The interdependence of frequency, predictability, and informativity in the segmental domain

Cohen Priva, Uriel
Brown University
Cognitive, Linguistic, and Psychological Sciences

Jaeger, T. Florian
University of Rochester
Brain and Cognitive Sciences

Abstract

It has long been noted that language production seems to reflect a correlation between message redundancy and signal reduction. More frequent words and contextually predictable instances of words, for example, tend to be produced with shorter and less clear signals. The same tendency is observed in the language code (e.g., the phonological lexicon), where more frequent words and words that are typically contextually predictable tend to have fewer segments or syllables. Average predictability in context (informativity) also seems to be an important factor in understanding phonological alternations. What has received little attention so far is the relation between various information-theoretic indices – such as frequency, contextual predictability, and informativity. Although each of these indices has been associated with different theories about the source of the redundancy-reduction link, different indices tend to be highly correlated in natural language, making it difficult to tease apart their effects. We present a computational approach to this problem. We assess the correlations between frequency, predictability, and informativity, and assess when these correlations are likely to create spurious (null or non-null) effects depending on, for example, the amount of data available to the researcher. Keywords: frequency, predictability, informativity, simulations

1 Introduction

Implicit knowledge about the statistics of linguistics input affects how we produce language. For example, both the frequency and the contextual predictability of linguistic units, such as phonological segments or words, have been linked to their realization (for review, see Jaeger & Buz 2017). While the existence of probabilistic effects is now well-documented, debates about the mechanisms that cause the effects are ongoing (for review, see Jaeger & Buz 2017; Ernestus 2014; Shaw & Kawahara 2018). Three probabilistic indices have received particular attention in these debates: frequency, contextual predictability (henceforth *predictability*), and average contextual predictability across context (henceforth *informativity*).[1] In natural language data, these three indices tend to be highly correlated, making it difficult to distinguish between their effects. Here we present a novel type of computational analysis that can shed light on the relative contributions of the three indices.

Effects of frequency on pronunciation have been considered at least since Sibawayhi, an Arabic grammarian of the 8th century (Al-Nassir 1993; Carter 2004). Frequent words tend to have fewer syllables and phonemes, which led to the formulation of the famous "law of abbreviation" (Zipf 1949). Even when controlling for segment count, frequent words tend to be produced with shorter duration (Gahl 2008). Similarly, more contextually predictable instances of a word tend to be realized with more reduced signals (e.g., Aylett & Turk 2004; Bell et al. 2003; Gahl et al. 2012; Pluymaekers et al. 2005). Both the effect of frequency and the effect of predictability have been attributed to practice effects, contributing to the ease with which segments or words are produced (e.g., Arnold et al. 2012; Bell et al. 2009; Bybee 2002). Alternative proposals attribute the same effects to communicative efficiency, trading off production ease against the ease, expected accuracy, or rate of message transmission (Aylett & Turk 2004; Jaeger 2013; Lindblom 1990; van Son & Pols 2003; Zipf 1935). Yet other accounts attribute the effects to filtering mechanisms during language comprehension, where more frequent or more predictable elements are understood more easily and accurately even when they are phonetically reduced, making it more likely for reduced exemplars to be stored and reproduced (Pierrehumbert 2001; Wedel 2006).

More recently, a number of researchers have argued that the *average* value of an element's predictability across all the contexts it occurs in during language use – sometimes referred to as informativity – affects its realization (Cohen Priva 2008; Seyfarth 2014). Informativity has been argued to be of particular relevance to phonology, potentially because informativity might reflect the cumulative conventionalized effect of predictability across many contexts (Hall et al. 2016; Hall et al. 2018; Seyfarth 2014). Indeed, there is evidence that processes such as lenition are described better by informativity, than by frequency or predictability (Cohen Priva 2008, 2015). Informativity also seems to capture variance in word-specific phonetic representations, beyond effects of frequency and predictability (Seyfarth 2014). In fact, frequency no longer seems to be a significant predictor of word duration after informativity is controlled for (however, Seyfarth 2014, shows that predictability remains a significant predictor of word duration). Similarly, a recent cross-linguistic analysis across eleven languages suggests that Zipf's law of abbreviation is, in fact, better captured by informativity (Piantadosi et al. 2011).

Findings like these raise questions about accounts that predict strong effects of frequency on phonology (such as the production ease or automatization accounts described above). More generally, these findings raise questions about the relation between frequency, predictability, and informativity. However, beyond the handful of studies described above, a systematic exploration of these three probabilistic indices has so far been lacking.

As mentioned above, the three indices are expected to be correlated, and there is a valid concern whether an effect that is found for one factor should really be attributed to either of the other two. Making matters worse,

[1] We note that the term informativity is also used with a different meaning than the one intended here. In the literature on learning and prediction, cue informativeness or cue informativity refers to the amount of information a cue carries about a specific outcome (e.g., Bates & MacWhinney 1987, and references thereto). These two notions are related, but not identical.

informativity is the expected value of (negative log transformed[2]) predictability, and frequency is the expected value of both informativity and (negative log transformed) predictability. In the absence of sufficient data, estimates for frequency are thus expected to be more reliable than estimates for informativity, and estimates for informativity are expected to be more reliable than estimates for predictability. This hierarchical relation between the three indices exacerbates the problem: effects that depend on a difficult-to-estimate factor may seem to depend on easier-to-estimate expected factors instead.

The goal of this paper is to estimate the degree to which the three probabilistic indices might have been confounded in previous works. To this end, we conduct a series of computational simulations in order to estimate how often one would by chance observe an effect of, for example, frequency even if the phenomenon under investigation is in reality only sensitive to, for example, predictability. Similarly, we test how likely it is that a true effect of, for example, predictability is not detected. Study 1 addresses this question for work on segment frequency, predictability, and informativity. Study 2 specifically asks how likely studies which do not control for effects of informativity are to wrongly conclude that there are frequency or predictability effects. Our simulations are based on linguistic distributions that are observed in actual language use. Beyond the specific goals we pursue here, the proposed simulation approach also allows researchers to assess Type I error (spurious results) and Type II error (missed results) rates for their studies on any of the three or other probabilistic indices.

1.1 Computational approach

We first describe the databases we employed to obtain natural word and segment distributions. Then we define the three probabilistic indices. Finally, we describe the simulation approach employed by all our studies.

1.1.1 Corpora

The studies described below require information about the frequency of word unigrams. To this end we combined three corpora of unscripted speech, namely the Switchboard (Godfrey & Holliman 1997), Buckeye (Pitt et al. 2007), and Fisher (Cieri et al. 2004, 2005) corpora. The Switchboard corpus contains ~2500 conversations between several hundred speakers from all over the United States, who did not know one another, and who were asked to discuss one topic from a list of a few dozen topics. The Buckeye corpus was collected at the Ohio State University and contains 40 relatively long interviews with 40 residents of Ohio. The Fisher corpus contains ~11,700 telephone conversations between speakers in the United States. This results in around 24 million word tokens and around 70,000 word types (including names, numbers, words followed by apostrophes, etc.). Figure 1 shows the (smoothed) frequency-rank plots for word frequencies from the combined corpora. Both axes are shown on log scale.

The studies reported below additionally require information about the sequences of phonological segments contained in each word. We used the CMU dictionary (Weide 2008), which provides pronunciations for around 120,000 unique words of English. When multiple pronunciations were available, only the first was used.

1.1.2 Quantifying probabilistic indices

For the three indices we used maximum likelihood estimates for frequency (1), predictability (2), and informativity (3). To put all measures on the same (bit) scale, we used the logarithm-transform (to base 2) of the reciprocal

[2] For both theoretical and empirical reasons, research on frequency and predictability usually has employed logarithm-transformed values of these indices.

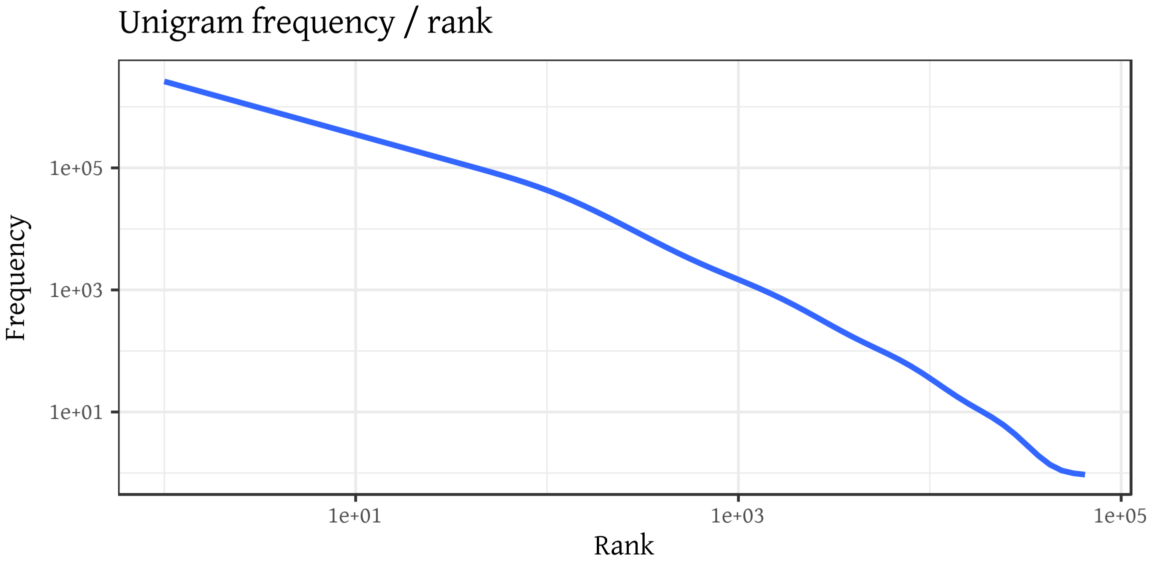

Figure 1: The plot shows the frequency-rank relationship for word frequencies in the corpus. Both axes are shown on log scale.

of each measure. The use of logarithm-transformed measures is the standard in research on language production (e.g., Frank & Jaeger 2008; Jaeger 2006, 2010). In particular, the research on phonetics and phonology we are interested in here has almost exclusively employed logarithm-transformed measures (e.g., Aylett & Turk 2004; Bell et al. 2003, 2009; Cohen Priva 2008, 2012, 2015; Gahl 2008; Gahl et al. 2012; Kuperman & Bresnan 2012; Pluymaekers et al. 2005; Seyfarth 2014; van Son & Pols 2003; van Son & van Santen 2005). The transform we use is also theoretically motivated as it relates all three measures to Shannon information (Shannon 1948; for discussion, see Cohen Priva 2015; Pate & Goldwater 2015).

(1) Frequency(element):

$$-\log_2 \Pr(\text{element})$$

A maximum likelihood estimate for unigram probability in this context means that the observed frequency of the element was divided by the frequency of all elements:

$$-\log_2 \frac{\text{element count}}{\text{all elements count}}$$

For instance, in a study in which words are the linguistic element being investigated, if *give* appears once, and there are 1024 words in the corpus, the frequency of *give* would be 10: $-\log_2(1/1024)$. This is the number of bits the element provides if no context is assumed.

(2) Predictability(element, context):

$$-\log_2 \Pr(\text{element}|\text{context})$$

A maximum likelihood for conditional probability in this context means that the observed frequency of the element when it appeared in the context was divided by the frequency of the context:

$$-\log_2 \frac{\text{element with context count}}{\text{context count}}$$

For instance, in a study in which words are the linguistic element being investigated, and the context is the previous word, if *give* appears once after *will*, and all other words appear 31 times after *will*, the predictability of *give* in the context of *will* would be 5: $-\log_2(1/32)$. This is the number of bits the element provides in the context in which it appears.

(3) Informativity(element):

$$-\sum_{\text{contexts}} \Pr\left(\text{context}|\text{element}\right) \log_2 \Pr(\text{element}|\text{context})$$

The (negative log) predictability of the element in every context in which it appears is averaged, weighted by the frequency of the cases the element appeared in each context. A maximum likelihood estimation of informativity assumes that the predictability of each element in each context is as provided in (2), and that the conditional predictability of contexts given elements is:

$$\frac{\text{element with context count}}{\text{element count}}$$

For instance, in a study in which words are the linguistic element being investigated, and the context is the previous word, if *give* appeared once after *will*, and its predictability were 5 bits of information in this context, and three times after *would*, and its predictability were 1 bit in this context, its informativity would be 2 bits: $\frac{1 \cdot 5 + 3 \cdot 1}{1+3}$. This is the expected Shannon information the element provides in the contexts in which it appears.

1.1.3 Data samples and models

All the studies below rely on generating samples from "true" distributions. What we are interested in is how a researcher's estimate of these true distributions might affect the researcher's conclusion about the effect of any particular index. To this end we calculate the correlations between the true values of an index (e.g., the true segment frequency) and the researcher's estimates of the same index (segment frequency) and the other indices (e.g., predictability and informativity) based on the sample the researcher has access to. The process is illustrated in Figure 2, and described next.

For the present purpose, we take the lexical and phonological corpora described above in Section 1.1.1 to provide the "true" distributions. These corpora contain n elements (e.g., for words, n=24 million tokens). Subsequently, varying k word tokens are sampled from the true multinomial distribution and used to calculate the sample distributions. These samples represent the *researcher's lexical database*: the lexical information the researcher collected from all available data on the language. To approximate the situation a researcher tends to be in, we mostly use $k \ll n$ – i.e., we consider lexical databases that are much smaller than the language experience of an average adult speakers. Finally, m=5,000 word tokens were sampled from each researcher's lexical database.[3] These represent the *researcher's behavioral database*: the more restricted dataset that contains, e.g., higher-quality annotated data that the researcher aims to test their hypotheses against. For example, a researcher might use the entire 2 million word corpus of Switchboard conversations to estimate unigram and bigram frequencies, but then use the 40,000 word manually phonetically annotated subsets of Switchboard to test how bigram predictability affects word durations (cf. Bell et al. 2003, 2009).

[3] We note that our simulations thus assume that the behavioral observations are drawn from the same underlying distribution as the lexical database. This is arguably the case when researchers employ the same corpus to obtain the behavioral measures and estimates of the probabilistic indices. However, when researchers choose lexical databases that differ in register, genre, topic, or other factors known to affect ngram distributions, this assumption is violated. The present simulations thus arguably provide estimates of the *lower* bound of the problems caused by collinearity between the probabilistic indices.

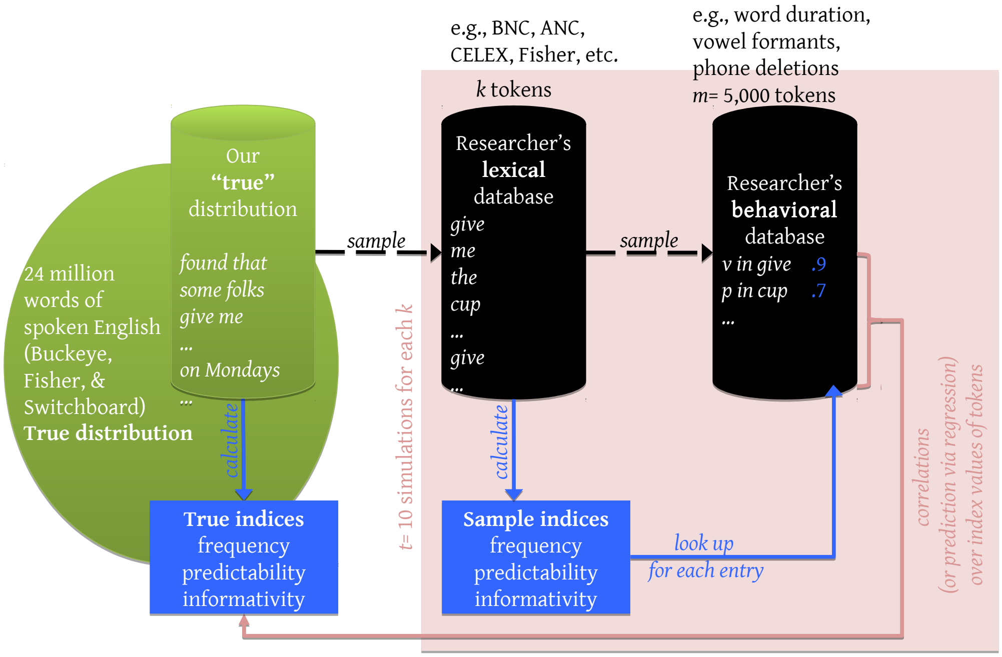

Figure 2: An illustration of the sampling process

For each sample drawn for each value of k, we then investigate the relation between the three indices across all m *tokens* (see Figure 2). Specifically, we investigate the simple and partial correlations between the "true" value for each index and the *estimated* values of all indices based on the sample. Further details about the method are reported in each study.

It is important to keep in mind that our studies aim to estimate (indirectly for now) how the nature of actual word and segment distributions affects how likely a researcher's estimate of an index is to (wrongly) emerge as a significant predictor of human behavior (e.g., word durations) when in reality another index is the predictor of that behavior. It is thus not critical whether our "true" distributions reflect every aspect of actual word distributions as experienced by an average language user. Rather, there is some subset of properties of the word and segment distributions that affect the strength of correlations between true indices and the researcher's estimates of indices (e.g., one plausible candidate for such a property is the distribution of tokens over types for word and segment distributions, cf. Figure 1). For our purpose, it is sufficient to assume that our database for the "true" distributions approximates actual distribution (as experienced by an average language user) in terms of those properties. This assumption strikes us as highly plausible.

2 Study 1: Segment frequency, predictability, and informativity

The goal of this set of studies is to estimate whether at the segmental level, using all the previous segments within the same word as context, some of the indices can appear to explain the effect that should be attributed to another index. This context has been used in information-theoretic linguistics (e.g., van Son & Pols 2003; van Son & van Santen 2005; Cohen Priva 2008, 2017a), though see Daland & Zuraw (2018).

We report our results separately for the (true) frequency (Study 1a), predictability (Study 1b), and informativity (Study 1c) indices. For each index, we first report the correlation between that (true) index and the *sample* estimates for each of the three indices. We then report the result of a linear regression that predicts the true index from all three sample indices combined. This allows us to assess the *partial* correlations between the sample's indices and the true index. This resembles the situation researchers are faced with when they include estimates of all three indices in a model predicting a behavioral outcome (such as phoneme deletion or phonetic reduction). However, rather than assessing the *consequences* of collinearity between the three indices on their respective effects on a behavioral outcome, our simulations aim to assess the degree of collinearity between the three indices that is to be expected given the sample size and the natural (power law) distributions of linguistic elements. This choice is made partly as a simplifying assumption (it means that we do not need to simulate behavioral data), and partly because it allows us to better understand the covariance between the three indices.

2.1 Methods and materials

Ten samples each were generated using exposure to each of k=5,000, 10,000, 20,000, 40,000, 60,000, ..., 200,000 words. Subsequently, m=5,000 words were sampled from the lexicon, as described in 1.1.3, and each of their segments was used as a data point. The *t*-values in each figure were scaled using $f(x) := \mathrm{sign}(x)\log(|x|+1)$.[4] Sample sizes were relatively even-spaced and were not scaled for presentation. The studies report t-values, as these are the values used in linear regressions.

The predictability model used in Study 1 was of segments in words, with all the preceding segments within the same word serving as context. That is, for the /t/ in *bat* the context would be /$[_w$bæ/, while for the /t/ in *talk*, the context would be /$[_w$/ (word initial), following van Son & Pols (2003) and Cohen Priva (2008). Since these are word-based models, the sampling procedure sampled *words* based on their unigram frequency, and probabilistic indices were computed for all segments within each word. In the example above, if *bat* was sampled, then the probabilistic indices were calculated for each of its segments (word initial /b/, /æ/ given /$[_w$b/, and /t/ given /$[_w$bæ/, following the approach described in Study 1b).

2.2 Results

2.2.1 Study 1a: Predicting segment frequency

For segments' true frequency (negative log unigram probability), the raw Pearson correlation between the predicted value and the sample indices was highest for sample frequency ($\hat{\mu}$=0.999, range=0.996 to 1), and much lower for sample informativity ($\hat{\mu}$=0.499, range=0.425 to 0.535) and predictability ($\hat{\mu}$=0.243, range=0.204 to 0.277).

In the regression models, the *t* value of frequency ($\hat{\mu}$=4942.0, range=1207.6 to 9464.8) was higher than that of predictability ($\hat{\mu}$=-0.094, range=-3.6 to 4.8). Informativity was lower than both ($\hat{\mu}$=-0.46, range=-30.6 to 35.7), and for most samples its *t* value was *negative*, signifying that it suppressed the effect of other indices (Lewis & Escobar 1986).

The relationship between the sample size and each index's correlation with true frequency as well as the *t*-values in the regression models are visualized in Figure 3. Sample size did not have a significant effect on the role of predictability and informativity in predicting frequency, likely because the estimates for frequency are highly accurate using fairly little data. The *t*-values for predictability and informativity, even when significantly positive or negative, were three orders of magnitude smaller than the *t*-values of frequency, making it seem highly unlikely that spurious effects of informativity or predictability would be found in the segmental domain, even with relatively small lexical databases.

[4] For high |x| values, log(|x|+1) is nearly indistinguishable from log(|x|). However, the function is better-behaved for plotting purposes. sign(x) returns 0 for x=0, x/|x| otherwise.

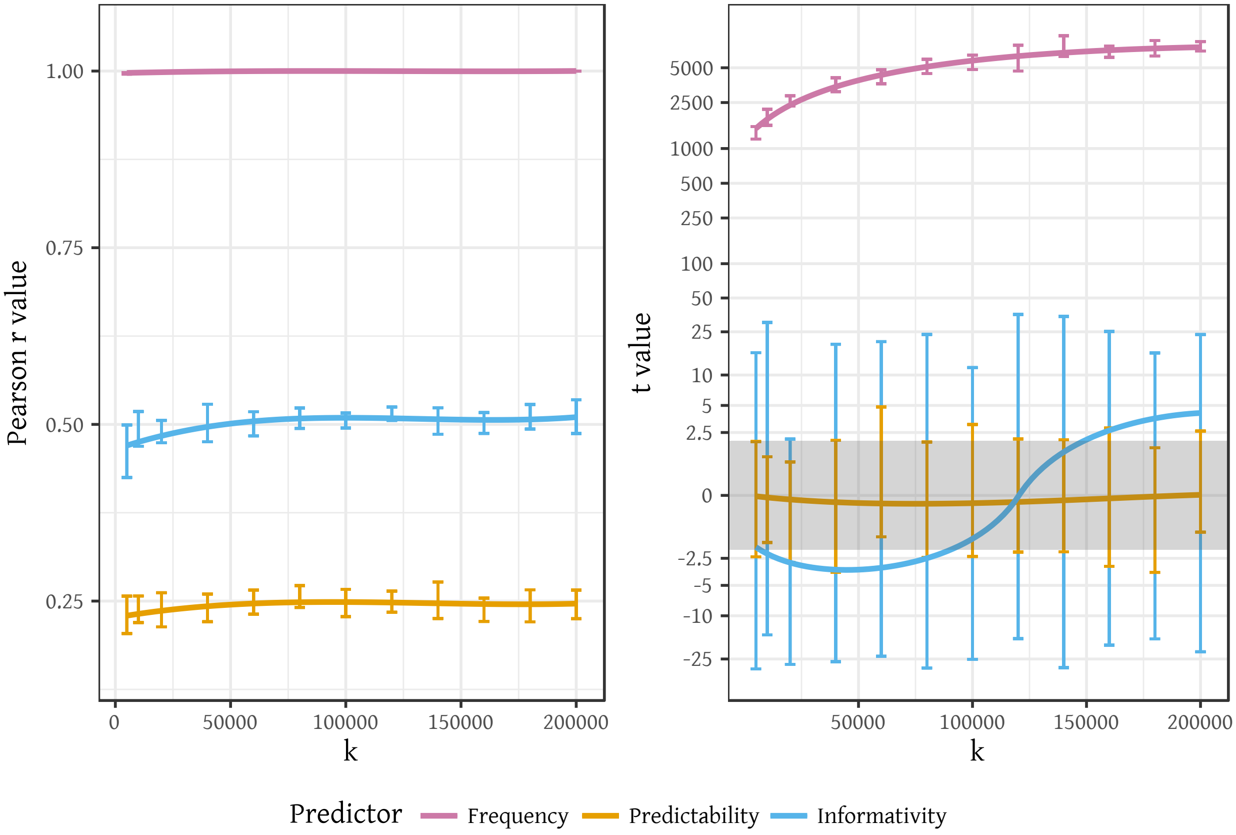

Figure 3: Left: Correlations between true segment frequency and researcher's estimates of frequency. Correlations are shown depending on the size of the researcher's lexical database. Right: *t* values of sample frequency, predictability, and informativity when all predictors are entered into a multiple regression predicting true segment frequency. The grey area signals absence of significance (|*t*|<1.96). Significant negative *t*-values are likely due to suppression. The y axis is in log scale, and the line visualizing the trend was fitted using a cubic function. In both figures the error bars mark the full range of values (10 each).

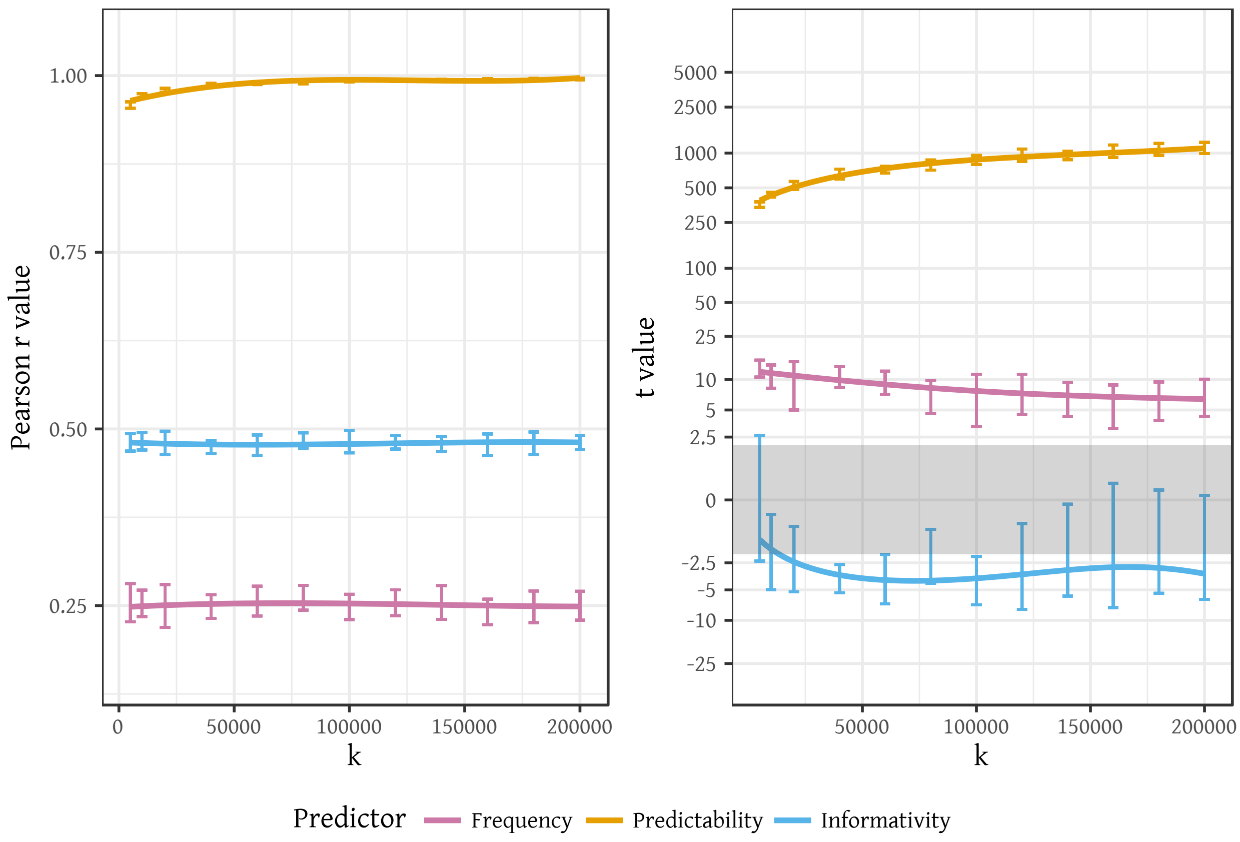

Figure 4: Left: Correlations between true segment predictability and researcher's estimates of predictability. Right: *t*-values of the predictors in a multiple regression predicting true segment predictability. Axes, scaling, and shading follow Figure 3.

2.2.2 Study 1b: Predicting segment predictability given all previous segments

For segments' true predictability, the raw Pearson correlation between the predicted value and the sample values was highest for sample predictability (µ̂=0.986, range=0.954 to 0.996), lower for sample informativity (µ̂=0.48, range=0.462 to 0.497), and lower still for sample frequency (µ̂=0.251, range=0.219 to 0.281).

In the regression models, the *t*-value of predictability (µ̂=786.5, range=338.2 to 1236.9) was higher than that of frequency (µ̂=8.6, range=3.1 to 15.2). Informativity was lower than both (µ̂=-3.0, range=-7.9 to 2.6), and for most samples its *t*-value was negative signifying that it suppressed the effect of other indices.

The relationship between the sample size and each index's correlation with true predictability as well as the *t*-values in the regression models are visualized in Figure 4. Sample predictability was consistently the strongest predictor of the true predictability effect, followed by frequency, though the *t*-values of frequency were still at least 20 times smaller than those of predictability, suggesting that spurious frequency effects are only likely in the presence of strong predictability effects. The effects of informativity were mostly negative, suggesting that informativity suppresses some other index, and is highly unlikely to falsely appear to be significant (in the expected direction). Sample sizes had a modest effect on the relationship among the three indices.

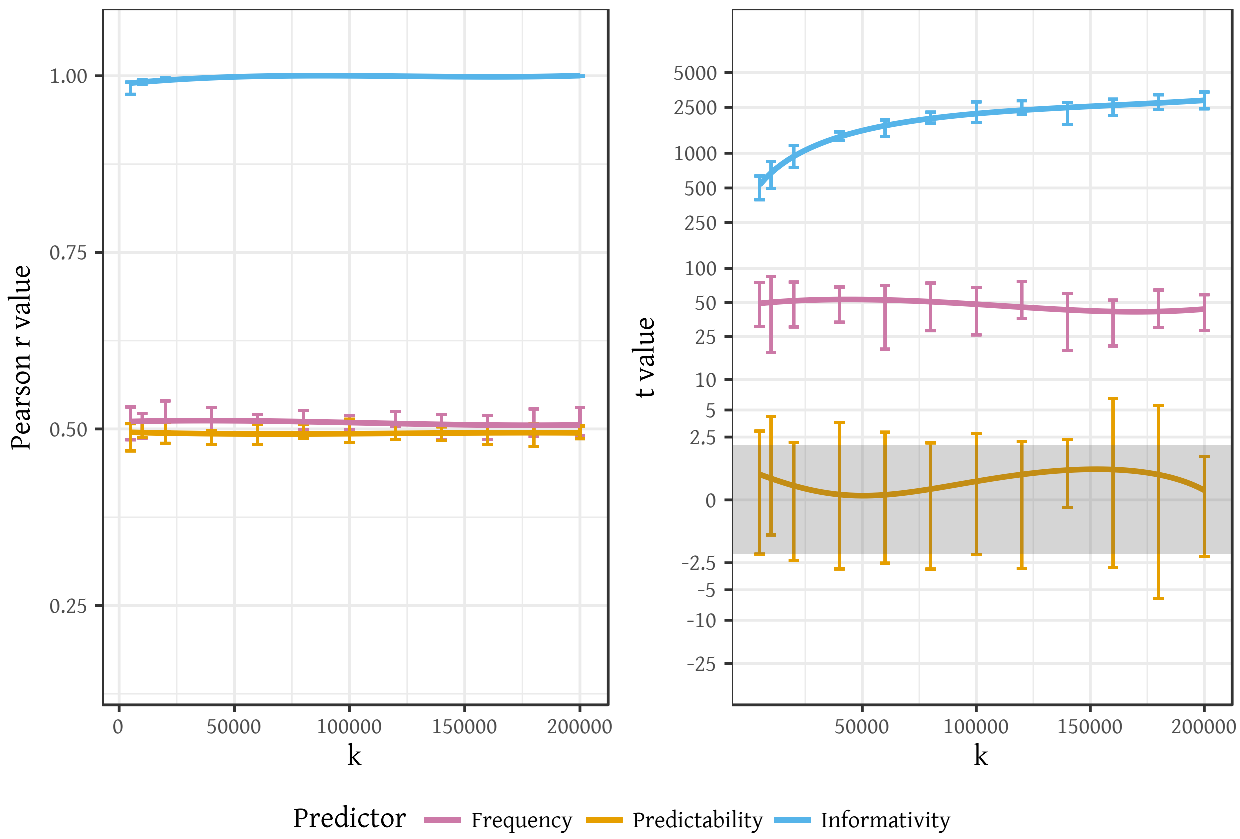

Figure 5: Left: Correlations between true segment informativity and researcher's estimates of informativity. Right: *t*-values of the predictors in a multiple regression predicting true segment informativity. Axes, scaling, and shading follow Figure 3.

2.2.3 Study 1c: Predicting segment informativity given all previous segments

For segments' true informativity, the raw Pearson correlation between the predicted value and the sample indices was highest for sample informativity (µ̂=0.997, range=0.974 to 1), and much lower for sample frequency (µ̂=0.509, range=0.485 to 0.54) and predictability (µ̂=0.494, range=0.469 to 0.514).

In the regression models, the *t*-value of informativity (µ̂=1876.5, range=394.3 to 3389.8) was higher than that of frequency (µ̂=47.7, range=17.9 to 84.3), which was higher than that of predictability (µ̂=0.47, range=-6.2 to 6.6).

The relationship between the sample size and each index's correlation with true informativity as well as the *t*-values in the regression models are visualized in Figure 5. Informativity was consistently the strongest predictor of the true informativity effect, followed by frequency, though the *t*-values of frequency were still at least an order of magnitude smaller than those of informativity, suggesting that false frequency effects are only likely in the presence of strong informativity effects. The *t*-values of predictability were rarely significant in any direction. It therefore seems highly unlikely that some index would spuriously seem significant when the true effect is due to informativity, provided that all three are used in the model.

2.3 Discussion

Even for lexical databases with fairly few word tokens (k=10,000) and a *behavioral* database containing about m=5,000 words, the estimates for all three probabilistic indices are well-behaved in the sense that each index is the best estimator of its "true" value, and that, with more data, the other indices lose explanatory power. This is not surprising, given that the number of values that need to be estimated is not as big as it can be in e.g. the word-level domain. At the segmental domain, frequency and informativity will typically be calculated for less than 50 unique segments, and predictability for hundreds of thousands of segments in context. However, even in the segmental domain, the sample frequency index reliably emerged spuriously as the second-strongest predictor of true segment predictability, and even more strongly for true segment informativity. In Study 2, we investigate the correlations of (true) informativity and both sample frequency and predictability in more depth.

3 Study 2: When informativity is not controlled for in the analysis

Informativity as defined here is a relatively new concept in linguistics. A remaining question is therefore whether it is possible that effects that were caused by informativity could have been erroneously attributed to other factors because informativity was not included in the models. Study 2 therefore used the model of Study 1c to predict true informativity with sample frequency and predictability as predictors, but without sample informativity.

3.1 Results

The correlations for predictability and frequency were the same as in Study 1c (same indices and true effect, and the correlation does not depend on the inclusion of other variables). In the regression models, the *t*-values of both frequency ($\hat{\mu}$=66.7, range=61.8 to 73.6) and predictability ($\hat{\mu}$=58.8, range=55.5 to 63.3) were consistently high.

The relationship between the sample size and each variable's correlation with true informativity as well as the *t*-values in the regression models are visualized in Figure 6. Though frequency is assigned higher *t*-values than predictability, they are in the same rough range, and both consistently seem to predict informativity when informativity is not included in the model.

3.2 Discussion

In Studies 1a-c informativity seemed unlikely to emerge as an artifact of frequency and predictability. However, when informativity is not controlled for, effects that should be attributed to informativity may falsely appear to be predicted by frequency and predictability. This raises the need to include informativity in analyses in linguistics alongside frequency and predictability.

4 General discussion

It is now well-established that language production, including phonological and phonetic encoding, is affected by probabilistic knowledge. Many studies have found that more frequent and more predictable word forms are

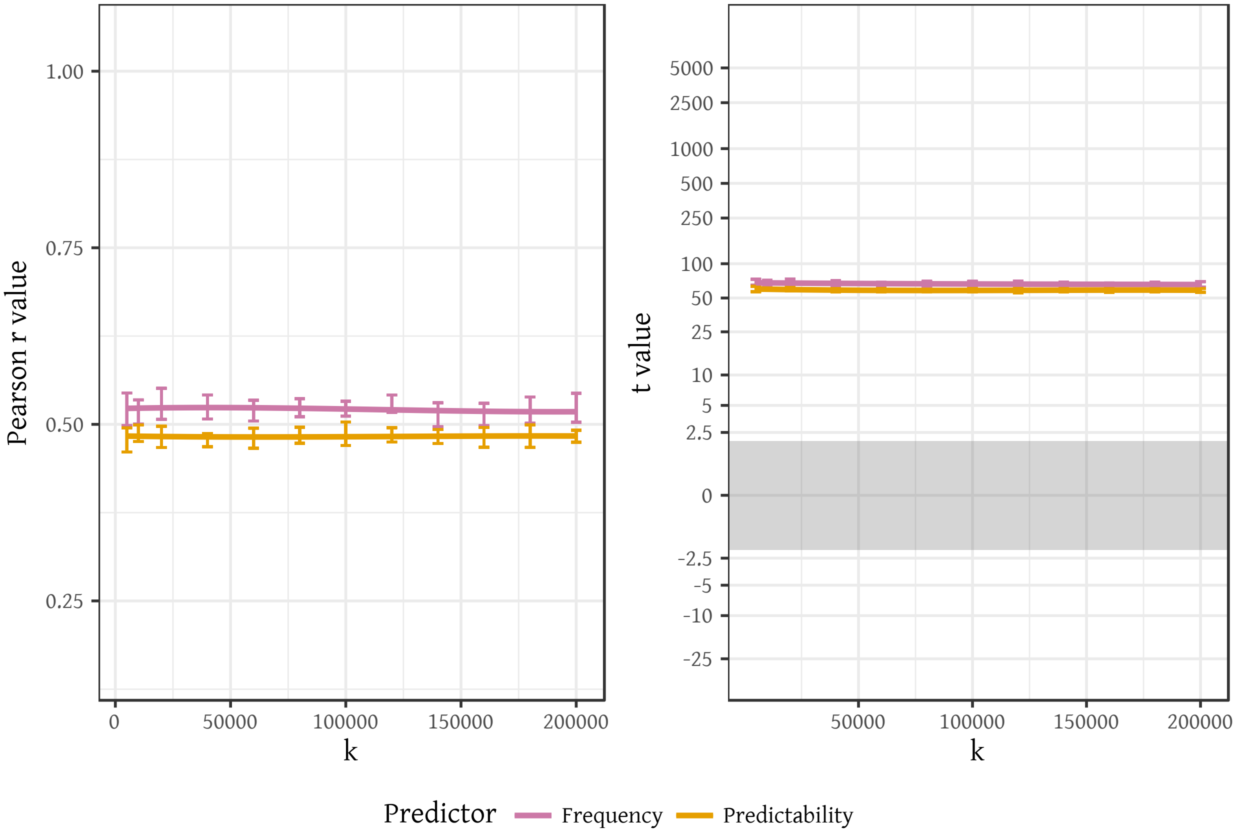

Figure 6: Left: Correlations between true segment informativity and researcher's estimates of informativity. Right: *t*-values of frequency and predictability, but without informativity, in a multiple regression predicting true segment informativity. Axes, scaling, and shading follow Figure 3.

more likely to be realized with reduced signals (for a recent review, see Jaeger & Buz, 2017). A similar link is observed between (phonetic and phonological) word forms and their frequency or average predictability. These and related findings highlight the importance of probabilistic indices in understanding phonology, including both synchronic and diachronic sound patterns (for review, see Cohen Priva 2017a; Hall et al. 2018). While there is now increasing empirical agreement that probabilistic indices are correlated with phonetic realization and phonological form (signal reduction), questions remain about *which* indices best predict signal reduction, and why (i.e., what are the mechanisms underlying these effects?). Here we have presented a computational approach to the first question, or rather, we have presented a first step toward understanding how different indices of probability are related, and how this affects the likelihood of spurious effects. Our simulations, based on the natural statistics of words and segments, suggest that there are clear risks of spurious effects – i.e., effects arising purely as a by-product of the high correlations between probabilistic indices. Most strikingly, we found that spuriously significant frequency effects are likely to emerge as a side-product of predictability and informativity effects. This risk is particularly high when the underlying factor is informativity (as recent work suggests; Cohen Priva 2008, 2015; Piantadosi et al. 2011; Seyfarth 2014). This calls for caution when assessing frequency effects in the absence of controls for predictability and informativity, as continues to be the case in many usage-based studies, and more broadly in psycholinguistics (e.g., Bybee & Scheibman 1999; Bybee et al. 2016; Fenk-Oczlon et al. 2010; Pierrehumbert 2001, 2003). The inclusion of multiple predictability indices and replication across mutiple corpora may address some of the concerns raised in Clopper et al. (2018) and Foulkes et al. (2018).

In separate studies (not reported here due to space limitations), we have replicated these results in the word-level domain using bigrams, and conditioning either on the previous or the following words. The results of those additional studies were similar, except that the problems observed here were observed more strongly and even for larger lexical databases when words, instead of segments, were analyzed. This is presumably due to the overall much larger number of word types, compared to segment types. For example, spurious frequency effects were found when predicting word predictability or word informativity, even for lexical databases with 10,000,000 words.

In its current form, our approach makes several simplifying assumptions. We briefly discuss three that we consider particularly relevant for future extension of our approach. First, all our studies simulated a behavioral database with m=5,000 words (roughly 15,000 segments). This is a bit smaller than in previous studies on segment articulation (e.g., 50,000 syllables in Aylett & Turk 2006; 50,000 words in van Son & Pols 2003; 13,000 segments in van Son & van Santen 2005) or insertion (e.g., 3,000 words in Tily & Kuperman 2012), and omission (e.g., 27,000 and 35,000 segments in Cohen Priva 2015). Still, it is not representative of *all* studies, and the value of m is expected to affect the severity of the problems we report: for larger m, the issues we discussed above will decrease; for smaller m, the issues may increase. Second, the regressions used in this paper are not a direct test of behavioral studies, in which multiple predictors can affect the outcome variable. Such predictors include multiple probabilistic indices and other predictors that might be correlated with probabilistic indices (e.g., function words are typically frequent). Moreover, such studies would likely be modeled using mixed effects models, which may diminish some of the effects we observe. Cohen Priva (2015) reports that word frequency effects disappeared after the inclusion of word as a random intercept. Finally, the methods used here were only applied to American English datasets. While we believe that these datasets represent relevant linguistic properties, it is still necessary to replicate the results in languages that have sparser (or denser) domains, such as languages with richer (or poorer) morphology (Turnbull 2018; Kawahara & Lee 2018). Future research may also extend to other probabilistic indices (e.g. global lexical entropy: Cohen Priva & Gleason 2016; Cohen Priva 2017b; mutual information: Bell et al. 2003; Gries 2010).

5 Acknowledgements

We would like to thank Emily Gleason and Elinor Amit for their feedback.

References

Al-Nassir, Abdulmunim A. 1993. *Sibawayh the phonologist: a critical study of the phonetic and phonological theory of sibawayh as presented in his treatise al-kitab.* London [u.a.]: Kegan Paul Internat.

Arnold, Jennifer E., Jason M. Kahn & Giulia C. Pancani. 2012. Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review* 19(3). 505–512. doi:10.3758/s13423-012-0233-y.

Aylett, Matthew & Alice Turk. 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.

Aylett, Matthew & Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Acoustical Society of America Journal* 119. 3048–3058.

Bates, Elizabeth & Brian MacWhinney. 1987. Competition, Variation, and Language Learning. In Brian MacWhinney (ed.), *Mechanisms of language acquisition*, chap. 6, 157–194.

Bell, Alan, Jason Brenier, Michelle Gregory, Cynthia Girand & Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1). 92–111.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2). 1001–1024. doi:10.1121/1.1534836.

Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(03). 261–290. doi:10.1017/S0954394502143018.

Bybee, Joan, Richard J. File-Muriel & Ricardo Napoleão De Souza. 2016. Special reduction: a usage-based approach. *Language and Cognition* 8(3). 421–446. doi:10.1017/langcog.2016.19.

Bybee, Joan & Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of don't in english. *Linguistics* 37(4). doi:10.1515/ling.37.4.575.

Carter, M. G. 2004. *Sibawayhi*. London; New York: I.B. Tauris.

Cieri, Christopher, David Graff, Owen Kimball, Dave Miller & Kevin Walker. 2005. Fisher English training part 2, transcripts. Linguistic Data Consortium, Philadelphia.

Cieri, Christopher, David Miller & Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the 4th international conference on language resources and evaluation*, 69–71.

Clopper, Cynthia G., Rory Turnbull & Rachel Steindel Burdin. 2018. Assessing predictability effects in connected read speech. *Linguistics Vanguard* 4(S2).

Cohen Priva, Uriel. 2008. Using information content to predict phone deletion. In Natasha Abner & Jason Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*, 90–98. Somerville, MA: Cascadilla Proceedings Project.

Cohen Priva, Uriel. 2012. *Sign and signal: deriving linguistic generalizations from information utility*: Stanford University dissertation. http://purl.stanford.edu/wg646gh4444.

Cohen Priva, Uriel. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology* 6(2). 243–278. doi:10.1515/lp-2015-0008.

Cohen Priva, Uriel. 2017a. Informativity and the actuation of lenition. *Language* 93(3). 569–597. doi:10.1353/lan.2017.0037. https://muse.jhu.edu/article/669547.

Cohen Priva, Uriel. 2017b. Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition* 160. 27–34. doi:10.1016/j.cognition.2016.12.002.

Cohen Priva, Uriel & Emily Gleason. 2016. Simpler structure for more informative words: a longitudinal study. In A. Papafragou, D. Grodner, D. Mirman & J.C. Trueswell (eds.), *Proceedings of the 38th annual conference of the cognitive science society*, 1895–1900. Austin, TX: Cognitive Science Society. https://mindmodeling.org/cogsci2016/papers/0331/index.html.

Daland, Robert & Kie Zuraw. 2018. Loci and locality of informational effects on phonetic implementation. *Linguistics Vanguard* 4(S2).

Ernestus, Mirjam. 2014. Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua* 142. 27–41. doi:10.1016/j.lingua.2012.12.006.

Fenk-Oczlon, Gertraud, August Fenk & Pamela Faber. 2010. Frequency effects on the emergence of polysemy and homophony. *International Journal of Information Technologies and Knowledge* 4(2). 103–109.

Foulkes, P., G. Docherty, S. Shattuck-Hufnagel & V. Hughes. 2018. Three steps forward for predictability. consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. *Linguistics Vanguard* 4(S2).

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production. In B. C. In Love, K. McRae & V. M. Sloutsky (eds.), *Proceedings of the 30th annual meeting of the cognitive science society (cogsci08)*, 939–944. Austin, TX: Cognitive Science Society.

Gahl, Susanne. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3). 474–496.

Gahl, Susanne, Yao Yao & Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4). 789 – 806. doi:10.1016/j.jml.2011.11.006.

Godfrey, John J. & Edward Holliman. 1997. Switchboard-1 release 2. Linguistic Data Consortium, Philadelphia.

Gries, Stefan Th. 2010. Useful statistics for corpus linguistics. In Aquilinio Sánchez Pérez & Moisés Almela Sánchez (eds.), *A mosaic of corpus linguistics*, 269–291. Bern, Switzerland: Peter Lang.

Hall, Kathleen, Elizabeth Hume, T. Florian Jaeger & Andrew B. Wedel. 2016. The message drives phonology .

Hall, Kathleen, Elizabeth Hume, T. Florian Jaeger & Andrew B. Wedel. 2018. The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4(S2).

Jaeger, T. Florian. 2006. *Redundancy and Syntactic Reduction in Spontaneous Speech*: Stanford University Doctoral.

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

Jaeger, T. Florian. 2013. Production preferences cannot be understood without reference to communication. *Frontiers in Psychology* 4(230). doi:10.3389/fpsyg.2013.00230.

Jaeger, T. Florian & Esteban Buz. 2017. Signal reduction and linguistic encoding. In Fernández Eva M. & Helen Smith Cairns (eds.), *Handbook of psycholinguistics*, Wiley-Blackwell.

Kawahara, Shigeto & S. Lee. 2018. Truncation in message-oriented phonology: A case study using korean vocative truncation. *Linguistics Vanguard* 4(S2).

Kuperman, Victor & Joan Bresnan. 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.

Lewis, Jerry W. & Luis A. Escobar. 1986. Suppression and enhancement in bivariate regression. *Journal of the Royal Statistical Society. Series D (The Statistician)* 35(1). 17–26. doi:10.2307/2988294.

Lindblom, Björn. 1990. Explaining phonetic variation: A sketch of the H&H theory. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modeling* (NATO ASI Series 55), 403–439. Springer Netherlands.

Pate, John K. & Sharon Goldwater. 2015. Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language* 78. 1–17. doi:10.1016/j.jml.2014.10.003.

Piantadosi, Steven T., Harry J Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* .

Pierrehumbert, Janet. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. John Benjamins Publishing Company.

Pierrehumbert, Janet B. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46(2-3). 115–154. doi:10.1177/00238309030460020501.

Pitt, M.A., L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume & E. Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release). Department of Psychology, Ohio State University.

Pluymaekers, Mark, Mirjam Ernestus & R. Harald Baayen. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62. 146–159.

Seyfarth, Scott. 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, Claude Elwood. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27. 379–423.

Shaw, Jason & Shigeto Kawahara. 2018. Predictability and phonology: Past, present & future. *Linguistics Vanguard* 4(S2).

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

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

Tily, Harry & Victor Kuperman. 2012. Rational phonological lengthening in spoken Dutch. *The Journal of the Acoustical Society of America* 132(6). 3935–3940. doi:10.1121/1.4765071.

Turnbull, Rory. 2018. Patterns of probabilistic segment deletion/reduction in English and Japanese. *Linguistics Vanguard* 4(S2).

Wedel, Andrew B. 2006. Exemplar models, evolution and language change. *The Linguistic Review* 23(3). 247–274.

Weide, R. 2008. The CMU pronunciation dictionary, release 0.7a. Carnegie Mellon University.

Zipf, George Kingsley. 1935. *The psycho-biology of language: an introduction to dynamic philology*. Houghton, Mifflin.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: an introduction to human ecology*. New York: Hafner Publisher Company.