Noisy-context surprisal as a human sentence processing cost model

Richard Futrell and Roger Levy
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{futrell, rplevy}@mit.edu

Abstract

We use the noisy-channel theory of human sentence comprehension to develop an incremental processing cost model that unifies and extends key features of expectation-based and memory-based models. In this model, which we call **noisy-context surprisal**, the processing cost of a word is the surprisal of the word given a noisy representation of the preceding context. We show that this model accounts for an outstanding puzzle in sentence comprehension, language-dependent structural forgetting effects (Gibson and Thomas, 1999; Vasishth et al., 2010; Frank et al., 2016), which are previously not well modeled by either expectation-based or memory-based approaches. Additionally, we show that this model derives and generalizes locality effects (Gibson, 1998; Demberg and Keller, 2008), a signature prediction of memory-based models. We give corpus-based evidence for a key assumption in this derivation.

1 Introduction

Models of human sentence processing difficulty can be divided into two kinds, **expectation-based** and **memory-based**. Expectation-based models predict the processing difficulty of a word from the word's surprisal given previous material in the sentence (Hale, 2001; Levy, 2008a). These models have good coverage: they can account for effects of syntactic construction frequency and resolution of ambiguity on incremental processing difficulty. Memory-based models, on the other hand, explain difficulty resulting from working memory limitations during incremental parsing (Gibson, 1998; Lewis and Vasishth, 2005); a major prediction of these models is **locality effects**, where processing a word is difficult when it is far from other words with which it must be syntactically integrated. Expectation-based models do not intrinsically capture this difficulty.

Integrating these two approaches at a high level has proven challenging. A major hurdle is that the theories are typically stated at different levels of analysis: expectation-based theories are computational-level theories (Marr, 1982) specifying what computational problem the human sentence processing system is solving—the problem of how update one's belief about a sentence given a new word—without specifying implementation details. Memory-based theories such as Lewis and Vasishth (2005) are for the most part based in mechanistic algorithmic-level theories describing the actions of a specific incremental parser.

Previous theories that capture both surprisal and locality effects have typically done so by augmenting parsing models with a special prediction-verification operation to capture surprisal effects (Demberg and Keller, 2009; Demberg et al., 2013), or by combining surprisal and memory-based cost derived from a parsing model as separate factors in a linear model (Shain et al., 2016). These models capture surprisal and locality effects at the same time, but they do not clearly capture phenomena involving the interaction of memory and probabilistic expectations such as language-dependent structural forgetting (see Section 3).

Here we develop a computational-level model capturing both memory and expectation effects from a single set of principles, without reference to a specific parsing algorithm. In our model, the processing cost of a word is a function of its surprisal given a *noisy* representation of previous context (Section 2). We show that the model can reproduce structural forgetting effects, including

the difference between English and German (Section 3), a phenomenon not previously captured by memory-based or expectation-based models in isolation. We also give a derivation of the existence of locality effects in the model; these effects were previously accounted for only in mechanistic memory-based models (Section 4). The derivation yields a generalization of classic locality effects which we call **information locality**: sentences are predicted to be easier to process when words with high mutual information are close. We give corpus-based evidence that words in syntactic dependencies have high mutual information, meaning that classical dependency locality effects can be seen as a subset of information locality effects.

2 Noisy-Context Surprisal

In surprisal theory, the processing cost of a word is asserted to be proportional to extent to which one must change one's beliefs given that word (Hale, 2001; Smith and Levy, 2013). So the cost of a word is (up to proportionality):

$$C_{\text{surprisal}}(w_i|w_{1:i-1}) = -\log p_L(w_i|w_{1:i-1}), \quad (1)$$

where $p_L(\cdot|\cdot)$ is the conditional probability of a word in context in a probabilistic language L.

Standard surprisal assumes that the comprehender has perfect access to a representation of w_i's full context, including the words preceding it in the sentence ($w_{1:i-1}$) and also extra-sentential context (which we leave implicit). But given that human working memory is limited, the assumption of perfect access is unrealistic. We propose that processing cost at a word is better modeled as the cost of belief updates given a *noisy representation* of the previous input. The probability of a word given a noisy context is modeled as the noisy channel probability of the word, assuming that people do noisy channel inference on their context representation (Levy, 2008b; Gibson et al., 2013). Given this model, the expected processing cost of a word is its expected surprisal over the possible noisy representations of its context.

The noisy-context surprisal processing cost function is thus:[1]

$$C(w_i|w_{1:i-1}) = \mathop{\mathbb{E}}_{V|w_{1:i-1}} [-\log p_L^{\text{NC}}(w_i|V)] \quad (2)$$

$$= -\sum_V p_N(V|w_{1:i-1}) \log p_L^{\text{NC}}(w_i|V) \quad (3)$$

[1]Neglecting the implicit proportionality term in Equation 1.

where V is the noisy representation of the previous material $w_{1:i-1}$, the **noise distribution** p_N characterizes how memory of previous material may be corrupted, and $p_L^{\text{NC}}(\cdot|\cdot)$ is the noisy-channel probability of a word given a noisy context, computed via marginalization:

$$p_L^{\text{NC}}(w_i|V) = \sum_{w_{1:i-1}} p_L(w_i|w_{1:i-1}) p^{\text{NC}}(w_{1:i-1}|V)$$

with $p^{\text{NC}}(w_{1:i-1}|V)$ computed via Bayes Rule:

$$p^{\text{NC}}(w_{1:i-1}|V) \propto p_N(V|w_{1:i-1}) p_L(w_{1:i-1}).$$

Note here that w_i's cost is computed using its true identity but a noisy representation of the context: from the incremental perspective, w_i is observed now, but context is stored and retrieved in a potentially noisy storage medium. This asymmetry between noise levels for proximal versus distal input differs from the noisy-channel surprisal model of Levy (2011), and is crucial to the derivation of information locality we present in Section 4.

Here we use two types of noise distributions for p_N: erasure noise and deletion noise. In **erasure noise**, a symbol in the context is probabilistically erased and replaced with a special symbol `E` with probability e. In **deletion noise**, a symbol is erased from the sequence completely, leaving no trace. Given deletion noise, a comprehender does not know how many symbols were in the original context; with erasure noise, the comprehender knows exactly which symbols were affected by noise. In both cases, we assume that the application or non-application of noise is probabilistically independent among elements in the context. We use these concrete noise distributions for convenience, but we believe our results should generalize to larger classes of noise distributions.

3 Structural Forgetting Effects

Here we show that noisy-context surprisal as a processing cost model can reproduce effects that were not previously well-explained by either expectation-based or memory-based theories. In particular, we take up the puzzle of **structural forgetting effects**, where comprehenders seem to forget structural elements of a sentence prefix when predicting the rest of the sentence. The result is that some ungrammatical sentences have lower processing cost and higher acceptability than some complex grammatical sentences: with

doubly nested relative clauses, for instance, subjects rate ungrammatical sentence (1) as more acceptable than sentence (2), forgetting about the VP predicted by the second noun (Gibson and Thomas, 1999).

(1) *The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_1$ well-decorated.

(2) The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_2$ cleaning every week was$_1$ well-decorated.

Vasishth et al. (2010) show this same effect in reading times at the last verb: in English native speakers are more surprised to encounter a third VP than not to. However, this effect is language-specific: the same authors find that in German, native speakers are more surprised when a third VP is missing than when it is present. Frank et al. (2016) show further that native speakers do not show the effect in Dutch, but Dutch-native L2 speakers of English do show the effect in English. The result shows that the memory resources taxed by these structures are themselves meaningfully shaped by the distributional statistics of the language.

The verb forgetting effect is a challenge for both expectation-based and memory-based models. Pure expectation-based models cannot reproduce the effect: they have no mechanism for forgetting an established VP prediction and thus they assign small or zero probability to ungrammatical sentences. On the other hand, memory-based models will have to account for why the same structures are forgotten in English but not in German. Here we show that noisy-context surprisal provides the first purely computational-level account for the language-dependent verb forgetting effect. The essential mechanism is that when verb-final nested structures are more probable in a language, then they will be better preserved in a noisy memory representation.

3.1 Model of Verb Forgetting

Table 1 presents a toy probabilistic context-free grammar for the constructions involved in verb forgetting. The grammar generates strings over the alphabet of N (noun), V (verb), C (complementizer), P (preposition). We apply deletion noise with by-symbol deletion probability d. So for example, given a prefix NCNCNVV, the prefix can be corrupted to NCNNVV with probability proportional to d, representing one deletion. In that case a noisy-channel comprehender might incorrectly infer that the original prefix was in fact NCNPNVV, and thus fail to predict a third verb.

Rule	Probability
S $\rightarrow$ NP V	1
NP $\rightarrow$ N	$1-m$
NP $\rightarrow$ N RC	mr
NP $\rightarrow$ N PP	$m(1-r)$
PP $\rightarrow$ P NP	1
RC $\rightarrow$ C V NP	s
RC $\rightarrow$ C NP V	$1-s$

Table 1: Toy grammar used to demonstrate verb forgetting. Nouns are postmodified with probability m; a postmodifier is a relative clause with probability r, and a relative clause is V-initial with probability s. For practical reasons we bound nonterminal rewrites of NP at 2.

To illustrate that noisy surprisal can account for language-dependent verb forgetting, we show in Figure 1 the differences between noisy surprisal values for grammatical (V) and ungrammatical (end-of-sentence) continuations of prefixes NCNCNVV under parameter settings reflecting the difference between English and German, and compare these differences with self-paced reading times observed after the final verb by Vasishth et al. (2010). Noisy surprisal qualitatively reproduces language-dependent verb forgetting: in English the ungrammatical continuation is higher surprisal, but in German the grammatical continuation is higher surprisal. The English–German difference in the model is entirely accounted for by the parameter s, which determines the proportion of relative clauses that are verb-initial. In English, most relative clauses are subject-extracted and those are verb-initial, so for English $s \approx .8$ (Roland et al., 2007). German, in contrast, has $s \approx 0$, since its relative clauses are obligatorily verb-final. When verb-final relative clauses have higher prior probability, a doubly-nested RC prefix NCNCVV is more likely to be preserved by a rational noisy-channel comprehender.

The results of Figure 1 do not speak, however, to the generality of the model's predictions regarding verb forgetting. To explore this matter, we partition the model's four-dimensional parameter space into regions distinguishing whether noisy-context surprisal is lower for (G) grammatical continuations or (U) ungrammatical contin-

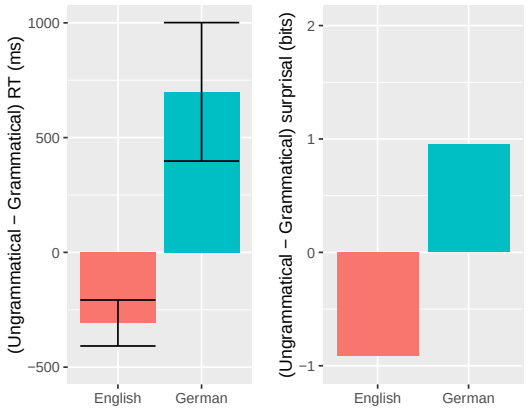

Figure 1: Differences in reaction times for ungrammatical continuations minus grammatical continuations, compared to noisy surprisal differences. RT data comes from self-paced reading experiments in Vasishth et al. (2010) in the post-VP region. The noisy surprisal predictions are produced with $d = .2$, $m = .5$, $r = .5$ fixed, and $s = .8$ for English and $s = 0$ for German.

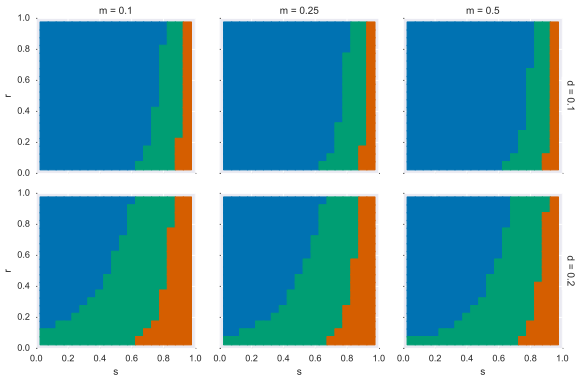

Figure 2: Regions of different model behavior with respect to parameters r, s, m, and d (see Table 1). Blue: $\mathrm{G}_1\mathrm{G}_2$; red: $\mathrm{U}_1\mathrm{U}_2$; green: $\mathrm{G}_1\mathrm{U}_2$ (see text).

uations for (1) singly-embedded NCNV and (2) doubly-embedded NCNCNVV contexts. Figure 2 shows this partition for a range of r, s, m, and d. In the blue region, grammatical continuations are lower-cost than ungrammatical continuations for both singly and doubly embedded contexts, as in German ($\mathrm{G}_1\mathrm{G}_2$); in the red region, the ungrammatical continuation is lower-cost for both contexts ($\mathrm{U}_1\mathrm{U}_2$). In the green region, the grammatical continuation is lower cost for single embedding, but higher cost for double embedding, as in English ($\mathrm{G}_1\mathrm{U}_2$). No combination of parameter values instantiates $\mathrm{U}_1\mathrm{G}_2$ (for either the depicted or other possible values of m and d). Thus both the English and German behavioral patterns are quite generally predicted by the model. Furthermore, each language's statistics place it in a region of parameter space plausibly corresponding to its behavioral pattern: the English-type forgetting effect is predicted mostly for high s, the German-type for low s.

The only previous formalized account of language-specific verb forgetting, Frank et al. (2016), showed that Simple Recurrent Networks (SRNs) trained on English and Dutch data partly reproduce the verb forgetting effect in the surprisals they assign to the final verb. Our model provides an explanation of the SRN's behavior. When an SRN predicts words, it effectively uses a lossily compressed representation of the previous words. This lossy compression is analogous to the noisy representation posited here.

4 Information Locality

Here we show how, given an appropriate noise distribution, noisy surprisal gives rise to locality effects. Standard locality effects are related to syntactic dependencies: the claim is that processing is difficult when the parser must make a syntactic connection with an element that has been in memory for a long time. In Section 4.1, we derive a more general prediction: that processing is difficult when any elements with high mutual information are far from one another. The effect arises under noisy surprisal because contexts that would have been helpful for predicting a word might have been forgotten. We call this principle **information locality**. In Section 4.3, we argue that words in syntactic dependencies have higher mutual information than other word pairs, which leads to a view of dependency locality effects as a special case of information locality effects.

4.1 Derivation of Information Locality

Viewing processing cost as a function of word order, noisy surprisal gives rise to the generalization that cost is minimized when elements with high mutual information are close. We show this by decomposing the noisy surprisal cost of a word into many terms of higher-order mutual information with the context, then showing that applying a certain kind of erasure noise to the context causes these terms to be downweighted based on their dis-

tance to the word. Thus the best word order puts the words that have high mutual information with a word close to that word.

4.1.1 Noise Distribution

Noisy surprisal gives rise to information locality under a family of noise distributions which we call **progressive erasure noise**, which is any noise function that erases discrete elements of a sequence with increasing probability the earlier those elements are in the sequence. Formally, in progressive erasure noise, the ith element in a sequence X with length $|X|$ is erased with probability proportional to some monotonically increasing function of how far left that element is in the sequence: $f(|X| - i)$. As a concrete example of progressive erasure noise, consider an exponential decay function, such that the probability that an element i in X remains unerased is $(1-e)^{|X|-i}$ for some probability e. The exponential decay function corresponds to a noise model where the context sequence is hit with erasure noise successively as each word is processed. Any progressive erasure noise distribution suffices for the derivation here to go through.

4.1.2 Decomposing Surprisal Cost

In noisy surprisal theory, the cost of a word w_i in context $w_{1:i-1}$ is:

$$\begin{aligned} C(w_i|w_{1:i-1}) &= \mathop{\mathbb{E}}_{V|w_{1:i-1}}[-\log p(w_i|V)] \\ &= \mathop{\mathbb{E}}_{V|w_{1:i-1}}[h(w_i) - \text{pmi}(w_i;V)] \\ &= h(w_i) - \mathop{\mathbb{E}}_{V|w_{1:i-1}}[\text{pmi}(w_i;V)], \end{aligned} \quad (4)$$

where $h(\cdot)$ is surprisal (here unconditional, equivalent to log inverse-frequency) and $\text{pmi}(\cdot;\cdot)$ is **pointwise mutual information** between two values under a joint distribution:

$$\text{pmi}(x;y) = h(x) + h(y) - h(x,y). \quad (5)$$

Essentially, each word has an inherent cost determined by its log inverse probability, mitigated to the extent that it is predictable from context ($\text{pmi}(w_i; w_{1:i-1})$).

Now define the **interaction information** between a sequence of m values $\{a\}$ drawn from a sequence of m random variables $\{\alpha\}$ (McGill, 1955; Bell, 2003) as:[2]

$$i(a_1;...;a_m) = \sum_{n=1}^{m} \sum_{I \in \binom{1:m}{n}} (-1)^{m-n-1} h(a_{I_1},...,a_{I_n}),$$

where the notation $\binom{1:m}{n}$ means all cardinality-n subsets of the set of integers 1 through m. The equation amounts to alternately adding and subtracting the joint surprisals of all subsets of values. For $m = 2$, expanding the equation reveals that mutual information is a special case of interaction information.

Supposing that the noisy representation of context V is the result of running the veridical context $w_{1:i-1}$ through progressive erasure noise, we can see V as a sequence of values $v_{1:i-1}$, where each v_i is equal to either w_i or the erasure symbol $\texttt{E}$. Rewriting $\text{pmi}(w_i;V)$ as $\text{pmi}(w_i; v_{1:i-1})$, we can decompose it into interaction informations as follows:

$$\text{pmi}(w_i; v_{1:i-1}) = \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(w_i; v_{I_1};...;v_{I_n}). \quad (6)$$

The equation expresses a sum of interaction informations between the current word w_i and all subsets of the context values.[3]

[2] Higher-order information terms are typically defined using a different sign convention and referred to as **coinformation** or **multivariate mutual information** (Bell, 2003). For even orders, interaction information is equal to coinformation. For odd orders, interaction information is equal to negative coinformation. We adopt our particular sign convention to make the generalization of information locality easier to express.

[3] To see that this is true, first note that we can express joint surprisal in terms of interaction information:

$$h(a_1,...,a_m) = -\sum_{n=1}^{m} \sum_{I \in \binom{1:m}{n}} i(a_{I_1};...;a_{I_n}).$$

Now consider the pmi of a value a_i with a sequence $a_{1:i-1}$. Using the decomposition of joint surprisal to expand the definition of pmi in Equation 5, we get:

$$\begin{aligned} \text{pmi}(a_i; a_{1:i-1}) &= h(a_i) + h(a_{1:i-1}) - h(a_i, a_{1:i-1}) \\ &= h(a_i) + h(a_{1:i-1}) - h(a_{1:i}) \\ &= h(a_i) - \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(a_{I_1};...;a_{I_n}) \\ &\quad + \sum_{n=1}^{i} \sum_{I \in \binom{1:i}{n}} i(a_{I_1};...;a_{I_n}) \end{aligned}$$

In the final expression, all the terms that do not contain a_i

Now combining Equations 4 and 6, we get:

$$C(w_i|w_{1:i-1}) = h(w_i) - \mathop{\mathbb{E}}_{v|w_{1:i-1}} \left[\sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(w_i; v_{I_1}; ...; v_{I_n}) \right]$$

$$= h(w_i) - \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} \sum_{v} p_N(v|w_{i:i-1}) i(w_i; v_{I_1}; ...; v_{I_n}).$$

Now if any element of an interaction information term is E, then that whole interaction information term is equal to 0. This happens because the probability that an element is erased is independent of the identity of other elements in the sequence, and thus E has no interaction information with any subset of those elements. That is, $i(w_i; v_{I_1}; ...; v_{I_n}) = 0$ unless $v_{I_j} = w_{I_j}$ for all j. This allows us to write:

$$C(w_i|w_{1:i-1}) = h(w_i) - \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(w_i; w_{I_1}; ...; w_{I_n}) \sum_{m \in \{0,1\}^{i-1}} p_N(m) m_I$$

where the variable m ranges over bit-masks of length $i-1$, and m_I is equal to 1 when all indices I in m are equal to 1, and 0 otherwise. Now $\sum_{m \in \{0,1\}^{i-1}} p_N(m) m_I$ is the total probability that all of a set of indices I survives erasure. Thus, informally:

$$C(w_i|w_{1:i-1}) = h(w_i) - \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} p_N(I \text{ survives}) i(w_i; w_{I_1}; ...; w_{I_n}). \quad (7)$$

That is, the cost of a word is its inherent cost minus its interaction informations with context, which are weighted by the probability that all elements of those interactions survive erasure.

cancel out, leaving:

$$\text{pmi}(a_i; a_{1:i-1}) = h(a_i) + \sum_{n=0}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(a_i; a_{I_1}; ...; a_{I_n})$$

$$= h(a_i) + \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(a_i; a_{I_1}; ...; a_{I_n}) - h(a_i)$$

$$= \sum_{n=1}^{i-1} \sum_{I \in \binom{1:i-1}{n}} i(a_i; a_{I_1}; ...; a_{I_n}),$$

which gives Equation 6 when applied to w_i and $v_{1:i-1}$.

Under progressive erasure noise, the probability that a subset of variables is erased increases the farther left those variables are in the context. Therefore, Equation 7 expresses information locality: context elements which are predictive of w_i will only get to mitigate the cost of processing w_i if they are close to it. The surprisal-mitigating effect of a context element on a word w_i decreases as that element gets farther from w_i.

4.2 Noisy Surprisal and Dependency Locality

Memory-based models of sentence processing account for apparent **dependency locality effects**, which is processing cost apparently arising from two words linked in a syntactic dependency appearing far from one another (Gibson, 1998). Dependency length has been proposed as a rough measure of comprehension and production difficulty, and studied as a predictor of reaction times (Grodner and Gibson, 2005; Demberg and Keller, 2008; Mitchell et al., 2010; Shain et al., 2016), and also as a theory of production preferences and linguistic typology, under the assumption that people prefer to produce sentences with short dependencies (dependency length minimization) (Hawkins, 1994; Gildea and Temperley, 2010; Futrell et al., 2015; Rajkumar et al., 2016).

Dependency locality follows from information locality if words linked in a syntactic dependency have particularly high mutual information. To see this, consider only the lowest-order interaction information terms in Equation 7, truncating the summation over n at 1. We can write

$$C(w_i|w_{1:i-1}) = h(w_i) - \sum_{j=1}^{i-1} f(i-j)\text{pmi}(w_i; w_j) + R,$$

where R collects all the interaction information terms of order greater than 2, and $f(d)$ is the monotonically decreasing survival probability of a d-back word, described in Section 4.1.1. The effects of R are bounded because higher-order mutual information terms are more penalized by erasure noise than lower-order terms, simply because large sets of context items are more likely to experience at least one erasure.

If the effects of R are negligible, then the cost of a whole utterance w as a function of word order is determined only by pairwise information locality:

$$C(w) \approx \sum_{i=1}^{|w|} h(w_i) - \sum_{i=2}^{|w|} \sum_{j=1}^{i-1} f(i-j)\text{pmi}(w_i; w_j).$$

If words linked in a dependency have higher mutual information than words that are not, then the processing cost as a function of word order is a monotonically increasing function of dependency length. Under this assumption, for which we provide evidence below, dependency locality effects can be seen as a special case of information locality effects. As a theory of production preferences or typology, processing cost as a monotonically increasing function of dependency length suffices to derive the predictions of dependency length minimization (Ferrer i Cancho, 2015).

4.3 Mutual Information and Syntactic Dependency

We have shown that noisy-context surprisal derives information locality, and argued that dependency locality can be seen as a special case of information locality. However, deriving dependency locality requires a crucial assumption that words linked in a dependency have higher mutual information than those words that are not.

To test this assumption, we calculated mutual information between wordforms in various dependency relations in the Google Syntactic n-gram corpus (Goldberg and Orwant, 2013). We compared the mutual information of content words in a direct dependency relationship to content words in grandparent–grandchild and sister–sister dependency relationships. Mutual information was estimated using maximum likelihood estimation from frequencies, treating the corpus as samples from a distribution over (head, dependent) pairs. In order to exclude nonlinguistic forms, we only included wordforms if they were among the top 10000 most frequent wordforms in the corpus. The direct head–dependent frequencies were calculated from the same corpus as the grandparent-grandchild frequencies, so that all mutual information estimates are affected by the same frequency cutoff. The results are shown in Table 2: direct head–dependent pairs indeed have the highest mutual information.

To test the crosslinguistic validity of this generalization about syntactic dependency and mutual information, we calculated mutual information between the distributions over POS tags for dependency pairs of 43 languages in the Universal Dependencies corpus (Nivre et al., 2016). For this calculation, we used mutual information over POS tags rather than wordforms to avoid data sparsity issues. The results are shown in Figure 3.

Relation	MI (bits)
Head–dependent	1.79
Grandparent–dependent	1.34
Sister–sister	1.19

Table 2: Mutual information over wordforms in different dependency relations in the Syntactic n-gram corpus. The pairwise comparison of head–dependent and grandparent–dependent MI is significant at $p < 0.005$ by Monte Carlo permutation tests over n-grams with 500 samples. The comparison of head–dependent and sister–sister MI is not significant.

Again, we find that mutual information is highest for direct head–dependency pairs, and falls off for more distant relations. These results show that two words in a syntactic dependency relationship are more predictive of each other than two words in some other kinds of relationship.

We also compared the mutual information of word pairs in and out of dependency relationships while controlling for distance. This test has a dual purpose. First, it allows us to control for distance when claiming that words in dependency relationships have high mutual information. Second, it allows us to test a simple prediction of information locality as applied to language production: that words with high mutual information should be close together. For pairs of words (w_i, w_{i+k}), we calculated the pmi values among POS tags of the words. Figure 4 shows the average pmi of all words at each distance compared with the average pmi of the subset of words in a direct dependency relationship at that distance. In all languages, we find that words in a dependency relationship have higher pmi than the baseline, especially at close distances. Furthermore, we find that words at close distances tend to have higher pmi, regardless of whether they are in a dependency relationship.

4.4 Discussion

Information locality can be seen as a decay in the effectiveness of contextual cues for predicting words. Precisely such a decay in cue effectiveness was found to be effective for predicting entropy distributions across sentences in Qian and Jaeger (2012), although that work did not distinguish between an inherent, noise-based decay in cue effectiveness or optimized placement of cues.

Figure 3: Mutual information over POS tags for dependency relations in the Universal Dependencies 1.4 corpus, for languages with over 500 sentences. All pairwise MI comparisons are significant at $p < 0.005$ by Monte Carlo permutation tests over dependency observations with 500 samples.

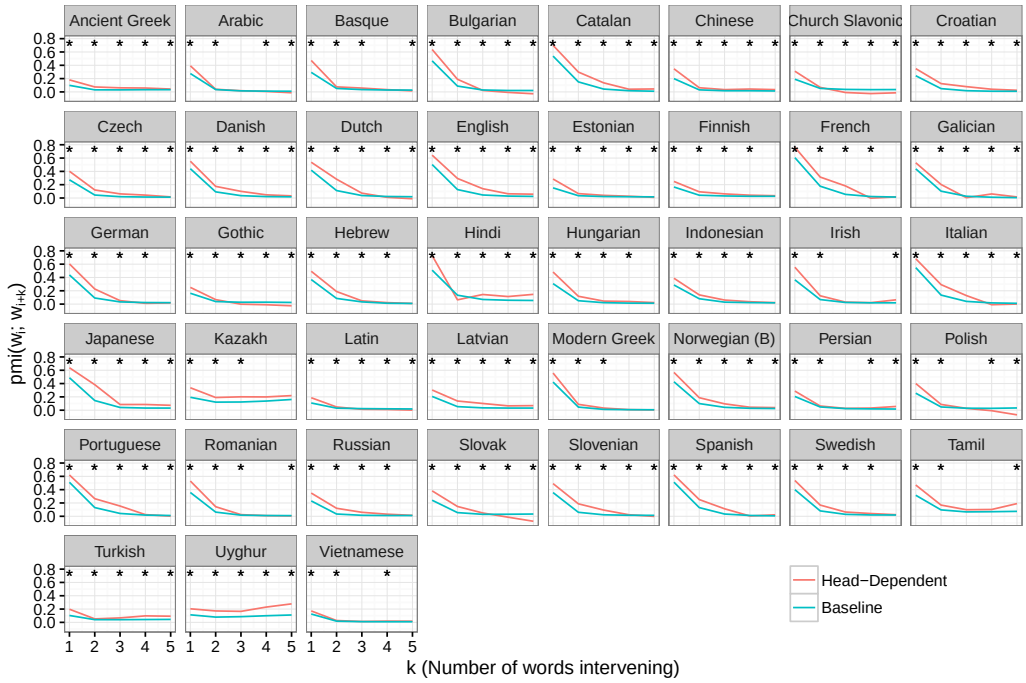

Figure 4: Average pointwise mutual information over POS tags for word pairs with k words intervening, for all words (baseline) and for words in a direct dependency relationship. Asterisks mark distances where the difference between the baseline and words in a dependency relationship is significant at $p < 0.005$ by Monte Carlo permutation tests over word pair observations with 500 samples.

The result of Gildea and Jaeger (2015), which shows that word orders in languages are optimized to minimize trigram surprisal of words, can be taken to show maximization of information locality under the noise distribution where context is truncated deterministically at length 2. Whereas Gildea and Jaeger (2015) treat dependency length minimization and trigram surprisal minimization as separate factors, under the view in this paper these two phenomena emerge as two aspects of information locality. In general, the mutual information of linguistic elements has been found to decrease with distance (Li, 1989; Lin and Tegmark, 2016), although this claim has only been tested for letters, not for larger linguistic units such as morphemes. The fact that linguistic units that are close typically have high mutual information could result from optimization of word order for information locality.

The idea that syntactically dependent words have high mutual information is also ubiquitously implicit in probabilistic models of language and in practical NLP models. For example, it is implied by head-outward generative models (Eisner, 1996; Eisner, 1997; Klein and Manning, 2004), the first successful models for grammar induction. Mutual information has been used directly for unsupervised discovery of syntactic dependencies (Yuret, 1998) and evaluation of dependency parses (de Paiva Alves, 1996), as well as commonly for collocation detection (Church and Hanks, 1990). In addition to providing evidence for a crucial assumption in the derivation of information locality, our results also give evidence backing up the theoretical validity of such models and methods.

The derivation of information locality given here assumed progressive erasure noise for concreteness, but we believe it should be possible to derive this generalization for a large family of noise distributions.

5 Conclusion

We have introduced a computational-level model of incremental sentence processing difficulty based on the principle that comprehenders have uncertainty about the previous input and act rationally on that uncertainty. Noisy-context surprisal accounts for key effects predicted by expectation-based and memory-based models, in addition to providing the first computational-level explanation of language-specific structural forgetting, which involves subtle interactions between memory and probabilistic expectations. Noisy-context surprisal also leads to a general principle of information locality offering a new interpretation of syntactic locality effects, and leading to broader and potentially different predictions than purely memory-based models.

Here we have used qualitative arguments and have used different specific noise distributions to make different points. Our aim has been to argue for the theoretical viability of noisy-context surprisal, without committing the theory to a particular noise distribution. We believe our predictions will be derivable under very general classes of noise distributions, and we plan to pursue these more general derivations in future work.

A more psychologically accurate model will likely use a more nuanced noise distribution than the simple decay functions in this paper, which do not capture the subtleties of human memory. In particular, simple decay functions to not capture memory retrieval effects of the kind described in Anderson and Schooler (1991), where different items in a sequence have different propensities to be forgotten, in accordance with rational allocation of resources for retrieval. Seen as a noise distribution, this memory model implies that the erasure probability of a word is a function of the word's identity, and not only the word's position in the sequence as in Section 4.1.1. Including such noise distributions in the noisy-context surprisal model could provide a rich set of predictions to test the model more extensively.

Acknowledgments

We would like to thank members of Tedlab and the Computational Psycholinguistics Lab at MIT for helpful comments. R.F. was supported by NSF grant #1551543.

References

John R. Anderson and Lael J. Schooler. 1991. Reflections of the environment in memory. *Psychological Science*, 2(6):396.

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

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

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

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

Vera Demberg and Frank Keller. 2009. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, The Netherlands. Cognitive Science Society.

Vera Demberg, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

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

Jason M. Eisner. 1997. An empirical comparison of probability models for dependency grammar. Technical report, IRCS Report 96–11, University of Pennsylvania.

Ramon Ferrer i Cancho. 2015. The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, 5(1):114–137.

Stefan L. Frank, Thijs Trompenaars, Richard L. Lewis, and Shravan Vasishth. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, 40:554–578.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*.

Edward Gibson and James Thomas. 1999. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248.

Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological science*, 24(7):1079–1088.

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

Daniel Gildea and T. Florian Jaeger. 2015. Human languages order information efficiently. *arXiv*, abs/1510.02823.

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

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

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

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

John A. Hawkins. 1994. *A performance theory of order and constituency.* Cambridge University Press, Cambridge.

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

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

Roger Levy. 2008b. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 234–243.

Roger Levy. 2011. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *ACL*, pages 1055–1065.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Wentian Li. 1989. Mutual information functions of natural language texts. Technical report, Santa Fe Institute Working Paper #1989-10-008.

Henry W. Lin and Max Tegmark. 2016. Critical behavior from deep dynamics: A hidden dimension in natural language. *arXiv*, abs/1606.06737.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman & Company.

William J. McGill. 1955. Multivariate information transmission. *IEEE Transactions on Information Theory*, 4(4):93–111.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague.

Ting Qian and T. Florian Jaeger. 2012. Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, 36:1312–1336.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, 155:204–232.

Douglas Roland, Frederic Dick, and Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka, Japan.

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

Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Deniz Yuret. 1998. Discovery of linguistic relations using lexical attraction. *arXiv preprint cmp-lg/9805009*.