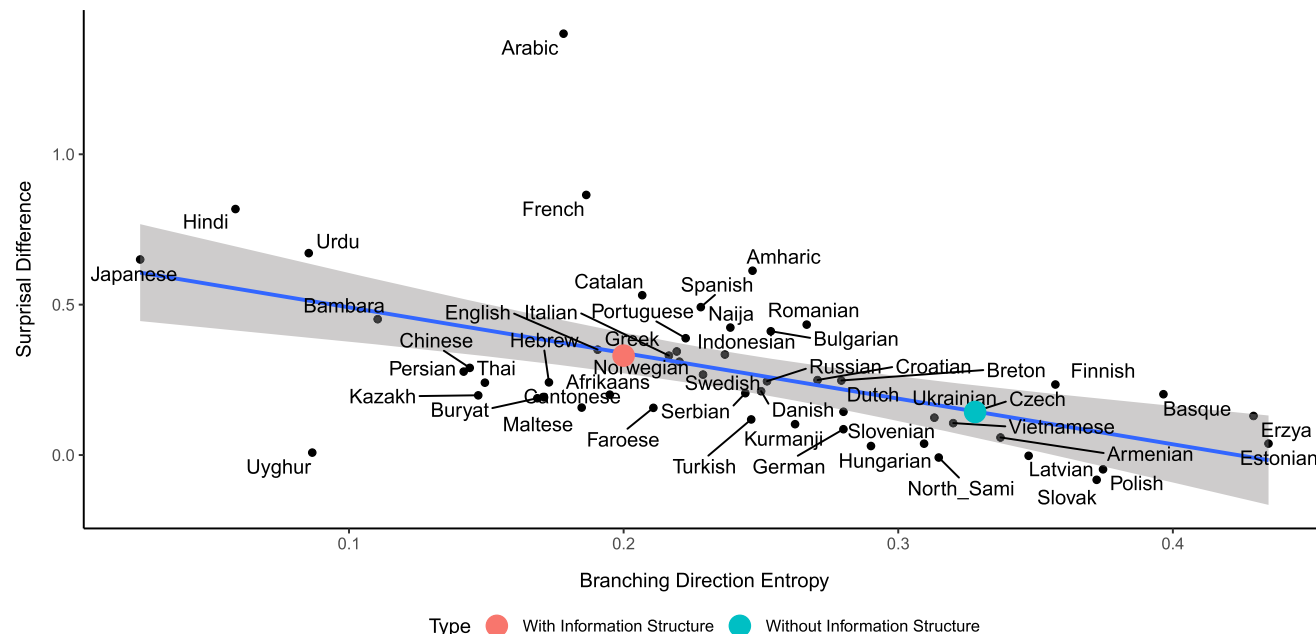
In this section, we address the question about word order freedom raised in the last section. We draw on a corpus of Czech with information structure annotation to determine whether real orders are more optimized when comparing to baselines taking information structure into account.

Languages with flexible word order often show a strong influence of information structure on word order (Givón, 1988; Jacobs, 1988; Neeleman & van de Koot, 2016). Due to the difficulty of annotating information structure, only relatively few data sets have annotations for information structure, and even fewer data sets have both syntactic and information structure annotation. We draw on the Prague Dependency Treebank of Czech (Böhmová et al., 2003; Mikulová et al., 2006), which has both types of annotation. Czech is a language with relatively high degree of word order freedom, which is generally thought to be strongly impacted by information structure (Firbas, 1966, 1974).

About one third of the Prague Dependency Treebank has annotation for topic-focus articulation (Mikulová et al., 2006). Constituents are annotated for contrastiveness and for contextual

**Figure 13**

*Word Order Freedom and Strength of Optimization: for Each of the 54 Languages, We Show Word Order Freedom as Measured by Branching Entropy, and the Difference Between the Real Order's Surprisal and the Median Surprisal of the Baseline Grammars, at the Maximum Memory Value (See Figure 11). Languages With Higher Branching Direction Entropy Show a Smaller Reduction in Surprisal Compared to Baseline Orders. for Czech, We Also Provide an Estimate Accounting for Information Structure (Red Dot), See Below, "Controlling for Information Structure," for More Information*



*Note.* See the online article for the color version of this figure.

boundedness, that is, givenness. Contextually bound expressions are presumed as given in context so that their referent is uniquely determined by the context; contextually bound expressions are contrastive if they choose from a contextually given set of alternatives (Mikulová et al., 2006, Section 10.2). Three labels are used as follows: “c” for contrastive and contextually bound, “f” for contextually nonbound, and “t” for noncontrastive contextually bound. These labels were diagnosed based on constituent order and intonation. Some constituents remain unmarked, the vast majority of which are function words such as adpositions, conjunctions, and auxiliaries; we introduce a label “NA” for these. To define baselines, we extend the word order grammar formalism by defining separate weights for each combination of the 37 syntactic relations and these four information structure labels.

We obtained 38,727 training sentences and 5,228 held-out sentences. We created 20 baseline grammars with information structure, 20 baseline grammars without it.

## Results

We show estimated trade-offs and the distributions over AUC values in Figure 14. As this experiment was conducted only on the subset of the Prague Dependency Treebank that has information structure annotation, the numerical values are slightly different from those in Figure 11. We compare the real orders both with the same baselines as above, and with the baselines taking information structure into account. Baselines show a larger gap in efficiency between real and baseline grammars when the baselines condition

word order on information structure. This suggests that, among word orders that encode information structure, the real order of Czech provides a very efficient memory–surprisal trade-off, and that the strength of optimization is underestimated when comparing against baselines that do not take information structure into account.

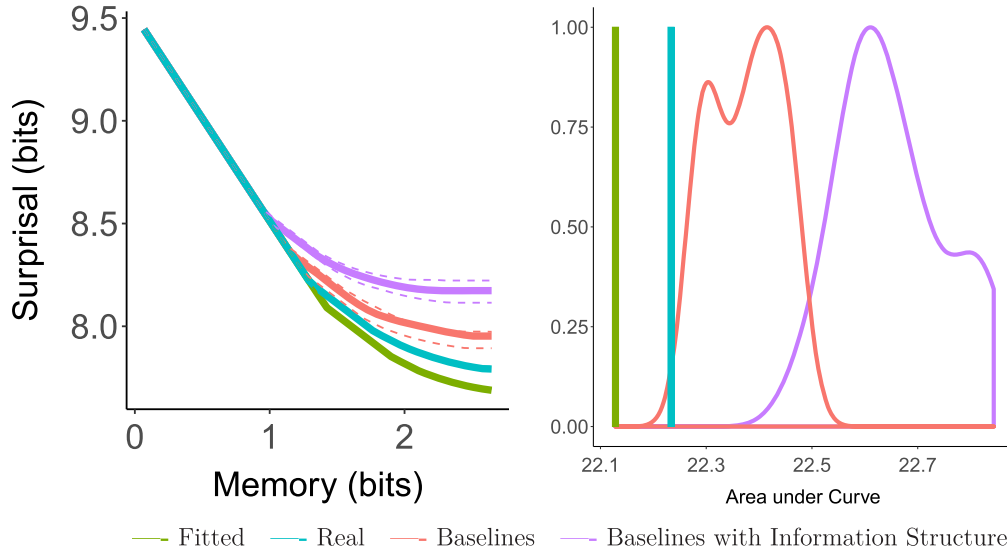
In Figure 13, we show how the data point for Czech changes when including information structure in the word order modeling. When modeling information structure, branching direction entropy decreases, while the surprisal difference between real and baseline orders increases. This suggests that the weaker optimization in free word order languages observed in Study 2 might in part be because ordering grammars do not take information structure into account. In general, we expect that conditioning word order on more sources of information will increase the set of possible word orders, and thus decrease predictability and increase surprisal. As more corpora become available, it will be important to reproduce this finding on data from further languages. If this finding replicates, then this would mean that the impact of order freedom on the strength of optimization observed in Study 2 is an artifact of the fact that languages differ in the degree to which their word order encodes information structure, and that similar degrees of optimization might actually hold across such different languages.

## Discussion

Using data from Czech, we found that the difference between the memory–surprisal trade-offs of real and baseline orders increases if we choose baseline orderings that encode information structure, as real orders do. We hypothesize that this in part explains why the

**Figure 14**

*Left: Memory-Surprisal Trade-Off for Czech With Information Structure. Right: AUC for Czech, for Baselines Without Information Structure and Baselines With Information Structure. Optimization of Real Orders Is Stronger When Considering Information Structure in Baselines*



*Note.* See the online article for the color version of this figure.

observed strength of memory efficiency optimization is negatively correlated with the degree of word order freedom: Languages with flexible word order typically encode information structure in word order, which increases average surprisal. This does not mean that conditioning word order on information structure makes language less efficient in general. Rather, encoding information structure in word order may increase the information content transmitted to the listener, which may in turn balance an increase in surprisal processing effort (Hahn et al., 2020). Due to the difficulty and cost of annotating information structure, we could only evaluate this hypothesis on data from one language. As more annotated data becomes available, this should be replicated on data from further languages.

### Interim Summary

In this section, we tested the Efficient Trade-off Hypothesis on dependency corpora from 54 languages, comparing observed word orders to hypothetical baseline grammars. We found that, in 50 out of 54 languages, real orders provide more efficient memory–surprisal trade-offs than most baseline grammars. This result also held when comparing real and baseline orderings within a single grammar formalism. These results suggest that, across languages, word order favors information locality more strongly than most possible alternative orders.

We also found that the degree of optimization was weaker in languages with high degrees of word order freedom. Using data from Czech that is annotated for both syntax and information structure, we provided evidence that this dependence on word order freedom is an artifact of the fact that languages with flexible word order tend to encode information structure in word order.

Taken together, Studies 1 and 2 suggest that crosslinguistic word orders are in part impacted by a pressure towards efficient memory–surprisal trade-offs, and thus, information locality. To test whether

the Efficient Trade-off Hypothesis holds at different levels of representation, we consider morpheme order in Study 3.

### Study 3: Morpheme Order

The Efficient Trade-off Hypothesis should apply not just at the level of words, but at the level of any linguistic element. For instance, just as observed word orders exhibit information locality, the order of morphemes within words should also be structured so that morphemes which predict each other are close to each other. Here, we apply the Efficient Trade-off Hypothesis to predict the order of morphemes within morphologically complex words in two agglutinative languages. We study two agglutinative languages for which extensive corpora with hand-annotated morphological segmentation and labeling are available: Japanese and Sesotho. We compare the memory–surprisal trade-off of the actual morpheme orders in these languages with hypothetical baseline orderings. Furthermore, we construct hypothetical orderings that are optimized for the efficiency of the memory–surprisal trade-off, and compare these to the actual morpheme orderings, to investigate whether morpheme order in these languages can be predicted by optimization of trade-off efficiency. Below, we first give brief sketches of the morphological patterns in these languages.

#### Verb Suffixes in Japanese

In Japanese, verbs are marked with an extensive number of suffixes. For example, the following verb forms are marked with multiple suffixes:

(1) a. mi rare mash yoo

see PASSIVE POLITENESS FUTURE

“will be seen.”



b. mi taku nakat ta

see DESIDERATIVE NEGATION PAST

“did not wish to see”

Based on corpus data and the linguistic literature on Japanese, we identified the following frequent verb suffixes, occurring in the following order outwards from the verb root (see SI for details).

1. *suru*: obligatory suffix after Sino-Japanese words when they are used as verbs
2. Valence: causative [-*ase*-; Hasegawa, (2014, p. 142), Kaiser et al., (2013, Chapter 13)]
3. Voice and Mood: passive [-*are*-, -*rare*-; Hasegawa (2014, p. 152), Kaiser et al. (2013, Chapter 12)] and potential (-*e*-, -*are*-, -*rare*-; Kaiser et al., 2013, p. 398)
4. Politeness (-*mas*-; Kaiser et al., 2013, p. 190).
5. Mood: desiderative (-*ta*-; Kaiser et al., 2013, p. 238)
6. Negation (-*n*-)
7. Tense, Aspect, Mood, and Finiteness: past (-*ta*), future/hortative (-*yoo*) (Kaiser et al., 2013, p. 229), nonfiniteness (-*te*) (Kaiser et al., 2013, p. 186)

## Verb Affixes in Sesotho

Sesotho (also known as Southern Sotho) is a Southern Bantu language spoken primarily in Lesotho and South Africa. Sesotho verbs are marked with both prefixes and suffixes (Demuth, 1992). Common prefixes include markers for agreement with subjects and objects; object prefixes always follow subject prefixes (2-a). Common suffixes include markers changing valence and voice, and a mood suffix (2-b).

(2) a. oa di rek a

SUBJECT.AGREEMENT      OBJECT.AGREEMENT      buy  
INDICATIVE

“(he) is buying (it)” (Demuth, 1992)

b. o pheh el a

SUBJECT.AGREEMENT cook APPLICATIVE INDICATIVE

“(he) cooks (food) for (him)” (Demuth, 1992)

We identified affix morphemes and their ordering based on the analysis in the study by Demuth, 1992, supplemented with information from grammars of Sesotho (Doke & Mofokeng, 1967; Guma, 1971). See SI for details. We identified the following prefixes:

1. Subject agreement: This morpheme encodes agreement with the subject, for person, number, and noun class (the latter only in the 3rd person) Doke & Mofokeng, 1967, 395. The annotation provided by Demuth, 1992 distinguishes between ordinary subject agreement prefixes and agreement prefixes used in relative clauses; we distinguish these morpheme types here.
2. Negation (Doke & Mofokeng, 1967, §429)

3. Tense/aspect marker (Doke & Mofokeng, 1967, §400–424)
4. Object agreement or reflexive marker (Doke & Mofokeng, 1967, §459). Similar to subject agreement, object agreement denotes person, number, and noun class features of the object.

We identified the following suffixes:

1. Semantic derivation: reversive (e.g., “do” → “undo”; Doke & Mofokeng, 1967, §345)
2. Valence: Common valence-altering suffixes include causative, neuter/stative, applicative, and reciprocal (Doke & Mofokeng, 1967, §307–338). See SI for details on their meanings.
3. Voice: passive (Doke & Mofokeng, 1967, §300)
4. Tense (Doke & Mofokeng, 1967, §369)
5. Mood (Doke & Mofokeng, 1967, §386–445)
6. Interrogative and relative markers, appended to verbs in certain interrogative and relative clauses (Doke & Mofokeng, 1967, §160, 271, 320, 714, 793).

## Experiment

### Data Selection and Processing

For Japanese, we drew on Universal Dependencies data. In the tokenization scheme used for Japanese, most affixes are separated as individual tokens, effectively providing morpheme segmentations. We used the GSD corpus, Version 2.4, (Asahara et al., 2018; Tanaka et al., 2016), as it was the only corpus with a training set and freely available word forms. In the corpus, verb suffixes largely correspond to auxiliaries (with tag AUX); only a few morphemes tagged AUX are not standardly treated as suffixes (see SI), and one frequent suffix (-*te*) is labeled SCONJ. We selected verb forms by selecting all chains of a verb (tag VERB) followed by any number of auxiliaries (tag AUX) from the training set of the corpus. When the suffix -*te* (tag SCONJ) followed such a chain, we added this. We labeled suffixes for underlying morphemes with the help of the lemmatization provided for each suffix in the corpus (see SI Section 4.3 for details). The passive and potential (Slot 3) markers are formally indistinguishable for many verbs. As we cannot systematically distinguish them on the basis of the available corpus annotation, we merge these into a single underlying morpheme “Passive/Potential.”

We obtained 15,281 verb forms in the training set and 1,048 verb forms in the held-out set. Of the forms in the training set, 27% had two or more suffixes (modal group: two suffixes, accounting for 20% of forms; maximum seven suffixes). Although predicting order naturally focuses on datapoints with more than one suffix, we include the other datapoints for estimating conditional mutual information  $I_t$ .

For Sesotho, we used the Demuth Corpus (Demuth, 1992) of child and child-directed speech, containing about 13 K utterances with 500 K morphemes. The corpus has very extensive manual morphological segmentation and annotation; each verb form is segmented into morphemes, which are annotated for their function. Sesotho verbs carry both prefixes and suffixes. We extracted 37 K

verb forms (see SI 4.2 for details). We randomly selected 5% to serve as held-out data and used the remaining 95% as training data. Of note, 93% of forms had two or more affixes (modal group: three affixes, accounting for 36% of forms; maximum eight affixes).

### Estimating Memory–Surprisal Trade-Off

We modeled incremental prediction on the level of morpheme sequences. To do so, we represented each verb form as a sequence of a stem and affix morphemes, abstracting away from morphophonemic interactions between neighboring morphemes. As in many languages, affixes in Japanese and Sesotho show morphophonemic interactions between neighboring morphemes; for instance, the Japanese politeness morpheme *-mas-* takes the form *-masu* when it is word-final, while it has the allomorph *-mase-* when followed by the negation suffix *-n*. Modeling prediction on the level of morphemes, as opposed to phonemes, controls for these interactions.<sup>10</sup>

In analogy to Studies 1–2, we modeled verb forms as a stationary stochastic process by concatenating the verb forms from the corpus in random order.

We calculated  $I_t$  by estimating an  $n$ -gram model on the training set and then computing the average surprisal  $S_t$  as cross-entropy on the held-out set using Kneser–Ney smoothing. The model may overfit as the context size  $t$  increases, leading to higher cross-entropies for larger values of  $t$ . We mitigated overfitting for large  $t$  by estimating

$$\hat{S}_t := \min_{s \leq t} S_s, \quad (7)$$

where  $S_s$  is the cross-entropy of the  $S$ 'th order Markov model on held-out data. This procedure ensures that  $\hat{S}_t$  can only decrease as the context size  $t$  increases.

### Parameterizing Alternative Orderings

We parameterized alternative affix orderings by assigning a weight in  $[0, 1]$  to each morpheme. Given such a grammar, affixes are ordered by the values assigned to their underlying morphemes. We considered all morphemes annotated in the corpora, including low-frequency ones going beyond the ones identified above (see SI for details).

To verify that this formalism is appropriate for capturing morpheme order in Japanese and Sesotho, we fitted models parameterized in this way to the observed orders. Ordering morphemes according to these fitted models recovered the observed order for almost all forms (98.6 % for Japanese, 99.93% for Sesotho prefixes, and 97.4% for Sesotho suffixes). Exceptions largely concern low-frequency affixes beyond those considered here. We take this as confirmation that the formalism is generally suited to capture morpheme order.

For each language, we constructed 40 baseline grammars by randomly sampling weights.

### Creating Optimal Orders

To create optimal orders to compare real orders to, we optimized orderings for the AUC under the memory–surprisal trade-off curve with an adaptation of the hill climbing method that Gildea and Jaeger (2015) used to optimize word order grammars for the length of syntactic dependencies and trigram surprisal.

We randomly initialized the assignment of weights to morphemes, and then iteratively change the assignment to reduce AUC. In each iteration, we randomly chose one morpheme, and evaluate AUC for each way of ordering it between two other morphemes. We then updated the weights to the ordering that yields the lowest AUC. To speed up optimization, we restricted to morphemes occurring at least 10 times in the corpus for 95% of iterations, and to 10% of possible orderings in each step. These choices vastly reduced computation time by reducing time spent on low-frequency morphemes. This optimization method is approximate, as it only guarantees convergence to a local optimum (Gildea & Jaeger, 2015), not to a globally optimal assignment.

We ran this method for 1,000 iterations. Empirically, AUC converged after a few hundred iterations. To control for the randomness in initialization and the optimization steps, we ran this algorithm 10 times. Different runs achieved almost the same AUC values (SD 0.0051 in Japanese, 0.0036 in Sesotho). For Sesotho, we ran the algorithm separately for prefixes and suffixes due to computational efficiency considerations.<sup>11</sup>

## Results

In Figure 15, we compare the area under curve of the memory–surprisal trade-off for Japanese and Sesotho verb forms under different orderings. Both observed orders and the approximately optimized grammars show lower AUCs than all 40 random baseline samples, in both languages. For comparison, we also show AUC for the order resulting from *reversing* all suffix chains in the observed orders; this results in high AUC even exceeding most random grammars. These results show that Japanese and Sesotho affix orderings enable approximately optimal memory–surprisal trade-offs.

We now ask to what extent the observed morpheme ordering is predicted correctly by approximately optimized grammars. In Table 1, we give summary statistics about the accuracy of optimized grammars in predicting affix order in the corpus, together with random baseline figures. We evaluate accuracy using two methods: In one method (“Pairs”), we consider, for each verb form in the corpus, all pairs of prefixes (or suffixes). We report the proportion of these pairs in the corpus for which the relative order of the two affixes is as predicted by the grammar. In the other method, (“Full”), we report the proportion of verb forms in the corpus that has exactly the affix ordering predicted by the grammar. In both measures, we average over all 10 approximately optimized grammars for each language.

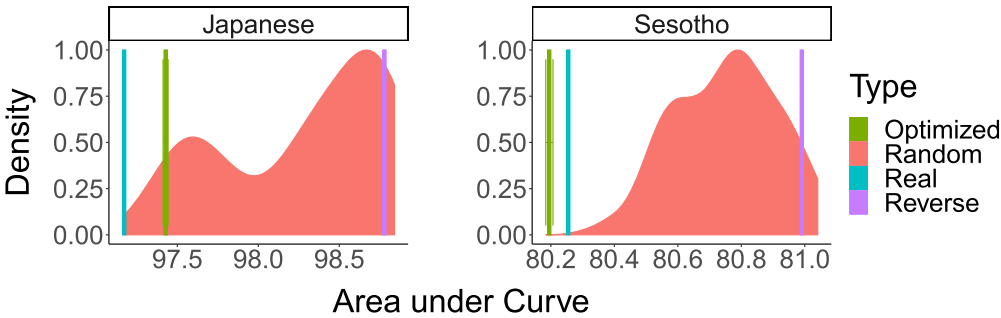
### Japanese Results

In Japanese, by both measures, optimized grammars recover the observed orders with high accuracy. We compare the real grammar with the approximately optimized grammar that achieved the lowest AUC value in Table 2. We conducted an error analysis comparing the real Japanese morpheme order against our approximately optimized orders. We extracted the pairs of morphemes whose relative

<sup>10</sup> See SI Section 4.3 for qualitatively similar results when modeling prediction at the phoneme level.

<sup>11</sup> With the exception of the tense/aspect markers, none of the morpheme types discussed above can occur both as prefixes and suffixes. Therefore, we do not expect this separation to impact results.

**Figure 15**  
*Areas Under the Curve (AUCs) for the Memory–Surprisal Trade-Off for Verb Affixes in Japanese (Left) and Sesotho (Right). for the Baseline Grammars, We Show a Kernel Density Estimate. In Both Japanese and Sesotho, Real and Optimized Orderings Lead to Lower AUC Than All of the 40 Baseline Samples*



*Note.* See the online article for the color version of this figure.

order is incorrectly predicted, excluding pairs involving low-frequency morphemes not discussed here. Results are shown in Table 3. The most frequent divergence for this grammar is that politeness and negation suffixes are consistently ordered incorrectly; this affects 74 corpus examples (out of 15 K total examples).

We also found that prediction was more accurate when modeling on the level of phonemes, suggesting that divergence between model predictions and actual order might be related to phonological pressure (see SI Section 4.3).

### Sesotho Results

We compare the real Sesotho grammar with the approximately optimized grammar that achieved the lowest AUC value in Table 4. In Sesotho, for prefixes, all optimized grammars almost exactly recover the ordering described above. The only divergence among the high-frequency morphemes is that negation and the tense/aspect prefix are ordered incorrectly; this accounts for only 12 occurrences in the data set, as the two prefixes rarely co-occur (Table 5, top).

For Sesotho suffixes, order is recovered at above-chance accuracies (Table 1, bottom), though with some divergences. The most common error (Table 5, bottom) is that relative and interrogative suffixes are consistently placed closer to the verb stem than the mood suffix. We conjecture that this happens

because all Sesotho verbs uniformly have a mood suffix, suggesting that there might be lower mutual information between the stem and the mood suffix than between the stem and these two suffixes. Furthermore, valence-changing suffixes are ordered farther away from the stem than various other suffixes, in contrast with the actual orders. Interestingly, we found that prediction was more accurate in this respect when estimating  $I_t$  naively on the training set (see SI Section 4.3), suggesting that the available corpus data does not sufficiently determine the optimal ordering.

### Discussion

We have found that the ordering of verb affixes in Japanese and Sesotho provides approximately optimal memory–surprisal trade-offs, close to the efficiency of orderings computationally optimized for efficiency. We further found that parts of these languages’ ordering rules can be derived from optimizing order for efficient trade-offs.

Here, we argue that the memory–surprisal trade-off provides an explanation of previously existing typological generalizations, and an operationalization of previous functionally motivated explanations for them; in particular, we argue that the notion of mutual information operationalizes the concept of “relevance.”

**Table 1**  
*Accuracy of Approximately Optimized Orderings, and of Random Baseline Orderings, in Predicting Verb Affix Order in Japanese and Sesotho. “Pairs” Denotes the Rate of Pairs of Morphemes That Are Ordered Correctly, and “Full” Denotes the Rate of Verb Forms Where Order Is Predicted Entirely Correctly. We Show Mean Values and Standard Deviations Over 10 Different Runs of the Optimization Algorithm (“Optimized”), and Over Different Random Orderings (“Random”)*

		Prefixes		Suffixes	
		Pairs	Full	Pairs	Full
Japanese	Optimized	—	—	0.953 ( <i>SD</i> 0.011)	0.943 ( <i>SD</i> 0.014)
	Baseline	—	—	0.497 ( <i>SD</i> 0.287)	0.425 ( <i>SD</i> 0.29)
Sesotho	Optimized	0.988 ( <i>SD</i> 0.0)	0.989 ( <i>SD</i> 0.0)	0.756 ( <i>SD</i> 0.014)	0.676 ( <i>SD</i> 0.017)
	Baseline	0.672 ( <i>SD</i> 0.305)	0.604 ( <i>SD</i> 0.338)	0.423 ( <i>SD</i> 0.204)	0.332 ( <i>SD</i> 0.211)

**Table 2**

*Comparing Order of Japanese Affixes in the Observed Orders (Left) and According to an Approximately Optimized Grammar (Right). We Organize the Affixes in the Real Order Into the Seven Slots Described Above*

	Real	Optimized
	Stem	Stem
1	suru	suru
2	causative	causative
3	passive/potential	passive/potential
4	desiderative	negation
5	politeness	future
6	negation	politeness
7	future	desiderative
	past	nonfinite
	nonfinite	past

One prominent typological generalization due to Bybee (1985, p. 24, 34–35) claims that there exists a universal ordering of verbal inflectional morphemes across languages:

verb stem    valence voice    aspect tense    mood subject agreement

Morphemes are claimed either to go in the order above (suffixes), or its reverse (prefixes). This hierarchy makes no statements as to which affixes are realized as prefixes or suffixes.

Japanese and Sesotho verb affixes are broadly in agreement with Bybee’s generalization. For instance, valence and voice suffixes are closer to the stem than tense/aspect/mood markers. Subject agreement in Sesotho is farther away from the verb than tense/aspect/mood prefixes. This ordering is reproduced closely by optimization in Japanese and for Sesotho prefixes, and to some extent also for Sesotho suffixes.

Bybee (1985, p. 37) argues further that morpheme order is determined by the degree of *relevance* between the affix and the stem, that is, the degree to which “the semantic content of the first [element] directly affects or modifies the semantic content of the second” (p. 13). She argues that elements whose meanings are more relevant to each other appear closer together. For instance, the meaning of a verb is impacted more strongly by a causative affix than by a tense affix: Combining a verb with a causative marker results in a form that denotes a different action, whereas a tense affix only locates the action in time.

**Table 3**

*Errors in Japanese: We Show Pairs of Morphemes That Are Ordered Incorrectly by the Approximately Optimized Grammar With the Lowest AUC Value. We Indicate the Number of Such Pairs Occurring in the Corpus. We Only Show Errors Where Both Morphemes Are Among the High-Frequency Ones Studied Here (AUC = Area Under the Curve)*

Error	Frequency
politeness    negation	74
desiderative    negation	14
politeness    future	9

**Table 4**

*Comparing Order of Sesotho Affixes in the Observed Orders (Left) and According to an Approximately Optimized Grammar (Right). Note That Order Was Separately Optimized for Prefixes and Suffixes*

	Real	Optimized
1	Subject	Subject
	Subject (rel.)	Subject (rel.)
2	Negation	Tense/aspect
3	Tense/aspect	Negation
4	Object	Object
	Stem	Stem
1	Reversive	Passive
2	Causative	Reciprocal
	Neuter	Tense/aspect
	Applicative	Neuter
	Reciprocal	Relative
3	Passive	Causative
4	Tense/aspect	Applicative
5	Mood	Interrogative
6	Interrogative	Reversive
	Relative	Mood

We conjecture that this notion of relevance is related to mutual information. If an affix has a stronger impact on the meaning of the verb, it will typically not be applicable to all verbs. For instance, causative markers will only attach to verbs whose semantics is compatible with causation. In contrast, a past tense marker can attach to all verbs that are compatible with actions that can have occurred in the past. Therefore, we expect that affixes that are more relevant to a verb stem will also tend to have higher mutual information with the verb stem. If they have higher mutual information with the verb stem, then the principle of information locality predicts that they will go close to the verb stem.

## General Discussion

We introduced a notion of memory efficiency in language processing: the memory–surprisal trade-off. We then tested the resulting Efficient Trade-off Hypothesis: Order of elements in

**Table 5**

*Errors in Sesotho Prefixes (Top) and Suffixes (Bottom). We Show the 10 Most Common Errors Where Both Morphemes Are Among the High-Frequency Ones Studied Here*

Error	Frequency
Negation    Tense/aspect	12
Mood    Interrogative	2204
Mood    Relative	858
Applicative    Tense/aspect	347
Causative    Tense/aspect	174
Neuter    Tense/aspect	155
Reversive    Causative	100
Applicative    Passive	81
Causative    Passive	61
Applicative    Relative	49
Causative    Relative	41



natural language is characterized by efficient memory–surprisal trade-offs, compared to other possible orders. In Study 1, we showed that the Efficient Trade-off Hypothesis predicts the known preference for short dependencies. In Study 2, we used corpus data from 54 languages to show that real word orders provide more efficient trade-offs than baseline order grammars. In Study 3, we showed that in two languages (Japanese and Sesotho) the order of verb affixes not only provides approximately optimal trade-offs but also can partly be predicted by optimizing for the efficiency of the memory–surprisal trade-off.

Here, we discuss the limitations of our results and the implications they have more broadly for the fields of psycholinguistics, typology, and information theory.

### Role of Comprehension, Production, and Acquisition

Our results leave open the causal mechanism leading to the observed optimization, in particular, whether optimization is the result of minimizing effort during comprehension, production, or acquisition. One possibility is that optimization reflects an effort on the side of the speaker to produce utterances that are easy to comprehend by listeners, a strategy known as *audience design* (Brennan & Williams, 1995; Clark & Murphy, 1982; Lindblom, 1990). More efficient memory–surprisal trade-offs are useful from the listener’s perspective because they allow for better prediction with lower memory investment than less efficient trade-offs.

Another possibility is that optimization reflects production-internal pressures to minimize effort on the speaker’s part during sentence planning (Bock & Warren, 1985; Fedzechkina & Jaeger, 2020; Ferreira & Dell, 2000; MacDonald, 2013). That is, instead of speakers optimizing for the benefit of listeners, the iterated application of production-internal heuristics that reduce speaker effort may result in more efficient trade-offs (MacDonald, 2013). Although our theory is stated in terms of the efficiency of language processing for a comprehender of language, we can show that an analogous memory–surprisal trade-off exists in language production, and that speakers with bounded memory capacity can minimize production errors when the language has stronger information locality. For discussion including mathematical proofs, see SI Section 1.4. Depending on the precise formalization of the production problem, the production-oriented version of the memory–surprisal trade-off may or may not be identical the comprehension-oriented version we have presented here. We leave the proper formulation of an information-theoretic model of production to future work.

Finally, optimization may reflect biases that come into play during language learning. It is possible that memory efficiency makes languages more learnable, as learning should require less memory resources for languages with more efficient memory–surprisal trade-offs. Evidence from artificial language learning experiments suggests that language acquisition is biased toward efficiency in communication and processing (e.g., Fedzechkina et al., 2012, 2017).

### Relation to Models of Sentence Processing

There is a substantial literature proposing sentence processing models and quantitative memory metrics for sentence processing. In

this section, we discuss how our theoretical results relate to and generalize these previously proposed models. We do not view our model as competing with or replacing any of these models; instead, our information-theoretic analysis captures aspects that are common to most of these models and shows how they arise from very general modeling assumptions.

In the Information Locality Bound Theorem, we proved a formal relationship between the entropy of memory  $H_M$  and average surprisal  $S_M$ . We made no assumptions about the architecture of incremental memory, and so our result is general across all such architectures. Memory representations do not have to be rational or optimal for our bound in Theorem 1 to hold. There is no physically realizable memory architecture that can violate this bound.

However, psycholinguistic theories may differ on whether the entropy of memory  $H_M$  really is the right measure of memory load, and on whether average surprisal  $S_M$  really is the right predictor of processing difficulty for humans. Therefore, to establish that our information-theoretic processing model generalizes previous theories, we will establish two links:

- Our measure of memory usage generalizes theories that are based on counting numbers of objects stored in incremental memory (e.g., De Santo, 2020; Frazier, 1985; Gerth, 2015; Gibson, 1998; Graf et al., 2015, 2017; Graf & Marcinek, 2014; Kobele et al., 2013; Miller & Chomsky, 1963; Yngve, 1960). Furthermore, for theories where memory is constrained in its capacity for *retrieval* rather than storage (e.g., Lewis & Vasisht, 2005; McElree et al., 2003), the information locality bound will still hold.
- Our predictor of processing difficulty (i.e., average surprisal) reflects at least a *component* of the predicted processing difficulty under other theories.

Below, we discuss the connections between our theory and existing theories of human sentence processing with regard to the points above.

### Storage-Based Theories

There is a long tradition of models of human language processing in which difficulty is attributed to high working memory load. These models go back to Yngve (1960)’s production model, where difficulty was associated with moments when a large number of items have to be kept on a parser stack; this model correctly predicted the difficulty of center-embedded clauses, but problematically predicted that left-branching structures should be hard (Kimball, 1973). Other early examples include Miller and Chomsky (1963) and Frazier (1985)’s measure of syntactic complexity based on counting the number of local nonterminal nodes. More recently, a line of literature has formulated complexity metrics based on how many nodes are kept in incremental memory for how long during parsing, and used linear or ranked combinations of these metrics to predict acceptability differences in complex embeddings (De Santo, 2020; Gerth, 2015; Graf et al., 2015, 2017; Graf & Marcinek, 2014; Kobele et al., 2013; Rambow & Joshi, 1994).

Our measure of memory complexity—that is, the memory entropy  $H_M$ —straightforwardly generalizes measures based on



counting items stored in memory. If each item stored in memory requires  $k$  bits of storage, then storing  $n$  items would require a capacity of  $nk$  bits in terms of memory entropy  $H_M$ . In general, if memory entropy is  $H_M$  and all items stored in memory take  $k$  bits each to store, then we can store  $H_M/k$  items. However, the memory entropy  $H_M$  is more general as a measure of storage cost because it allows that different items stored in memory might take different numbers of bits to store, and also that the memory representation might be able to compress the representations of multiple items when they are stored together, so that the capacity required to store two items might be less than the sum of the capacity required to store each individual item. Previous work has argued that visual working memory is characterized by an information-theoretic capacity limit (Brady et al., 2008; Sims et al., 2012); we extend this idea to incremental memory as used in language processing.

### The Dependency Locality Theory

The connection with the DLT is particularly interesting. Our lower bound on memory usage, described in Theorem 1 Eq. 4, is formally similar to Storage Cost in the DLT (Gibson, 1998, 2000). In that theory, storage cost at a given timestep is defined as the *number of predictions* that are held in memory. Our bound on memory usage is stated in terms of mutual information, which indicates the *amount of predictive information* extracted from the previous context and stored in memory. As the notion of “number of predictions” is subsumed by the notion of “amount of predictive information,” our measure generalizes DLT storage cost.

The other component of the DLT is integration cost, the amount of difficulty incurred by establishing a long-term syntactic dependency. In our framework, DLT integration cost corresponds to surprisal given an imperfect memory representation, following Futrell, Gibson, et al., 2020.

There is one remaining missing link between our theory of processing difficulty and theories such as the Dependency Locality Theory: Our information locality theorem says that *statistical* dependencies should be short term, whereas psycholinguistic theories of locality have typically focused on the time span of *syntactic dependencies*: Words which depend on each other to determine the meaning or the well-formedness of a sentence. Statistical dependencies, in contrast, mean that whenever one element of a sequence determines or predicts another element *in any way*, those two elements should be close to each other in time.

If statistical dependencies, as measured using mutual information, can be identified with syntactic dependencies, then that would mean that information locality is straightforwardly a generalization of dependency locality. Futrell et al., (2019) give theoretical and empirical arguments that this is so. They show that syntactic dependencies as annotated in dependency treebanks identify word pairs with especially high mutual information, and give a derivation showing that this is to be expected according to a formalization of the postulates of dependency grammar. The connection between mutual information and syntactic dependency has also been explored in the literature on grammar induction and unsupervised chunking (Clark & Fijalkow, 2020; de Paiva Alves, 1996; Harris, 1955; McCauley & Christiansen, 2019; Yuret, 1998).

### Cue-Based Retrieval Models

Work within cue-based retrieval frameworks has suggested that working memory is not characterized by a decay in information over time, but rather an accumulation of interference among similar items stored in memory (Lewis & Vasishth, 2005, p. 408). In contrast, the formula for memory usage in Eq. 4 might appear to suggest that boundedness of memory entails that representations have to decay over time. However, this is not the case: Our theorem does not imply that a listener forgets words beyond some amount of time  $T$  in the past. An optimal listener may well decide to remember information about words more distant than  $T$ , but in order to stay within the bounds of memory, she can only do so at the cost of forgetting some information about words closer than  $T$ . The Information Locality Lower Bound still holds, in the sense that the long-term dependency will cause processing difficulty, even if the long-term dependency is not itself forgotten. See SI Section 2.1–2.2 for a mathematical example illustrating this phenomenon.

The ACT-R model of Lewis & Vasishth, 2005 also does not have an explicit surprisal cost. Instead, surprisal effects are interpreted as arising because, in less constraining contexts, the parser is more likely to make decisions that then turn out to be incorrect, leading to additional correcting steps. We view this as an algorithmic-level implementation of a surprisal cost: If a word  $w_i$  is unexpected given the current state of the working memory, then its current state must provide insufficient information to constrain the actual syntactic state of the sentence, meaning that the parsing steps made to integrate  $w_i$  are likely to include more backtracking and correction steps. Thus, we argue that cue-based retrieval models predict that the surprisal  $-\log P(w_i|m_i)$  will be part of the cost of processing word  $w_i$ .

### The Role of Surprisal

There are more general reasons to believe that any realistic theory of sentence processing must include surprisal as at least a *component* of the cost of processing a word, even if it is not explicitly stated as such. There are both empirical and theoretical grounds for this claim. Empirically, surprisal makes a well-documented and robust contribution to processing difficulty in empirical studies of reading times and event-related potentials (Frank, Otten, et al., 2015; Smith & Levy, 2013). Theoretically, surprisal may represent an irreducible thermodynamic cost incurred by any information processing system (Landauer, 1961; Still et al., 2012; Zénou et al., 2019), and there are multiple converging theoretical arguments for why it should hold as a cost in human language processing in particular (see Levy, 2013, for a review).

A few prior models explicitly include both surprisal and memory components (Demberg & Keller, 2009; Rasmussen & Schuler, 2018). The model proposed by Demberg and Keller (2009) assumes that processing cost is composed of surprisal and a verification cost term similar to DLT integration cost. According to this term, processing of a new word costs more effort when the relevant prediction has not been accessed for a longer time, or has low prior probability. Although this model has separate costs for surprisal and for memory access, their overall effect is similar to surprisal conditioned on memory representations generated by an encoding function  $M$  that stores predictions made from prior words and which

decay over time: Processing cost is dominated by surprisal when a word is predicted by information from the recent past, while processing cost is increased when the relevant prediction stored in memory has been affected by memory decay. In the model of Rasmussen and Schuler (2018), memory effects arise from interference in a distributed model of memory, whereas surprisal effects arise from the need to renormalize distributed representations of possible parse trees in proportion to their probability. The explanation of memory effects can be viewed as a specific type of capacity constraint, forcing  $M$  to take values in a fixed-dimensional vector space.

### Previous Information Locality Results

Previous work has attempted to derive the principle of information locality from incremental processing models. Gibson and Levy (2020) describe a processing model where listeners make predictions (and incur surprisal) based on lossy memory representations. In particular, they consider loss models that delete, erase, or replace words in the past. Within this model, they were able to establish a similar information locality result, by showing that the theoretical processing difficulty increases when words with high *pointwise mutual information* are separated by large distances. Pointwise mutual information is the extent to which a *particular value* predicts another value in a joint probability distribution. For example, if we have words  $w_1$  and  $w_2$  in a sentence, their pointwise mutual information is as follows:

$$\text{pmi}(w_1; w_2) \equiv \log \frac{P(w_2|w_1)}{P(w_2)}.$$

Mutual information, as we defined it in Eq. 3, is the *average* pointwise mutual information over an entire probability distribution.

Our information locality bound theorem differs from this previous result in three ways:

1. Futrell, Gibson, and Levy (2020) required an assumption that incremental memory is subject to decay over time. In contrast, we do not require any assumptions about incremental memory except that it has bounded capacity (or that retrieval operations have bounded capacity; see above).
2. Our result is a precise bound, whereas the previous result was an approximation based on neglecting higher-order interactions among words.
3. Our result is about the falloff of the mutual information between words, *conditional on the intervening words*. The previous result was about the falloff of *pointwise* mutual information between specific words, without conditioning on the intervening words.

We would like to emphasize the last point: Previous work defined information locality in terms of the *unconditional* mutual information between linguistic elements. In contrast, we advocate that *conditional* mutual information is more relevant for measuring memory usage than unconditional mutual information. Although the decay of conditional mutual information provably provides a lower bound on memory entropy, the decay of unconditional

mutual information does not. In SI Section 2.3, we provide an example of a stochastic process where unconditional mutual information does not decay with distance, but memory requirements remain low.

### Experience-Based and Connectionist Models

Our model and results are compatible with work arguing that memory strategies adapt to language structure and language statistics, and that experience shapes memory performance in syntactic processing (e.g., MacDonald & Christiansen, 2002; Wells et al., 2009). For instance, MacDonald and Christiansen (2002) argue for a connectionist model in which network structure and language experience account for processing capacity. Such models use recurrent neural networks with some fixed number of neurons, which can be understood as a specific kind of constrained memory. A case in point is the observation that forgetting effects in nested head-final dependencies are reduced or absent in head-final structures (Frank & Ernst, 2019; Frank, Trompenaars, & Vasishth, 2015; Vasishth et al., 2010), which has been modeled using connectionist models (Engelmann & Vasishth, 2009; Frank, Trompenaars & Vasishth, 2015), which can be interpreted as modeling surprisal conditioned on imperfect memory (Futrell, Gibson, et al., 2020).

### Limitations

#### Finiteness of Data

As corpora are finite, estimates for  $I_t$  may not be reliable for larger values of  $t$ . In particular, we expect that models will underestimate  $I_t$  for large  $t$ , as models will not be able to extract and utilize all available information over longer distances. This means that we might not be able to consistently estimate the asymptotic values of the average surprisal  $S_M$  as the memory capacity goes to infinity, that is, the entropy rate  $S_\infty$ . We specifically expect this to happen in languages where less data are available (see SI Section 3.1 for corpus sizes). We expect this bias to be roughly equal in magnitude across real and baseline languages for a given  $t$ , enabling us to compare across these languages at a given  $t$ .

The finiteness of data also has implications for the interpretation of the memory-surprisal trade-offs at higher values of memory entropy  $H_M$ . In Study 2, the lowest achieved surprisals are different for real and baseline orderings. This does not necessarily mean that these orderings really have different entropy rates  $S_\infty$ . It is logically possible that real and baseline languages actually have the same entropy rate  $S_\infty$ , but that baseline orderings spread the same amount of predictive information over a larger distance, making it harder for models to extract given finite corpus data. What our results do imply is that real languages provide lower surprisals in the setting of relatively small memory budgets. This result only depends on the estimates of  $I_t$  for small values of  $t$ , which are most trustworthy. To the extent that  $I_t$  is underestimated even for small values of  $t$ , such a bias equally applies to different ordering grammars. We therefore expect that estimating the relative efficiency of different orderings at the same level of memory is still reliable (see SI Section 3.6 for supporting experiments comparing estimation with different sample sizes).

## Nature of the Bound

Our theoretical result provides a lower bound on the trade-off curve that holds across all ways of physically realizing a memory representation obeying the postulates (1–3). However, this bound may be loose in two ways.

First, architectural properties of human memory might introduce additional constraints on possible representations. Depending on the role played by factors other than information-theoretic capacity, the trade-offs achieved by these human memory representations need not be close to achieving the theoretical bounds.

Second, depending on properties of the stochastic process, the bound might be loose across all models, that is, there are processes where the bound is not attainable by any memory architecture. This can happen if there is strong uncertainty as to which aspects of the past observations will be relevant to the future. We provide an artificial example with analytical calculations in SI Section 2.1, but this example does not seem linguistically natural.

## Extralinguistic Context

Comprehension Postulate 1 states that the memory state after receiving a word is determined by that word and the memory state before receiving this word. The assumption about information flow disregards the role of information sources that are external to the linguistic material in the sentence. For instance, the interlocutors might have common knowledge of the weather, and the listener might use this to construct predictions for the speaker's utterances, even if no relevant information has been mentioned in the prior discourse. Such sources of information are disregarded in our model. They are also disregarded in many other models of memory in sentence processing. Taking extralinguistic context into account would likely result in more efficient trade-offs, as this can introduce additional cues helping to predict the future better.

## Limitations of Baseline Language Grammar Model

In Study 2, baseline grammars are constructed in a formalism that cannot fully express some word order regularities found in languages. For instance, it cannot express orders that differ in main clauses and embedded clauses (see discussion there for further limitations). These limitations are common to most other order grammar formalisms considered in the literature; despite these limitations, such word order models have demonstrated reasonably good fits to corpus data and human judgments of fluency (Futrell & Gibson, 2015; Wang & Eisner, 2016). These limitations do not affect the estimated trade-offs of real orders. However, the grammar model determines the baseline distribution, and thus impacts their comparison with real orders. For example, to the extent that strict word order decreases surprisal, this baseline distribution will put more weight on relatively efficient baselines, potentially resulting in a smaller difference with real orders than for baseline distributions that allow more flexibility. This limitation does not hold in Study 3, where the formalism provides very close fit to observed morpheme orders.

## Relation to Linguistic Typology

As a theory of linguistic typology, our Efficient Trade-off Hypothesis aims to explain universals in terms of functional

efficiency (Haspelmath, 2008). We have shown that it derives two previous typological principles—dependency length minimization and the Proximity Principle—which have been claimed to explain typological patterns such as Greenberg's harmonic word order correlations (Dryer, 1992; Greenberg, 1963), universal tendencies to order phrases with respect to their length (Behaghel, 1909; Chang, 2009; Wasow & Arnold, 2003), and the order of morphemes within words (Bybee, 1985; Givón, 1985). The Efficient Trade-off Hypothesis explains these apparently disparate phenomena via a simple and easily operationalizable principle of information locality: Elements with high mutual information are expected to be close to each other.

The idea of information locality goes beyond the idea of dependency length minimization by claiming that the strength of the pressure for words to be close to each other varies in proportion to their mutual information. This allows information locality to make predictions where dependency length minimization does not, for example, in the order of elements with the noun phrase, including adjective ordering. These predictions have met with empirical success (Futrell, 2019; Futrell, Dyer & Scontras, 2020; Hahn et al., 2018) (cf. Kirby et al., 2018).

Given the success of the memory–surprisal trade-off in capturing previous generalizations and in making new ones, further work on using the trade-off to predict more properties of languages seems promising. In this connection, we note that the memory–surprisal trade-off is mathematically nontrivial, and its properties have not yet been fully explored. We have provided only a lower bound on the trade-off and shown that it derives a principle of information locality. A fuller mathematical treatment may reveal further predictions to be tested, perhaps expanding the empirical coverage of the theory.

One limitation of our current treatment of the memory–surprisal trade-off is that its predictions are invariant with respect to word order reversal.<sup>12</sup> That is, it does not make any direct predictions about what elements should go earlier or later in a sentence; rather, it only predicts what elements should be relatively close or far from each other. This limitation means that the theory might not capture widespread universals which are *not* invariant to word order reversal, for example, the fact that suffixes are generally preferred over prefixes in morphology (Cutler et al., 1985), or the fact that elements which are animate, given, definite, and frequent tend to go earlier in sentences (Bock & Warren, 1985). Similarly, any asymmetries between head-final and head-initial constructions and languages are beyond the reach of our treatment. These order-asymmetrical universals have been explained in previous work using principles such as easy-first production (e.g., Bock & Warren, 1985; MacDonald, 2013) and the principle of Maximize Online Processing (MaxOP: Hawkins, 2004, 2014). However, this invariance to reversal applies only to our *lower bound* on the memory–surprisal trade-off curve; the true curve may not generally be invariant to word order reversal (cf., Crutchfield et al., 2009). Therefore, a more complete mathematical treatment might make predictions that are not invariant to word order reversal. We leave it to future work to derive these predictions and to determine if they match the typological data and the intuitions underlying theories such as MaxOP.

<sup>12</sup> For a mathematical proof, see SI Section 1.5.



## Relation to Information-Theoretic Studies of Language

Our work opens up a connection between psycholinguistics, linguistic typology, and statistical studies of language. Here, we survey the connections between our work and previous statistical studies. The average surprisal of real and counterfactual word orders has been studied by Gildea and Jaeger (2015) and Hahn et al. (2020). Gildea and Jaeger (2015) found that, in five languages, real orders provide lower trigram surprisal than baseline languages. This work can be viewed as instantiating our model in the case where the encoding function  $M$  records exactly the past two words, and showing that these five languages show optimization for surprisal under this encoding function. Hahn et al. (2020) compared surprisal and parseability for real and baseline orders as estimated using neural network models, arguing that word orders optimize a trade-off between these quantities. The results of Experiment 2 complement this by showing that real word orders optimize surprisal across possible memory capacities and memory encoding functions. Although we define information locality in terms of *conditional* mutual information, prior work has studied how *unconditional* mutual information decays with distance in natural language texts, at the level of orthographic characters (Ebeling & Pöschel, 1994; Lin & Tegmark, 2017) and words (Futrell et al., 2019). The link between memory and information locality provided by our Theorem 1 appears to be a novel contribution. The closest existing result is by Sharan et al., (2016) who show a link between excess entropy and approximability by  $n$ 'th order Markov models, noting that processes with low excess entropy can be approximated well with Markov models of low order. Our formalization of memory is related to studies of dynamic systems in the physics literature. Our memory-surprisal curve is closely related to the *predictive information bottleneck* introduced by Still (2014) and studied by Marzen and Crutchfield (2016); in particular, it is a version of the *recursive information bottleneck* (Still, 2014, §4). Hahn and Futrell (2019) empirically estimate the predictive information bottleneck trade-off of natural language using neural variational inference, providing an upper bound on the trade-off, whereas the current article provides a lower bound. In the limit of optimal prediction, our formalization of memory cost is equivalent to the notion of *statistical complexity* (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001); in our terminology, the statistical complexity of a stochastic process is the minimum value of  $H_M$  that achieves  $S_M = S_\infty$ . Furthermore, in the limit  $T \rightarrow \infty$ , the quantity in Eq. 4 is equal to another quantity from the theory of statistical complexity: *excess entropy* (Crutchfield & Young, 1989), the mutual information between the past and future of a sequence.

Our results are also closely related to information-theoretic scaling laws that characterize natural language, and in particular, the Relaxed Hilberg Conjecture (Debowski, 2015, 2020; Hilberg, 1990). The Relaxed Hilberg Conjecture is the claim that the average surprisal of a  $t$ 'th-order Markov approximation to language decays as a power law in  $t$ :

$$S_t \approx kt^{-\alpha} + S_\infty,$$

with the Hilberg exponent  $\alpha \approx \frac{1}{2}$ , and  $k$  a scaling factor. The Relaxed Hilberg Conjecture implies that conditional mutual information  $I_t$  falls off with distance as

$$I_t = S_t - S_{t+1} \\ \propto t^{-\alpha} - (t+1)^{-\alpha}.$$

The steepness of the falloff of mutual information depends on the value of the Hilberg exponent  $\alpha$ . As  $\alpha$  gets small, the falloff of mutual information is more rapid, corresponding to more information locality. Therefore, our Efficient Trade-off Hypothesis can be read as a claim about the Hilberg exponent  $\alpha$  for natural language: that it is lower than would be expected in a comparable system not constrained by incremental memory.

## Conclusion

In this work, we have provided evidence that human languages order elements in a way that reduces cognitive resource requirements, in particular memory effort. We provided an information-theoretic formalization of memory requirements as a trade-off of memory and surprisal. We showed theoretically that languages have more efficient trade-offs when they show stronger degrees of information locality. Information locality provides a formalization of various locality principles from the linguistic literature, including dependency locality (Gibson, 1998), domain minimization (Hawkins, 2004), and the proximity principle (Givón, 1985). Using this result, we provided evidence that languages order words and morphemes in such a way as to provide efficient memory-surprisal trade-offs. Therefore, the memory-surprisal trade-off simultaneously provides (a) a unified explanation of diverse typological phenomena which is rigorously grounded in the psycholinguistics literature, (b) a theory which makes new successful quantitative predictions about word and morpheme order within and across languages, and (c) a mathematical framework relating universals of language to principles of efficient coding from information theory. Our result shows that wide-ranging principles of order in natural language can be explained from highly generic cognitively motivated information-theoretic principles. The locality properties we have discussed are some of the most characteristic properties of natural language, setting natural language apart from other codes studied in information theory. Therefore, our result raises the question of whether other distinctive characteristics of language—for example, mildly context-sensitive syntax, duality of patterning, and compositionality—might also be explained in terms of information-theoretic resource constraints.

## References

- Abney, S. P. & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3), 233–250.
- Altmann, G. T. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1), 28–55.
- Asahara, M., Kanayama, H., Tanaka, T., Miyao, Y., Uematsu, S., Mori, S., Matsumoto, Y., Omura, M., & Murawaki, Y. (2018). *Universal dependencies version 2 for Japanese* [Conference session]. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan (Irec 2018).
- Aurnhammer, C. & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134, Article 107198.

- Balling, L. W. & Kizach, J. (2017). Effects of surprisal and locality on danish sentence processing: An eye-tracking investigation. *Journal of Psycholinguistic Research*, 46(5), 1119–1136.
- Bartek, B., Lewis, R. L., Vasishth, S., & 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.
- Behaghel, O. (1909). Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, 25, 110–142.
- Behaghel, O. (1932). *Deutsche syntax* (Vol. 4). Winter.
- Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6), Article 275.
- Berger, T. (2003). *Rate-distortion theory*. Wiley Encyclopedia of Telecommunications.
- Berwick, R. C. & Weinberg, A. (1984). *The grammatical basis of linguistic performance*. MIT Press.
- Bock, J. K. & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1), 47–67.
- Böhmová, A., Hajic, J., Hajicová, E., & Hladká, B. (2003). The Prague dependency treebank. In A. Abeillé (Ed.), *Treebanks* (pp. 103–127). Springer.
- Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.
- Brady, T. F., Konkle, T., & Alvarez, G. A. (2008). *Efficient coding in visual short-term memory: Evidence for an information-limited capacity*. Proceedings of the Annual Meeting of the Cognitive Science Society, Washington, DC.
- Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138(4), 487–502.
- Brennan, S. E. & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the meta-cognitive states of speakers. *Journal of Memory and Language*, 34(3), 383–398.
- Bybee, J. L. (1985). *Morphology : A study of the relation between meaning and form*. John Benjamins.
- Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61(3), 374–397.
- Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, 36(1), 1–61.
- Clark, A. & Fijalkow, N. (2020). Consistent unsupervised estimators for anchored PCFGs. *Transactions of the Association for Computational Linguistics*, 8, 409–422. [https://doi.org/10.1162/tac1\\_a\\_00323](https://doi.org/10.1162/tac1_a_00323)
- Clark, H. H. & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology* (Vol. 9, pp. 287–299). Elsevier. <https://www.sciencedirect.com/bookseries/advances-in-psychology>
- Corbett, G. G., Fraser, N. M., & McGlashan, S. (1993). *Heads in grammatical theory*. Cambridge University Press.
- Cover, T. M. & Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.
- Crutchfield, J. P., Ellison, C. J., & Mahoney, J. R. (2009). Time's barbed arrow: Irreversibility, crypticity, and stored information. *Physical Review Letters*, 103(9), Article 094101.
- Crutchfield, J. P. & Feldman, D. P. (2003). Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1), 25–54.
- Crutchfield, J. P. & Young, K. (1989). Inferring statistical complexity. *Physical Review Letters*, 63(2), 105–108. <https://doi.org/10.1103/PhysRevLett.63.105>
- Cutler, A., Hawkins, J. A., & Gilligan, G. (1985). The suffixing preference: A processing explanation. *Linguistics*, 23(5), 723–758.
- Daniluk, M., Rocktäschel, T., Welbl, J., & Riedel, S. (2017, April 24–26). *Frustratingly short attention spans in neural language modeling* [Conference session]. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, Conference Track Proceedings. <https://openreview.net/forum?id=ByIAPUcee>
- de Paiva Alves, E. (1996). *The selection of the most probable dependency structure in Japanese using mutual information*. 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, California.
- De Santo, A. (2020). *MG parsing as a model of gradient acceptability in syntactic Islands*. Proceedings of the Society for Computation in Linguistics, New Orleans, Louisiana. <https://doi.org/10.7275/srck-2j50>
- Debowski, Ł. (2015). The relaxed hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4), 311–337.
- Debowski, Ł. (2020). *Information theory meets power laws: Stochastic processes and language models*. John Wiley & Sons.
- Debowski, Ł. (2011). Excess entropy in natural language: Present state and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3), Article 037105.
- Demberg, V. & Keller, F. (2009). *A computational model of prediction in human parsing: Unifying locality and surprisal effects*. Proceedings of the Annual Meeting of the Cognitive Science Society, Amsterdam, Netherlands.
- Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-joining grammar. *Computational Linguistics*, 39(4), 1025–1066. [https://doi.org/10.1162/COLI\\_a\\_00160](https://doi.org/10.1162/COLI_a_00160)
- Demuth, K. (1992). Acquisition of sesotho. In D. Slobin (Ed.), *The cross-linguistic study of language acquisition* (pp. 557–638). Lawrence Erlbaum Associates.
- Doke, C. M. & Mofokeng, S. M. (1967). *Textbook of southern Sotho grammar*. Longmans.
- Doob, J. L. (1953). *Stochastic processes*. New York Wiley.
- Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68(1), 81–138. <https://doi.org/10.2307/416370>
- Ebeling, W. & Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters (EPL)*, 26(4), 241–246. <https://doi.org/10.1209/0295-5075/26/4/001>
- Engelmann, F. & Vasishth, S. (2009). *Processing grammatical and ungrammatical center embeddings in English and German: A computational model* [Conference Session]. Proceedings of the Ninth International Conference on Cognitive Modeling, Manchester, U.K. (pp. 240–45).
- Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, 37(2), 378–394.
- Fedzechkina, M., Chu, B., & Jaeger, T. F. (2017). *Human information processing shapes language change*. [http://mfedzech.github.io/docs/FedzechkinaChuJaeger%5C\\_submitted.pdf](http://mfedzech.github.io/docs/FedzechkinaChuJaeger%5C_submitted.pdf)
- Fedzechkina, M. & Jaeger, T. F. (2020). Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, 196, Article 104115.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44), 17897–17902.
- Ferreira, V. S. & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, 40(4), 296–340.
- Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70, Article 056135. <https://doi.org/10.1103/PhysRevE.70.056135>
- Firbas, J. (1966). On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, 1, 267–280.
- Firbas, J. (1974). Some aspects of the Czechoslovak approach to problems of functional sentence perspective. In F. Danes (Ed.), *Papers on functional sentence perspective* (pp. 11–37).



- Frank, S. L. & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6), 829–834. <https://doi.org/10.1177/0956797611409589>
- Frank, S. L. & Ernst, P. (2019). Judgements about double-embedded relative clauses differ between languages. *Psychological research*, 83(7), 1581–1593.
- Frank, S. L. & Hoeks, J. C. J. (2019). *The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times* [Conference session]. In Proceedings of the 41st Annual Conference of the Cognitive Science Society, Montreal, Canada.
- Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11.
- Frank, S. L., Trompenaars, T., & Vasishth, S. (2015). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*. <http://onlinelibrary.wiley.com/doi/10.1111/cogs.12247/full>
- Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives* (pp. 129–189), Cambridge University Press.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, 11(2), 127–138.
- Futrell, R. (2019). *Information-theoretic locality properties of natural language*. Proceedings of the First Workshop on Quantitative Syntax, Paris, France (Quasy, Syntaxfest 2019).
- Futrell, R., Dyer, W., & Scontras, G. (2020, July 5–10). What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (ACL 2020, Online, pp. 2003–2012). Association for Computational Linguistics. <https://www.aclweb.org/anthology/2020.acl-main.181/>
- Futrell, R. & Gibson, E. (2015). Experiments with generative models for dependency tree linearization. In L. Márquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing*, (pp. 1978–1983). Association for Computational Linguistics. <http://aclweb.org/anthology/D15-1231>
- Futrell, R., Gibson, E., & Levy, R. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive science*, 44(3), e12814.
- Futrell, R., Mahowald, K., & Gibson, E. (2015a). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33), 10336–10341. <https://doi.org/10.1073/pnas.1502134112>
- Futrell, R., Mahowald, K., & Gibson, E. (2015b). *Quantifying word order freedom in dependency corpora* [Conference session]. Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), Uppsala, Sweden. <http://www.aclweb.org/anthology/W15-21#page=101>
- Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). *Syntactic dependencies correspond to word pairs with high mutual information* [Conference session]. Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, Syntaxfest 2019), Paris, France.
- Gershman, S. J. (2020). *Origin of perseveration in the trade-off between reward and complexity*. BioRxiv. <https://www.biorxiv.org/content/10.1101/2020.01.16.903476v1>
- Gerth, S. (2015). *Memory limitations in sentence comprehension: A structural-based complexity metric of processing difficulty*. Universitätsverlag Potsdam.
- Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown* [Doctoral dissertation]. Carnegie Mellon University Pittsburgh, PA.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.
- Gibson, E. (2000). Chapter 5: The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O’Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126), MIT Press.
- Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407.
- Gibson, E. & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3), 225–248.
- Gildea, D. & Jaeger, T. F. (2015). *Human languages order information efficiently*. arXiv:1510.02823 [cs]. arXiv: 1510.02823. <http://arxiv.org/abs/1510.02823>
- Gildea, D. & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 184–191). Association for Computational Linguistics. <http://www.aclweb.org/anthology/P07-1024>
- Gildea, D. & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34(2), 286–310. <https://doi.org/10.1111/j.1551-6709.2009.01073.x>
- Givón, T. (1985). Iconicity, isomorphism and non-arbitrary coding in syntax. In J. Haiman (Ed.), *Iconicity in Syntax* (pp. 187–219). John Benjamins.
- Givón, T. (1988). The pragmatics of word-order: Predictability, importance and attention. In M. Hammond, E. Moravcsik, & J. Wirth (Eds.), *Studies in syntactic typology* (pp. 243–284). John Benjamins.
- Givón, T. (1991). Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language*, 15, 335–370.
- Goodkind, A. & Bicknell, K. (2018). *Predictive power of word surprisal for reading times is a linear function of language model quality*. Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics, (CMCL 2018), Salt Lake City, Utah.
- Goodman, J. (1999). Semiring parsing. *Computational Linguistics*, 25(4), 573–605.
- Graf, T., Fodor, B., Monette, J., Rachiele, G., Warren, A., & Zhang, C. (2015). A refined notion of memory usage for minimalist parsing. In *Proceedings of the 14th meeting on the mathematics of language (MOL 2015)*, (pp. 1–14). Association for Computational Linguistics. <http://www.aclweb.org/anthology/W15-2301>
- Graf, T. & Marcinek, B. (2014). *Evaluating evaluation metrics for minimalist parsing*. Proceedings of the fifth workshop on cognitive modeling and computational linguistics, Baltimore, Maryland.
- Graf, T., Monette, J., & Zhang, C. (2017). Relative clauses as a benchmark for minimalist parsing. *Journal of Language Modelling*, 5, 57–106. <https://doi.org/10.15398/jlm.v5i1.157>
- Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*, 2, 73–113.
- Grodner, D. & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2), 261–290.
- Guma, S. M. (1971). *An outline structure of southern Sotho*. Shuter and Shooter.
- Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2020). *Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content*. BioRxiv.
- Hahn, M., Degen, J., Goodman, N., Jurafsky, D., & Futrell, R. (2018). *An information-theoretic explanation of adjective ordering preferences* [Paper presentation]. Proceedings of the 40th Annual Meeting of the Cognitive Science Society (CogSci).
- Hahn, M. & Futrell, R. (2019). Estimating predictive rate-distortion curves via neural variational inference. *Entropy*, 21(7), Article 640. <https://www.mdpi.com/1099-4300/21/7/640>

- Hahn, M., Jurafsky, D., & Futrell, R. (2020). Universals of word order reflect optimization for communicative efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, 117(5), 2347–2353.
- Hale, J. T. (2001). *A probabilistic Earley parser as a psycholinguistic model*. Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies, Pittsburgh, PA.
- Harris, Z. (1955). From phonemes to morphemes. *Language*, 31, 190–222.
- Hasegawa, Y. (2014). *Japanese: A linguistic introduction*. Cambridge University Press.
- Haspelmath, M. (2008). Parametric versus functional explanations of syntactic universals. In T. Biberauer (Ed.), *The limits of syntactic variation* (pp. 75–107). John Benjamins.
- Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.
- Hawkins, J. A. (2003). Efficiency and complexity in grammars: Three general principles. In M. Polinsky, & J. Moore (Eds.), *The Nature of Explanation in Linguistic Theory* (pp. 121–152). Center for the Study of Language and Information.
- Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.
- Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency*. Oxford University Press.
- Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, 40, 511–525.
- Hilberg, W. (1990). Der bekannte grenzwert der redundanzfreien information in texten—eine fehlinterpretation der shannonschen experimente? *Frequenz*, 44, 243–248.
- Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800–802.
- Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Hudson, R. A. (1984). *Word grammar*. Blackwell Oxford.
- Jacobs, J. (1988). Probleme der freien wortstellung im deutschen. *Sprache und Pragmatik. Arbeitsberichte*, 5, 8–37.
- Jaeger, T. F. & Tily, H. J. (2011). On language “utility”: Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 323–335.
- Jelinek, F. & Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3), 315–323.
- Kaiser, S., Ichikawa, Y., Kobayashi, N., & Yamamoto, H. (2013). *Japanese: A comprehensive grammar*. Routledge.
- Kim, Y., Dyer, C., & Rush, A. M. (2019, July 29–August 2). Compound probabilistic context-free grammars for grammar induction. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics (ACL 2019)*, Florence, Italy, Long papers, Vol. 1, pp. 2369–2385. Association for Computational Linguistics. <https://doi.org/10.18653/v1/p19-1228>
- Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, 2(1), 15–47.
- Kirby, S., Culbertson, J., & Schouwstra, M. (2018). *The origins of word order universals: Evidence from corpus statistics and silent gesture* [Conference session]. Proceedings of the 12th International Conference on the Evolution of Language (evolang12), Torun, Poland.
- Kneser, R. & Ney, H. (1995). *Improved backing-off for m-gram language modeling*. Proceedings of the 1995 International conference on acoustics, speech, and signal processing (ICASSP-95), Detroit, Michigan.
- Kobe, G. M., Gerth, S., & Hale, J. (2013). Memory resource allocation in top-down minimalist parsing. In M. J. Nederhof (Ed.), *Formal grammar* (pp. 32–51). Springer.
- Kuperberg, G. R. & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59.
- Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.
- Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). Psychology Press.
- Lewis, R. L. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419. [https://doi.org/10.1207/s15516709cog0000\\_25](https://doi.org/10.1207/s15516709cog0000_25)
- Lin, H. W. & Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7), Article 299. <https://doi.org/10.3390/e19070299>
- Lindblom, B. (1990). On the communication process: Speaker–listener interaction and the development of speech. *Augmentative and Alternative Communication*, 6(4), 220–230.
- Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2), 159–191.
- Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*. <https://doi.org/10.1016/j.plrev.2017.03.002>
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, 4, Article 226.
- MacDonald, M. C. & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1), 35–54.
- Marzen, S. E. & Crutchfield, J. P. (2016). Predictive rate-distortion for infinite-order markov processes. *Journal of Statistical Physics*, 163(6), 1312–1338. <https://doi.org/10.1007/s10955-016-1520-1>
- McCauley, S. M. & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, 126(1), 1–51.
- McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29(2), 111–123.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48(1), 67–91. [https://doi.org/10.1016/S0749-596X\(02\)00515-6](https://doi.org/10.1016/S0749-596X(02)00515-6)
- Melcuk, I. A. (1988). *Dependency syntax: Theory and practice*. SUNY Press.
- Mikolov, T., Karafiát, M., Burget, L., ěernocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model*. Proceedings of INTER-SPEECH, Makuhari, Japan.
- Mikulová, M., Bémová, A., Hajic, J., Hajicová, E., Havelka, J., Kolárová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razimová, M., Sgall, P., Štěpánek, J., Uřešová, Z., Veselá, K., & Žabokrtský, Z. (2006). *Annotation on the tectogrammatical layer in the Prague dependency treebank, annotation manual*. <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf>
- Miller, G. A. & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–492). John Wiley
- Myhill, J. (1985). Pragmatic and categorial correlates of vs word order. *Lingua*, 66, 177–200.
- Nederhof, M. & Satta, G. (2011, July 27–31). Computation of infix probabilities for probabilistic context-free grammars. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP 2011)*, John McIntyre Conference Centre, Edinburgh, U.K., a meeting of Sigdat, a special interest group of the ACL pp. 1213–1221. ACL. <https://www.aclweb.org/anthology/D11-1112/>
- Neeleman, A. & van de Koot, H. (2016). Word order and information structure. In C. Féry & S. Ishihara (Eds.), *The oxford handbook of information structure* (pp. 383–401). Oxford University Press.

- Neubig, G. & Mori, S. (2010, May 17–23). Word-based partial annotation for efficient corpus construction. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the international conference on language resources and evaluation (LREC 2010, Valletta, Malta)* (pp. 2723–2727). European Language Resources Association. <http://www.lrec-conf.org/proceedings/lrec2010/summaries/408.html>
- Neubig, G., Nakata, Y., & Mori, S. (2011, June 19–24). Pointwise prediction for robust, adaptable Japanese morphological analysis. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *The 49th annual meeting of the association for computational linguistics: Human language technologies, proceedings of the conference* (Portland, Oregon, USA—short papers, pp. 529–533). The Association for Computer Linguistics. <https://www.aclweb.org/anthology/P11-2093/>
- Newmeyer, F. J. (1992). Iconicity and generative grammar. *Language*, 756–796.
- Nicenboim, B. & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, 99, 1–34.
- Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, 6, Article 312.
- Nivre, J., Agic, Z., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bosco, C., Bouma, G., & Zhu, H. (2017). *Universal dependencies 2.0—CoNLL 2017 shared task development and test data*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2184>
- Petrov, S. & Klein, D. (2007, July 22–26). Learning and inference for hierarchically split PCFGs. In R. Holte, & A. Howe (Eds.), *Proceedings of the twenty-second AAAI conference on artificial intelligence* (Vancouver, British Columbia, Canada, pp. 1663–1666). AAAI Press. <http://www.aaai.org/Library/AAAI/2007/aaai07-268.php>
- Rambow, O. & Joshi, A. K. (1994). A processing model for free word-order languages. In C. Clifton, Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on Sentence Processing*, (pp. 267–301). Lawrence Erlbaum Associates, Inc.
- Rasmussen, N. E. & Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive science*, 42, 1009–1042.
- Resnik, P. (1992). Left-corner parsing and psychological plausibility. In Ch. Boitet (Ed.), *Proceedings of the 14th conference on computational linguistics* (Vol. 1, pp. 191–197). Association for Computational Linguistics.
- Rijkhoff, J. (1990). Explaining word order in the noun phrase. *Linguistics*, 28(1), 5–42.
- Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Comput. Linguistics*, 27(2), 249–276. <https://doi.org/10.1162/089120101750300526>
- Ross, J. R. (1967). *Constraints on variables in syntax* [Dissertation], Institute of Technology, Massachusetts.
- Schach, S., Gottwald, S., & Braun, D. A. (2018). Quantifying motor task performance by bounded rational decision theory. *Frontiers in neuroscience*, 12, Article 932.
- Schadeberg, T. (2003). Derivation. In D. Nurse & G. Philippon (Eds.), *The bantu languages* (pp. 71–89). Routledge.
- Schijndel, M. V., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *topiCS*, 5(3), 522–540. <https://doi.org/10.1111/tops.12034>
- Shalizi, C. R. & 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.
- Sharan, V., Kakade, S., Liang, P., & Valiant, G. (2016). *Prediction with a short memory*. arXiv:1612.02526 [cs, stat]. arXiv: 1612.02526. <http://arxiv.org/abs/1612.02526>
- Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389), 652–656.
- Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, 119(4), 807–830.
- Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems* (pp. 2951–2959). <http://papers.nips.cc/paper/4522-practical>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1), 1929–1958. <http://dl.acm.org/citation.cfm?id=2670313>
- Stallings, L. M., & MacDonald, M. C. (2011). It's not just the “heavy NP”: Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research*, 40(3), 177–187.
- Staub, A., Clifton C., Jr. & Frazier, L. (2006). Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language*, 54(3), 389–406.
- Staub, A. & Clifton C., Jr. (2006). Syntactic prediction in language comprehension: Evidence from either ... or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2), 425–436.
- Still, S. (2014). Information bottleneck approach to predictive inference. *Entropy*, 16(2), 968–989. <https://doi.org/10.3390/e16020968>
- Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of prediction. *Physical review letters*, 109(12), Article 120604.
- Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2), 165–201.
- Takahashi, S. & Tanaka-Ishii, K. (2018). Cross entropy of neural language models at infinity—A new bound of the entropy rate. *Entropy*, 20(11), Article 839.
- Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., & Matsumoto, Y. (2016). *Universal dependencies for Japanese* [Conference session]. Proceedings of the Tenth International Conference on Language Resources and Evaluation (lrec'16), (pp. 1651–1658).
- Temperley, D. & Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4, 1–15.
- Tesnière, L. & Kahane, S. (2015). *Elements of structural syntax*. John Benjamins Publishing Company.
- Vaccari, O. & Vaccari, E. E. (1938). *Complete course of Japanese conversation-grammar*. Maruzen in Komm.
- Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23(11), 968–982.
- Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4), 533–567.
- Wang, D. & Eisner, J. (2016). The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4, 491–505.
- Wasow, T. & Arnold, J. (2003). Post-verbal constituent ordering in English. *Determinants of grammatical variation in English*, 119–154.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2), 250–271.

- Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). *On the predictive power of neural language models for human real-time comprehension behavior*. *Proceedings of CogSci 2020* (pp. 1707–1713).
- Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2017). Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.
- Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5), 444–466.
- Yuret, D. (1998). Discovery of linguistic relations using lexical attraction. *arXiv preprint cmp-lg/9805009*.
- Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942.
- Zénon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123, 5–18.

Received June 5, 2020

Revision received September 14, 2020

Accepted October 21, 2020 ■