Close = Relevant? The Role of Context in Efficient Language Production

Ting Qian and T. Florian Jaeger
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627 United States
{tqian,fjaeger}@bcs.rochester.edu

Abstract

We formally derive a mathematical model for evaluating the effect of context relevance in language production. The model is based on the principle that distant contextual cues tend to gradually lose their relevance for predicting upcoming linguistic signals. We evaluate our model against a hypothesis of efficient communication (Genzel and Charniak's Constant Entropy Rate hypothesis). We show that the development of entropy throughout discourses is described significantly better by a model with cue relevance decay than by previous models that do not consider context effects.

1 Introduction

In this paper, we present a study on the effect of context relevance decay on the entropy of linguistic signals in natural discourses. Context relevance decay refers to the phenomenon that contextual cues that are distant from an upcoming event (e.g. production of a new linguistic signal) are less likely to be relevant to the event, as discourse contents that are close to one another are likely to be semantically related. One can also view the words and sentences in a discourse as time steps, where distant context becomes less relevant simply due to normal forgetting over time (e.g. activation decay in memory). The present study investigates how this decaying property of discourse context might affect the development of entropy of linguistic signals in discourses. We first introduce the background on efficient language production and then propose our hypothesis.

1.1 Background on Efficient Language Production

The metaphor "communication channel", borrowed from Shannon's information theory (Shannon, 1948), can be conceived of as an abstract entity that defines the constraints of language communication (e.g. ambient noise, distortions in articulation). For error free communication to occur, the ensemble of messages that a speaker may utter must be encoded in a system of signals whose entropy is under the capacity of the communication channel. Entropy of these signals, in this context, correlates with the average number of upcoming messages that the speaker can choose from for a particular signal (e.g. a word to be spoken) given preceding discourse context. In other words, if the average number of choices given any linguistic signal exceeds the channel capacity, it cannot be guaranteed that the receiver can correctly infer the originally intended message. Such transmission errors will reduce the efficiency of language communication.

Keeping the entropy of linguistic signals below the channel capacity alone is not efficient, for one can devise a code where each signal corresponds to a distinct message. With a unique choice per signal, this encoding achieves an entropy of zero at the cost of requiring a look-up table that is too large to be possible (cf. Zipf (1935), who makes a similar argument for meaning and form). In fact, the most efficient code requires language users to encode messages into signals of the entropy bounded by the capacity of the channel. One implication of this efficient encoding is that over time, the entropy of the signals is constant. One of the first studies to investigate such constancy is Genzel and Charniak (2002), in which the authors proposed the Constant Entropy Rate (CER) hypothesis: in written text, the entropy per signal symbol is constant across sentence positions in discourses. That is, if we view sentence positions as a measure of time steps, then the entropy per word at each step should be the same in order to achieve efficient communication (word is selected as the unit of signal, although it does not have to

Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics, ACL 2010, pages 45–53,
Uppsala, Sweden, 15 July 2010.

be case; cf. Qian and Jaeger (2009)).

The difficulty in testing this *direct* prediction is computationally specifying the code used by human speakers to obtain a context-sensitive estimate of the entropy per word. An *n*gram model overestimates the entropy of upcoming messages by relying on only the preceding *n-1* words within a sentence, while in reality the upcoming message is also constrained by extra-sentential context that accumulates within a discourse. The more extra-sentential context that the *n*gram model ignores, the higher estimate for entropy will be. Hence, the CER hypothesis indirectly predicts that the entropy of signals, as estimated by *n*grams, will increase across sentence positions. While some studies have found the predicted positive correlation between sentence position and the per-word entropy of signals estimated by *n*grams, most of them assumed the correlation to be linear (Genzel and Charniak, 2002; Genzel and Charniak, 2003; Keller, 2004; Piantadosi and Gibson, 2008). However, in previous work, we found that a log-linear regression model was a better fit for empirical data than a simple linear regression model based on data of 12 languages (Qian and Jaeger, under review). Why this would be case remained a puzzle.

Our research question is closely related to this *indirect* prediction of the Constant Entropy Rate hypothesis. Intuitively, the number of possible messages that a speaker can choose from for an upcoming signal in a discourse is often restricted by the presence of discourse context. Contextual cues in the preceding discourse can make the upcoming content more predictable and thus effectively reduces signal entropy. As previously mentioned, however, different contextual cues, depending on how long ago they were provided, have various degrees of effectiveness in reducing signal entropy. Thus we ask the question whether the decay of context relevance could explain the sublinear relation between entropy and discourse progress that has been observed in previous studies.

We formally derive two nonlinear models for testing our Relevance Decay Hypothesis (introduced next). In addition to the constant entropy assumption in CER, our model assumed that the relevance of early sentences in the discourse systematically decays as a function of discourse progress. Our models provide the best fit to the distribution of entropy of signals, suggesting the availability of discourse context can affect the planning of the rest of a discourse.

1.2 Relevance Decay Hypothesis

We hypothesize the sublinear relation between the entropy of signals, when estimated out of discourse context (hereafter, *out-of-context* entropy of signals) using an *n*gram model, and sentence position (Piantadosi and Gibson, 2008; Qian and Jaeger, under review) is due to the role of discourse context (hereafter, *context*). Consider the following example. Assume that context at the kth sentence position comes from the $1 \dots k-1$ sentences in the past. If k is large enough, context from the early sentences $1 \dots i$ ($i \ll k$) is essentially no longer relevant. Rather, the nearby $k-i$ sentences are contributing most of the discourse context. As a result, the constraint on the entropy of signals at sentence position k is *mostly* due to the nearby window of $k-i$ sentences. Then if we look ahead to the $(k+1)$th sentence position and follow the same steps of reasoning, context at that point also mostly comes from the nearby window of $k-i$ sentences (i.e. $(k+1)-(i+1) = k-i$). Hence, for later sentence positions, the difference in available context is minimal. Consequently, their out-of-context entropy of signals increases at a very small rate. On the other hand, when k is fairly small, to the extent that the $k-i$ window covers the entire preceding discourse, all of the $1 \dots k-1$ sentences are contributing relevant context. As k increases, the number of preceding sentences increases, which results in a more significant change in relevant context, but the relevance of each individual sentence decreases with its distance to k, which results in a sublinear pattern of relevant context with respect to sentence position overall. As we will show, the relation of out-of-context entropy of signals to sentence position follows from the relation of relevant context to sentence position, exhibiting a sublinear form as well.

The problem of interest here is to specify how quickly the relevance of a preceding sentence decays as a function of its distance to a target sentence position k. We experimented with two forms of decay functions – power law decay and exponential decay. It has been established that many types of human behaviors can be well described by the power function (Wixted and Ebbesen, 1991), so we mainly focus on building a model under the

Language	Training Data		Test Data		
	in words	*in sentences*	*in words*	*in sentences*	*per position*
Danish	154,514	5,640	8,048	270	18
Dutch	50,309	3,255	2,105	90	6
English	597,698	23,295	31,276	1155	77
French	229,461	9,300	11,371	435	29
Italian	97,198	4,245	4,524	225	15
Mandarin Chinese	145,127	4,875	4,310	150	10
Norwegian	89,724	4,125	2,973	150	10
Portuguese	170,342	5,340	9,044	240	16
Russian	398,786	18,075	20,668	930	62
Spanish (Latin-American)	1,363,560	41,160	67,870	2,070	138
Spanish (European)	255,366	7,485	8,653	240	16
Swedish	266,348	11,535	13,369	555	37

Table 1: Number of words and sentences in the training and test data for each of the twelve languages. The last column gives the number of sentences at each sentence position (which is identical to the number of documents contained in the corpora).

power law, and examine if the model under the exponential law yields any difference. Under the assumptions of true entropy rate is constant across sentences, we predict that our models will better characterize the changes in estimated entropy of signals than general regression models that are blind to the role of context.

2 Methods

2.1 Data

We used the Reuters Corpus Volume 1 and 2 (Lewis et al., 2004). The corpus contains about 810,000 English news articles and over 487,000 news articles in thirteen languages. Because of inconsistent annotation, we excluded the data from three languages, Chinese, German, and Japanese. For Chinese, we substituted the Treebank Corpus (Xue et al., 2005) for the Reuters data, leaving us with twelve languages: Danish, Dutch, English, French, Italian, Mandarin Chinese, Norwegian, Portuguese, Russian, European Spanish, Latin-American Spanish, and Swedish. In order to estimate out-of-context entropy per word (i.e. per signal symbol) for each sentence position, articles were divided into a training set (95% of all stories) for training language models and a test set (the remaining 5%) for analysis (see Table 1 for details). Out-of-context entropy per word was estimated by computing the average log probability of sentences at that position, normalized by their lengths in words (i.e. for an individual sentence token s, the term to be averaged is $\frac{-\log p(s)}{length(s)}$ *bits* per word). Standard trigram language models were used to compute these probabilities (Clarkson and Rosenfeld, 1997). The majority of the 12 languages belong to the Indo-European family, while Mandarin Chinese is a Sino-Tibetan language.

2.2 Modeling Relevance Decay of Context

Formally, we define the relevance of context in the same unit as entropy of signals – bits per word. Let r_0 denote the entropy of signals that efficiently encode the ensemble of messages a speaker can choose from for any sentence position, a constant under the assumption of CER. According to Information Theory, r_0 is equivalent to the uncertainty associated with any sentence position if context is considered. Thus, in error free communication, linguistic signals presented at the kth sentence position are said to have resolved the uncertainty at k and therefore are r_0-bit relevant at the kth sentence position. Then, at the $(k+i)$th sentence position, these linguistic signals have become context by definition and their relevance has decayed to some r bits. Our models start from defining the value of r as a function of the distance between context and a target sentence position.

2.2.1 Power-law Decay Model

If the relevance of a cue q (e.g. a preceding sentence), which is originally r_0-bit relevant at position k_q, decays at the rate following the power function, its remaining relevance at target sentence position k is:

$$relevance_{pow}(k, q) = r_0(k - k_q + 1)^{-\lambda} \quad (1)$$

In Equation (1), $k > k_q$ and λ is the decay rate. This means at position k, the relevance of the cue

from the $(k-1)$th sentence is $r_0 * 2^{-\lambda}$-bit relevant; the relevance of the cue from the $(k-2)$th sentence is $r_0 * 3^{-\lambda}$-bit relevant, and so on. As a result, the relevance of discourse-specific context at position k is the marginalization of all cues up to q_{k-1}:

$$context_{pow}(k) = r_0 \sum_{q_i \in \{q_1 \cdots q_{k-1}\}} (k - k_{q_i} + 1)^{-\lambda} \quad (2)$$

The general trend predicted by Equation (2) is that discourse-specific context increases more rapidly at the beginning of a discourse and much more slowly towards the end due to the relevance decay of distant cues. Rewriting Equation (2) in a closed-form formula so that a model can be fitted to data is not a trivial task without knowing the rate λ, but the paradox is that λ has to be estimated from the data. As a workaround, we approximated the value of Equation (2) by computing a definite integral of Equation (1), where Δi is a shorthand for $k - k_q + 1$:

$$\begin{aligned} context_{pow}(k) &\approx \int_1^k r_0 \Delta i^{-\lambda} d\Delta i \\ &= r_0 (\frac{k^{1-\lambda} - 1}{1 - \lambda}) \end{aligned} \quad (3)$$

Equation (3) uses an integral to approximate the sum of a series defined as a function. The result is usually acceptable as long as λ is greater than 1 so that the series defined by Equation 1 is convergent (this assumption is empirically supported; see Figure 5). Note that Equation (3) produces the desirable effect that upon encountering the first sentence of a discourse, no discourse-specific contextual cues are available to the speaker (i.e. $context(1) = 0$).

Now that we know the maximum relevance of context at sentence position k, we can predict the amount of out-of-context entropy of signals $r(k)$ based on the idea of uncertainty again. There are new linguistic signals that are r_0-bit relevant *in context* at any sentence position. In addition, we now know $context(k)$ bits of relevant context are also available. Thus, the sum of r_0 and $context(k)$ defines the maximum amount of out-of-context uncertainty that can be resolved at sentence position k. Therefore, the out-of-context entropy of signals at k is at most:

$$\begin{aligned} r_{pow}(k) &= context(k) + r_0 \\ &= r_0 \frac{k^{1-\lambda} - 1}{1 - \lambda} + r_0 \end{aligned} \quad (4)$$

Whether speakers will utilize all available context as predicted by Equation (4) is another debate. Here we adopt the view that speakers are maximally efficient in that they do make use of all available context. Thus, we make the prediction that out-of-context entropy of signals, as observed empirically from data, can be described by this model. Figure 1 shows the behavior of this function with various parameter sets.

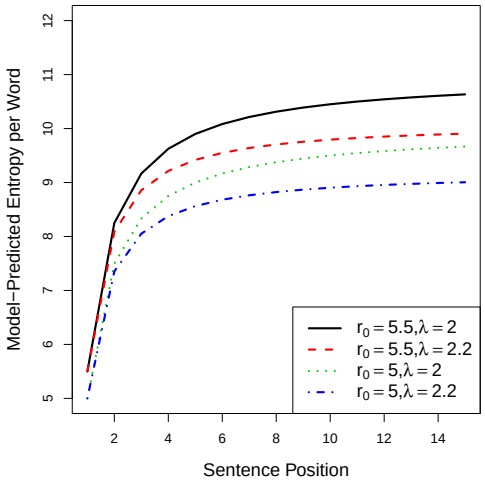

Figure 1: Schematic plots of the behavior of out-of-context entropy of signals assuming the decay of the relevance of context is a power function.

2.2.2 Exponential Decay Model

The second model assumes the relevance of context decays exponentially. Following the same notations as before, the relevance of a cue q at position k is:

$$relevance_{exp}(k, q) = r_0 e^{-\lambda(k - k_q)} \quad (5)$$

The major difference between the power function and the exponential one is that the relevance of a contextual cue drops more slowly in the exponential case (Anderson, 1995). The relevance of all discourse-specific context for a speaker at k is:

$$context_{exp}(k) = r_0 \sum_{i=1}^{k-1} e^{-\lambda i} \quad (6)$$

Equation (6) is the sum of a geometric progression series. We can write Equation (6) in a closed-form:

$$context_{exp}(k) = \frac{r_0}{e^{\lambda} - 1}(1 - e^{-(k-1)\lambda}) \quad (7)$$

As a result, the out-of-context entropy of signals is:

$$r_{exp}(k) = \frac{r_0}{e^{\lambda} - 1}(1 - e^{-(k-1)\lambda}) + r_0 \quad (8)$$

Figure 2 schematically shows the behavior of this function. One can notice this function converges against a ceiling more quickly than the power function. Thus, this model makes a slightly different prediction from the power law model.

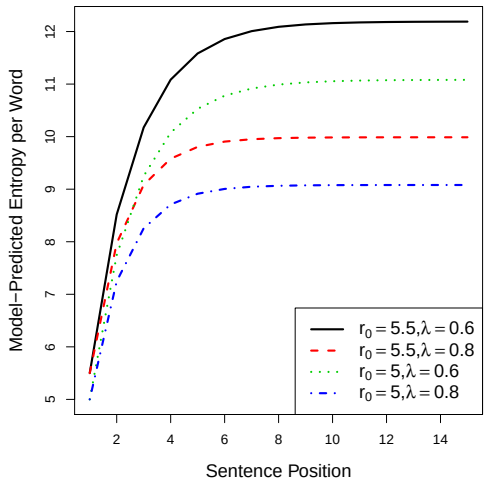

Figure 2: Schematic plots of the behavior of out-of-context entropy of signals assuming the decay of the relevance of context is an exponential function.

2.3 Nonlinear Regression Analysis

To test whether the proposed models (i.e. Equations 4 and 8) better characterize the data, we built nonlinear regression models with document-specific random effects, where the out-of-context entropy of signals, r_{ij}, is regressed on sentence position, k_j. Based on the power law model, we have

$$r_{ij} = (\beta_1 + b_{1i})\frac{{k_j}^{1-\beta_2} - 1}{1 - \beta_2} + (\beta_1 + b_{1i}) + \epsilon_{ij} \quad (9)$$

where β_1 corresponds to r_0, the theoretical constant entropy of signals under an ideal encoding. b_{1i} represents the document-specific deviations from the overall mean. β_2 corresponds to λ, the mean rate at which the relevance of a past cue decays, which is unfortunately not considered for random effects for the practical purpose of making computation feasible in the current work. Finally, ϵ_{ij} represents the errors independently distributed as $\mathcal{N}(0, \sigma^2)$, orthogonal to document specific deviations.

For the exponential model, the nonlinear model is the following (symbols have the same interpretations as in Equation 9):

$$r_{ij} = \frac{(\beta_1 + b_{1i})}{e^{\beta_2} - 1}(1 - e^{-(k_j - 1)\beta_2}) + (\beta_1 + b_{1i}) + \epsilon_{ij} \quad (10)$$

Fitting data with the above nonlinear models requires starting estimates for fixed-effect coefficients (i.e. β_1s and β_2s). Unfortunately, there are no principled methods for selecting these values. We heuristically selected 6 for β_1 and 2 for β_2 as starting values for the power law model, and 4 and 0.5 as starting values for the exponential model.

3 Results

We examined the quality of the models and the parameters in the models: r_0, the within-context entropy rate, and λ, the rate of context decay.

3.1 Model Quality Comparison

The CER hypothesis indirectly predicts that out-of-context entropy of signals of sentence positions (bits per word) should increase throughout a discourse. The two models go one step further to predict specific sublinear increase patterns, based on the speaker's considerations of the relevance of past contextual cues. We compared the quality of models in terms of Bayesian Information Criterion (BIC) *within languages*. A lower BIC score indicates a better fit. As shown by Figure 3, we find our models best explain the data in 9 out of the 12 languages, reporting lower BIC scores than both the linear and log-linear models as reported in our previous work (Qian and Jaeger, under review). For Danish, English and Italian, although neither of our models produced a better score than the log-linear model, the relative difference is small: 0.54 on average (comparing to BIC scores on the order of 10^2 to 10^3).

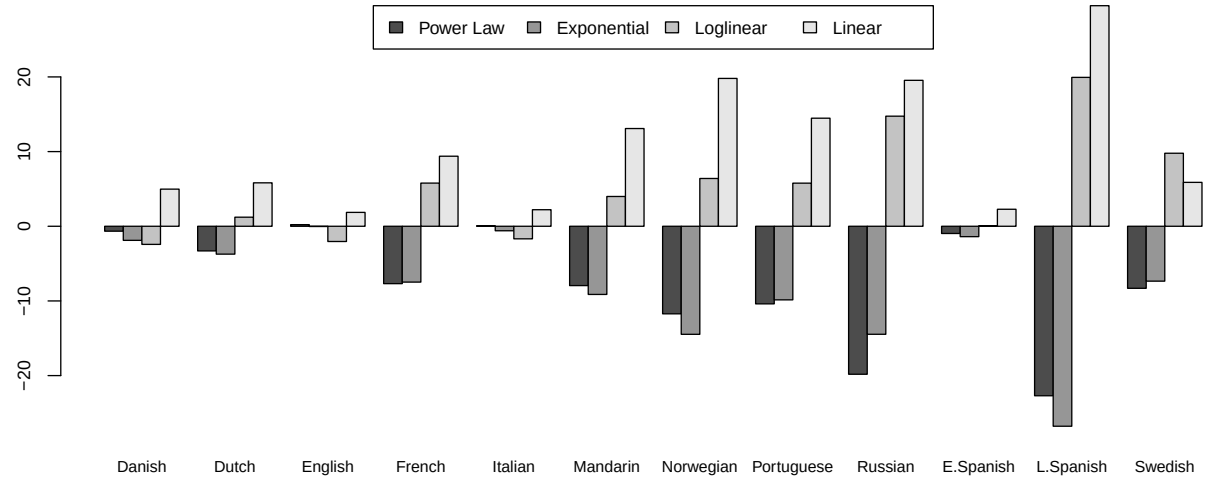

Figure 3: Our models yield superior BIC scores in most languages. The y-axis shows the differences between BIC scores of individual models for a language and mean BIC of the models for that language (*E.Spanish* = European Spanish; *L.Spanish* = Latin-American Spanish).

Specifically, in terms of BIC scores, the power-law model is better than the linear model ($t(11) = -3.98$, $p < 0.01$), and the log-linear model ($t(11) = -3.10$, $p < 0.05$). The exponential model is also better than the linear model ($t(11) = -3.98$, $p < 0.01$), and the log-linear model ($t(11) = -3.18$, $p < 0.01$). The power-law model and the exponential model are not significantly different from each other ($t(11) = 0.5$, $p > 0.5$).

3.2 Interpretation of Parameters

Constant Entropy of Signals r_0. Both models are constructed in such a way that the first parameter r_0, in theory, corresponds to the theoretical within-context entropy of signals of sentence positions. This parameter refers to how many bits per word are needed to encode the ensemble of messages at a sentence position when context is taken into account. The CER hypothesis directly predicts that this rate should be constant throughout a discourse. Although we are unable to test this prediction directly, it is nevertheless interesting to compare whether these two independently developed models yield the same estimates for this parameter in each language.

Figure 4 shows encouraging results. Not only the estimates made by the power model are well correlated with those by the exponential model, but also the slope of this correlation is equal to 1 ($t(10) = 1.01$, $p < 0.0001$). Since there are no reasons *a priori* to suspect that these two models would give the same estimates, this is a first step to confirming the entropy per word in sentence production is indeed a tractable constant throughout discourses.

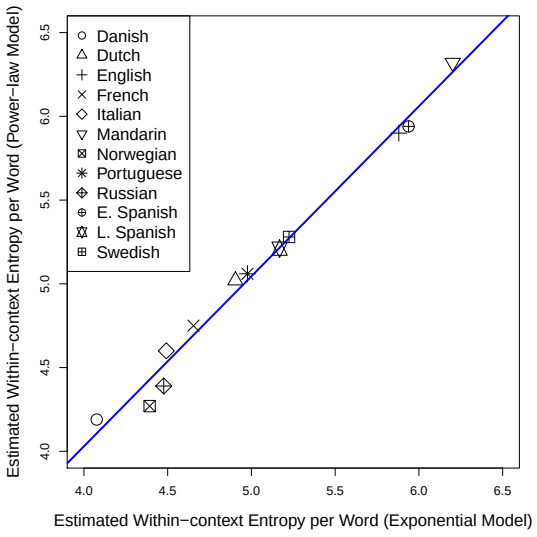

Figure 4: Estimates of r_0 correlate between both models with a slope of 1.

Among all languages, r_0 has a mean of 5.0 bits in both models, and a variance of 0.46 in the power-law model and 0.48 in the exponential model, both remarkably small. 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. On the other hand,

the differences in r_0 may reveal the specific properties of different languages. Meanwhile, precautions need to be taken in interpreting those estimates given that the corpora are of different sizes, and the *n*gram model is simplistic in nature.

Decay Rate λ. The second parameter λ corresponds to the rate of relevance decay in both models. Since the base relevance r_0 varies between languages, λ can be more intuitively interpreted as to indicate the percentage of the original relevance of a contextual cue still remains in n positions. In the power-law model, for example, the context information from a previous sentence in Danish, on average, is only 11.6% ($2^{-3.10} = 0.116$) as relevant. Hence, the relevance of a contextual cue decreases rather quickly for Danish. Table 2 shows this is in fact the general picture for all languages we tested.

Language	**Relevance of Context in Discourse** (%)		
	1 pos. before	*2 pos. before*	*3 pos. before*
Danish	11.6	3.3	1.4
Dutch	10.4	2.8	1.1
English	0.1	0.0	0.0
French	8.5	2.0	0.7
Italian	10.2	2.7	1.0
Mandarin	7.7	1.7	0.6
Norwegian	18.9	7.1	3.6
Portuguese	5.5	1.0	0.3
Russian	12.7	3.8	1.6
E. Spanish	0.8	0.0	0.0
L. Spanish	2.7	0.3	0.1
Swedish	5.8	1.1	0.3

Table 2: In the power model, relevance of a contextual cue decays rather quickly for each language.

The picture of λ looks a little different in the exponential model. The relevance percentage on average is significantly higher, which confirms an earlier point that the power function decreases more quickly than the exponential function. Table 3 shows a summary for the 12 languages.

One may note that the decay rate varies greatly between languages under the prediction of both models. However, these number are only approximations since the entropy estimated by the *n*gram language model is far from psychological reality. Furthermore, it is unlikely that speakers of one language would exhibit the same decay rate of context relevance in their production, let alone speakers of different languages, who may be subject to language-specific constraints during production. Therefore, the variation in estimates of λ seems reasonable.

Language	**Relevance of Context in Discourse** (%)		
	1 pos. before	*2 pos. before*	*3 pos. before*
Danish	30.1	9.1	2.7
Dutch	28.7	8.2	2.4
English	9.6	0.9	0.1
French	26.7	7.1	1.9
Italian	28.7	8.2	2.4
Mandarin	25.7	6.6	1.7
Norwegian	42.3	17.9	7.6
Portuguese	22.5	5.1	1.1
Russian	34.6	12.0	4.2
E. Spanish	14.2	2.0	0.3
L. Spanish	18.6	3.5	0.6
Swedish	23.7	5.6	1.3

Table 3: In the exponential model, relevance of a contextual cue decays more slowly.

Correlation between r_0 and λ. Interestingly, r_0 and λ are highly correlated ($r^2 = 0.39, p < 0.05$ in the power model, Figure 5; $r^2 = 0.47, p < 0.01$ in the exponential model, Figure 6): a high relevance decay rate tends to be coupled with high within-context entropy of signals. This unanticipated observation is in fact compatible with the account of efficient language production: a high within-context entropy of signals indicates the base relevance of a contextual cue (i.e. r_0) is high. It is then useful for its relevance to decay more quickly to allow the speaker to integrate context from other cues. Otherwise, the total amount of relevant context may presumably overload working memory. However, our current results come from only cross-linguistic samples. Cross-validation in within-language samples is needed for confirming this hypothesis.

3.3 The Bigger Picture

Having obtained the estimates for r_0 and λ, we are now in a position to examine how out-of-context entropy of signals increases as a function of sentence positions, given the estimates of these two parameters. As shown in Figure 7, the predictions from both models are qualitative similar except that 1) when the decay rate in the power-law model is low, out-of-context entropy of signals converges more slowly than in the exponential model (Figure 7, right panel); 2) when the decay rate in the power model is high, it almost converges as quickly as the exponential model, and only minor differences exist in their predictions (Figure 7, left panel).

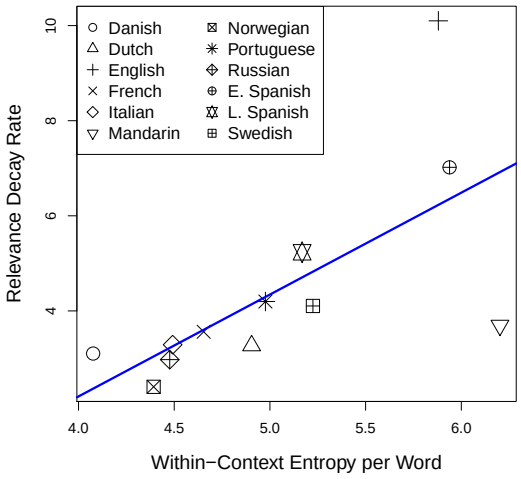

Figure 5: The rate of relevance decay is correlated with within-context entropy of signals in the power-law model.

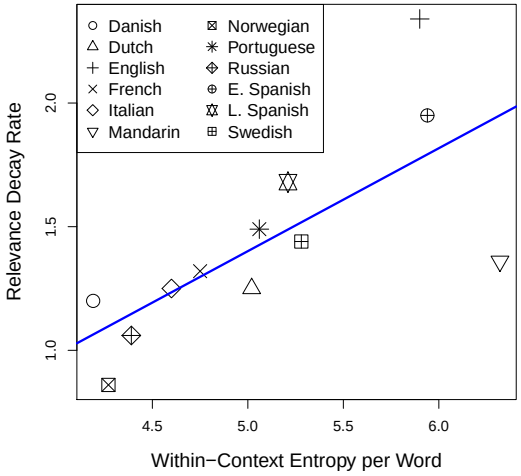

Figure 6: The rate of relevance decay is correlated with within-context entropy of signals in the exponential model.

Because of the nonlinearity in our models, it is not possible to report the results in an intuitive manner as in "an increase in sentence position corresponds to an increase of X bits of out-of-context entropy per word". Instead, we can analytically solve for the derivative of the predicted out-of-context entropy of signals with respect to sentence position (Equation 4 and 8). This gives us:

$$r_{power}(k)' = r_0 k^{-\lambda} \qquad (11)$$

for the power-law model, showing the rate of increase in predicted out-of-context entropy of signals is a monotonically decreasing power function, and

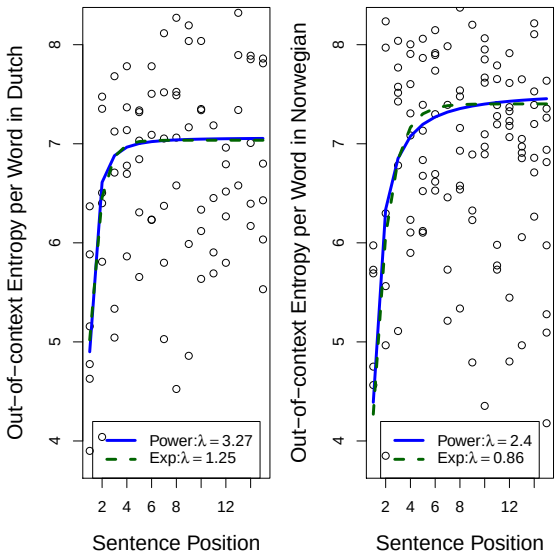

Figure 7: Predicted out-of-context entropy of signals by the power-law model (solid) and the exponential model (dashed) in Dutch and Norwegian, with the actual distributions plotted on the background.

$$r_{exp}(k)' = \frac{r_0 \lambda}{e^{\lambda} - 1}(e^{-(k-1)\lambda}) \qquad (12)$$

for the exponential model, showing the rate of increase is a monotonically decreasing exponential function. These mathematical properties indeed match our observations in Figure 7.

4 Discussion and Future Work

The models introduced in this paper try to answer this question: if the relevance of a contextual cue for predicting an upcoming linguistic signal decays over the course of a discourse, how much uncertainty (entropy) is associated with each individual sentence position? We have shown under that models that incorporate (power law or exponential) cue relevance decay in most cases describe the relation of out-of-context entropy of signals to sentence position are better accounted for than previously suggested models.

We are continuing to investigate along this line. Specifically, we are interested in finding the role of semantic memory in affecting the relevance decay of context. To test that, we plan to implement a probabilistic topic model, in which topic continuity between a preceding sentence and an upcoming sentence is quantitatively measured. Thus, the decay of contextual cues can be based on the esti-

mated semantic relatedness between sentences, in addition to the abstract notion of *rate* as used in this paper.

Finally, our relevance decay model can be applied to the domain of language processing as well. For instance, the distance between a contextual cue and the target word may affect how quickly a comprehender can process the information conveyed by the word. We plan to address these question in future work.

5 Conclusion

We have presented a new approach for examining the distribution of entropy of linguistic signals in discourses, showing that not only the out-of-context entropy of signals increases sublinearly with sentence position, but also the sublinear trend is better explained by our nonlinear models than by log-linear models of previous work. Our models are built on the assumption that the relevance of a contextual cue for predicting a linguistic signal in the future decays with its distance to the target, and predict the relation of out-of-context entropy of signals to sentence position in discourses. These results indirectly lend support to the hypothesis that speakers maintain a constant entropy of signals across sentence positions in a discourse.

Acknowledgements

We wish to thank Meredith Brown, Alex Fine and three anonymous reviewers for their helpful comments on this paper. This work was supported by NSF grant BCS-0845059 to TFJ.

References

John R. Anderson. 1995. *Learning and Memory: An integrated approach*. John Wiley & Sons.

Philip R. Clarkson and Roni Rosenfeld. 1997. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of ESCA Eurospeech*.

Dimitry Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *ACL*, pages 199–206.

Dimitry Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. in. In *EMNLP*, pages 65–72.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *EMNLP*, pages 317–324.

D. D. Lewis, Y. Yang, T. Rose, and F Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J Mach Learn Res*, 5:361–397.

Steve Piantadosi and Edwards Gibson. 2008. Uniform information density in discourse: a cross-corpus analysis of syntactic and lexical predictability. In *CUNY*.

Ting Qian and T. Florian Jaeger. 2009. Evidence for efficient language production in chinese. In *CogSci09*, pages 851–856.

Ting Qian and T. Florian Jaeger. under review. Entropy profiles in language: A cross-linguistic investigation.

C. E. Shannon. 1948. A mathematical theory of communications. *Bell Labs Tech J*, 27(4):623–656.

J. T. Wixted and E. B. Ebbesen. 1991. On the form of forgetting. *Psychological Science*, 2:409–415.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Nat Lang Eng*, 11:207–238.

G. K. Zipf. 1935. *Psycho-Biology of Languages*. Houghton-Mifflin.