Topic Shift in Efficient Discourse Production

Ting Qian (tqian@bcs.rochester.edu)
T. Florian Jaeger (fjaeger@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627 USA

Abstract

Speakers have been hypothesized to organize discourse content so as to achieve communicative efficiency. Previous work has focused on indirect tests of the hypothesis that speakers aim to keep per-word entropy constant across discourses to achieve communicative efficiency (Genzel & Charniak, 2002). We present novel and more direct evidence by examining the role of topic shift in discourse planning. If speakers aim for constant per-word entropy, they should encode less unconditional per-word entropy (as estimated based on only sentence-internal cues) following topic shifts, as there is less relevant context to condition on. Applying latent topic modeling to a large set of English texts, we find that speakers are indeed sensitive to the recent topic structure in the predicted way.

Keywords: discourse production; topic shift; communicative efficiency

Introduction

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 (Aylett & Turk, 2004; Ferrer i Cancho & Díaz-Guilera, 2007; Genzel & Charniak, 2002; Jaeger, 2010; Levy & Jaeger, 2007). Here we focus on evidence from discourse production that speakers distribute information across sentence so as to hold the conditional entropy associated with a word constant, which would facilitate efficient information transfer (Genzel & Charniak, 2002). As language production unfolds over time, information needs to be computed with reference to conditional (contextualized) probabilities. This raises the question as to what cues are integrated into contextualized probabilities, and how they are integrated. This issue so far has received little to no attention. We investigate whether shifts in the latent topics of a discourse influence the amount of information encoded in each sentence.

Previous research on efficiency in discourse production has revealed an interesting relation between the information content of a sentence and its position in a discourse: on average, sentences that occur later in a discourse tend to contain more *unconditional* information per word than earlier ones (Genzel & Charniak, 2002, 2003; Keller, 2004; Piantadosi & Gibson, 2008; Qian & Jaeger, 2009). Unconditional information refers to the information a word (or sentence) carries if only sentence-internal cues are considered (i.e. without consideration of preceding discourse context). Why would speakers distribute discourse information in such a way? Information theoretic considerations about efficient communication provide a possible explanation. Shannon's noisy channel theorem implies that an efficient communication system should transmit information at a constant rate close to the channel capacity (Shannon, 1948). The information of a sentence is defined as the negative log-probability of a sentence. The difference between unconditional and conditional information relates to whether sentence information reflects the effect of discourse context. These considerations led Genzel and Charniak (2002) to propose the Constant Entropy Rate hypothesis, according to which unconditional sentence information is expected to increase over time if conditional sentence information stays more or less uniform (for more detail, see Qian and Jaeger (submitted)). Thus, the finding of a positive correlation between unconditional information and sentence position can be taken as evidence for communicative efficiency of language.

However, the observed positive correlation is a rather weak confirmation of communicative efficiency, primarily because it is only a necessary, but not sufficient, condition of the hypothesized uniform distribution of conditional information in discourses. One way of obtaining stronger results is to estimate the conditional information of sentences and then tests whether those estimates indeed form a uniform distribution. This would require one to obtain a discourse-sensitive language model, from which conditional information estimates can be derived. One can also work with unconditional information estimates and try to identify variables that maximally correlate with discourse context. The effect of sentence position on unconditional information is expected to be subsumed by such predictors, since they have essentially compensated for the lack of discourse context in the unconditional estimates of information. Here we present a series of studies that apply both methods through the use of topic modeling. We derived two partially conditioned estimators of sentence information by estimating the topics in the preceding discourse (a *fully* conditioned estimator would at least incorporate world knowledge that is relevant to the discourse, which is almost implausible). At the same time, topic modeling also allowed us to measure *topic shifts* in a discourse. When a discourse undergoes a large topic shift, the previously mentioned materials are less predictive of the upcoming materials, leading to higher information in those sentences. This intuitively suggests that topic shift will be a good predictor of sentence information. We test this hypothesis in both studies.

We begin with a review of previous work that motivates the approach taken here.

Estimating Sentence Information

Previous studies mostly used *n*gram to estimate how much information a sentence contains independent of discourse context. Under those models, the probability of a word is conditioned on certain within-sentence elements (e.g. the *n-1*

preceding words in a sentence for an *n*gram model). The probability estimate of a sentence is simply a product of the probability estimates of its words.

Besides the obvious problem that these estimators of sentence information do not consider the effect of discourse context at all, it is difficult to intuitively understand whether these estimates represent how much information the speaker has planned or how much information the listener may perceive. A hypothesis of efficient language production should distinguish between these two possibilities.

We approach this problem by first considering what information may be privileged to the speaker in the process of discourse production. One such factor is the topics of a discourse. In discourse production, the speaker typically has more information about the intended topics than the listener before producing the corresponding utterances. The listener, on the other hand, has to infer the topics of the discourse after observing the utterances. Thus, the amount of information that is in a sentence may appear differently to the speaker and to the listener. This raises the question whether speakers distribute information, taking into consideration the listener's uncertainty about the topic.

We present two studies designed to address this question. Study 1 proposes a *topic conditional estimator* of sentence information, which is an attempt to estimate how much information a speaker has planned in each sentence given the current discourse topic. Study 2 proposes a *latent bigram estimator* of sentence information, which marginalizes over all possible topics to preserve maximal uncertainty – arguably a rational strategy that the listener might adopt (although an interpretation in terms of production is also possible, since speakers may not have perfect certainty about what they said and how they intend to convey their current message). Because the topics of a discourse usually stretch over a few sentences, these estimators are implicitly sensitive to discourse context in a limited fashion (see below for more detail).

Explaining Nonlinear Patterns

One additional issue is that many early studies have implicitly assumed a linear correlation between unconditional sentence information per word and sentence position (Genzel & Charniak, 2002, 2003; Keller, 2004). This assumption is challenged by recent studies that found sublinear relations between a sentence's position in the discourse and its unconditional information (Piantadosi & Gibson, 2008; Qian & Jaeger, 2009, submitted). Qian and Jaeger (2010) derive this pattern from the assumption that the informativity of contextual cues on average decays with increasing distance (cue weight decay hypothesis). This assumption is based on the intuition that discourses typically consist of several topics so that contextual cues that have been introduced under Topic A may have little predictive power over the content of Topic B. To test this hypothesis, we applied generic decay functions which lower the predictive power of distant discourse contextual cues, and found that the resulting predictions about the average unconditional per-word entropy of sentences based on their position in the discourse were a better fit.

However, the decay functions in Qian and Jaeger (2010) only superimposed a general nonlinear pattern onto data based on the idea of topic shift. The actual sizes of topic shifts were not measured. Our current studies adopt a more direct approach by estimating topic shifts throughout the discourse. If topic shift correlates with discourse context, the positive correlation between sentence position and unconditional estimates of sentence information is expected to disappear. In addition, topic shift itself is expected to be a significant predictor of sentence information. When there is a large topic shift, the preceding discourse context may not be so useful in predicting the upcoming sentence, and a rational speaker should encode less information in the upcoming sentence. Therefore, a negative correlation between topic shift and sentence information is expected.

Methods

Data

We used the Brown corpus in the form provided by the Python Natural Language Toolkit (Bird, Loper, & Klein, 2009). The data set consists of 500 English articles. We divided the corpus into a training set of 400 articles for building the topic model, a development set of 50 articles for monitoring the quality of the trained model, and a testing set of 50 articles for conducting the studies. Each group has a random mixture of topic categories as labeled in the corpus. To normalize the lengths of articles, only the first 50 sentences of each article are included in data sets. We excluded all function words and content words that appeared less than 15 times or more than 450 times in the training data. This exclusion criterion aimed to keep only the semantically significant content words in sentences.

Modeling Topics

The topic of a sentence can influence the predictability of words. For example, the two words "the wall" are almost certainly followed by "street" in an article that discusses the financial market; whereas in a fairy tale, they are much more likely to be a complete noun phrase. An *n*gram model, which only considers the surface dependency of word tokens, will predict "street" is equally likely in both contexts.

Work in natural language processing uses topic models to estimate the latent topics of a collection of texts, for example, in order to find "hot" research topics published in journals (Griffiths & Steyvers, 2004). Here, we adopt the generative approach described in Blei, Ng, and Jordan (2003). Each topic t is defined by a multinomial probability distribution ϕ over words w, and each text d is a multinomial distribution θ of topics:

$$w|t,\phi \sim Discrete(\phi)$$
$$d|\theta \sim Discrete(\theta)$$

The generative assumption entails that a speaker produces the content of a discourse by first selecting a mixture of topics θ_i that they want to convey to their audience, and then sampling from topic distributions ϕ_i for words. In training the topic model, each individual sentence in the training set is provided as a single training "document" to the model for discovering latent topics. The Python package `deltaLDA` (Andrzejewski, Mulhern, Liblit, & Zhu, 2007), which implements a Gibbs sampler as described in Griffiths and Steyvers (2004), was used. To determine the optimal number of latent topics, we monitored the cross-entropy on the development set as the number of topics was varied. The number was determined to be 80. When the topic model was trained, two probability distribution were obtained: $\mathbf{P}(W|T)$, which gives the conditional probability of content words given topics, and $\mathbf{P}(T|D)$, which gives the conditional probability of topics given training documents. The latter distribution will be useful in computing the marginal probability of a topic:

$$p(T=t) = \sum_d p(T=t|D=d) \tag{1}$$

which allows us to compute an informative prior on the probabilities of topics.

Estimating Sentence Information

Sentence information refers to the negative log-probability of a sentence:

$$I(s) = -\log p(s) \tag{2}$$

Estimates of *per-word* sentence information were obtained by dividing sentence information estimates by sentence lengths, yielding a normalized quantity that can be compared across sentences. We describe the detailed methods as part of the studies.

Measuring Topic Shifts

We model the changes in discourse topics using Bayesian belief updating. At the beginning of a discourse, either the speaker or the listener is assumed to have an initial belief about which topics may be likely and which may not based on their knowledge of probable topics. As the discourse continues, this belief is updated at the end of each sentence to reflect the most likely topic to date. According to the Bayes' rule, the posterior belief about the distribution of topics is proportional to the product of the likelihood function and the prior:

$$\mathbf{P}(T|s) \propto p(s|T)\mathbf{P}(T) \tag{3}$$

The likelihood function $p(s|T)$ refers to the probability of the sentence given a particular topic. To simplify the model, we assume that all words are conditionally independent given a topic. Thus, the probability of a sentence is the product of the probabilities of its words:

$$p(s|T) = \prod_{w \in s} p(w|T) \tag{4}$$

$\mathbf{P}(T)$ is the prior belief about the distribution of topics; results of using a flat prior and an informative prior will be presented. In either case, the belief about the distribution of topics is updated on-line so that the posterior belief carries over to be the prior belief at the next sentence.

As a result, topic shifts can be conveniently quantified by measuring the distance between the posterior and prior belief distributions of topics. One metric for such a distance is Kullback-Leibler (KL) divergence. The KL divergence between the two belief distributions of topics is computed as:

$$KL = \sum_t \log_2 p(t|s) \frac{p(t|s)}{p(t)} \tag{5}$$

In other words, it's the logarithmic difference between the posterior and the prior belief distributions of topics, weighted by the posterior distribution. This distance intuitively represents the amount of topic shifts between sentences.

Study 1

Methods

We assume that the speaker plans the topic of a sentence $t_{intended}$ before actually producing the content of that sentence $s_i = \{w_{i1} \dots w_{in}\}$ (i.e. we assume that speakers know with certainty which topic they intend to talk about). Thus, the amount of information in that sentence, to the speaker, is an *a posteriori* estimate, depending on the intended topic: $\hat{I}(s_i) = I(s_i|t_{intended})$.

The first task is then to infer the topic of a sentence s_i. Unfortunately, we as researchers (as opposed to the speaker herself) can only estimate this topic in a *post-hoc* fashion since we are unable to foretell the speaker's plan before observing the production results. Therefore, we take the most probable topic given the sentence content as the closest estimate of the speaker's intended topic:

$$\hat{t}_{intended} = \underset{t}{\operatorname{argmax}}\, p(t|s = \{w_1 \dots w_n\}) \tag{6}$$

Equation (6) is a maximum *a posteriori* (MAP) estimate of the topic of a sentence, where

$$p(t|s = \{w_1 \dots w_n\}) = \frac{p(s = \{w_1 \dots w_n\}|t)p(t)}{p(s)} \tag{7}$$

The first term of the numerator in Equation (7) is the probability of a sentence given a topic t, which follows the same conditional independence assumption as shown in Equation (4). The second term is the prior probability of a topic. We used a flat prior for estimating the posterior distribution of topics for the first sentence in a discourse. When an informative prior was used, differences were minimal and thus were not reported separately. For each subsequent sentence position, an on-line updating procedure ensues. In some sense, the probability estimate of a sentence is implicitly sensitive to the entire preceding discourse, since the belief distribution of topics has been carried over consecutively.

We refer to the corresponding estimator of sentence information (shown below in Equation (8)) as a *topic conditional estimator* (TCE) of sentence information. The information estimate for sentence s based on the TCE is:

$$\hat{I}_{tce}(s;t) = -\log_2 p(s|t) = -\sum_{w \in s} \log_2 p(w|\hat{t}_{intended}) \qquad (8)$$

This sentence information estimate represents the amount of information that the speaker has *planned* for the utterance.

Mixed Model Analysis To investigate how speakers organize the content of a discourse, we conducted linear mixed model analyses.[1] Mixed models are regression models that provide ways to account for potential clusters in the data, which would otherwise lead to inflated Type I errors (Pinheiro & Bates, 2000). The **dependent variable** is per-word sentence information as estimated by the topic conditional estimator. The **independent variables** (i.e. predictors) are sentence position and topic shift. In addition to these fixed effects, we also included random intercepts for individual differences between writers. As introduced above, we predict that topic shift will be negatively correlated with sentence information, and such an effect may subsume the effect of sentence position.

To determine the significance of independent variables, we compared models with and without a given variable in terms of the χ^2 distributed differences in deviance. We compared the model with two different models, one of which has only sentence position as the independent variable, the other of which only topic shift. If the model without sentence position was significantly different from the model with both independent variables, we could conclude sentence position is significant predictor. We would then look at its coefficient in the regression model to determine if the effect was in the predicted direction. The same method also applies to the analysis of topic shift. We report χ^2-based p-values instead of those associated with error-based t-values in parameter estimation. This approach avoids problems with inflated standard errors for collinear predictions.

Results

The prediction that sentence information should be negatively correlated with topic shift is confirmed in the current study. We found that when the planned content of the upcoming sentence shifts the prior belief distribution of topics by 1 bit, the content will be encoded with 0.43 bits of information *per word* less than the case where there is no topic shift (models with a flat prior for the belief distribution of topics: $\chi^2 = 301.31, p < 0.0001$; informative prior: $\chi^2 = 273.29, p < 0.0001$). At the same time, the effect of sentence position is also significant. Speakers were found to encode 0.001 more bits of information per word in each subsequent sentence, after the effect of topic shift is controlled for (flat prior: $\chi^2 = 62.41, p < 0.0001$; informative prior: $\chi^2 = 65.6, p < 0.0001$).

[1]Following Qian and Jaeger (2010), we also conducted nonlinear mixed model analysis, which yielded qualitatively similar results. Those results are not reported here separately due to space limit.

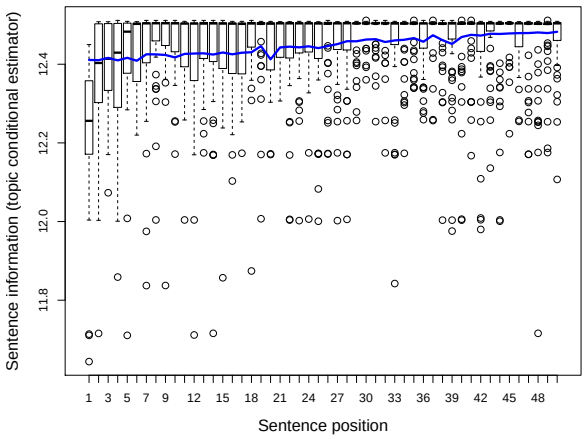


Figure 1: Boxplot shows the actual distribution of sentence information against sentence position. Blue curve shows the combined predicted effects of both predictors (random effects are not shown).

Figure 1 summarizes the results graphically. Note that the predictor topic shift allows our model to fit a high level of nonlinearities in the profile of sentence information. Previous studies have attempted to fit these nonlinear patterns with log- or restricted cubic spline (RCS) transformed sentence position (Qian & Jaeger, 2010). A post-hoc test revealed that topic shift remained a significant predictor even after these transformations were applied (for log-transformed sentence position: $\chi^2 = 303, p < 0.0001$; for RCS-transformed sentence position: $\chi^2 = 320.47, p < 0.0001$), indicating the nonlinear patterns captured by topic shift are additional to these general nonlinear functions.

Discussion

The results of Study 1 suggest that speakers are sensitive to topic shift. When the content of an upcoming sentence leads to a significant change in the belief of likely topics, the speaker typically plans less information in that sentence. This shows that the speaker is aware of the fact that the predictive power of discourse context is not as powerful as it would have been if there were no topic shift.

The effect of sentence position can be interpreted in two ways. One possibility is that topic shift is not a perfect correlate of discourse context. Thus, when topic shift is controlled for, not all of the effect of sentence position is subsumed. An alternative view is that the topic conditional estimator of sentence information do not reflect the actual measure of information that speakers use to organize a discourse. In the next study, we used a different estimator to explore this question.

Study 2

In Study 1, the topic conditional estimator calculates the information of a sentence by conditioning sentence content on the topic (that we as researchers assume to be) intended by the

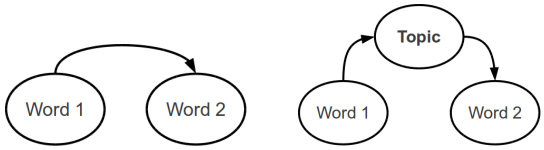
speaker. Since the intended topic of a sentence can only be known to the speaker, a listener must adopt a different strategy in estimating sentence information. One possible strategy is to infer the topics of the discourse word by word, and preserve the uncertainty in the inference process by performing marginalization over all possible topics. We refer to it as a *latent bigram estimator*, the details of which are described below.

Methods

The latent bigram estimator of sentence information estimates the information of a sentence s as follows:

$$\begin{aligned}\hat{I}_{lbe}(s) &= \sum_{w_i \in s} -\log_2 p(w_i|w_{i-1}) \\ &= \sum_{w_i \in s} \sum_{t} -\log_2 p(w_i|T=t)p(T=t|w_{i-1}) \end{aligned} \quad (9)$$

The last step of Equation (9) shows that the probability of a word w_i is obtained by marginalizing over a full distribution of latent topics, which is inferred from the context word w_{i-1}. Thus, the latent bigram estimator is different from a traditional bigram estimator that examines only the correlations between surface tokens. An illustration of the difference is shown in Figure 2.



(a) Bigrams (b) Latent bigrams

Figure 2: Regular bigram models vs. latent bigram models.

Inferring the posterior probability of a topic $p(T=t|w_{i-1})$ requires a simple manipulation of the Bayes' rule:

$$p(T=t|w_{i-1}) = \frac{p(w_{i-1}|T=t)p(T=t)}{p(w_{i-1})} \quad (10)$$

Terms in Equations (9) and (10) were directly obtained from the training results $\mathbf{P}(W|T)$. The prior distribution of topics $\mathbf{P}(T)$ was computed using Equation (1). The probability of the first word of a sentence is conditioned on the last word of the previous sentence. This sentence information estimates represents how much information that faces a rational listener in *a sentence*.

Results

We used the same statistical procedure as in Study 1. Topic shift remained a highly significant predictor of unconditional per-word sentence information (as estimated using latent bigrams). On average, speakers encode 0.05 bits less information in a sentence for every 1 bit of topic shift (model with a flat prior: $\chi^2 = 252.77$, $p < 0.0001$; informative prior: $\chi^2 = 257.56$, $p < 0.0001$; see Figure 3). Interestingly, the effect of sentence position is no longer significant. The estimated slope is close to 0 and is non-significant (flat prior: $p > 0.5$; informative prior: $p > 0.6$). Importantly, when topic shift is removed from the regression model, sentence position is a significant predictor on its own. Taken together, these results showed that the effect of topic shift subsumes the effect of sentence position.

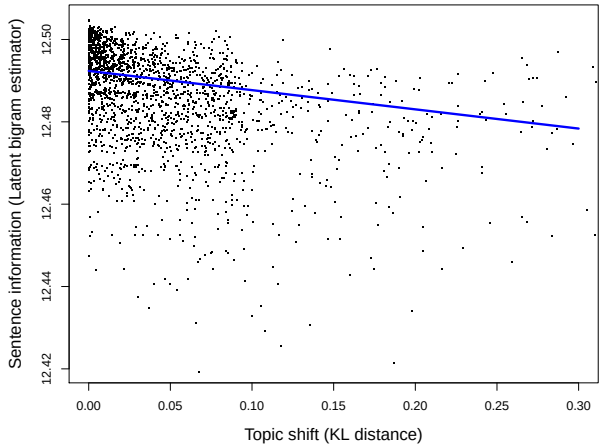


Figure 3: Sentence information (estimated by latent bigrams) is negatively correlated with topic shift. Scatterplot shows the actual distribution. Blue line shows the predicted effect. Random effects are not shown.

Discussion

Study 2 replicated the predicted negative correlation between topic shift and sentence information found in Study 1. With sentence information estimates reflecting how much information that the listener perceives, it shows that the speaker organizes discourse content in such a way that the listener would perceive less information when the discourse undergoes a topic shift.

	TCE (Study 1)	LBE (Study 2)
Partial – Topic shift	13.69%	10.32%
Partial – Sent position	2.89%	0.01%
Total R^2	17.10%	21.44%

Table 1: Partial and total R^2 of the both models.

The fact that the effect of sentence position becomes non-significant when topic shift is controlled for implies that topic shift accounts for the variance that is originally predicted by sentence position. This might be taken to mean that marginalization over topics is the more appropriate model of *what* speakers try to hold constant during discourse production. This is, however, potentially misleading. While overall more variance is accounted for in Study 2 (Table 1), the proportion accounted for by the topic shift predictor out of the overall variance is actually smaller in Study 2 (48.1%) than in Study 1 (80.1%). In other words, the higher R^2 of the model in

Study 2 is mostly due to more variance being captured by the random intercepts (which adjust for individual differences between writers). It is hence unclear which of the two models presented in this paper is more appropriate.

General Discussion

Unlike the test for a positive correlation between unconditional information and sentence position, the topic modeling approach employed here directly tests whether unconditional sentence information is uniform once discourse context has been taken into account. It is thus a stronger test of the hypothesis that language production is communicative efficient (Genzel & Charniak, 2002; Jaeger, 2010; Levy & Jaeger, 2007). The results of current studies showed a clear negative correlation between sentence information (based on content words only) and topic shift, no matter which estimator was used for sentence information. When sentence information is estimated from the listener's perspective, topic shift can even account for all the variances in sentence information that is originally predicted by sentence position.[2] These results suggest that speakers distribute less information in parts of a discourse where discourse context is less relevant.

Study 2 suggested that the speaker may be optimizing discourse content for the ease of comprehension. If confirmed by further studies (e.g. on other languages) and under the assumption that the latent topic model employed Study 2 cannot be motivated by production-internal considerations alone, this finding may be taken to speak to the question as to whether production is adapted to comprehension (but see the caveats discussed above). Models of audience design have emphasized the importance of a speaker taking their listener's knowledge and belief into consideration during production (Clark & Marshal, 1981). There is also evidence that speakers do not adapt to a specific listener, but to the way how language is comprehended in general (Brown & Dell, 1987). Our modeling results do not distinguish between these two views. To test a strong hypothesis of audience design, it would be necessary to first pick a *maximum a posteriori* estimate of the listener's perceived topic, like the speaker's MAP topic in Study 1, and then examine if sentence information estimates derived under such conditions follow the predicted pattern. Our approach deviates from such a method since it remains agnostic about what specific topics that the listener is committed to. We would like to pursue this direction with behavioral methods in future work.

Conclusion

We provide evidence that speakers adjust the amount of unconditional information encoded in a sentence according to topic shifts in a discourse. When sentence information was estimated from a rational listener's perspective (Study 2), this effect subsumes previously reported effects. This finding is compatible with the hypothesis that speakers plan utterances from the listeners' perspective to achieve communicative efficiency.

[2]Results were compared with an *n*gram estimator. The effect of topic shift holds. The effect of sentence position is only significant for the first 11 sentences. This closely matches previous studies where the effect tends to be reported only for early discourse.

References

Andrzejewski, D., Mulhern, A., Liblit, B., & Zhu, X. (2007). Statistical debugging using latent topic models. In *ECML* (p. 6-17).

Aylett, M. P., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang Speech*, *47*(1), 31-56.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python.* O'Reilly.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J Mach Learn Res*, *3*, 993-1022.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*, 441-472.

Clark, H. H., & Marshal, C. R. (1981). Elements of discourse understanding. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), (chap. Definite reference and mutual knowledge). Cambridge University Press.

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *J Stat Mech-Theory E*. (P06009)

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *ACL* (pp. 199–206).

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In *EMNLP* (p. 65-72).

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *PNAS*, 5228-5235.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychol*, *61*, 23-62.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *EMNLP* (p. 317-324).

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *NIPS* (p. 849-856).

Piantadosi, S., & Gibson, E. (2008). Uniform information density in discourse: a cross-corpus analysis of syntactic and lexical predictability. In *CUNY*.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* New York, NY: Springer-Verlag.

Qian, T., & Jaeger, T. F. (2009). Evidence for efficient language production in chinese. In *CogSci09* (p. 851-856).

Qian, T., & Jaeger, T. F. (2010). Close = relevant? the role of context in efficient language production. In *ACL 2010, CMCL Workshop.*

Qian, T., & Jaeger, T. F. (submitted). *Cue weight decay in communicatively efficient discourse production.*

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