Cue Effectiveness in Communicatively Efficient Discourse Production

Ting Qian, a T. Florian Jaeger a,b

a Department of Brain and Cognitive Sciences, University of Rochester
b Department of Computer Science, University of Rochester

Received 25 December 2010; received in revised form 21 September 2011; accepted 8 November 2011

Abstract

Recent years have seen a surge in accounts motivated by information theory that consider language production to be partially driven by a preference for communicative efficiency. Evidence from discourse production (i.e., production beyond the sentence level) has been argued to suggest that speakers distribute information across discourse so as to hold the conditional per-word entropy associated with each word constant, which would facilitate efficient information transfer (Genzel & Charniak, 2002). This hypothesis implies that the conditional (contextualized) probabilities of linguistic units affect speakers’ preferences during production. Here, we extend this work in two ways. First, we explore how preceding cues are integrated into contextualized probabilities, a question which so far has received little to no attention. Specifically, we investigate how a cue’s maximal informativity about upcoming words (the cue’s effectiveness) decays as a function of the cue’s recency. Based on properties of linguistic discourses as well as properties of human memory, we analytically derive a model of cue effectiveness decay and evaluate it against cross-linguistic data from 12 languages. Second, we relate the information theoretic accounts of discourse production to well-established mechanistic (activation-based) accounts: We relate contextualized probability distributions over words to their relative activation in a lexical network given preceding discourse.

Keywords: Language; Cue informativity; Cue effectiveness decay; Entropy; Communicative efficiency; Constant entropy rate; Cross-linguistic

1. Introduction

Language can be viewed as a shared code used to communicate with each other. Researchers have noted for some time that human languages have properties that are particularly suited for communication. For example, Zipf found that more frequently used words

Correspondence should be sent to Ting Qian, Department of Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14627-0268. E-mail: tqian@bcs.rochester.edu
had on average fewer syllables and fewer phonemes (e.g., Kaeding, 1898 for German; Zipf, 1935 for Chinese, English, and Latin). One interpretation of this correlation—the one favored by Zipf himself—is that high frequency causes word brevity. In other words, the frequently used word forms in the mental lexicon become shorter over time so that speakers will spend less effort in uttering them and communication will thus be more efficient (for support, see e.g., Bybee, 2002; Fidelholtz, 1975; Phillips, 1984; Schuchardt, 1885; Zipf, 1929, 1935; see Kuperman, Pluymaekers, Ernestus, & Baayen, 2007 for relevant discussion). If indeed language usage shapes the diachronic trajectory of natural languages, we might expect language to exhibit properties of an ideal code, defined according to the principles of efficiency as Zipf suggested. Recent accounts in this spirit have drawn on Shannon’s information theory to develop models of such an ideal code based on first principles. Focusing on discourse production, Genzel and Charniak (2002) presented an account in this vein. Based on the famous noisy channel theorem (Shannon, 1948), which states that information generated by a symbolic source can be transferred through a noisy channel at an arbitrarily low error rate up to a maximum transmission rate defined as the channel’s capacity, Genzel and Charniak (2002) derived that an ideal code should keep the average amount of information conveyed per word constant across discourses (for details, see below).

This has spawned a line of work over the last decade, which has provided preliminary evidence compatible with the hypothesis. However, this line of work (including our own previous work) has arguably suffered from two major shortcomings. First, previous investigations implicitly assumed that preceding discourse context can be arbitrarily informative about upcoming words. This assumption, we argue below, is at odds with what is known about human memory as well as the topic structure of discourses. We propose an alternative model of cue effectiveness and test it against data from 12 languages. Second, it has so far remained unclear how the concepts evoked in the information theoretic account by Genzel and Charniak (2002) relate to the activation-based architecture underlying most contemporary mechanistic models of language production (Dell, 1986; Dell, Chang, & Griffin, 1999; Levelt, Roelofs, & Meyer, 1999; Mackay, 1982; Roelofs, 2002; Stemberger, 1985). Here, we aim to address these two points. We start by laying out the motivation for Genzel and Charniak’s so-called Constant Entropy Rate (CER) hypothesis.

1.1. Information theory and ideal codes

Shannon characterized an efficient communication system as one under which the rate of information transmission is maximized (i.e., information transmission per unit time is maximized). In the case of language communication, maximal efficiency can be achieved when the transmission rate of linguistic signals is relatively uniform and close to the channel capacity. A number of studies have tested to what extent language and language use exhibit properties that are predicted by this theorem (e.g., Aylett & Turk, 2004; Fenk-Oczlon, 2001; Ferrer i Cancho & Díaz-Guilera, 2007; Genzel & Charniak, 2002; Jaeger, 2006, 2010; Levy & Jaeger, 2007; Manin, 2006; Piantadosi, Tily, & Gibson, 2011; van Son & Pols, 2003). In fact, the aforementioned findings from Zipf can be restated in information theoretic terms. Consider the unconditional (out of context) information of a word, I(word), defined
by Shannon as the logarithm of the reciprocal of its probability: \( I(\text{word}) = \log \frac{1}{p(\text{word})} = -\log p(\text{word}) \) (Shannon, 1948). Hence, what Zipf observed is the phenomenon that longer words, on average, carry more unconditional information. Put differently, if the mental lexicon is viewed as a mixture of word distributions based on word length, then the distributions of long words tend to have higher unconditional per-word entropy than those of short words (the unconditional per-word entropy is the expected unconditional information of all words in a distribution, \( \Sigma p(\text{word}) I(\text{word}) \)). As a result, the entropy per letter (or per phoneme for speech) across words in the mental lexicon could be relatively constant, matching the pattern predicted by the noisy channel theorem.

While Zipf’s work and many others following have focused on words (e.g., Ferrer i Cancho, 2005a,b; Manin, 2006; Piantadosi, Tily, & Gibson, 2011), research over recent years has begun to investigate whether a principle of constant entropy affects language production beyond the word. Genzel and Charniak (2002) propose that constant entropy may be observed across sentences throughout a discourse. According to their CER hypothesis, the per-word entropy of the first sentence position \( H(S_1) \) should equal that of the second sentence position in the same discourse \( H(S_2|c_2 = \{s_1\}) \), which is conditioned on the content of the preceding discourse \( c_2 = \{s_1\} \). In other words, the discourse content at the second sentence position is expected to be as informative as that at the first one per time unit as measured by words. In general, for sentence positions \( i \) and \( j \), the hypothesis predicts that the conditional per-word entropies of both sentence positions are the same:

\[
H(S_i|c_i = \{s_1 \ldots s_{i-1}\}) = H(S_j|c_j = \{s_1 \ldots s_{j-1}\})
\]  

Unfortunately, a direct test of the prediction as outlined in Eq. 1 is difficult with current technology and available language resources. The difficulty lies in deriving sentence probability estimates that are conditioned on the preceding discourse context as well as other knowledge that the speaker has access to (e.g., world knowledge). Although considerable progress has been made to tackle this notoriously hard problem (Blei, Ng, & Jordan, 2003; Gildea & Hofmann, 1999; Griffiths & Steyvers, 2004; Kuhn & Mori, 1990; Seymore & Rosenfeld, 1997) models that would be sufficient to reliably test the direct prediction of the CER hypothesis are still difficult to obtain. A weaker but more feasible indirect test of the CER hypothesis has received support from a number of studies (Genzel & Charniak, 2002, 2003; Keller, 2004; Piantadosi & Gibson, 2008; Qian, 2009; Qian & Jaeger, 2009, 2010). If the true per-word entropy conditioned on world knowledge, contextual knowledge, and the preceding linguistic discourse stays constant throughout discourse, then estimates of the per-word entropy conditioned on only sentence-internal cues (henceforth unconditional per-word entropy estimates) should increase throughout discourse:

\[
H(S_j) > H(S_i), \quad \text{when} \quad j > i
\]  

The prediction in Eq. 2 follows from Eq. 1 given the well-known fact that the probabilities of linguistic units (e.g., sentences) depend on the preceding discourse and this dependency persists between arbitrarily distant linguistic units (i.e., language exhibits long-range
correlations, cf. Moscoso Del Prado Martín, submitted). Thus, the conditional per-word entropy of a sentence position, which measures how much uncertainty is associated with that position given preceding cues, must be lower than its unconditional per-word entropy that is \( H(S_k) > H(S_k|c_k) \) for any sentence position \( k \). In other words, the unconditional per-word entropy of sentences must increase with discourse progress.

Indeed, this prediction has received support from a number of studies on written and spoken language (Genzel & Charniak, 2002, 2003; Keller, 2004; Piantadosi & Gibson, 2008; Qian, 2009; Qian & Jaeger, 2009, 2010). However, most of this work has tested a more specific, and in the limit ill-formed, variant of the indirect prediction of CER: This work has employed linear regression to test the indirect prediction (Genzel & Charniak, 2002, 2003; Keller, 2004). By assuming a linear relation between unconditional per-word entropy and sentence position, those studies effectively state that each past contextual cue (i.e., each past sentence in the current case) is equally informative in predicting the upcoming discourse content, and thus the cumulative discourse informativity increases linearly. However, as we detail next, this assumption, which we will refer to as the no-decay assumption, is unrealistic.

### 1.2. Bounds on the cumulative informativeness of preceding context

One of the reasons that the no-decay assumption is unrealistic is that the unconditional per-word entropy of a sentence position must be bounded by \( \log \lambda \), where \( \lambda \) is the number of sentences in the collected sample (assuming every sentence is equiprobable, which maximizes the total entropy) (for a similar argument, see Moscoso Del Prado Martín, submitted). Based on this bound on the unconditional per-word entropy and the fact that language exhibits long-range correlations, the indirect prediction of the CER hypothesis is that unconditional per-word entropy should increase with increasing sentence position, while never exceeding a certain bound.

Aside from this formal argument, there are at least two additional a priori reasons to expect that the unconditional per-word entropy should increase at a rate that is less than linear (i.e., sublinear) over longer stretches of discourse. The first argument is built on considerations about human memory, the second on considerations about discourse structure. Considerations about human memory suggest cognitive limitations might cause speakers to rely less on more distant cues. It is well known that memory is subject to decay and interference (e.g., Baddeley, 2000). Processing models that incorporate decay and interference have been shown to provide a good fit against data from incremental sentence production and comprehension (Badecker & Lewis, 2007; Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006). It is possible that memory decay and interference reduces the effectiveness of distant cues differentially as a function of distance such that more distance cues are much less effective. As a result, the cumulative informativity of preceding discourse is expected to increase only sublinearly, possibly converging against an upper bound. If so, the CER hypothesis predicts a sublinear increase in the unconditional per-word entropy of sentence positions.

Second, sublinearly increasing cumulative informativity is also expected based on considerations about the structure of natural discourses. Linguistic discourses typically span several topics or subtopics (or questions under discussion, Roberts, 1996). Information
provided during the discussion of one topic A will usually be less effective in predicting words that are uttered as part of a later topic B. Given that speakers typically create locally coherent discourses, where adjacent sentences provide information about the same topic, more recent cues are on average likely to be more effective than distant cues in predicting the next sentence (see also Piantadosi & Gibson, 2008). Hence, even if memory does not constrain the effectiveness of preceding cues, discourse structure usually will. Both discourse structure and possible limitations of human memory hence lead us to expect a sublinear increase in the cumulative discourse informativity as well as in the unconditional per-word entropy of sentence positions.

In an attempt to put the CER hypothesis to a cognitively more plausible test, we revise its prediction based on the Cue Effectiveness Decay hypothesis, which states that more distant cues, on average, are less informative about the identity of upcoming words. We analytically derive the expected pattern of unconditional discourse entropy under the assumption of power-law decay, which is independently motivated by previous work (e.g., Anderson & Paulson, 1977; Squire, 1989; Wixted & Ebbesen, 1991). We find that this cognitively more plausible model describes the distribution of information across discourse positions significantly better than models based on the no-decay assumption presented in previous work. This finding has consequences for work on efficient communication, such as the CER hypothesis, but also for work on cue integration during language production.

1.3. Cue effectiveness decay

Each linguistic unit in a discourse can be considered a cue to upcoming linguistic units. To make our results comparable to previous work on CER, we consider words as both the unit we want to predict and the unit that constitutes a cue. We can then think of the cumulative informativity of a preceding linguistic discourse (the cumulative discourse informativity), as the sum of all the information provided by the words in that discourse. For the reasons given above, we hypothesize that words that have been uttered further in the past are on average less effective in predicting upcoming material than more recent words. We will refer to this as the Cue Effectiveness Decay hypothesis.

Consider the following example. The cumulative discourse informativity at the \(k\)th sentence position is attributable to the \(1 \ldots k - 1\) sentences in the preceding discourse. However, these sentences are likely to vary in their effectiveness in predicting the \(k\)th sentence. If \(k\) is large enough, context from the early sentences \(1 \ldots i\) \((i \ll k)\) bears little-to-none weight. This could be due to the loss of topic informativity or due to memory limitations, as discussed above. Meanwhile, the nearby \(k - i\) sentences are more effective in predicting the \(k\)th sentence. In other words, the average uncertainty (i.e., entropy) associated with the upcoming content at sentence position \(k\) is mostly resolved by the nearby window of \(k - i\) sentences. Then if we look ahead to the \((k + 1)\)th sentence position, there are more of the early sentences that have become less effective in predicting upcoming content, and effective cues at that point also come from the nearby window of \(k - i\) sentences (i.e., \((k + 1) - (i + 1) = k - i\)). Therefore, for later sentence positions, on average, the amount of effective context stays relatively the same. On the other hand, when \(k\) is fairly small, to the extent that the
If the cumulative effectiveness of discourse cues increases sublinearly, we should also expect that the rate of increase in unconditional per-word entropy is sublinear as well. This is because speakers can only introduce as much new information as allowed by discourse context (a formal argument is presented in the next section and in Appendix A). Fig. 1 illustrates the expected sublinear correlation between sentence position and unconditional per-word entropy (we will use the term entropy profile to refer to the distribution of unconditional per-word entropy throughout the discourse).

In previous work, we have found initial evidence consistent with the cue effectiveness decay hypothesis (Qian & Jaeger, 2009; see also Piantadosi & Gibson, 2008). Below, we derive a precisely quantified prediction about the shape of this sublinear increase based on the assumption that the effectiveness of cues decreases as a power-law with increasing distance (see Piantadosi & Gibson, 2008, for similar idea). The power-law function has been found to describe behaviors ranging from learning (e.g., Newell & Rosenbloom, 1981; Pirolli & Anderson, 1985) to memory retention (e.g., Anderson & Paulson, 1977; Squire, 1989; Wixted & Ebbesen, 1991). One study of interest is conducted by Anderson and Paulson (1977), who exposed participants to stimuli and tested how retrieval times changed depending on the amount of time that had passed between exposure to the stimuli and the recognition test. They found that retrieval times were well described by a power-law function. Since we assume that memory limitations are operative in discourse production, it is natural to use the power-law decay function as a first step in the

Fig. 1. Unconditional per-word entropy profile predicted by CER hypothesis under the hypothesis of cue effectiveness decay. Unconditional per-word entropy is expected to increase sublinearly throughout the discourse.
exploration. Another finding of interest is that the re-occurrence probabilities of words is well described by power-law decay over the time since the last mention (Altmann, Pierrehumbert, & Motter, 2009). In other words, the self-cuing of a word is well described by a power-law function. Altmann and colleagues show that this holds across words with a wide range of dispersion indices, across different genres, and across several orders of magnitude in word distances. This provides additional motivation for the power-law decay function for the current purpose.

1.4. A power-law decay model of cue effectiveness

We assume that each preceding sentence constitutes a contextual cue to upcoming material. The informativity of a discourse contextual cue is defined in the same unit as entropy—bits per word. Let \( r_0 \) denote the conditional per-word entropy of sentence positions, a constant under the assumption of CER. At the same time, \( r_0 \) also describes how much uncertainty is associated with any sentence position if context is considered. In successful communication, the speaker’s utterances at the \( k \)th sentence position must resolve the uncertainty of position \( k \). In other words, the informativity of any discourse content is always \( r_0 \) bits at the sentence position where it is first uttered. At the forthcoming sentence positions, this discourse content has become context by definition and their predictive effectiveness decays. We consider the power-law decay function based on previous work on cue effectiveness decay in memory. The claim here is not that power-law decay is required to obtain the results reported below (we also modeled the data with exponential decay functions, which yield qualitatively similar data). Rather, the hypothesis we set out to test is whether a model of entropy profiles fit under reasonable assumptions about cue decay provides a better fit to language data than a model with constant cue effectiveness. This would provide evidence that cue effectiveness does decay.

If a contextual cue \( q \) is originally at position \( k_q \), and its effectiveness decays at the rate following the power function, its remaining effectiveness at an upcoming sentence position \( k \) is:

\[
effectiveness_{pow}(k, k_q) = (k - k_q + 1)^{-\lambda}
\]

In Eq. 3, \( \lambda \) is the decay rate. This function can be most intuitively understood as a function of \( k_q \): For instance, holding position \( k \) fixed, the cue from the \((k - 1)\)th sentence has \(2^{-\lambda}\) percent of its original amount of information (i.e., informativity). Given that all cues have \( r_0 \) bits of information at their original positions, the cumulative discourse informativity at position \( k \) is the weighted sum of context contributed by all cues up to sentence position \( k - 1 \):

\[
context_{pow}(k) = r_0 \sum_{q_i \in \{q_1, \ldots, q_{k-1}\}} (k - k_{q_i} + 1)^{-\lambda}
\]

Eq. 4 is a sum of power-law functions. To convert Eq. 4 to a nonlinear model that can be automatically fit, we need to rewrite it in a closed form as a function of \( k \). However, without knowing the rate \( \lambda \), which in turn has to be estimated from the data after the closed form is
derived, this is not a trivial task. As a workaround, we approximated the value of Eq. 4 by computing a definite integral of Eq. 3, where \( \Delta i \) is a shorthand for \( k - k_q + 1 \):

\[
\text{context}^{\text{pow}}(k) \approx \int_1^k r_0 \Delta i^{-\lambda} d\Delta i
\]

\[
= r_0 \left( \frac{k^{1-\lambda} - 1}{1-\lambda} \right)
\]  

Eq. 5 measures the cumulative discourse informativity at sentence position \( k \). This quantity has an interesting interpretation (proof in Appendix A): When averaged over all sentence tokens in our data (see Method), the cumulative discourse informativity, plus the conditional per-word entropy of sentence positions, defines the upper bound for the unconditional per-word entropy for the \( k \)th sentence position:

\[
\text{r}^{\text{pow}}(k) = \text{context}(k) + r_0
\]

\[
= r_0 \left( \frac{k^{1-\lambda} - 1}{1-\lambda} \right) + r_0
\]  

Whether speakers will utilize all available context as predicted by Eq. 6 is another question. Here we hypothesize that speakers are maximally efficient in that they do take advantage of all available relevant context. Thus, the prediction is that the unconditional per-word entropy of sentence positions, as observed empirically from data, can be described by Eq. 6.

We evaluate the power-law decay model against 12 languages, including genetically and areally unrelated languages. To facilitate comparison with previous work, we also fit a no-decay model to the same data. By evaluating the CER hypothesis under assumptions about cue effectiveness decay (no decay vs. power law), this study simultaneously (a) assesses the validity of the CER hypothesis and (b) explores cue effectiveness decay as part of the computational nature of cue integration in language production.

2. Method

2.1. Data

To facilitate comparison across the 12 languages in our data set, we aimed to hold genre as constant as possible. We chose to use Reuters Corpus Volumes 1 and 2 as our data set. The corpus contains news stories in 14 languages. Because of lacking or inconsistent annotation, we excluded the data from three languages: Chinese, German, and Japanese. For Chinese, we substituted the Treebank Corpus (Xue, Xia, Chiou, & Palmer, 2005) for the Reuters data, leaving us with 12 languages, as shown in Table 1.
Following standards in research on natural language processing, the selected stories were divided into a training set (95% of all stories) for fitting language models and a test set (the remaining 5%) for estimating and analyzing entropy profiles. We selected only stories that were at least 15 sentences in length. Previous studies either did not control for article length (Genzel & Charniak, 2002; Keller, 2004) or used a smaller cut-off value, typically the first 10 sentences of an article (Genzel & Charniak, 2003; Qian & Jaeger, 2009). We chose a larger cut-off value to be able to investigate in more detail the development of unconditional per-word entropy throughout discourses, and in particular the shape of unconditional per-word entropy profiles. We also balanced the data by using only the first 15 sentences in either training or testing sets.

The language corpora differed significantly in size. For the larger corpora (English and Latin-American Spanish), we used only a subset of the available data for the sake of feasible computation. Table 1 lists the number of words and sentences in the training and test data for each of the 12 languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Training Data</th>
<th>Test Data</th>
<th>Per Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In Words</td>
<td>In Sentences</td>
<td>In Words</td>
</tr>
<tr>
<td>Danish</td>
<td>154,514</td>
<td>5,640</td>
<td>8,048</td>
</tr>
<tr>
<td>Dutch</td>
<td>50,309</td>
<td>3,255</td>
<td>2,105</td>
</tr>
<tr>
<td>English</td>
<td>597,698</td>
<td>23,295</td>
<td>31,276</td>
</tr>
<tr>
<td>French</td>
<td>229,461</td>
<td>9,300</td>
<td>11,371</td>
</tr>
<tr>
<td>Italian</td>
<td>97,198</td>
<td>4,245</td>
<td>4,524</td>
</tr>
<tr>
<td>Mandarin Chinese</td>
<td>145,127</td>
<td>4,875</td>
<td>4,310</td>
</tr>
<tr>
<td>Norwegian</td>
<td>89,724</td>
<td>4,125</td>
<td>2,973</td>
</tr>
<tr>
<td>Portuguese</td>
<td>170,342</td>
<td>5,340</td>
<td>9,044</td>
</tr>
<tr>
<td>Russian</td>
<td>398,786</td>
<td>18,075</td>
<td>20,668</td>
</tr>
<tr>
<td>Spanish (Latin-American)</td>
<td>613,874</td>
<td>22,160</td>
<td>27,802</td>
</tr>
<tr>
<td>Spanish (European)</td>
<td>255,366</td>
<td>7,485</td>
<td>8,653</td>
</tr>
<tr>
<td>Swedish</td>
<td>266,348</td>
<td>11,535</td>
<td>13,369</td>
</tr>
</tbody>
</table>

Note. The last column gives the number of sentences at each sentence position (which is identical to the number of documents contained in the language corpus).

<table>
<thead>
<tr>
<th></th>
<th>Training Data</th>
<th>Test Data</th>
<th>Per Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In Words</td>
<td>In Sentences</td>
<td>In Words</td>
</tr>
</tbody>
</table>

Following standards in research on natural language processing, the selected stories were divided into a training set (95% of all stories) for fitting language models and a test set (the remaining 5%) for estimating and analyzing entropy profiles. We selected only stories that were at least 15 sentences in length. Previous studies either did not control for article length (Genzel & Charniak, 2002; Keller, 2004) or used a smaller cut-off value, typically the first 10 sentences of an article (Genzel & Charniak, 2003; Qian & Jaeger, 2009). We chose a larger cut-off value to be able to investigate in more detail the development of unconditional per-word entropy throughout discourses, and in particular the shape of unconditional per-word entropy profiles. We also balanced the data by using only the first 15 sentences in either training or testing sets.

The language corpora differed significantly in size. For the larger corpora (English and Latin-American Spanish), we used only a subset of the available data for the sake of feasible computation. Table 1 lists the number of words and sentences in the training and test data for each of the 12 languages.

2.2. Per-word entropy of sentence position

To estimate the unconditional per-word entropy of each sentence position, we built separate ngram language models on each of the 12 training sets. Ngram models can estimate the probability of a word given the \( n - 1 \) preceding words. Trigram models were used in the current work, with Good-Turing smoothing and Katz back-off discounting (Jurafsky & Martin, 2008; Katz, 1987). Under the Markov assumption that trigrams of a sentence are independent of each other, we obtained the probability of a complete sentence \( s \) by multiplying all its trigram probabilities, where \( w_i \) is the \( i \)th word of \( s \):
\[ p(s) = \prod_{w_j \in s} p(w_i | w_{i-2}, w_{i-1}) \] (7)

This estimate is *unconditional* in the sense that no context information outside the scope of the sentence \( s \) is considered in computing its probability. To study how per-word entropy of a sentence position is predicted by discourse progress, we need to examine the relation between a sentence position and per-word information estimates of sentences at that position. The information of a sentence \( s \) in bits is:

\[ \text{info}(s) = - \log_2 p(s) = - \sum_{w_j \in s} \log_2 p(w_i | w_{i-2}, w_{i-1}) \] (8)

To get *per-word* information of a sentence \( s \), normalization is necessary.\(^2\) This is done by dividing sentence information by sentence length (measured in words):

\[ \text{PerWordInfo}(s) = \frac{1}{\text{length}(s)} \times \text{info}(s) \] (9)

Finally, we estimate a correlate of unconditional per-word entropy of a sentence position—the average per-word information of all sentence tokens at that position. This estimate will approach the true unconditional per-word entropy of that position in the limit of an infinite number of sentences in the database. This estimate of unconditional per-word entropy is implicitly calculated in regression analyses, which we describe as part of the studies.

### 2.3. Nonlinear mixed model analysis

To evaluate how well the power-law cue effectiveness decay model (i.e., Eq. 6 above) characterizes the relation between unconditional per-word entropy and sentence position, we built nonlinear mixed models with document-specific random effects, where the estimates of unconditional per-word entropy of sentence positions is regressed on sentence position (see Appendix B for the detailed specifications of the model). The model was fit to data with the *nlme* package from R (Pinheiro, Bates, DebRoy, & Sarkar, 2009). Known methods of fitting a nonlinear mixed model can be trapped in local maxima and are sensitive to the chosen starting values for the model’s parameters (i.e., \( r_0 \)s and \( \lambda \)s). There are also no principled methods for selecting these values optimally. We selected 6 for \( r_0 \) and 2 for \( \lambda \) as starting values for the power-law decay model, based on estimates observed in previous work. We also explored a range of other starting values to ensure that the results reported here are robust and guarantee convergence for our data sets.

### 3. Results

To understand how unconditional per-word entropy changes throughout discourses from the modeling results, we first need to solve for the derivative of the predicted unconditional
per-word entropy with respect to sentence position (the derivative of Eq. 6 with respect to \( k \)). This gives us:

\[
\hat{r}_{\text{power}}(k) = \hat{r}_0 k^{-\hat{\lambda}}
\] (10)

Therefore, as long as the estimates of \( \hat{\lambda} \) are significantly larger than 0, predicted unconditional per-word entropy will follow a negative acceleration pattern as its derivative is a monotonically decreasing power-law function. Table 2 lists the estimated decay rates of cue effectiveness for each language. Clearly, all decay rate estimates are significantly above 0 (on average, \( p < .01 \)), confirming the sublinear prediction of the cue effectiveness decay hypothesis. Fig. 2 illustrates the fit of the power-law decay model against Danish, Dutch, Mandarin Chinese, and Norwegian data.

Among all languages, the conditional per-word entropy of sentences, \( r_0 \), has a mean of 5.0 bits and a variance of 0.46 across languages. The similarity in \( r_0 \) between languages may lead one to speculate whether the amount of uncertainty per word in discourses is largely the same regardless of the actual language used by the speakers. However, precautions need to be taken in interpreting those estimates given that the corpora are of different sizes, and the ngram model is simplistic in nature.

### 3.1. Comparing decay and no-decay models

The results reported in Table 2 are promising, but to assess whether the cue effectiveness decay hypothesis indeed provides a better fit against the data, it is necessary to compare it

<table>
<thead>
<tr>
<th>Language</th>
<th>Avg. Number of Words (per Sentence)</th>
<th>Entropy ( (r_0) )</th>
<th>Rate ( (\hat{\lambda}) )</th>
<th>Remaining Cue Effectiveness at (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Next One Position Further Two Positions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Position Further</td>
</tr>
<tr>
<td>Danish</td>
<td>30</td>
<td>4.08</td>
<td>3.11</td>
<td>11.6 3.3 1.4</td>
</tr>
<tr>
<td>Dutch</td>
<td>23</td>
<td>4.90</td>
<td>3.27</td>
<td>10.4 2.8 1.1</td>
</tr>
<tr>
<td>English</td>
<td>27</td>
<td>5.88</td>
<td>9.97</td>
<td>0.1 0.0 0.0</td>
</tr>
<tr>
<td>French</td>
<td>26</td>
<td>4.65</td>
<td>3.56</td>
<td>8.5 2.0 0.7</td>
</tr>
<tr>
<td>Italian</td>
<td>20</td>
<td>4.49</td>
<td>3.29</td>
<td>10.2 2.7 1.0</td>
</tr>
<tr>
<td>Mandarin</td>
<td>29</td>
<td>6.20</td>
<td>3.70</td>
<td>7.7 1.7 0.6</td>
</tr>
<tr>
<td>Norwegian</td>
<td>20</td>
<td>4.39</td>
<td>2.40</td>
<td>18.9 7.1 3.6</td>
</tr>
<tr>
<td>Portuguese</td>
<td>38</td>
<td>4.98</td>
<td>4.18</td>
<td>5.5 1.0 0.3</td>
</tr>
<tr>
<td>Russian</td>
<td>22</td>
<td>4.48</td>
<td>2.98</td>
<td>12.7 3.8 1.6</td>
</tr>
<tr>
<td>E. Spanish</td>
<td>36</td>
<td>5.94</td>
<td>6.97</td>
<td>0.8 0.0 0.0</td>
</tr>
<tr>
<td>L. Spanish</td>
<td>33</td>
<td>5.17</td>
<td>5.21</td>
<td>2.7 0.3 0.1</td>
</tr>
<tr>
<td>Swedish</td>
<td>24</td>
<td>5.22</td>
<td>4.11</td>
<td>5.8 1.1 0.3</td>
</tr>
</tbody>
</table>

\textit{Note.} For example, in Danish, only 11.6% of the information contained in a contextual cue (i.e., a preceding sentence in this case) is effective in predicting the upcoming material at the next sentence position. The effectiveness further decays to 3.3% when the material to be predicted is one sentence position further. Estimates of conditional per-word entropy of sentences have a mean of 5.0 bits and vary relatively little between languages. The average number of words per sentence is also listed for each language.
directly against the no-decay model employed in most previous work (Genzel & Charniak, 2002, 2003; Keller, 2004), in which per-word information of sentences is linearly regressed on sentence position.

We note that both types of models have only two free parameters (the intercept and slope/decay parameters), so that the better fit of the power-law decay model is not attributable simply to differences in complexity between the models. Still, it is possible that a potential better fit of the decay model could be driven by any type of sublinear pattern in the entropy profiles. To address this possibility, we fit an alternative non-linear model to the data from each language that allows both sub- and superlinear patterns (or a mix of these patterns) without committing to any specific function. The nonlinear control model used here is a restricted cubic spline model (henceforth the RCS model). Unlike our power-law decay model, restricted cubic splines allow for local nonlinearities (e.g., local quadratic or cubic trends). Although the RCS model constitutes a relatively efficient way to model high degrees of nonlinearity, it requires more degrees of freedom than the linear predictors.³ Since more complex models require more data to avoid problems with overfitting, we did not fit the RCS model to five of the 12 languages (Dutch, Italian, Mandarin Chinese, and Norwegian).

Fig. 2. Predicted unconditional per-word entropy per word as a complex nonlinear function of sentence position based on the power-law decay model for Danish, Dutch, Mandarin Chinese, and Norwegian (blue curves). The actual distributions are shown by the barplots.
To account for the difference in complexity between the linear model, the power-law decay model, and the RCS model, the Bayesian Information Criterion (BIC) was selected as the measure for model comparison. The BIC adjusts models’ data likelihood (quality of fit) by complexity (number of parameters). Fig. 3 shows the difference BIC scores of the power-law decay, no-decay, and RCS models, separately by language, where smaller values of the BIC are better. Pair-wise comparisons over these BIC scores found that the power-law decay model is better than the no-decay model (paired Wilcoxon test: $V = 78, p < .001$), as shown in Fig. 3 (a table of raw BIC scores is provided in Supplementary Materials).

For seven out of the eight languages for which we were able to fit an RCS model, the power-law decay model results in lower BIC scores than the RCS models. The qualitative advantage of the power-law decay model does not reach significance (paired Wilcoxon test: $V = 7, p = .14$), probably partly due to the smaller set of languages that enter the comparison. Still, the results suggest that mere nonlinearity does not guarantee a better fit (see Supplementary Materials). There is one exception to this generalization (Swedish), which we discuss below.

The BIC weighs quality of fit against generalizability. However, the BIC assesses generalizability solely in terms of the number of a model’s parameters (the idea being that more complex models are less likely to generalize). This can be problematic since models with the same number of parameters can differ hugely in terms of their generalizability due to differences in their functional form (for discussion see Navarro, Pitt, & Myung, 2004). For this reason, we also compared the three types of models using leave-one-cluster-out cross-validation (reported in the Supplementary Materials). Cross-validation provides a computationally

---

Fig. 3. Models based on cue effectiveness decay yield superior BIC scores in most languages. Lower values are better (although BIC scores listed here can only be compared within but not between languages). The y-axis shows the centered BIC scores ($E.Spanish$ = European Spanish; $L.Spanish$ = Latin-American Spanish). BIC scores were centered within-language to facilitate comparison across languages. Three types of models are compared for each language: a linear no-decay model, the sublinear power-law decay model, and a restricted cubic splines (RCS) model capable of fitting both sub- and superlinear patterns. For four languages (Dutch, Italian, Mandarin, and Norwegian), we did not have enough data to fit an RCS model.
cheap way to consider differences in functional form during model comparison. The cross-validation results reported in the Supplementary Materials suggest that the functional forms of the power-law decay model is partially responsible for its superior performance effect. However, the cross-validation results also confirm that the power-law model is overall the best model for our data in terms of fit and generalizability.

4. General discussion

We have explored the development of unconditional per-word entropy throughout discourses (entropy profiles) across 12 languages and found a sublinear relation between unconditional per-word entropy and sentence position. This sublinear pattern is motivated by the assumption of the cue effectiveness decay hypothesis—the informativity of discourse contextual cues decays as a power function. The observed pattern is also consistent with the indirect prediction of the CER hypothesis (Genzel & Charniak, 2002), which assumes that speakers keep the amount of conditional per-word entropy associated with each word constant throughout discourses, thereby facilitating efficient communication.

Previous work on the CER hypothesis has focused on empirical tests of the hypothesis (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian, 2009; Qian & Jaeger, 2010, 2011). This work has accumulated a considerable body of evidence in favor of the CER hypothesis, and thereby also support for the hypothesis that the language production system is organized to facilitate efficient communication (e.g., Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). Previous work (including our own) has, however, not related these findings to models of language production. Before we discuss two possible interpretations of the observed entropy profiles in terms of theories of language production, we briefly address three important caveats that apply to our study.

4.1. Caveats

Several caveats apply to our studies. First, we have relied on written language data. Specifically, most of our data come from news reports, which are usually edited for clarity by professional journalists. This choice was made because at the time of this research we had no access to a comparable, typologically diverse but stylistically homogeneous spoken data set. The question thus arises whether such a writing style is representative of human language production in general. Encouraging evidence comes from previous work showing that the predicted increase in unconditional per-word entropy throughout discourses is observed in English and Chinese speech (Piantadosi & Gibson, 2008; Qian & Jaeger, 2009). Interestingly, both studies also found a sublinear trend in the observed increase in unconditional per-word entropy. The cue effectiveness decay hypothesis discussed here offers a tentative account for these speech data as well.

Second, like previous work, the cue effectiveness decay hypothesis is also an indirect prediction of CER (more broadly, of efficient communication). A stronger test of whether information is transmitted at a constant rate in language communication is a desirable next step.
Advances in natural language processing and machine learning can be used to derive improved estimates of conditional per-word entropy (e.g., topic models using Latent Dirichlet Allocation and other Bayesian methods: Blei, Griffiths, & Jordan, 2010; Blei et al., 2003; Canini, Shi, & Griffiths, 2009; Gildea & Hofmann, 1999; Tam & Schultz, 2005). While these language models outperform previous approaches, they are still unlikely to yield a sufficiently close estimate of conditional per-word entropy. Still, the prediction of CER is that conditional per-word entropy profiles based on such language models will have a smaller, but still positive, slope, as the effect of cue effectiveness decay is partially accounted for by these language models (see Qian & Jaeger, 2011 for preliminary evidence consistent with this prediction).

Finally, one of the 12 languages we investigated, Swedish, does not seem to follow the predictions of CER and the cue effectiveness decay hypothesis. Fig. 4 shows the restricted cubic spline fit against the Swedish data. We note that, as all other 11 language, Swedish exhibits the expected sublinear increase in unconditional per-word entropy for the first five sentence positions predicted by the cue effectiveness decay hypothesis. However, the data also clearly show an unexpected negative trend for sentence positions 10–15. Additional tests confirmed that this unexpected trend is significant.

Based on our previous work (Qian, 2009; Qian & Jaeger, 2011), we hypothesized that this deviation is due to an unusually high proportion of topic shifts over that stretch of discourse. In Qian (2009), we found that the beginning of paragraphs, which frequently mark relatively large topic shifts, are associated with lower than otherwise expected unconditional per-word entropy. As outlined in the Introduction, under the reasonable assumption that sentences about topic A are less informative about upcoming material once a new topic B has been opened, this pattern is expected in communicatively efficient language production if Topic A and B are significantly different. For the same reasons that sentence positions at the
beginning of a discourse are expected to have lower unconditional per-word entropy, sentence positions following paragraph breaks also should have lower entropy (see also Genzel & Charniak, 2003). The fluctuations in unconditional per-word entropy observed for Swedish data, as a result, could be driven by non-uniformly distributed paragraph breaks.

In Qian and Jaeger (2011), we employed latent topic modeling (Blei et al., 2003) to provide estimates of topic shifts. We found that these topic shifts correlated in the predicted direction with changes in unconditional per-word entropy in the English Brown corpus. We applied the approach described in Qian and Jaeger (2011) to our Swedish data. We found that the unexpected trend in the Swedish data could not be entirely reduced to topic shifts, although topic shift does account for some variance in unconditional per-word entropy. It is possible that the latent topic model we employed failed to provide an adequate model of topic shifts in the Swedish news text corpus. We are forced to leave this question to future research and tentative conclude that 11 out of 12 languages showed the predicted effect.

4.2. Relating the constant entropy hypothesis to language production

Turning now to the relation between the CER hypothesis and theories of language production, we see two mutually compatible explanations for the observed entropy profiles. First, it is helpful to restate the result that the unconditional per-word entropy of sentences and words tends to increase throughout discourses: Words that a priori (out of context) have lower probability of occurrence become relatively more likely to occur later in the discourse. Put yet differently, on average, the probability for a word that has a priori low probability of being selected for pronunciation increases throughout discourse.

This formulation suggests a tentative link between the observed increase in unconditional per-word entropy and mechanistic accounts of lexical access. There is broad agreement that the probability with which a word or structure will be selected is determined by its relative activation compared to other candidates. For example, models of lexical access typically describe the probability that a word (also known as lexical, lemma, or L-level) node is selected as function of the target lemma’s relative activation and the relative activation of all other lemma nodes (Dell, 1986; Dell et al., 1999; Levelt, Roelofs, & Meyer, 1999; Mackay, 1982; Roelofs, 2002; Stemberger, 1985). The relative activation of each node is the result of both its relative “starting activation” and top–down activation that spreads from concepts to lexical nodes. We use the term “starting activation” to refer to the activation a node carries at the beginning of the lexical selection process due to prior experience. The starting activation of a node at any point in a discourse is due to both life-long experience (e.g., the frequency of that word) and the immediately preceding discourse context. For example, lexical nodes might have increased relative activation because they have recently been selected or because activation from recently selected related nodes has spread to them (the assumption that activation spreads directly or indirectly between semantically related words is shared by most models of language production, e.g., Dell, 1986; Levelt, 1989; Roelofs, 1997; see also Levelt, 2001). In this way, preceding context can differentially affect the activation of lexical nodes. Similarly, it is commonly assumed that the selection of syntactic frames is affected by their differential activation in context. For example, the well-known phenomenon of
Syntactic priming is usually attributed to increased activation of recently processed syntactic structures—either due to longer-term implicit learning or short-term boosts (Bock, 1986b; Chang, Dell, & Bock, 2006; Pickering & Branigan, 1998; Reitter, Keller, & Moore, 2011). The data we report here present an opportunity for future work to tease apart competing models of lexical and syntactic priming effects (Chang et al., 2006; Malhotra, 2009; Reitter et al., 2011). If the observed entropy profiles were to fall out of any of these accounts without further stipulation, this would provide support for that model.

We can also interpret the finding that shifts in topic are associated with a decrease in unconditional word entropy (Qian & Jaeger, 2011) within a spreading activation network. Works in natural language processing view latent topics as distributions over the vocabulary (for an overview, see Bellegarda, 2004). These probability distributions can be thought of as describing the relative (i.e., normalized) starting activation of all nodes in a lexical network. With this in mind, we can compare the production of a topic-continuing sentence to the production of a topic-shifting sentence. In both cases, top–down spreading activation is sent to a subset of lexical nodes that are suited for the encoding of the meaning the speaker intends to convey. However, the two scenarios differ as regard to what extent the relative starting activation is due to the most recent discourse. If the recent discourse supports the same subset of words that received top–down activation due to the intended meaning, the sentence is a topic-continuing one and topic-based support will overlap with the subset of lexical nodes receiving top–down support. This contrasts with the production of topic-shifting sentences. The more the sentence will constitute a topic-shift, the fewer of its words will have increased starting activation from the most recent discourse. This is illustrated in Fig. 5.

In short, sentences that constitute topic-shifts are less likely to result in the production of a priori less probable words compared to sentences that continue a topic. More specifically, the account outlined here makes an interesting prediction to be pursued in future work: It should be the amount of preceding discourse that is relevant to the current topic that

![Fig. 5. When the current sentence is a continuation of the same topic as the preceding discourse (in Fig. 5a, the speaker intends to continue talking about food), the overlap in context helps further increase the activation levels of words that are related to the topic (e.g., bagel). When there is a topic shift (in Fig. 5b, the speaker changes the topic from food to politics), the activation levels of lexical nodes remain relatively the same due to the previous topic. The speaker may experience difficulty in selecting the lexical node senate in such a situation. Red background color refers to high levels of activation; blue refers to low levels of activation.](image-url)
determines to what extent words with a priori low starting activation will be selected for production. This prediction is not shared by accounts like the centering theory (Grosz, Joshi, & Weinstein, 1983, 1995), which view topic-shift as a categorical notion, although the prediction is certainly compatible with the general idea advanced by the centering theory.

Another, mutually compatible, explanation for the observed entropy profiles is that speakers learn to avoid words and structures that would carry too much (or too little) information given the preceding discourse because this would be disadvantageous for communication (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). This would be consistent with recent findings that speakers have a preference for phonetic, phonological, morphological and syntactic variants that distribute information more uniformly across the linguistic signal (e.g., Aylett & Turk, 2004, 2006; Bell, Brenier, Gregory, Girard, & Jurafsky, 2009; Bell et al., 2003; Frank & Jaeger, 2008; Gómez Gallo, Jaeger, & Smyth, 2008; Jaeger, 2010, 2011; Maurits, Perfors, & Navarro, 2010; van Son & Pols, 2003; van Son & van Santen, 2005). For example, speakers tend to produce words with longer duration and more articulatory detail when they carry more information (i.e., when they are less excepted in their context, Aylett & Turk, 2004, 2006; Bell et al., 2003, 2009; Gahl & Garnsey, 2004; Pluymaekers, Ernestus, & Baayen, 2005; Tily et al., 2009). Similarly, speakers are more likely to reduce syntactic structures where afforded by grammar when that structure carries less information in its context (Jaeger, 2010, 2011; Wasow, Jaeger, & Orr, 2011). Intriguingly, there is also evidence that speakers show a similar preference to avoid peaks and troughs in information density when selecting between different ways to realize a referring expression (Tily & Piantadosi, 2009; see also Arnold, 2001) and when selecting between possible word orders (Maurits et al., 2010).

It is possible that this preference for uniform information density affects local choices during incremental production (e.g., referring expression choice) in such a way that results in the observed entropy profiles. While relatively little is known about the link between local choices and the distribution of information across discourses, there is substantially no lack of evidence that previous discourse affects lexical and syntactic preferences during sentence production. For example, previous mention of a referent or semantically related referents affects speakers’ word order preferences (Arnold, Wasow, Losongco, & Ginstrom, 2000; Bock, 1986a; Bock & Irwin, 1980; Ferreira & Yoshita, 2003). This in turn would affect the per-word entropy estimates employed above, which are based on ngrams. Similarly, it is well known that referring expression choice depends on whether the referent has previously been mentioned (e.g., Brennan, 1995; Fukumura, van Gompel, & Pickering, 2010; Gundel, Hedberg, & Zacharski, 1993; see also Gundel, 1995 and references therein, for the effects of topicality). While these effects of previous mention might be due to differential starting activation (as outlined above), there is evidence that the previous discourse can affect referring expression choice in ways that are less obviously due to differences in starting activation. For example, Arnold and Griffin (2007) show that speakers are sensitivity to the co-presence of multiple potentially compatible referents in the preceding discourse in that they prefer referring expressions that unambiguously refer to the intended referent (see also Tily & Piantadosi, 2009).

Findings like these are highly compatible with the hypothesis that local decisions during language production preferences are affected by a preference for efficient communication.
How these local decisions result in the observed entropy profiles remains, however, an open question to be addressed by future work. The current work serves as a starting point for further investigations that will help to distinguish between and elaborate on the alternative accounts outlined here in broad strokes only.

Notes

1. The term “sentence position” refers to the position of a sentence in a discourse.
2. In particular, in the type of language model employed here and in previous work on the CER hypothesis, probability estimates are always larger than 0. Therefore, the estimated information gain associated with an additional word is always larger than 0. Note, however, that normalization is not necessarily sufficient to address the effect of additional words. It is possible that additional words add in a nonlinear way to the information of a sentence. In previous work, we have shown that this indeed sometimes the case (Qian, 2009; Qian & Jaeger, 2009). In the same work, we show that this problem can be addressed by including nonlinear effects of sentence length in the regression model used to analyze the relation between sentence position and sentence entropy. All results presented here hold if the data are analyzed that way.
3. For the restricted cubic spline, sentence position was partitioned by \( k \) knots (here \( k = 5 \)), and four functions are fit on the intervals \([1, 5), [5, 8), [8, 11), [11, 15]\), respectively. Hence, modeling document-specific random effects of sentence position introduces 14 parameters: four parameters to model the document-specific variances of the four coefficients of the spline, as well as \( 4 + 3 + 2 + 1 = 10 \) parameters to model all covariances between the random intercept and each of the variances associated with the spline.
4. The BIC is defined as \(-2\hat{\lambda} + k \ln n\), where \( \hat{\lambda} \) is the data likelihood, \( k \) is the number parameters in the model, and \( n \) the number of data points.
5. While the original spreading activation and related accounts mostly focused on isolated word recognition and production and hence remained agnostic as to whether and how activation decays over time, recent work on language processing within the ACT-R framework holds that retrieval of a word results in an activation boost that exhibits power-law decays over time (Lewis et al., 2006). This idea could be combined with activation spreading accounts to also account for increased activation of nodes related to the retrieved node. We leave a more detailed discussion of these ideas to future work. Similarly, we intentionally abstract away from additional conditions on the linking function between activation and selection of probability that would be required to obtain the observed shape of unconditional entropy profiles.

Acknowledgments

This work was partially supported by an Alfred P. Sloan Research Fellowship, the University of Rochester’s Provost Award for Multidisciplinary Research, and NSF grant
BCS-0845059 to TFJ. We thank Alex Fine, Camber Hansen-Karr, Thomas Farmer, and three anonymous reviewers for feedback on an earlier version of the manuscript. We are particularly grateful for discussions and exchange of ideas with Steven Piantadosi early in the development of the cue effectiveness decay ideas presented here.

References


Appendix A

Cumulative discourse informativity

This proof intends to show that the sum of conditional per-word entropy and cumulative discourse informativity (the latter as defined in Eq. 5) corresponds to the upper bound of the unconditional per-word entropy at sentence position $k$.

For a given sentence position, the probability of a specific sentence $s$ being the outcome of random variable $S_k$ is

$$p(S_k = s) = p(S_k = s|c_k)p(c_k)$$

(11)

Taking the negative logarithm of both sides, we get

$$-\log p(S_k = s) = -\log p(S_k = s|c_k) - \log p(c_k)$$

(12)

The left-hand side of Eq. 12 is the unconditional information of sentence $s$; the right-hand side is the sum of the conditional information of that sentence and the cumulative informativity of context. In the limit, when we have an infinitely large sample, the average of the left hand is the unconditional per-word entropy and the average of the right hand is the sum of conditional per-word entropy and cumulative discourse context.

Appendix B

Specifications of the nonlinear regression model

Based on the power-law decay model (i.e., Eq. 6), we have

$$r_{ij} = (r_0 + r_{0i}) \frac{k_j^{1-\lambda} - 1}{1-\lambda} + (r_0 + r_{0i}) + \epsilon_{ij}$$

(13)
\[ r_{0i} \sim N(0, \sigma^2), r_{0i} \perp \epsilon_{ij} \]

where \( r_{ij} \) is unconditional per-word entropy and \( k_j \) is sentence position. In addition, \( r_{0i} \) represents the document-specific deviations from the overall mean. Random effects were not considered for the decay rate parameter in order for the estimation process to converge successfully. Finally, \( \epsilon_{ij} \) represents the errors independently distributed as \( \mathcal{N}(0, \sigma^2) \), orthogonal to document specific deviations.

### Supporting Information

Additional Supporting Information may be found in the online version of this article on Wiley Online Library:

**Data S1.** Further details on model comparisons.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.