Evidence for Efficient Language Production in Chinese

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

Abstract

Recent work proposes that language production is organized to facilitate efficient communication by means of transmitting information at a constant rate. However, evidence has almost exclusively come from English. We present new results from Mandarin Chinese supporting the hypothesis that *Constant Entropy Rate* is observed cross-linguistically, and may be a universal property of the language production system. We show that this result holds even if several important confounds that previous work failed to address are controlled for. Finally, we present evidence that Constant Entropy Rate is observed at the syllable level as well as the word level, suggesting findings do not depend on the chosen unit of observation.

Keywords: constant entropy rate; efficient language production; Chinese; cross-linguistic study; information theory.

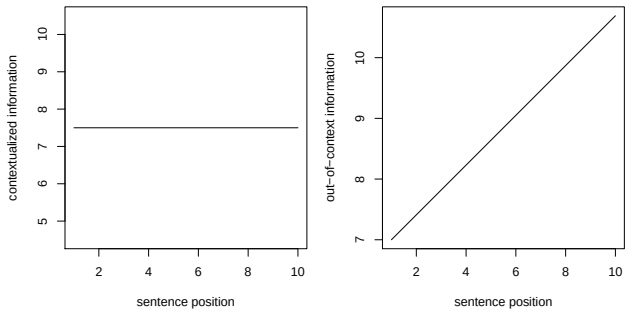
Introduction

The idea that language can be mathematically described just like any other communication system goes back at least to Shannon (1951), who suggests that anyone speaking a language should also possess a statistical knowledge of that language. According to Shannon, this statistical knowledge enables us to use language probabilistically, as evidenced by our ability to to fill in missing or incorrect letters in proofreading, to complete an unfinished phrase in a conversation, or to perform other common tasks.

Recent work has proposed that language users exploit this statistical knowledge for efficient language production (Genzel & Charniak, 2002; Aylett & Turk, 2004; Jaeger, 2006; Levy & Jaeger, 2007; van Son et al., 1998). Genzel and Charniak (2002) hypothesize that speakers use their probabilistic knowledge of language to maintain a constant entropy rate in language production. According to the information theory, transmitting information at a constant rate through a noisy channel is communicatively optimal (Shannon, 1948).

If speaker follow the principle of Constant Entropy Rate (hereafter, CER, Genzel and Charniak, 2002), we should observe that the sentences they produce carry on average the same amount of information. This direct prediction of CER is illustrated in Figure 1a. However, the direct prediction of CER is difficult to examine because it is difficult to derive estimate of sentences' information content in context. Current natural language processing techniques (e.g. *n*-grams, probabilistic context-free grammar models) only assess the a priori, or out-of-context, information of a sentence. To circumvent this problem, Genzel and Charniak tested an indirect prediction of CER: out-of-context sentence information should *increase* throughout discourse. This indirect prediction of CER is illustrated in Figure 1b.

To understand why out-of-context information ought to increase throughout discourse, one needs to look at the context-dependent nature of human communication. Utterances in a discourse build on each other. The information encoded in a string of words (or a stream of sounds) is co-determined by its context. In a situation where context is not properly provided, such as sentences that are randomly extracted from a well-structured discourse, the content will seem surprising without context. If speakers are efficient, they should thus encode less out-of-context information in sentence early in discourse than late in discourse, where more preceding discourse will (on average) lower the actual (contextualized) information content. However, a reverse pattern – too much information at the beginning and too little at the end – may turn a discourse overwhelmingly difficult to understand at the beginning and barely informative toward the end. This is hardly efficient from the speaker's perspective since it is likely to result in unsuccessful communication to the listener.



(a) direct prediction (b) indirect prediction

Figure 1: The direct and indirect predictions of CER

Corpus studies have provided evidence for the indirect prediction of CER. For articles in the Wall Street Journal corpus, Genzel and Charniak (2002) found that average out-of-context sentence information increases throughout the discourse (see also Keller, 2004). Piantadosi and Gibson (2008) found that data from spoken English also follow the prediction of CER.

In this paper, we build on these previous findings. We address certain methodological shortcomings and extend the scope of the empirical investigation of CER in three important ways. First, previous studies reported only gross correlation between sentence information and sentence position. We use a linear mixed model to analyze the relation between out-of-context information content and sentence position, *while controlling for possible confounds such as potentially non-linear effects of sentence length*. **Study 1** replicates Genzel

and Charniak's studies with this new method to validate CER in English.

Second, if CER is a result of rational language production, its effects should be widely observable in any language. **Study 2** presents the first test of CER on a non-Indo-European language, Mandarin Chinese. Mandarin Chinese differs typologically from previously studied languages. If the prediction of CER can be observed in Mandarin Chinese, this would provide cross-linguistic support for CER as a principle of efficient language production.

Third, previous studies are limited to study CER at the *word* level. That is, they have tested whether out-of-context per-word entropy is correlated with the sentence position in discourse. However, nothing about the hypothesis of Constant Entropy Rate attributes a special status to words compared to other linguistic units. If constant entropy rate is a pervasive effect in language production, it should be observed even if entropy is calculated over units other than words. **Study 3** provides a test of this prediction by extending Study 2 over toned syllables.

Methods

This section describes the computational model used to derive sentence information based on a language model. First, an *n*-gram model is used to compute probability estimates of sentences. Then, Shannon information (Shannon, 1948) estimates, which are used as the measure for sentence information, are derived from probability values. Finally, the regression model used in current studies is described.

Language Model An *n*-gram language model can be used to make predictions about the probability of a word given the *n-1* preceding words. Here, we use a trigram model. That is, the probability of a word is conditioned on two preceding words. While a trigram model is unlikely to reflect human probability estimates, there is no reason to believe that this simplification introduces a confound. Additionally, Genzel and Charniak (2002) showed that the results derived by trigram models closely matched those derived by probabilistic phrase structure grammars. Good-Turing discounting and Katz back-off (Katz, 1987) were used to smooth probability estimates. Katz discount coefficients i.e. α values are also derived from training data and used for backing off to bigrams or unigrams when a specific trigram is not in the model. This methods yield more reliable estimates for low count events and provide probabilities even for words that have not been observed in the training corpus (out-of-vocabulary words, OOVs).

$$P_{katz}(w_3|w_1,w_2) = \begin{cases} P(w_3|w_1,w_2) & \text{if } c(w_1,w_2,w_3) > 0 \\ \alpha(w_2|w_1) * P_{katz}(w_3|w_2) & \text{if } c(w_1,w_2) > 0 \\ P(w_3) & \end{cases} \tag{1}$$

where

$$P_{katz}(w_2|w_1) = \begin{cases} P(w_2|w_1) & \text{if } c(w_1,w_2) > 0 \\ \alpha(w_1) * P(w_2) & \end{cases} \tag{2}$$

In (1) and (2), $c()$ refers to the joint count (the number of collocations), and P refers to unsmoothed probability estimates.

Sentence Information Estimates The Shannon information content of a word is defined as the logarithm of the reciprocal of the word's probability (i.e. $\log_2 \frac{1}{p(w)}$), a quantity whose unit is called *bits* when the logarithmic base is 2. The information content of a sentence S is the sum of information of all its words:

$$I(S) = \sum_{w_i \in S} \log_2 \frac{1}{P_{katz}(w_i|w_{i-2},w_{i-1})}. \tag{3}$$

Since the a sentence's information content increases with additional words, previous work (Genzel & Charniak, 2002, 2003; Keller, 2004) has tested CER by calculating *entropy rates* for all sentence in the k-th position of a discourse:

$$H(S_k) = \frac{1}{N} \sum_j \frac{1}{|s_j|} I(s_j) \tag{4}$$

In Equation (4), $H(S_k)$ is the entropy rate, N is the total number of sentences at Position k, $|s_j|$ is the length of a sentence j, and $I(s)$ is its information content (as in equation 3). Computing entropy rates for sentence positions conveniently allows for correlation tests between sentence position and entropy rates reported by early studies.

However, correlations over averages do not control for possible *non-linear* effects of sentence length on sentence information. This approach also does not extend well to the investigation of additional controls. Here, we use linear mixed regression models to circumvent this problem.

Regression Analysis Individual sentence information estimates are regressed against sentence position (the position of a sentence in discourse) and several additional control predictors.

1. Sentence length is an important predictor of sentence information, because, mathematically, sentence contains more information when they have more words (for an additional word w, $\log_2 \frac{1}{p(w)} > 0$ is always true). This idea is also intuitive since each word must serve to convey some information in a sentence. **2. Out-of-vocabulary words** (OOVs) are words that have not been observed by the language model during the training phase. Since it is unlikely to find a corpus large enough to include all words of a language for the language model to learn, OOVs are a common problem in fitting *n*-gram models. We include a separate control for OOVs in our model to ascertain that any effect of sentence position is not purely driven by OOVs. Due to the smoothing algorithm (see above), these words may be assigned a random (but uniform) probability value and thus have superficially high information content. While OOV words are typically low frequency (and hence high information words), sometimes even relatively frequent words may not be observed in the training data.

To test for **super- or sublinear relations**, we model all controll effects and the effect of sentence position using 2nd-order polynomials. In future work, we plan to explore additional means of modeling non-linearities.

The **random effects** of document-level differences and agency differences, if such information is available, are also controlled for to make sure any potential effects hold beyond the particular texts observed in the sample.

Study 1: Evidence for CER in English

The purpose of Study 1 is to test whether the results reported in Genzel and Charniak (2002) hold after all potential confounds discussed above are addressed.

Data

The trigram model is trained on Sections 0-20 of the Wall Street Journal subset of the Penn Treebank, and Sections 21-24 are used for hypothesis testing. We apply the same cut-off value for sentence position as in Genzel and Charniak (2002) – only the first 25 sentence positions are considered. There are a total of 42,075 sentences in the training set, and 7,133 sentences in the test set.

Prediction and Results

If English speakers produce language efficiently, according to CER, they should encode more information in late sentences than in early ones. This is indeed observed: on average, the sentence information increases by 0.29 bits for each increase in sentence position (linear: $\beta = 0.29$, $t_{(5456)} = 3.87$, $p < 0.001$, quadratic: $\beta = -0.03$, $t_{(5456)} = -3.03$ $p < 0.003$). Unsurprisingly, longer sentences encode more information: each additional word corresponds to 7.76 bits of information[1]. Interestingly, the effect contains a significant quadratic component (linear: $\beta = 7.76$, $t_{(5456)} = 169.83$, $p < 0.0001$; quadratic: $\beta = 0.01$, $t_{(5456)} = 9.13$, $p < 0.0001$). The number of OOV words did not reach significance ($p > 0.1$) (we will discuss the role of OOV in detail in the next study, where it has a significant effect on sentence information).

Discussion

Our results provide converging evidence to findings in previous studies (Genzel & Charniak, 2002, 2003; Keller, 2004). Keller (2004) discovered that each increase in sentence position correlates with an increase of 0.64 bits of information per word (using the same corpus data). However, our results show a smaller effect. English speakers add only 0.29 bits of information to each subsequent sentence. That is, the effect approximately corresponds to an increase of only 0.01 bits per word on average, given that the mean sentence length is 25 words in this data. This suggests that previous studies may have overestimated the effect of sentence position (and hence the role of CER in discourse planning).

Note that the data have shown an interesting sublinear effect of sentence position, suggesting that sentence information, even when it is measured out of context, does not increase indefinitely. This suggests that the average amount of a priori information speakers convey per word (entropy rate) converges against some as of yet unknown maximum (see Piantadosi & Gibson, 2008 for a similar observation). It is possible that information is distributed across natural language use in such a way that more local contextual cues are more informative (as hypothesized by Piantadosi, p.c.). This would indeed result in the observed convergence against a maximum entropy rate. Future work is necessary to test this hypothesis.

In summary, Study 1 provides evidence that CER holds for English, after further controls. We also find that the additional controls we introduced are justified (e.g. the weak, but highly significant superlinear effect of sentence length). Study 2 investigates whether CER also holds in Mandarin Chinese.

Study 2: Evidence for CER in Chinese

Only one previous study has investigated the extent to which CER holds cross-linguistically. Genzel and Charniak (2003) analyzed the novel *War and Peace* in both Russian and Spanish and found that sentence entropy correlates with sentence position in all three languages, as predicted by CER. However, Genzel and Charniak's results suffered from the methodological drawbacks mentioned above. We extending the study of CER to Mandarin Chinese, a language typologically and areally unrelated to previously studied languages.

Data

The Chinese Treebank v6.0 (Xue, Xia, Chiou, & Palmer, 2005) corpus is used in Study 2. The corpus consists of 2,036 documents, containing 28,295 sentences, and 781,351 words. The content of the corpus is a mixture of news articles from four news agencies: Xinhua news, Hong Kong news, Taiwan Sinorama, and ACE broadcast news. Corpus articles are predominantly in the form of news reports and differ little in style across agencies.

All headlines, author lines, and ending lines (i.e. the word "-fine-" typically written at the end of a Chinese news report) are ignored. Abstract-style news reports are also excluded from the corpus. Then, we create a balanced dataset with the following sampling method: we select articles with more than 10 sentences, and then extract the first 10 sentences from each (i.e. there are an equal number of sentences in each position). 6,240 sentences are used for training and 320 sentences for testing CER[2].

Our sampling method described as above is adopted from Genzel and Charniak (2003). It ensures that probability patterns of *n*-grams that are only typical of sentences after the cut-off position (i.e. 10 in our work) will *not* be learned by the language model. Therefore, it avoids another potential confound.

[1] This is also the average per-word information.

[2] Study 1 did not follow this method in order to be as close to Genzel and Charniak (2002) in design as possible.

Results

Due to the small size of the Chinese Treebank as well as our selection threshold (no. of sentences ≥ 10), testing of the CER hypothesis is limited to a relatively small dataset (n = 320). The hypothesized effect of sentence position fails to reach significance (linear: $p > 0.3$; no quadratic effects). The amount of information that a word contributes to a sentence is 9.35 bits (linear: $\beta = 9.35$, $t_{(282)} = 61.58$, $p < 0.0001$; quadratic: $\beta = -0.01$, $t_{(282)} = -2.82$, $p < 0.01$). Since sentence length is already controlled for, the slope parameter for the predictor OOV is really about how many bits of information will be changed if a known word in a sentence is made unknown. That is, it is the average difference between information of regular words and OOVs. In this case, each additional OOV word has 11.81 more bits than a known word on average (linear: $\beta = 11.81$, $t_{(282)} = 6.01$, $p < 0.0001$; no quadratic effect: $p > 0.6$; see Figure 2).

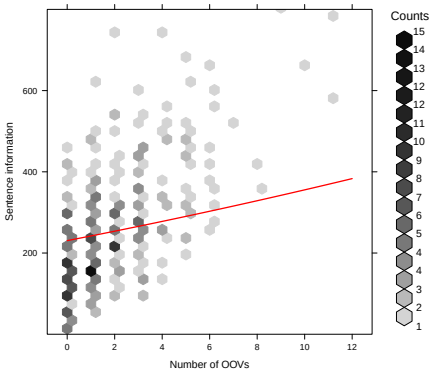


Figure 2: The number of OOV words has a significant effect on sentence information.

Discussion

These results seem to suggest that later sentences in discourse are not encoded with more information than earlier ones. The increase of sentence information seems to be related to the number of unknown words in a sentence rather than to an increase in sentence position. However, what the predictor of sentence position in this model really tests is whether information content of *known* words in a sentence correlates with sentence position. Although this hypothesis is not supported by results obtained in this model[3], it is also much more conservative than the original hypothesis of CER by viewing OOVs merely as a confounding factor.

[3]The Chinese data set is relatively small, which results in a language model that does not recognize 7.1% of the words in test data. To understand this number in context, the English language model built on the much larger WSJ corpus contained only 3.4% OOV words. In other words, probability estimates in the current study are likely to be more noisy, and OOV words have bigger influence on our result.

Out-of-vocabulary words are, as a matter of fact, low-frequency words and therefore should have higher-than-average information. Unfortunately, the n-gram model used here cannot distinguish them from each other and treat them as a group. Consider that the amount of information of each OOV word is the same under such a model. This obscures the real information those words would be encoded with by human speakers. This distinction is particularly problematic to the current study, since Chinese word boundaries marked in corpus data tend to make words seem more informative to a language model than they really are to speakers. For example, "红旗" means "red flag", and is typically tagged as a single word consisting of two characters in corpus data. "白旗(white flag)", a similar word whose meaning only differs in the color property of that flag being referred to (whiteness instead of redness), will nevertheless be marked as an OOV word if the language model has not seen "white flag" explicitly during training. On the contrary, this word is unlikely to be of high information to native Chinese speakers.

As a result, we adopt two alternative methods. **Method 1** tests the CER hypothesis on the training data (nine times as much data). Although the ideal scenario would be to have a language model that recognizes every word in the test data, training data do have an OOV rate close to 0. By definition, a trained language model is overfitted to its training data. Probability estimates will be superficially high, resulting in underestimation of words' information content. However, underestimated information content in itself does not lead to the conclusion that out-of-context sentence information will increase throughout discourse. When using training data to test the CER hypothesis (6,240 sentences), we find a strongly significant effect of sentence position ($\beta = 0.45$, $t_{(5612)} = 9.46$, $p < 0.0001$) and once again, a sublinear effect ($\beta = -0.08$, $t_{(5612)} = -4.33$, $p < 0.0001$). Interestingly, the sublinear effect coincides with the case of English: there is a limit of out-of-context sentence information yet to be investigated in future work.

Method 2 drops the control for OOV words in the linear mixed model to test whether the information content of *all* words (out-of-vocabulary as well as within-vocabulary) in a sentence correlates with sentence position. However, the trade-off is that the information content of all unknown words has to be approximated to a uniform quantity (21.16 bits). The predicted effect of sentence position reaches marginal significance in this model: Chinese speakers increase sentence information by 1.02 bits for each subsequent sentence (linear: $\beta = 1.02$, $t_{(284)} = 1.62$, $p = 0.10$; no quadratic effect; see Figure 3). Additionally, longer sentences contain significantly more information than short sentences (linear: $\beta = 10.22$, $t_{(284)} = 64.25$, $p < 0.0001$; quadratic: $\beta = -0.01$, $t_{(284)} = -1.91$, $p = 0.05$; see Figure 4).

From the results produced by Method 1 and the marginally significant linear effect reported using Method 2, we tentatively conclude that Chinese speakers also produce language efficiently.

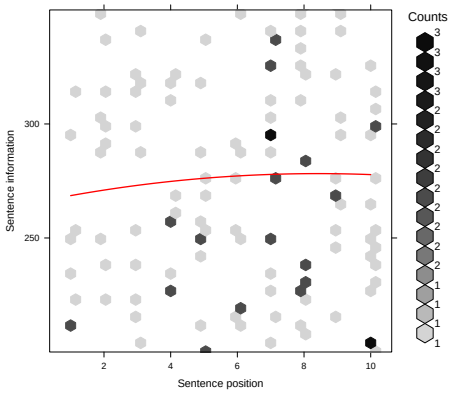


Figure 3: Effect of sentence position

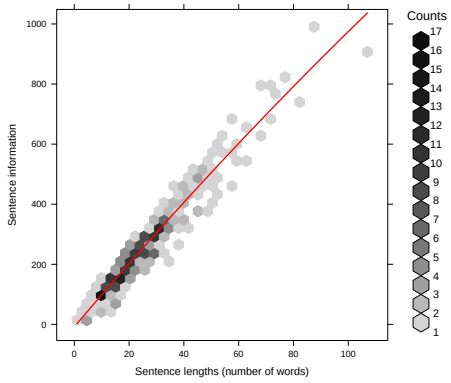


Figure 4: Effect of sentence length (number of words)

Study 3: Syllable-Level Evidence

This study presents yet another way of reducing sparsity in Chinese data: applying a syllabic transformation to the test part using the *Pinyin* system. The one-to-many mapping relation from Pinyin to characters is useful in reducing the dimensionality of the feature space needed to model the language.

Data

Study 3 uses the same dataset as in Study 2. However, corpus data are first converted to a syllabic representation before being used to build a language model. We use the Pinyin system to represent the syllabic characteristics of Chinese. Tones are coded with numerals 1(flat), 2(rising), 3(falling-rising), 4(falling), and 5(neutral). Furthermore, word boundaries are removed from the corpus so that this study effectively models the probabilistic relation between Chinese syllables.

Table 1 confirms our expectation that the syllabic model fare better than the word-level model used in Study 1 in terms of reducing the number of OOV words. The results are obtained by evaluating the same test data under different representations. While there are more than 7 unknown words for every hundred words in the word-level model, this percentage is close to 0 in the syllabic model. Improvement in OOV rate correlates with an increase in the number of observed trigrams as well as a reduction in entropy. In summary, the influence from OOV words on the effect of sentence position is minimized, and reduced entropy rate also suggests improved accuracy in estimating information content.

	3-gram coverage	OOV rate	Perplexity
Character	15.84%	7.14%	460.29
Pinyin	43.84%	0.02%	79.92

Table 1: The quality of Pinyin model is better than that of character model in terms of higher 3-gram coverage, and lower OOV rate as well as perplexity (perplexity=$2^{entropy}$).

Predictions and Results

Given the decrease in data sparsity, the hypothesized effect of sentence position is expected to reach significance, while the previously observed OOV effect may become non-significant. This is indeed observed. For each subsequent sentence in test documents, sentence information independently and significantly increases by 2.28 bits (linear: $\beta = 2.28$, $t_{(284)} = 12.04$, $p < 0.02$; no quadratic effects). The effect of OOV words is weakened (linear: $p > 0.4$; quadratic: $p > 0.6$). Again, we sentence length affects sentence information in the expected way, although this time only a linear effect is observed (linear: $\beta = 6.82$, $t_{(284)} = 237.13$, $p < 0.0001$; no quadratic effects).

Discussion

With improved information estimates, we are able to confirm that Chinese speakers seem to follow the predictions of CER. That is, Mandarin Chinese production distribute information across sentence in a way that is efficient for communication. The results also show that the predictions of CER hold not only for per-word information, but also at the syllabic level. Efficient language production on the syllabic level is not improbable since Chinese characters/syllables are generally meaningful by themselves. Although the meaning of a syllable is often ambiguous out of context, syllables still carry information from which succeeding characters are highly predictable. A more careful look at the results reveals that the increment of information between connected sentences is more than double the size in the word-level model (2.28 bits vs. 1.02 bits). In other words, more information appears to be added in each subsequent sentence when production efficiency is measured on the syllable level.

Conclusions and Open Questions

The present work has demonstrated through a series of three studies that language production of Mandarin Chinese is efficient independent of whether sentence information is estimated over words or syllables. We find that the principle of Constant Entropy Rate (CER) is observed by speakers of English (Study 1) and speakers of Mandarin Chinese (Studies 2

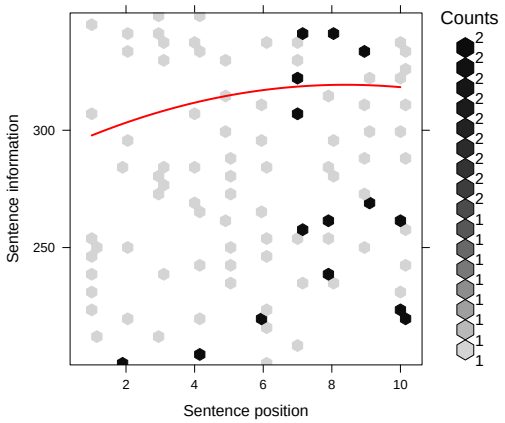


Figure 5: The effect of sentence position in Study 3.

and 3) even after several confounds not addressed by previous work (Genzel & Charniak, 2002, 2003; Keller, 2004) are accounted for: language producers seem to distribute information uniformly across the sentences they produce. Findings in this paper lend support to the rational cognition argument in general and its application in optimizing information distribution in language communication.

Note that the evidence for CER in Chinese found in above studies only indicates that Chinese speakers produce the language efficiently in writing. Do the same results hold for speech too? Another work of ours has affirmed this hypothesis by testing CER on a speech corpus of Chinese broadcast news transcripts with the same syllable-level modeling framework (part of Qian, 2009). Chinese speakers are shown to increase information for each subsequent sentences in spontaneous speech as well, just as English speakers do (Piantadosi & Gibson, 2008). Research on more languages is needed to further strengthen the argument that CER holds for both writing and speech.

Finally, in an independent line of research, one of us (TQ) has been investigating to what extent second language users' performance can be considered efficient (Qian, 2009). With regard to CER, it makes sense to ask the question whether non-native speakers who are sufficiently well-trained in using the target language are able to produce it efficiently. For example, when English L2 speakers whose native language is Chinese produce English, do they distribute less information in early discourse and more later on? On the one hand, they should do so because they are capable of producing language efficiently. On the other hand, they apparently have an imperfect knowledge of language statistics of English. Preliminary evidence leads us to believe it is a matter of perspective: non-native language production seems efficient if the language model is also trained on non-native corpus data; however, it violates the prediction of CER (i.e. in the form of putting more information at the beginning and less later on) if the language model is trained on native corpus data (Qian, 2009). In other words, non-native speakers also try to maximize the efficiency of language production according to their grammar and vocabulary of the target language. It is only that this knowledge of the target language differs from native speakers' version.

References

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. *Language and Speech*, *47*(1), 31-56.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 199–206). Philadelphia, PA.

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

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(3), 400–401.

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

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of NIPS*.

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

Qian, T. (2009). *Efficiency of language production in native and non-native speakers.* (University of Rochester, Unpublished thesis)

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423,623–656.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, *30*, 50-64.

van Son, R. J. J. H., Beinum, F. J. K., & Pols, L. C. W. (1998). Efficiency as an organizing principle of natural speech. In *ICSLP*. Sydney.

Xue, N., Xia, F., Chiou, F.-D., & Palmer, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, *11*, 207–238.