Valence Weakly Constrains the Information Density of Messages

David W. Vinson (dvinson@ucmerced.edu)
Rick Dale (rdale@ucmerced.edu)
Cognitive & Information Sciences
University of California, Merced
5200 N. Lake Rd., Merced, CA 95343

Abstract

Some recent analyses of language as a transmission medium have fruitfully applied information theory in various ways to sequences of words. In most cases, the information contained in a word is defined as a function of that word's local context (e.g., its probability conditioned on the preceding word). A central assumption in much of this work is the important role of context. For example, the hypothesis of uniform information density (Jaeger, 2010) requires some notion of context in order to be tested. We sought a structured corpus in order to extend and explore the potential role of a context in the observed information density of messages. Specifically, how might a language user's affective state influence their language use? We used a database of over one hundred thousand consumer reviews that includes an assortment of user-related variables. These user-related variables, such as the overall rating of a review used here as a proxy for a user's affect, appear to have an interesting relationship to basic information-theoretic measures such as the average amount and variability of observed information of a review's words. We discuss these results in terms of the broader context that may shape the information structure of messages, and relate these findings to existing theories.

Keywords: language; information theory; context; corpus analysis; word distribution; natural language processing

Introduction

Tools from information theory have allowed researchers to explore whether language use is, in some sense, optimal (e.g., Levy & Jaeger, 2007). At the production level, speakers may structure their utterances so as to optimize information density (Jaeger, 2010), while over longer timescales aspects of language such as word length, may be optimized according to information content (Piantadosi, Tily, & Gibson, 2011).

In most cases, factors beyond the lexical level that influence information density must be abstracted away. For example, "context" is often confined to a lexical definition, namely the immediate preceding word. In this case, the information encoded by a word can be expressed using the log of the probability that the word would occur in this lexical context:

$$I(w_i) = -log_2 p(w_i|w_{i-1})$$

Though easy to compute, this definition abstracts away a variety of other contextual factors such as a user's cognitive affective state, message content and intended audience that may help explain why a user chooses a given word. This simplification is justifiable, of course, because of the difficulty in defining other contextual factors (e.g., at a semantic level), and the complexity that seems endemic to high-level aspects of language (see Jaeger, 2010 for discussion).

More recently, studies have begun to show information density is influenced by factors at a variety of linguistic levels including syntactic variation and phonetic reduction (Aylett & Turk, 2004; Jaeger, 2010; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). Relatedly, the information density of a linguistic message may be subject to more social or cognitive constraints that help define the content of a message, reaching beyond phonological and syntactic levels. In other words, in abstract terms, the transfer of information may be subject to a variety of ubiquitous contextual constraints at a variety of levels.

One such constraint, and one of interest to the current paper, is the relative valence of a linguistic message (and potentially, the language user herself). If a message is intended to be a highly positive evaluation of some situation, does the language user seek different patterns of information density to convey it?

As we review below, there is reason to suspect that such a pervasive contextual variable — like that of intended message valence; one specific to the user while composing a message — may shape the information content of that message. Evidence for this would further support the idea that the information-theoretic properties of language are contextually modulated. We sought a corpus well suited to test this idea. We analyzed over 100,000 consumer reviews with associated information about a review's rating. Even after accounting for a variety other linguistic variables, findings show the valence of a user's message influences the amount of information transmitted. Crucially, specific findings depend on how the linguistic context and as a result, information, is quantified.

Lexical Constraints on Information

Studies that stem from an information-theoretic standpoint have only recently begun to theorize what contextual effects influence the transfer of information at a lexical level (Aylett, 1999; Genzel & Charniak, 2002). One such theory, known as Uniform Information Density (UID), states that language users will structure their utterances so as to optimize information transfer within a given context (Frank & Jaeger, 2008; Levy & Jaeger, 2007). That is, a speaker will communicate at a rate that is optimal for transferring the greatest amount of information within a specific (noisy) channel, without loss of information or miscommunication (Genzel & Charniak, 2002). Recent evidence supporting this notion shows speakers may be sensitive to linguistic probability distributions that help define the information density of a message (Fine & Jaeger, 2011; Fine, Qian, Jaeger & Jacobs, 2010).

In support of optimal information transfer theories Aylett (1999) found individuals take longer to communicate more information dense messages. In addition, Levy and Jaeger (2007) found that speakers' use of an optional “that” complementizer is dependent on the information density of their utterance. Further, Piantadosi, Tilly and Gibson (2011) show word length in general may be optimized to the amount of information transferred, in contrast to the well known *Zipf’s law* which posits that word length is optimized for the frequency of word use (Zipf, 1949). Each study shows information density optimization occurs in subtly different, but related ways.

Crucially, a word's lexical context stands as the primary constraint guiding one’s understanding of the amount of information present within any given message; even though other, higher-level visual and social constraints are known to influence language use and comprehension (see Vinson, Dale, Tabataebeian & Duran, in press, for review). Importantly, if individuals are sensitive to specific linguistic probability distributions, social or cognitive factors influencing these distributions may affect the density of information within a message.

Affect and Message Valence

The information density of a message is at least partially dependent on contextual constraints such as its local lexical context. However, lexical contexts may be further influenced by other, more global, constraints.

Several findings, especially in social cognition, recommend this hypothesis. In particular, past research suggests cognitive or affective states with more positive valence are likely to generate more flexible, open-ended behaviors (Cacioppo & Gardner, 1999; Diener & Diener, 1996; Fredrickson, 2001; Isen & Means, 1983). Consider an example study that shows this tendency. When primed to experience a positive affective feeling, doctors correctly diagnose patients faster than doctors not primed to have this experience (Estrada, Isen, & Young, 1997). Doctors were more likely to accept new information when in a positive affective state than when in a neutral state. Similarly, it may be that when experiencing a positive affective state, one might transmit a more information-dense message or one that is more variable or open, than when in a less positive state provided transmitting and receiving information is affected by similar contextual constraints.

This notion finds relevance in information-theoretic terms where positive valence may provide an appropriate context for transferring more or broader information. We speculate further on this relationship below, but one possibility is that particular affective states might increase the channel capacity for both sender and recipient. Though a provocative hypothesis, the corpus we use here provides a massive amount of text data where individuals label their experiences as positive or negative on a scale of 1-5 and briefly report on them. We speculate that one’s experiential rating provides a measure whereby the influence of one's affective valence on information density can be assessed; even if only weakly connected.

We hypothesize that when individuals experience a positive affective state, their use of language may be more informative, more lexically rich and differ in frequency of use compared to individuals in a more negative affective state. Provided this hypothesis, one’s affective state, may be predicted by their language use; acting as a constraint on the density of information transferred over the course of a single message or, in this case, a consumer business review.

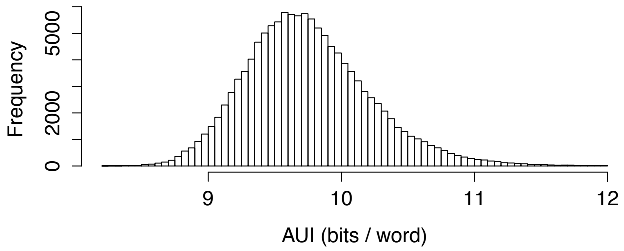
Current Study

The current study used a dataset from Yelp, Inc. consisting of over 100,000 consumer reviews of businesses throughout the city of Phoenix AZ. This dataset consists of written reviews associated with the reviewer’s explicit feelings specified by a rating from 1 (negative) to 5 (positive) stars about the business reviewed. Each review was subject to being labeled useful, funny, or cool by other reviewers. For the purpose of this study we assume the cognitive-affective state of an individual is in some way correlated to the number of stars associated with their review; more stars being correlated with a more positive affective state while fewer stars are correlated with a negative affective state. Because this corpus of reviews consists of both an explicit rating of the business and a linguistic message about the consumer’s experience, it is highly suitable for testing how various contextual factors, in particular cognitive-affective states, influence the information density of a linguistic message.

Measures and Method

Prior to testing how a reviewer’s cognitive state might influence the information density of their message, we must define measures of information that seem relevant. This has been done in a variety of ways. Here we define information in four very simple ways, commensurate with classic information-theoretic definitions. Each function defines the linguistic context of an utterance slightly differently. Importantly, such differences might reveal a unique relationship to message valence. Listed here are the four functions along with a brief definition of each:

(1) Review-internal entropy (RI-Ent). A review may simply be structured in distinct ways depending on how a language user decides to use lexical tokens more or less regularly in a way purely internally to a review itself. In other words, the frequency distribution over words may reflect a diverse selection of types (higher entropy), or it may be relatively more repetitive (lower entropy). This can be expressed in the following way:



Very low AUI: "This is a great place for lunch and dinner. The food is great, the price is good and the service is friendly and quick." [AUI = 7.4]

Very high AUI: "I don't know if this qualifies as an update. However 101 Bistro is now closed. Eighty sixed. Nada here anymora. Adios. Hasta la pasta." [AUI = 13.9]

Figure 1: The distribution of AUI. Note: we omit .05% of the data that is farther out on the details (72 of about 124,000 cases). Example (short) reviews are shown on the relevant side of the distribution. All measures exhibited unimodal, near-normal distributions, with some observations on distant tails (expected given the very large sample).

$$RI\text{-}Ent_j = -\sum_{i=1}^{N} p(w_i|R_j) log_2 p(w_i|R_j)$$

Here, *RI-Ent$_j$* denotes the j^{th} review, containing N words, as the probability of the i^{th} word occurring within that review (for notational convenience we treat it as a conditional probability, equivalent to restricting computations to a given review). This measure can be seen as a kind of lexical richness score, expressed as the expected number of bits required to encode a message given its unique internal word distribution. If information density is high, the text can be said to be lexically rich. Indeed, it can be easily shown that RI-Ent correlates with common measures of lexical richness, such as type-token frequency. Put simply, a review with higher entropy will have more unique tokens, thus being, in a sense, more "information dense."

(2) Average unigram information (AUI). This measure is computed from the lexical distribution over the entire set of Yelp reviews. As noted in the introduction, the information encoded in a word can be simply seen as the negative log of the probability of its occurrence (the less probable a word, the more informative). For any given review *j*:

$$AUI_j = -\frac{1}{N}\sum_{i=1}^{N} log_2 p(w_i)$$

This differs from the previous measure in that the probability of a word's occurrence is defined by a much larger distribution of words. If we regard the overall distribution of terms in the Yelp corpus as a simple but direct measure of how informative a word is, then a review may vary in its informational content depending on the language user's state.

(3) Average conditional information (ACI). A more common way of expressing the information encoded in a word is *relative* to some context (i.e., a second-order estimate). As noted in the introduction, this is commonly taken to be some immediate lexical context. In our case, we extract a very simple contextual information measure:

$$ACI_j = -\frac{1}{N-1}\sum_{i=1}^{N} log_2 p(w_i|w_{i-1})$$

Here the information in a word is the negative log of the probability of its occurrence given the previous word. This differs from RI-Ent and AUI in that it accounts for the most immediate or local context, namely, the previous word.

(4) Conditional information variability (CIV). ACI reflects the average information, but the work of Jaeger (2010) and Levy and Jaeger (2007) suggests that the uniformity, or variability, of this information measure may be interesting to explore.

$$CIV_j = \sigma(\mathbf{CI}_j)$$

Here, $\mathbf{CI}_j$ is the set of conditional information scores for each word of the j^{th} review; we compute the standard deviation of this set. Greater variability in information density would reflect an increase in the channel capacity. This would permit more variability in word choice allowing

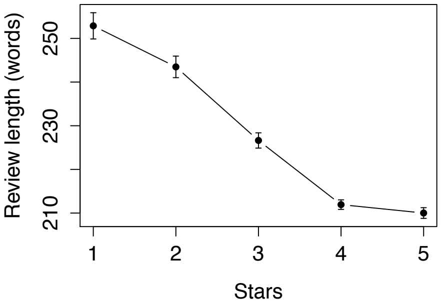


Figure 2: The number of words in a review (y-axis) by star rating (x-axis). Error bars reflect 95% confidence intervals over the whole filtered Yelp dataset.

[1] The full Yelp dataset contains about 229,000 reviews. We filtered this dataset by choosing reviews with 100 words or more so as to increase the reliability of our information measures.

[2] Due to the computation required in estimating models from so much data, we chose simple and multiple regression with `lm` in R; we also confirmed general patterns by centering scores relative to reviewers, and exploring linear mixed-effects models.

differences in the rate of information transmitted (i.e., by diminishing range restriction).

Measures 1-3 are derived from previous studies that focus on the probability of a word's occurrence. A word's probability is dependent on subtle differences in how its lexical context is defined. The fourth measure, CIV, is novel. According to UID the variability of information density should remain relatively constant across a message. CIV measures the variability of information across messages within a specific context. Differences in CIV dependent on messages' affective context would indicate fluctuation in the context's channel capacity. This would support a notion of optimal information transfer that is dependent on other contextual factors such as cognitive-affective states.

From these measures, reviews can be defined as more or less information dense dependent on their general and local linguistic context. For example, Fig. 1 shows the Average Unigram Information distribution over more than 100,000 reviews along with two example reviews. Using simple measures we tested if information encoded in a message is related to cognitive context: the intended valence of that message. To test this, we use star rating to predict information in regression models: Does variation in valence (rating) predict the level of information encoded?

124,622 Yelp reviews[1] were imported and processed in Python using `json`. We used `nltk` and `numpy`/`scipy` libraries to carry out most calculations. To calculate RI-Ent we used `nltk`'s MLE entropy function.

Results

At least one obvious measure may correlate with star ratings: review length (in number of words). We first test this variable and then include it as a covariate when testing our key information-theoretic measures.

Review length. It is well known that bin count can impact our key information-theoretic measures. In fact, review length indeed differed by star rating (see Fig. 2). We used a simple linear model to predict review length by stars.[2] There were significantly more words per review for lower stars ($r^2 = .01$, $t(124,621) = -38.7$, $p < .001$). This represents a small but significant effect—detectable thanks to the massive power of the large Yelp corpus. We used review length as a covariate in our subsequent analyses of information-theoretic measures.

(1) Review-internal entropy (RI-Ent). When not controlling for review length there was a small, but highly significant effect of stars in predicting RI-Ent ($r^2 = .009$, $t(124,621) = -34.01$, $p < .001$). Again, this shows a reliable but weak effect; stars account significantly for about 1% of the variance in RI-Ent (see Fig. 3A).

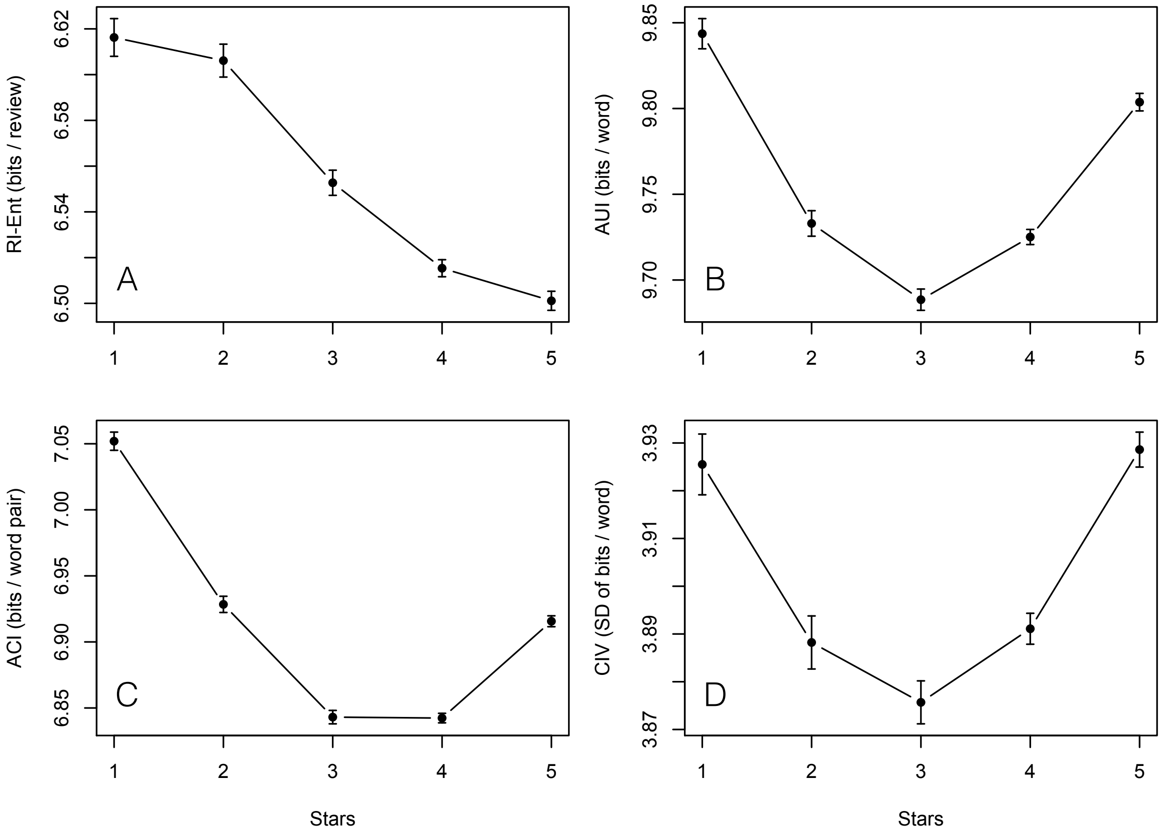


Figure 3: Initial relationships, without additional covariates, between star rating and (A) review-internal entropy (RI-Ent), (B) average unigram information (AUI), (C) average conditional information (ACI), and (D) conditional information variability (CIV).

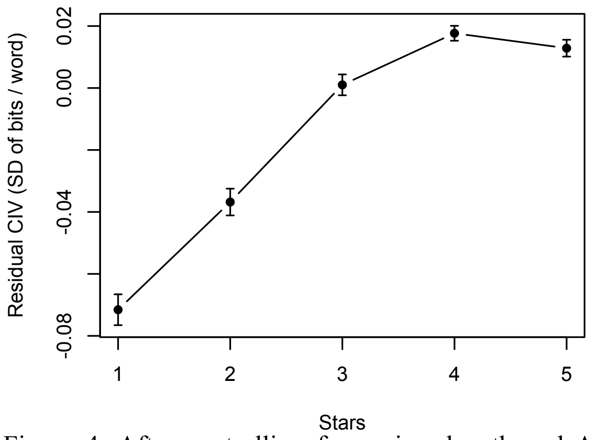


Figure 4: After controlling for review length and ACI, there is a small but significant relationship between CIV and stars.

We controlled for review length by fully residualizing RI-Ent in the following way: We predicted RI-Ent by review length, and stored the residuals as a new outcome variable for the linear model with stars as the predictor. Residuals would therefore reflect unique variance associated with star rating in predicting RI-Ent. When doing this, there is no longer a significant effect of rating ($r^2 \cong 0$, $t(124{,}621) = -0.43$, $p = .67$). It appears that the variability present in RI-Ent does not covary with message valence when review length is controlled.

(2) Average unigram information (AUI). Interestingly, and unexpectedly, AUI shows a quadratic relationship with stars (see also Hu, Pavlou & Zhang, 2006 for similar findings). This is plainly seen in Fig. 3B. To model this, we converted stars into a quadratic term ([1,2,3,4,5] = [4,1,0,1,4]). When not controlling for review length the raw analysis revealed a small but highly significant effect of stars in predicting AUI ($r^2 = .010$, $t(124{,}621) = 36.24$, $p < .001$), such that information density of a review increased as rating levels became more extreme.

When taking out review length, and running the regression with residuals, this effect remained ($r^2 = .011$, $t(124{,}621) = 36.90$, $p < .001$), suggesting it is highly independent of review length. Again, though a weak effect, there is a relationship between the average single-word information of reviews and star ratings. It appears that positive valence is not predictive of overall information; rather, the extremity of the valence predicts slightly more loading of reviews with greater information density (i.e., equivalently, lower frequency terms).

(3) Average conditiona l information (ACI). Interestingly, ACI also showed a nonlinear relationship with stars, shown in Fig. 3C. With star rating predicting ACI alone, there is a significant quadratic relationship ($r^2 = .013$, $t(124{,}621) = 39.89$, $p < .001$). The same pattern appears to hold: Extreme reviews seem to generate more information in bigrams patterns, though this seems to be more pronounced in the negative reviews. When residualizing ACI by review length and AUI (as an additional covariate), this relationship shrinks in effect, but remains statistically reliable ($r^2 = .003$, $t(124{,}621) = 20.01$, $p < .001$).

(4) Conditional information variability (CIV). At first blush, CIV also has a nonlinear relationship with stars (Fig. 3D). Again, the quadratic term for stars significantly predicts CIV scores ($r^2 = .004$, $t(124{,}621) = 21.35$, $p < .001$), though the effect is even smaller. However, we controlled for both review length and ACI, since the height of ACI will generally correlate with CIV's range (due to range restriction with true 0). When we do this, the relationship between CIV and stars completely changes (see Fig. 4). Now, greater informational variability appears to be related to increased positive valence. This relationship, between the residual of CIV and star rating, is statistically reliable but again very small ($r^2 = .009$, $t(124{,}621) = 32.86$, $p < .001$).

Other user-related variables. The Yelp dataset also allows us to explore information as it relates to the “listener” in this context. Users who read reviews have the option to rate them as useful, cool, or funny. Exploration with simple logistic regression finds that even using these simple, surface information-theoretic measures provides a boost in predicting whether a review will be categorized as funny or cool (useful is not predicted by information measures, surprisingly). Some details are shown in Table 1, detailing fully specified models with centered interaction terms for all information-theory values, review length, and comparison models.

Table 1: Basic results of logistic regression when categorizing reviews along certain “listener” dimensions.

	Full model	Length only	Intercept only
cool?	57.9% (168117)	55.8% (170419)	52.7% (172411)
useful?	68.5% (152015)	68.5% (152976)	68.5% (155378)
funny?	63.9% (159643)	62.4% (163005)	61.4% (166277)

Note: Categorization uses a 0.5 threshold in GLM predictions using `family=binomial(“logit”)`. AIC shown in parentheses. All full models have lower AIC, though performance difference is small.

General Discussion

Variance in information density is partially, if only weakly, captured by a review’s star rating. Though we obtain very small effects overall, we would argue that these remain theoretically intriguing. For example, we find a curious and unpredicted quadratic relationship between average lexical information and review rating. This suggests participants may be choosing lower frequent terms–greater lexical richness–when composing reviews at the extremes of the scale (in contrast to our hypothesis that positive reviews, specifically, would be of greater lexical richness).

Overall, the variability in information density is at least partially accounted for by contextual influences beyond the linguistic level. If star rating is an indication of a reviewer’s affective valence, then the cognitive state of a reviewer may stand as one contextual factor that can account for changes in the information density expressed in a message. This supports previous findings that show contextual factors affect the rate of information transfer (Jaeger, 2010; Genzel & Charniak, 2002). In light of current and previous findings, speakers may be sensitive to a variety of linguistic probability distributions well suited to convey messages under any variety of constraints. Perhaps it should be no surprise that the content of a message expressing intense joy is information-theoretically different, slightly, from one of mediocrity.

An underlying principle of Uniform Information Density suggests language users preference a uniform distribution of information across a message. If the variability of information across a message increases, information transfer would be less uniform. When controlling for ACI and the word length of a message the variability of information across messages increased as star rating increased. This suggests the optimal rate of information transfer across messages is dependent on context; in this case messages' affective valence. Positive affective states may result in more open ended and flexible behaviors (Cacioppo & Gardner, 1999; Diener & Diener, 1996; Fredrickson, 2001; Isen & Means, 1983) possibly moderating the optimal rate of information transfer within a specific context. This opens up the possibility that what is considered *optimal* may be subject to a variety of higher-level constraints (see also, Ferrer-i-Cancho, Debowski & Moscoso del Prado Martin, 2013 and Mahowald et al., 2013 for a more recent debate over the use of constant entropy rate measures in describing language use). To be sure, the optimal rate of information transfer over some context may be more or less uniform; the variability of information expanding and contracting depending on one's current affective state or intended message valence.

In summary, speakers may be sensitive to the rate of information in sequencing their message; adjusting their message according to a particular rate of information transfer (Jaeger, 2010; Fine & Jaeger, 2011). Our results are commensurate, in a way, with this intuition. The information density of a message, and the variability of that density are sensitive, at least weakly, to the message's affective valence.

References

Aylett, M. P. (1999). Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. *Proceedings of ICPhS–99, San Francisco*.

Aylett, M., & 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.

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual Review of Psychology*, *50*(1), 191–214.

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological Science*, *7*(3), 181–185.

Estrada, C. A., Isen, A. M., & Young, M. J. (1997). Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians. *Organizational Behavior and Human Decision Processes*, *72*(1), 117–135.

Ferrer-i-Cancho, R., Dębowski, Ł., & Martín, F. M. D. P. (2013). Constant conditional entropy and related hypotheses. arXiv preprint arXiv:1304.7359.

Fine, A. B., & Jaeger, T. F. (2011). Language comprehension is sensitive to changes in the reliability of lexical cues. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 925–930).

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics* (pp. 18–26). Association for Computational Linguistics.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, *56*(3), 218.

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

Hu, N., Pavlou, P. A., & Zhang, J. (2006) Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of Online word-of-mouth communication. *In Proceedings of the 7th ACM Conference on Electronic Commerce*, 324–330.

Isen, A. M., & Means, B. (1983). The influence of positive affect on decision-making strategy. *Social Cognition*, *2*(1), 18–31.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Levy, R. & Jaeger, T. F. (2007). Speakers optimize information density through syntactive reduction. In B. Schlokopf, J. Platt & T. Hoofman (Eds.), *Advances in neural information processing systems (NIPS) 19,* 849-856. Cambridge, MA:MIT Press.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition, 126*(2) 313-318

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Vinson, D. W., Dale, R., Tabatabaeian, M, & Duran, N.D. (in press). Seeing and believing: Social influences on language processing. In Mishra, R. K., Srinivasan, N., & Huettig, F. (Eds.) Attention and Vision in Language Processing. *Springer*.

Zipf, G. K. (1949). Human behavior and the principle of least effort.