Simpler structure for more informative words: a longitudinal study

Uriel Cohen Priva (uriel_cohen_priva@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, RI 02912, USA

Emily Gleason (emily_gleason@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, RI 02912, USA

Abstract

As new concepts and discoveries accumulate over time, the amount of information available to speakers increases as well. One would expect that an utterance today would be more informative than an utterance 100 years ago (basing information on surprisal; Shannon, 1948), given the increase in technology and scientific discoveries. This prediction, however, is at odds with recent theories regarding information in human language use, which suggest that speakers maintain a somewhat constant information rate over time. Using the Google Ngram corpus (Michel et al., 2011), we show for multiple languages that changes in lexical information (a unigram model) are actually negatively correlated with changes in structural information (a trigram model), supporting recent proposals on information theoretic constraints.

keywords: information rate, information theory, Google

Introduction

Most of Campbell's condensed soup cans in Andy Warhol's famous 1962 work show between four and seven words on the front of the can. Currently, a typical Campbell's can has more than ten. The "cream of mushroom with roasted garlic" soup is "great for cooking", and the can additionally specifies its net weight, as shown in Figure 1. Campbell now offers more soup varieties, more than 80, compared to only 32 in 1962.

Figure 1: Andy Warhol's Campbell's tomato soup (left), and modern Campbell's cream of mushroom with roasted garlic (right)

On every front, the development of human society is accompanied by information growth. There are more books to read today, objects to use, and apps to download. In information theory (Shannon, 1948), the information encoded in some event is its negative log probability – unpredictable events are more informative. Imagine if every exchange of words in English were recorded. In terms of information theory, some exchanges would be more informative than others. Predictable utterances provide less information than unpredictable ones. Repetitive "how are you"s do not contribute much information, but utterances such as Neil Armstrong's "small step for man" do. It is reasonable to expect that a random sample of exchanges collected today will contain more information than an equal-sized sample collected a hundred years ago: the modern sample may contain *unfollow*, *Higg's boson* and *politically correct* – lexical items, scientific discoveries and social concepts that were first used or discovered within the past 100 years.[1]

If a million word sample collected today contains more information than it did a hundred years ago, the expectation is that information rate (entropy), will be higher as well. However, the prediction that information rate has been rising is incompatible with recent findings in psycholinguistics. It has been proposed that speakers manipulate their speech so that they will not exceed or fall below acceptable information rates, such as by omitting, reducing, or hypo-articulating low-information linguistic material and expanding or hyper-articulating high-information linguistic material (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). Expansion and reduction have been demonstrated for individual segments (Cohen Priva, 2015; R. van Son & van Santen, 2005; R. J. J. H. van Son & Pols, 2003), syllables (Aylett & Turk, 2004), morphemes (Kuperman, Pluymaekers, Ernestus, & Baayen, 2007; Kurumada & Jaeger, 2015; Pluymaekers, Ernestus, & Baayen, 2005), and words (Arnon & Cohen Priva, 2014; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Jurafsky, Bell, Gregory, & Raymond, 2001; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 2011; Seyfarth, 2014). Such effects have even been demonstrated at the edge of clauses (Jaeger, 2010; Levy & Jaeger, 2007; Norcliffe & Jaeger, 2014), suggesting that information theoretic considerations are also driven by syntactic information. Other studies suggest that higher-level syntactic considerations affect the duration of individual words within that construction (Gahl & Garnsey, 2004; Kuperman & Bresnan, 2012). This trend is even suggested in the Campbell's soup example presented in the first paragraph– there are more possible choices of soup, and additional words are re-

[1] This is not to say that some words do not fall out of fashion– see Petersen, Tenenbaum, Havlin, and Stanley (2012).

quired to describe most choices, spreading out the amount of information per symbol.

What factors contribute to the limitations on information rate? Current research considers at least two. First, speakers may be unable to speak faster or provide more information due to speaker-internal limitations, such as time constraints for motor planning or a cap in cognitive ability (Bell et al., 2009; the within-speaker model in Jaeger, 2010, pp. 50–51). Alternatively, the limit may focus on the communication channel – even if speakers are able to produce high information rates, their listeners may be unable to follow what is being said at such rates (Jaeger, 2010; Pate & Goldwater, 2015). Both explanations predict that information rate will not exceed certain thresholds, even if more information does become available.

How can information rate be held constant given an increase in available information? Several factors affect the amount of information provided by speakers. Consider the phrase *it is raining*. The information provided by an utterance as a whole is the negative log probability of observing the utterance, combining the probability of the content and the probability of the structure. One aspect of this is *world knowledge*. If it is -10 degrees outside, then *it is raining* becomes a highly unlikely utterance, whereas *it is snowing* becomes far more likely. Studying this type of information is beyond the scope of this paper, but it is possible to measure *lexical* and *structural* information. Lexical information is derived from the frequency or probability of individual words. The word *precipitating* is less frequent than the word *raining* (despite denoting a larger set of events). The phrase *it is precipitating* is therefore lexically more informative than *it is raining*. Structural knowledge would tell us that *it is raining* is more common than *raining it is*, and so the untopicalized form is more probable and less informative than the second, topicalized form. Therefore, one of the ways language use can change to accommodate the rising amounts of information is by reducing structural complexity. Increase in available information tends to increase the information provided in any utterance, but using more probable (less informative) structures would balance this increase. Does language compensate for the rising availability of information by reducing structural complexity?

We test this hypothesis using a longitudinal study of language by contrasting the entropy of a three-word language model with the entropy of a single-word language model (unigram model, Jurafsky & Martin, 2000). A three-word (trigram) language model determines the probability of the appearance of a word using the expected negative log probability of observing a word given the two preceding words (1). This method is similar to the one used in Genzel and Charniak (2002). A unigram model determines the probability of a word using no context (2). Both models take yearly entropy as the weighted average surprisal of all words in the corpus for that year. If more information is available, the diversity of the lexicon will be higher, but if the language focuses on a more restricted subset of available information, the entropy of the single-word model would drop. In contrast, the three-word model factors in both available information and the structural complexity of the language. If these two values were independent, they should be positively correlated– if context were not available, then the best estimate for the trigram model (1) would be the unigram model (2), and therefore a rise in unigram entropy would predict a rise in trigram entropy. However, we instead expect that one would come at the expense of the other, and that increasing the amount of available information should lead to a reduction in trigram entropy to keep information rate within acceptable ranges.

(1) Trigram entropy

$$E\left[-\log \Pr(\text{word}|\text{two previous words})\right]$$

(2) Unigram entropy

$$E\left[-\log \Pr(\text{word})\right]$$

Another possible hypothesis is that information rate constraints would have no effect on textual data. After all, readers (and writers) can theoretically slow down and speed up as they will, in order to digest (or produce) denser or more informative words and structures (although, see Genzel & Charniak, 2002, who found evidence for *entropy rate constancy* in text). This expectation has the same prediction as the null hypothesis: increase in unigram information rate would lead to a rise in structural information rate. A negative correlation between unigram and trigram entropy would suggest that writers still tend not to exceed some level of information rate.

Methods and materials

The Google Ngram corpus

Historical spoken data was not systematically collected, but written data is available. The Google Ngram corpus (Lin et al., 2012; Michel et al., 2011) provides yearly frequency counts for sequences of words, and has previously been used to study related phenomena, such as the lifecycle of words (Petersen et al., 2012). The corpus contains several subsets that limit the type of word sequences to words that were published in a specific language, or a specific country. For example, the American English subset includes only word sequence counts of English books that were published in the United States. A typical datum in the Google Ngram corpus for the American English subset might contain a three-word sequence, such as "take aerial photographs", followed by two numbers, e.g. "1992 23". This would mean that the sequence *take aerial photographs* appeared 23 times in all the books scanned by Google that were published in 1992 in English in the United States.

We focus on data from the 20th century, for which data is available for the greatest number of languages. We exclude data from 2000 and onwards, as suggested by the authors of corpus (supplementary material of Michel et al., 2011). We

excluded languages that had too little data in the 20th century (Simplified Chinese, Hebrew),[2] and languages for which no single country is dominant (Spanish).[3] English data was split by the corpus into American English and British English, and English was therefore included. The exact same methodology, as detailed below, was replicated for each of the remaining languages: American English, British English, French, German, Italian, and Russian. For both trigrams and unigrams, we excluded words that mixed letters and numbers, as too many of those seemed like data from tables, rather than language use. [4]

Calculating trigram entropy

Trigram surprisal was estimated as the maximum likelihood estimate (MLE) of observing the third word in a three word sequence given all the possible words that could follow the previous two words. For example, to calculate the probability of the word *photographs* appearing in the context *take aerial*, the frequency of *take aerial photographs* is divided with the frequency of *take aerial* followed by any word. The negative log of the probability provides the number of bits the word *photographs* provided in that context. The average number of bits per word is the entropy of the corpus given the model.

MLE estimates were used rather than models incorporating smoothing or backoff (Jurafsky & Martin, 2000, ch. 4), as such methods explicitly integrate information from lower-order n-grams to the probability calculations of trigrams. Thus, they already factor out cases in which a word's frequency is biased by the context in which it appears (e.g. *Francisco* is frequent, but almost always preceded by *San*). The proposed account predicts that new words are likely to be structurally accommodated by facilitating (restrictive) contexts. Switching to smoothed models could mask this effect.

Calculating unigram entropy

For unigram entropy, the first words of each trigram in the trigram model were counted. The first word was chosen since trigrams in the 2012 version of the Google Ngram data do not span sentence boundaries (this is the version used here; Lin et al., 2012), and we did not want to bias the sample towards sentence-final words, which are likely to be less informative if our hypothesis is correct. The surprisal of observing a particular word in any context was taken to be the negative log of the number of times the word was observed, divided by the number of times each word was observed (MLE of word probability).

[2]Hebrew and Chinese had comparable results to the other languages when using only data for the last 30 years of the 20th century.

[3]We did run the study for Spanish, with comparable results to the other languages.

[4]Median total number of trigrams per year after exclusions (in millions): American English: 623.11; British English: 221.1; French: 202.8; German: 151.84; Italian: 54.97; Russian: 82.83. Median number of unique trigrams per year (in millions): American English: 80.22; British English: 35.75; French: 30.11; German: 26.97; Italian: 14.2; Russian: 19.29.

Statistical method

For each language we measured the relationship between trigram entropy and unigram entropy in a linear regression, with trigram entropy as the predicted variable and unigram entropy as the main predictor. The log number of unique trigrams per year, the log of the total number of trigrams in the corpus, and the log number of unique unigrams were used as controls, as well as the log number of volumes and log number of pages that were included in the original Google Ngram corpus. The total number of unigrams was identical to the total number of trigrams, and was not used (as unigrams were taken from the trigram dataset, see previous section). The greater the unique number of trigrams relative to the total number of trigrams, the higher the entropy is expected to be, everything else being equal (as the entropy of a uniform distribution over $n+1$ outcomes is higher than a uniform distribution over n outcomes). There is no concrete prediction for the total number of trigrams as a predictor, but it is expected to capture some of the variance that is associated with having more books (or topics of discussion). The number of unique unigrams expresses an alternative (though less accurate) estimate for the richness of the lexicon than unigram entropy. All the predictors and predicted values are *time series*, and are not considered to be independent from their previous values (e.g. the correlation between trigram entropy in year n and year n-1 is 0.99 for American English). Therefore, the regression used the differences between each pair of consecutive years for all variables. All counts-based controls were logged, as the logged values correlated better with trigram entropy (e.g. pearson r 0.88 for the correlation between log unique number of trigrams and trigram entropy, but only 0.81 for its unlogged counterpart in American English), and therefore consistute more appropriate controls.

Results and discussion

For all languages, changes to unigram entropy were strongly negatively correlated with changes to trigram entropy (American English: β=-0.44, SE=0.05, t=-8.753, $p<10^{-12}$; British English: β=-0.51, SE=0.034, t=-15.05, $p<10^{-15}$; French: β=-0.55, SE=0.055, t=-9.925, $p<10^{-15}$; German: β=-0.46, SE=0.07, t=-6.532, $p<10^{-8}$; Italian: β=-0.76, SE=0.058, t=-13.148, $p<10^{-15}$; Russian: β=-0.55, SE=0.042, t=-12.945, $p<10^{-15}$), suggesting that structural (or transitional) complexity is reduced when lexical complexity rises, even when the size of the corpus is controlled for. Figure 2 plots for American English the partial correlation between changes in unigram entropy and the residual changes in trigram entropy after other predictors were controlled for. Figure 3 shows the equivalent relationship for German, the language with the least significant relationship between unigram entropy and trigram entropy.

For all languages used, the number of unique trigrams was positively correlated with trigram entropy (American English: β=0.6, SE=0.049, t=12.26, $p<10^{-15}$; British English: β=0.71, SE=0.035, t=20.624, $p<10^{-15}$; French: β=0.76,

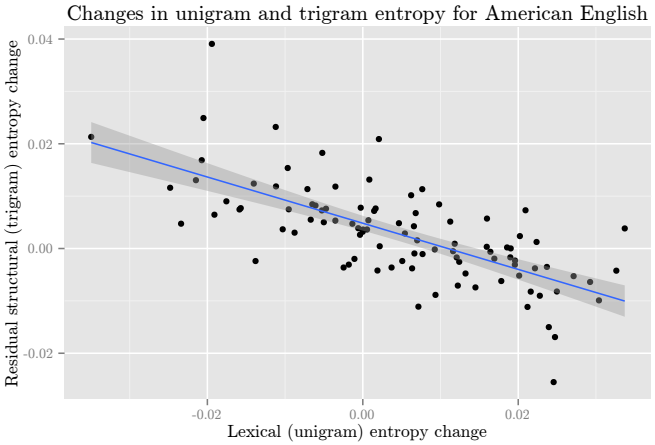

Figure 2: A plot showing the relationship in American English between changes in unigram entropy on the x-axis, and residual changes in trigram entropy on the y-axis, after other predictors are controlled for.

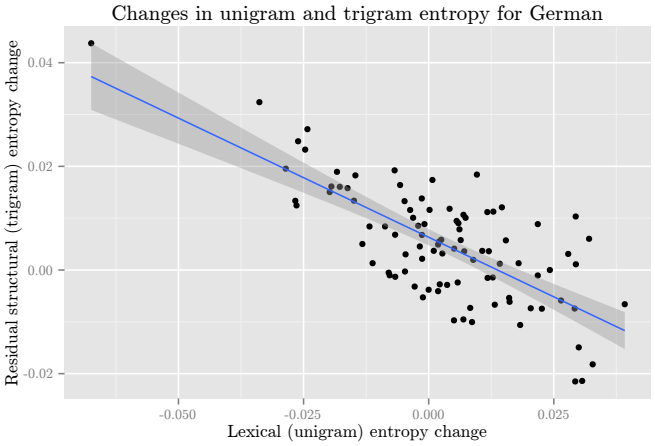

Figure 3: A plot showing the relationship in German between changes in unigram entropy on the x-axis, and residual changes in trigram entropy on the y-axis, after other predictors are controlled for.

SE=0.065, t=11.777, $p<10^{-15}$; German: β=0.86, SE=0.11, t=7.852, $p<10^{-11}$; Italian: β=0.84, SE=0.078, t=10.794, $p<10^{-15}$; Russian: β=1.161, SE=0.093, t=12.537, $p<10^{-15}$), as expected. The total number of trigrams was significantly negatively correlated with trigram entropy in every language except British English (American English: β=-0.24, SE=0.058, t=-4.142, $p<10^{-4}$; French: β=-0.23, SE=0.051, t=-4.531, $p<10^{-4}$; German: β=-0.22, SE=0.054, t=-3.983, $p<0.001$; Italian: β=-0.47, SE=0.05, t=-9.286, $p<10^{-14}$; Russian: β=-0.36, SE=0.057, t=-6.326, $p<10^{-8}$). In all languages except Italian and Russian, the number of unique unigrams had no effect on trigram entropy (positive correlation for Italian: β=0.28, SE=0.12, t=2.376, $p<0.05$; negative correlation for Russian: β=-0.31, SE=0.11, t=-2.687, $p<0.01$). The number of books was positively correlated with trigram entropy for Russian only (β=0.066, SE=0.03, t=2.19, $p<0.05$), and the number of pages was positively correlated with trigram entropy for only Italian (β=0.23, SE=0.065, t=3.454, $p<0.001$). No other languages were affected by book or page count.

In all languages, unigram and trigram entropy change over time. There are clear drops in trigram entropy, e.g. all Western world countries have a drop in trigram entropies in the 1970s, despite an increase in the size of the corpus. This in itself, prior to controlled analysis, is an interesting finding. It would suggest that the acceptable range for information rate may sometimes drop. Figure 4 plots the change in residual unigram entropy over time, controlling for the total number of unigrams and the number of unique unigrams in the corpus. Figure 5 plots the change in residual trigram entropy over time, controlling for the total number of trigrams and the number of unique trigrams in the corpus.

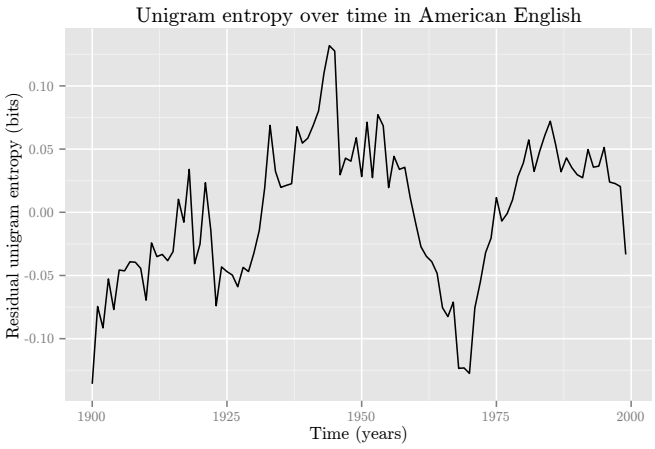

Figure 4: Residual values of unigram entropy in American English during the 20th century. The x-axis is years. The y-axis is the residual unigram (lexical) entropy after controlling for parameters signifying the size of the corpus: log number of unique unigrams, log number of unigrams.

Because of the individual language differences, we additionally combined all languages in a mixed-effects regres-

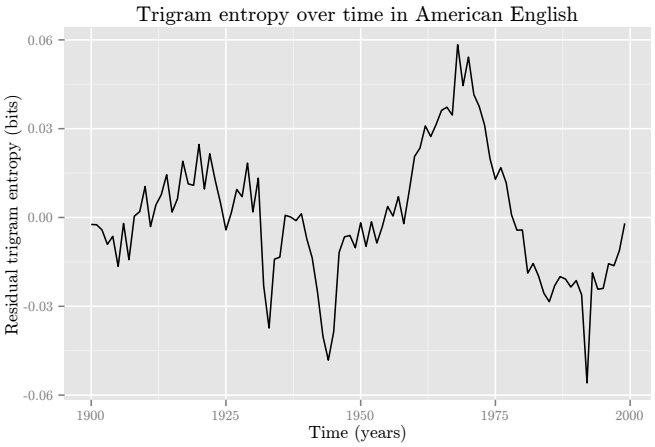

Figure 5: Residual values of trigram entropy in American English during the 20th century. The x-axis is years. The y-axis is the residual trigram (structural) entropy after controlling for parameters signifying the size of the corpus: log number of unique trigrams, log number of trigrams.

sion, post-hoc, using random effects for language, with trigram entropy as a random slope. The controls for this regression were identical for those used in the individual linear regressions, and the results were nearly identical. Critically, changes to unigram entropy were still strongly negatively correlated with changes to trigram entropy (β=-0.55, SE=0.0161, t=-34.359, $p<10^{-15}$). The number of unique trigrams and number of volumes were positively correlated with trigram entropy (β=0.96, SE=0.021, t=45.410, $p<10^{-15}$ and β=0.048, SE=0.0107, t=4.496, $p<10^{-5}$, respectively). The total number of trigrams and the number of unique unigrams were negatively correlated with trigram entropy (β=-0.29, SE=0.0191, t=-14.916, $p<10^{-15}$ and β=-0.088, SE=0.0238, t=-3.705, $p<0.001$, respectively). Number of pages was not significant.

Summary

Changes to unigram entropy and trigram entropy were negatively correlated, the opposite of what the null hypothesis expects: When amounts of lexical information rise, structural information drops. This constitutes strong evidence for accounts that expect language to restrict the amount of information provided at a given time (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). It is quite surprising that the transitional or structural properties of language should change in response to the increasing amount of information, as predicted by information theoretic accounts, and yet this is the case for all the languages studied here.

These findings open the door to studies of other trade-offs in long term information rate. We use transitional probabilities here as an estimate of structural complexity (more complex transitions would indicate more complex structure), but it would also be interesting to use parsed corpora to look at the frequency of different grammatical constructs. Does a rise in information also predict a decrease in complex grammatical structures, such as complex clauses? Our hypothesis would suggest so. The Google Ngram corpus does contain basic syntactic information, in the form of rough part-of-speech tags (e.g. noun vs. verb, but not preterite vs. participle; Lin et al., 2012), and those too can perhaps be used to infer complexity in future studies. In a separate study, Cohen Priva (under revision) shows that speech rate (another form of information rate) is negatively correlated with both lexical and syntactic information rates.

Both unigram and trigram information rates rose during certain years and fell in others. There were large scale dips in both unigram and trigram entropy in all languages at different times, suggesting the existence of additional factors that play a part in determining the acceptable information rate for a language at any particular time. The timing of the drops in information rate is potentially quite telling– in most of the languages analyzed here, there are large dips in trigram entropy around 1915-1920 and 1935-1945, perhaps corresponding to the two world wars. If information rate corresponds to large-scale societal attitudes, then this would give us a new tool to study societal change. Fluctuations in information rate may be used to track society-level mood, or even predict future events. For instance, does information rate rise before democratic revolutions or following them? Being allowed to express one's opinion will mean that more opinions will be expressed, but perhaps the expression of new opinions is what brings about democratic revolutions in the first place.

Acknowledgments

This work greatly benefited from discussions with Fiery Cushman, Dan Jurafsky, Elinor Amit, Scott AnderBois, Laura Kertz, and David Badre.

References

Arnon, I., & Cohen Priva, U. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon*, *9*(3), 377–400. doi: 10.1075/ml.9.3.01arn

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.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Cohen Priva, U. (under revision). *Not so fast: Fast speech correlates with lower lexical and structural information.* Brown University manuscript.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language*, *80*(4), 748–775.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the Association for Computational Linguistics* (pp. 199–206).

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

Jurafsky, D., Bell, A., Gregory, M. L., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: Benjamins.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: an introduction to Natural Language Processing, computational linguistics, and speech recognition*. New York: Prentice Hall.

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611. doi: 10.1016/j.jml.2012.04.003

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261–2271. doi: 10.1121/1.2537393

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178. doi: 10.1016/j.jml.2015.03.003

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Scholkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the acl 2012 system demonstrations* (pp. 169–174).

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. doi: 10.1016/j.cognition.2012.09.010

Michel, J.-B., Shen, Y. K., Presser Aiden, A., Veres, A., Gray, M. K., Pickett, J. P., ... Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Norcliffe, E., & Jaeger, T. F. (2014). Predicting head-marking variability in yucatec maya relative clause production. *Language and Cognition*, *FirstView*, 1–39. doi: 10.1017/langcog.2014.39

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17. doi: http://dx.doi.org/10.1016/j.jml.2014.10.003

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*.

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

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*, 146–159.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. doi: 10.1016/j.cognition.2014.06.013

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

van Son, R., & van Santen, J. (2005). Duration and spectral balance of intervocalic consonants: a case for efficient communication. *Speech Communication*, *47*, 100–123.

van Son, R. J. J. H., & Pols, L. C. W. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, *25*, 171–184.