Cynthia G. Clopper*, Rory Turnbull and Rachel Steindel Burdin

Assessing predictability effects in connected read speech

https://doi.org/10.1515/lingvan-2017-0044
Received April 28, 2017; accepted June 4, 2018

Abstract: A wide range of reduction phenomena have been described in the literature as predictability effects, in which more predictable units (i.e. words, syllables, vowels) are reduced in duration or other acoustic dimensions relative to less predictable units. The goal of the current study was to critically evaluate these predictability effects on vowel duration in read speech to explore the extent to which they reflect a single underlying phenomenon. The results revealed shorter vowel duration for words with high phonotactic probability, for high-frequency words (in clear speech only), and for words in plain lab speech relative to clear speech. However, the results also revealed qualitatively different effects of three measures of contextual probability (cloze probability, written trigram probability, and spoken trigram probability). Greater spoken trigram probability predicted longer vowel duration, contrary to expectations, and this effect was limited to high-frequency words in first mentions and in plain speech. Cloze probability and written trigram probability exhibited even more complex interactions with other predictability measures. These results provide evidence for fundamental differences in these measures of predictability, suggesting that a more nuanced perspective on predictability effects and the mechanisms underlying them is necessary to account for the complexity of the empirical data.

Keywords: predictability; cloze probability; trigram probability; vowel duration.

1 Introduction

The concept of predictability is invoked to explain a wide range of acoustic-phonetic reduction phenomena (see Clopper and Turnbull 2018). The primary claim is that more predictable units (e.g. words, syllables, vowels) are reduced in duration, f0, intensity, and/or spectral properties relative to less predictable units. Predictability is estimated in many ways, including unigram probability (i.e. lexical frequency), phonotactic probability, contextual probability (e.g. cloze probability, corpus-based n-gram probabilities), and discourse mention. Conceptual connections are also drawn between these various measures of predictability and measures of informativity (e.g. Seyfarth 2014; Cohen Priva 2015) and notions of accessibility or the givenness of referents in a discourse (Bard et al. 2000; Arnold 2008; Lam and Watson 2010).

Two theoretical accounts of predictability effects have been provided in the literature. The listener-oriented account defines predictability in terms of the trade-off between talker effort and communicative efficacy (Lindblom 1990): talkers reduce forms when listeners are expected to correctly access a word and talkers can therefore produce a shorter form without sacrificing intelligibility (Fowler and Housum 1987; Fowler 1988; Aylett and Turk 2004). The talker-oriented account defines predictability in terms of speech planning: talkers reduce forms because they are easier to access for the talker (Bard et al. 2000; Baese-Berk and Goldrick 2009; Bell et al. 2009; Lam and Watson 2010; Daland and Zuraw 2018; cf. Tomaschek et al. 2018). Thus, despite a critical difference in the proposed underlying mechanism, the two theories are aligned in their assumption that predictability effects arise from the relative availability of words in context.

***Corresponding author: Cynthia G. Clopper,** Department of Linguistics, Ohio State University, 1712 Neil Ave, Oxley Hall 100, Columbus, OH 43210, USA, E-mail: clopper.1@osu.edu
Rory Turnbull: Department of Linguistics, University of Hawaiʻi at Mānoa, 1890 East West Road, Moore Hall 569, Honolulu, HI 96822 USA, E-mail: rory.turnbull@hawaii.edu
Rachel Steindel Burdin: Department of English, Hamilton Smith Hall, University of New Hampshire, 95 Main Street, Durham, NH 03824 USA, E-mail: Rachel.Burdin@unh.edu

These theoretical models consider different measures of predictability to reflect approximations of the same fundamental phenomenon. For example, Aylett and Turk (2004) conceptualized lexical frequency, syllable trigram probability, and discourse mention as different measures of redundancy at different levels of linguistic structure. This treatment of these various measures as reflecting different aspects of the same underlying phenomenon is potentially sensible, given that these factors are often correlated with one another and appear to produce comparable effects on duration. However, these different measures of predictability are not interchangeable. For example, Tremblay and Tucker (2011) observed stronger effects of lexical frequency than n-gram probabilities on the duration of four-word sequences, and Turnbull (2017) observed stronger effects of contextual probability than discourse mention on word f0 peaks. In addition, measures of predictability have been shown to interact with one another, revealing limits on the magnitude of reduction that is produced in unpredictable contexts (Baker and Bradlow 2009; Bell et al. 2009). This variation in the magnitude of predictability effects on speech production means that the choice of predictability measure(s) may affect the outcome of a given study. Moreover, because predictability measures are not interchangeable and can exhibit interactions with one another, these measures may reflect different underlying phenomena, rather than different approximations of the same phenomenon.

The goal of the current study was to critically evaluate the contribution of different measures of predictability to vowel duration in read speech. First, to assess the potentially variable role of predictability at different levels of linguistic structure, we considered a number of different measures, including lexical frequency, phonotactic probability, contextual probability, and discourse mention, in a corpus of read speech. We examined the main effects and interactions of these measures on vowel duration to understand how these different aspects of predictability contribute individually and in combination to temporal reduction (cf. Bell et al. 2009). We expected to observe shorter vowel durations for higher-frequency words, words with higher mean phonotactic probability, second mentions, and relatively more probable words, consistent with the literature showing that greater predictability leads to greater temporal reduction (Fowler and Housum 1987; Bell et al. 2009). However, we also expected to observe interactions among these variables, as well as variation in the strength of their effects on vowel duration, given the variability that has been observed in previous work (Baker and Bradlow 2009; Lam and Watson 2010; Smith and Levy 2011).

Second, whereas lexical frequency, neighborhood density, discourse mention, and speaking style were orthogonally manipulated in the corpus that we analyzed, contextual probability was not defined in advance. Moreover, given that the corpus materials that we analyzed were read speech, it is not obvious what kind of estimate of contextual probability is most appropriate. We therefore assessed potential variation across three different measures of contextual probability: cloze probability, written trigram probability, and spoken trigram probability (cf. Smith and Levy 2011). These three measures each address a conceptually distinct, but equally plausible formulation of local predictability. Cloze probabilities could be argued to be the most ecologically valid measure of contextual probability (Smith and Levy 2011). Written trigram probabilities involve extensive genre and style coverage and may therefore best reflect contextual probabilities in English overall. Finally, the spoken trigram probabilities may be most reflective of contextual probability in naturally-produced speech. Thus, all three measures provide a potential approximation of contextual predictability in our materials.

2 Methods

2.1 Corpus

The materials comprise the Columbus subset of the Ohio State Stories Corpus (Burdin and Clopper 2015), which includes readings of 30 short stories by each of 30 talkers (10 females and 5 males each from the Midland and Northern dialects of American English). Each story was read by each talker first in a plain lab style as if talking to a friend or family member and then in a clear lab style as if talking to a non-native or hearing-impaired listener. Since the stories were more predictable overall in the clear speech reading than in the plain speech reading because the talkers had previously read them, the magnitude of the style effect on vowel duration may be underestimated in this corpus.

The stories were designed to include 236 mostly monosyllabic target words that vary orthogonally in lexical frequency and neighborhood density, as defined in the Hoosier Mental Lexicon (Nusbaum et al. 1984). Each target word occurs twice in its story, with the number of words intervening between the two mentions varying from 4 to 231. Thus, each target word was produced four times by each talker in a 2 (discourse mention) $\times$ 2 (speaking style) design. The local contextual probability of each mention in each story was estimated using a cloze task. In this task, an independent set of participants were visually presented with a sentence from a story with the target word removed and were asked to provide the missing word; the cloze probability estimate is the mean response accuracy in this task (see Burdin and Clopper 2015 for further details). Additionally, written trigram probability estimates were extracted from the Google Web 1T corpus (Brants and Franz 2006) using Get1T (Hawker et al. 2007) and spoken trigram probability estimates were extracted from a weighted mixture of the Buckeye (Pitt et al. 2007) and Fisher (Cieri et al. 2005) corpora of conversational speech using SRILM (Stolcke 2002; Stolcke et al. 2011). These trigram measures provide the conditional probability of the target word given the preceding two words. All three contextual probability measures are uncorrelated with discourse mention.

The corpus was aligned using the Penn Phonetics Lab Forced Aligner (Yuan and Liberman 2008) and the vowel boundaries of the target words were hand-corrected. The duration, F1, and F2 of the stressed vowel in each target word were extracted; only the duration measure was considered in the current study. Outliers (defined as tokens with duration, F1, and/or F2 estimates more than three standard deviations from the mean for that vowel and talker; 3%), disfluencies (2%), and tokens involving non-modal voice qualities that may have disrupted measurement estimation (13%) were excluded, leaving a total of 23,262 tokens for analysis.

2.2 Statistical analysis

Mixed-effects linear regression modeling was used to predict log-transformed and centered vowel duration from the log lexical frequency (Nusbaum et al. 1984) and mean biphone probability (Vitevitch and Luce 2004) of the target word,[1] discourse mention (first or second), speaking style (plain or clear), and the log probability of the target word using the cloze,[2] Google 1T, and Fisher/Buckeye estimates (R version 2.10.1; lme4 version 0.999375-32). All interactions among these fixed effects were also considered. Prior to the analysis, words with outlier contextual probability values, defined as raw values greater than 0.1 for the Google 1T ($N = 17$) and Fisher/Buckeye ($N = 11$) measures, were removed. Visual inspection of the data showed clear breaks in the probability distributions at these points and most of the outliers involved highly lexicalized phrases (e.g. *duffel bag* and *chief of staff*). Two target words were also inadvertently left out of the cloze task and were excluded. The exclusions of outliers and missing data resulted in a total of 20,660 observations in the analysis. As shown in Figure 1, the three contextual probability measures were significantly correlated with one another (all $p < 0.001$). Thus, each contextual probability measure was analyzed in a separate model with the other fixed factors. Variance inflation factors (VIFs) confirmed that the predictor variables were not problematically collinear in any of the three models (all VIF <2.40). The log frequency, phonotactic probability, and contextual probability measures were standardized (z-scored) to facilitate interpretation of effect sizes. Discourse mention and speaking style were sum-contrast coded. Models included maximal simple random effects for talkers and words. Statistical significance of the fixed factors was defined as $|t| > 2$ (Baayen et al. 2008). Comparisons across the models involving different measures of contextual probability were based on

1 Although neighborhood density was manipulated in the design of the corpus, we examined phonotactic probability in the current analysis, given our focus on measures of predictability. We acknowledge that these two measures are not necessary correlated (Vitevitch and Luce 1999), but both were orthogonal to lexical frequency in our materials.

2 For the cloze probability measure, values of 0 were replaced with 0.01 prior to log-transformation to avoid undefined values and achieve a distribution that approximates a normal distribution. All Google 1T and Fisher/Buckeye estimates were greater than 0 due to smoothing.

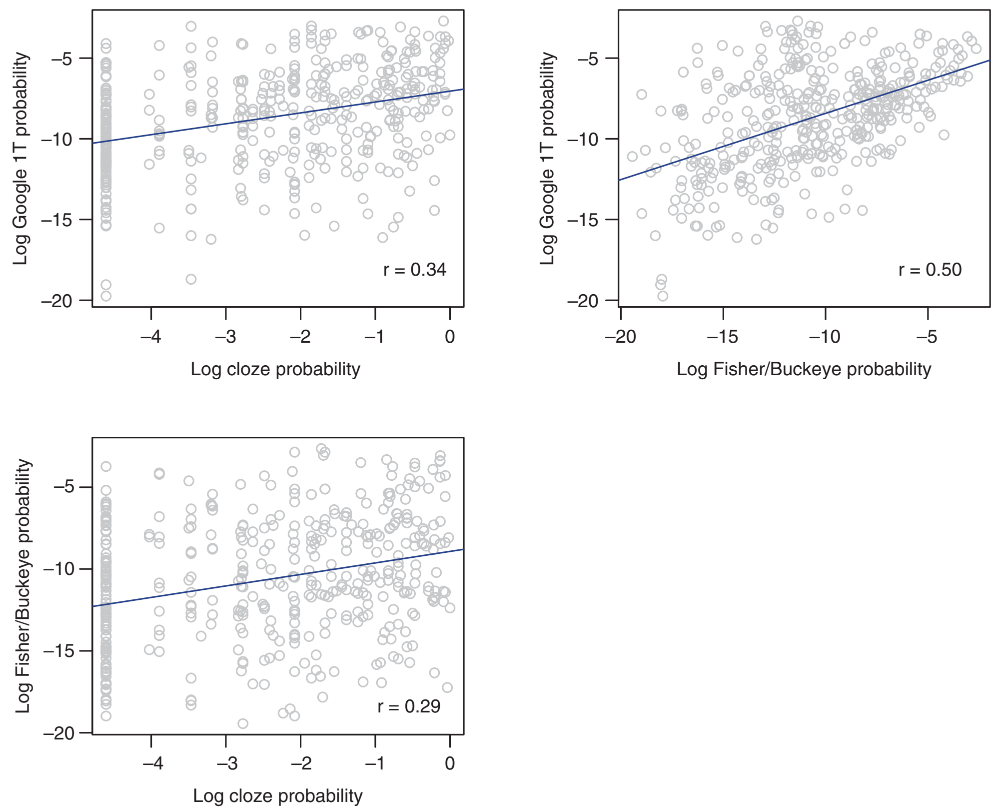

Figure 1: Correlations among the three contextual probability measures for each mention of each target word.

BICs and pseudo-r^2, defined as the correlation between the observed and fitted model values.[3] Simple correlations between each measure of contextual probability and mean vowel duration for each mention of each word were also calculated to assess the independent contribution of contextual probability to vowel duration in the corpus.

3 Results

3.1 Cloze probability

The BIC for the model involving cloze probability was −2005 and the pseudo-r^2 was 0.84. The full model output is provided in Table A1 in the Appendix. A significant negative main effect of phonotactic probability ($b = -0.058$, $t = -3.08$) and a significant positive main effect of speaking style ($b = 0.058$, $t = 4.71$) were observed. As expected, higher phonotactic probability led to shorter durations than lower phonotactic probability and clear speech led to longer durations than plain speech. The frequency × speaking style interaction was also significant ($b = -0.008$, $t = -3.25$); post-hoc models using treatment contrasts for speaking style revealed a significant negative effect of frequency in clear speech, but no significant effect of frequency in plain speech. Thus, in clear speech only, higher lexical frequency led to shorter durations than lower lexical frequency, as expected (Bell et al. 2009).

Three interactions involving cloze probability were also significant: frequency × cloze probability × discourse mention ($b = 0.020$, $t = 2.60$), phonotactic probability × cloze probability × speaking

3 To allow for comparisons of model fit across the three measures of contextual probability, the full models are reported. Backward selection procedures were applied to each analysis and resulted in the same patterns of significant main effects and interactions.

style ($b = 0.005$, $t = 2.29$), and the five-way interaction between frequency, phonotactic probability, cloze probability, discourse mention, and speaking style ($b = -0.004$, $t = -2.31$). The combined effect of the five predictors on vowel duration is shown in Figure 2. Although frequency and phonotactic probability were treated as continuous variables in the analysis, they are shown as binary variables based on a median split in Figure 2 to allow for visualization of the interaction. The figure suggests that words with low phonotactic probability (dotted lines) exhibit the expected effect of cloze probability, in which higher probability led to shorter duration, but only in second mentions in clear speech (top right panel). However, the figure also suggests an unexpected positive effect of cloze probability, in which higher probability led to longer duration for high-frequency words with high phonotactic probability (dashed green lines), especially in first mentions (left panels). However, all of the slopes in the figure are modest in magnitude and the interactions may therefore primarily reflect noise in the dataset. Unsurprisingly given this pattern of results, the simple correlation between cloze probability and mean vowel duration was very small ($r = -0.04$).

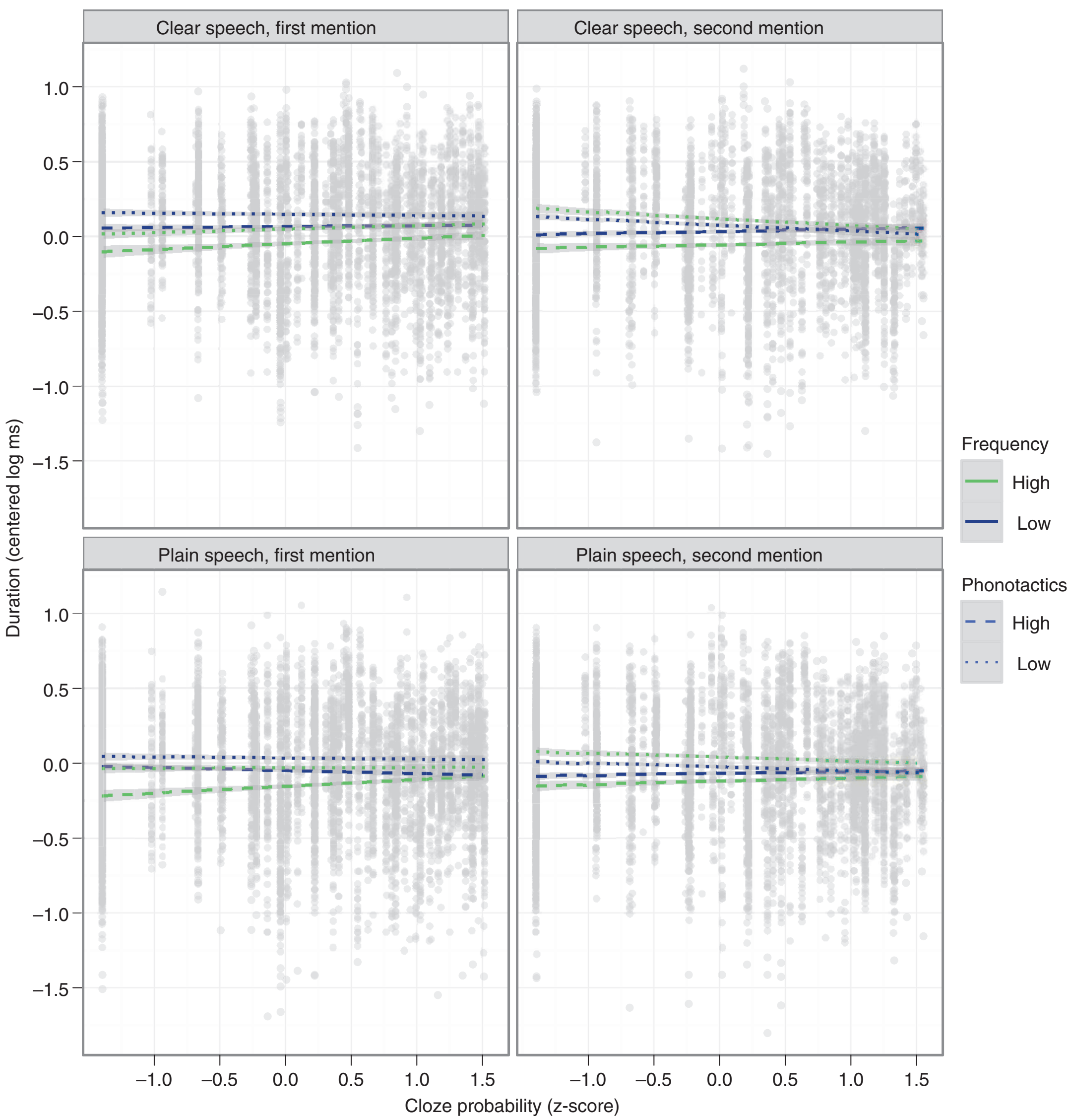

Figure 2: Combined effects of frequency, phonotactic probability, cloze probability, discourse mention, and speaking style. Lines represent frequency and phonotactic probability as binary variables with linear model estimate regression lines superimposed on the data points.

3.2 Google 1T probability

The BIC for the model involving Google 1T probability was −2015 and the pseudo-r^2 was 0.84. The full model output is provided in Table A2 in the Appendix. As in the cloze probability analysis, a significant negative main effect of phonotactic probability ($b = -0.055$, $t = -2.83$), a significant positive main effect of speaking style ($b = 0.059$, $t = 4.81$), and a significant interaction between frequency and style ($b = -0.010$, $t = -3.68$) were observed. Unlike in the cloze probability analysis, the five-way interaction was not significant, but two four-way interactions were: frequency × phonotactic probability × discourse mention × speaking style ($b = 0.005$, $t = 2.50$) and phonotactic probability × Google 1T probability × discourse mention × speaking style ($b = -0.007$, $t = -2.96$). Given the two significant four-way interactions, the combined effect of all five predictors on vowel duration is shown in Figure 3. The figure suggests that the effect of Google 1T probability was in the unexpected positive direction for first mention words with low phonotactic probability (dotted lines in left panels), and was more generally in the unexpected positive direction for second mention words of all types (right panels). The magnitude of the interaction between phonotactic probability and Google 1T probability was smaller overall in plain speech

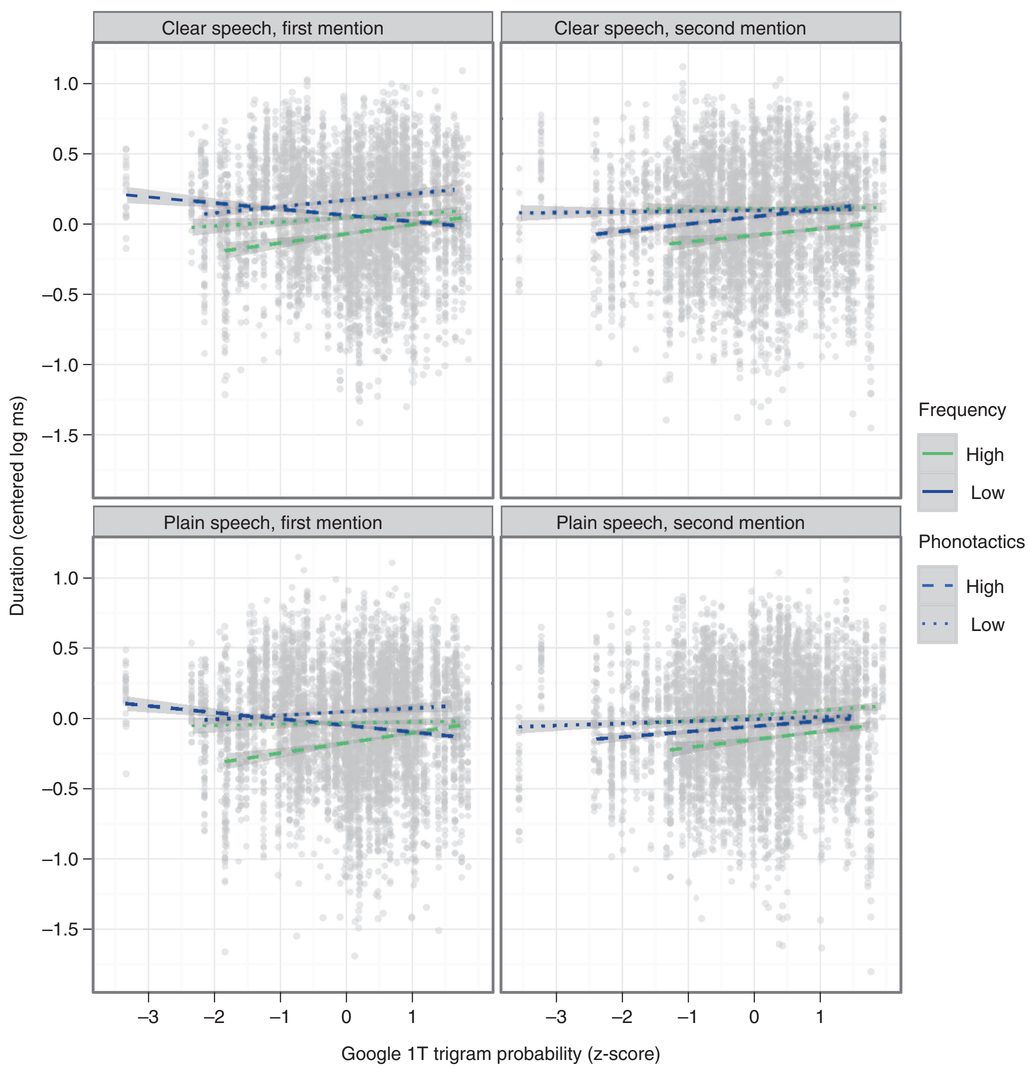

Figure 3: Combined effects of frequency, phonotactic probability, Google 1T probability, discourse mention, and speaking style. Lines represent frequency and phonotactic probability as binary variables with linear model estimate regression lines superimposed on the data points.

(bottom panels) than in clear speech (top panels), but was not significant in either style or for either mention in post-hoc models using treatment contrasts, suggesting that these interactions may primarily reflect noise in the dataset. The simple correlation between Google 1T probability and mean vowel duration was very small ($r = -0.01$), reflecting this mixed pattern of results.

3.3 Fisher/Buckeye probability

The BIC for the model involving Fisher/Buckeye probability was -2019 and the pseudo-r^2 was 0.84. The full model output is provided in Table A3 in the Appendix. As in the two previous analyses, a significant negative main effect of phonotactic probability ($b = -0.081$, $t = -3.80$), a significant positive main effect of speaking style ($b = 0.061$, $t = 4.92$), and a significant interaction between frequency and style ($b = -0.008$, $t = -2.76$) were observed. The model also revealed four additional significant interactions: frequency × Fisher/Buckeye probability ($b = 0.030$, $t = 2.07$), frequency × phonotactic probability × Fisher/Buckeye probability ($b = 0.037$, $t = 2.31$), frequency × Fisher/Buckeye probability × discourse mention ($b = 0.015$, $t = 2.30$), and frequency × Fisher/Buckeye probability × speaking style ($b = -0.005$, $t = -2.42$). Given that Fisher/Buckeye probability interacts with all other factors, the

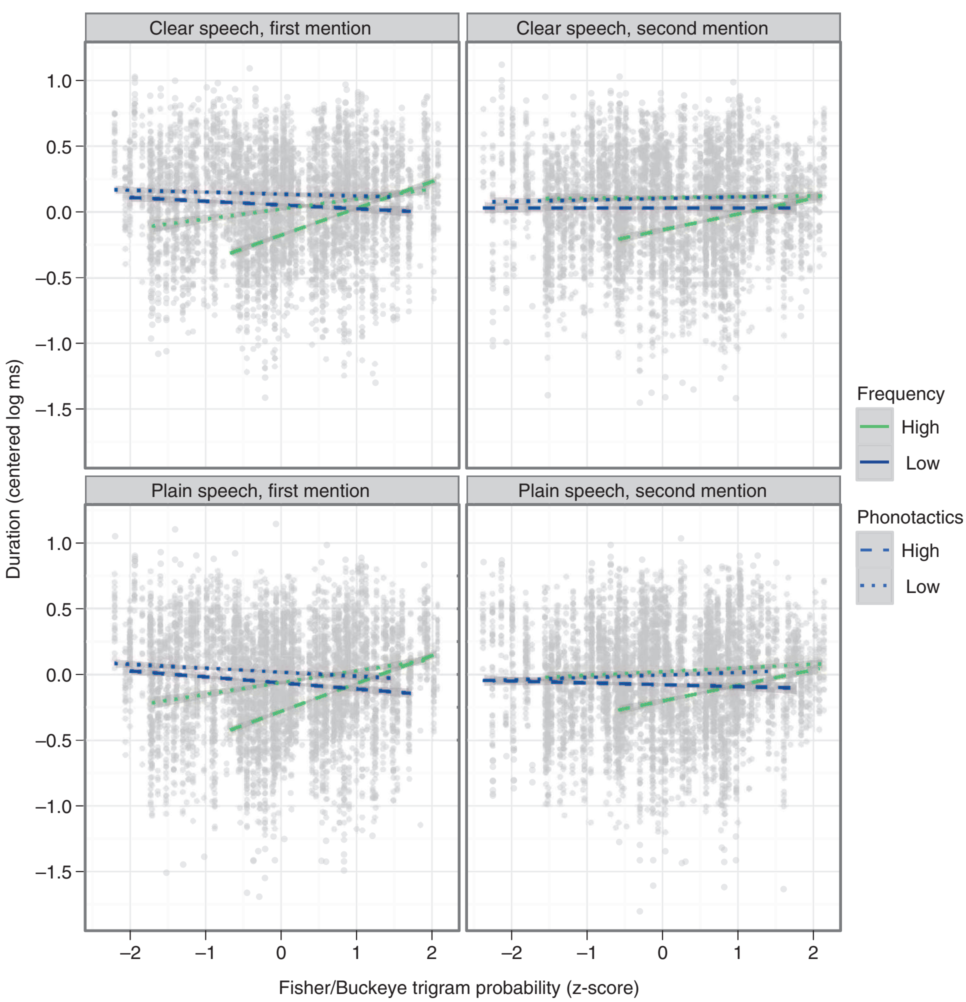

Figure 4: Combined effects of log frequency, phonotactic probability, log Fisher/Buckeye probability, discourse mention, and speaking style. Lines represent log frequency and phonotactic probability as binary variables with linear model estimate regression lines superimposed on the data points.

combined effect of all five predictors on vowel duration is shown in Figure 4. The figure reveals a significant positive effect of Fisher/Buckeye probability on vowel duration for high-frequency words (green lines), especially for first mentions (left panels) and plain speech (bottom panels); post-hoc models using treatment contrasts confirmed a significant frequency × Fisher/Buckeye probability interaction for both first mentions and plain speech. The simple correlation between Fisher/Buckeye probability and mean vowel duration was very small ($r = 0.01$), consistent with this mixed pattern of results.

4 Discussion

Together, the results of the three analyses provide additional evidence for differential effects of predictability, defined at different levels of linguistic structure, on vowel duration. Phonotactic probability and speaking style had consistent effects of similar magnitude in all three models, confirming both reduction of words with high phonotactic probability and in plain speech relative to clear speech. Although lexical frequency was not independently significant in any of the models, it emerged in a significant interaction with speaking style, as well as in higher-order interactions, in all three models. The effect of lexical frequency on vowel duration was smaller than the phonotactic probability and speaking style effects and only emerged as significant in clear speech, where higher-frequency words were shorter than lower-frequency words. The interaction is consistent with previous work suggesting a lower bound on reduction (e.g. Bell et al. 2009), such that high-frequency words are less reduced in plain speech than in clear speech because plain speech is reduced overall relative to clear speech. However, the same kind of interaction was not observed for phonotactic probability, despite its robust overall effect. Thus, predictability effects at the lexical level (i.e. frequency and phonotactics) differ in their magnitude and in their interaction with style, although both are consistent with the observation that more predictable units are shorter than less predictable units.

The predictability effects at higher levels of linguistic structure (i.e. contextual probability and discourse mention) differ substantially from the lexical-level effects. First, as discussed in more detail below, the effect of contextual probability varied depending on the specific measure that was considered, but when contextual probability was significant, the effect was in the unexpected direction. Greater Fisher/Buckeye probability led to longer vowel duration rather than shorter vowel duration for higher-frequency words in first mentions and in plain speech.[4] Second, discourse mention was not independently significant in any of the three analyses, although it emerged in significant interactions in all three models. This lack of an overall effect of discourse mention may be due to the length of the mostly monosyllabic target words, given that second mention reduction is greater for longer words than shorter words because longer words have more "room" to reduce (Fowler and Housum 1987). Alternatively, or in addition, second mention reduction may have been limited due to variability in the distance between mentions (4–231 words) or in the semantic relationship between the two mentions, given that second mention reduction is smaller following a change in the discourse scene or topic (Fowler et al. 1997) and for words with different referents (Bard et al. 1989).

These differences in the magnitude and direction of the effects of lexical-level and higher-level predictability effects suggest that these measures may not be capturing different aspects of a single phenomenon, but rather different phenomena altogether. This proposal that predictability is not a unitary phenomenon is also supported by the results of several recent studies showing that reduction in different acoustic domains is related to distinct measures of predictability. For example, reduction in duration is observed primarily as a function of discourse mention, whereas reduction in intensity and f0 is observed primarily as a function of contextual probability (Lam and Watson 2010; Turnbull 2017). This separation of predictability effects across

4 Parallels to this unexpected effect have been observed for phonemically long vowels in German, which are longer in high-frequency words than low-frequency words (Tomaschek et al. 2013), and for phonemically long /a/ in Japanese, which is relatively longer than short /a/ in more predictable contexts as opposed to less predictable contexts (Shaw and Kawahara In press). In both cases, the phonemic vowel length contrast is enhanced in the more predictable context. A consideration of a parallel effect of vowel length in American English is left for future research.

acoustic dimensions has been argued to reflect a distinction in the underlying phenomena (Watson 2010; Turnbull 2015). In particular, Watson (2010) proposed that contextual probability affects listener-oriented processes, whereas discourse mention affects talker-oriented processes. Bell et al. (2009) similarly concluded that predictability effects reflect a range of sources in the speech production process, including lexical access, lexical and phonological encoding, and articulatory planning.

Identifying these kinds of similarities and differences among predictability effects requires the consideration of a broad range of predictability measures at different levels of linguistic structure, so that the contributions of different factors can be examined independently of the contributions of other factors. This type of analysis requires carefully designed corpora where the factors of interest are orthogonally controlled (e.g. the current study; Lam and Watson 2010; Turnbull 2015) or very large corpora where the factors of interest are sufficiently varied and independent to allow the contributions of different factors to be examined (e.g. Bell et al. 2009; Turnbull 2018). The identification of the independent contributions of different types of predictability is critical for our understanding of predictability effects, given that these measures are often correlated with one another, leading to potentially misleading results if other relevant factors are not controlled (see also Cohen Priva and Jaeger 2018). Moreover, it is essential to consider not only main effects, but also interactions among predictor variables, so that unexpected or null results can be interpreted, as in the frequency $\times$ style interaction in the current study.

The results of the three analyses also provide evidence that different ways of estimating contextual probability lead to different patterns of results. Although the lexical frequency, phonotactic probability, and style effects were consistent across analyses, the three measures of contextual probability each had a qualitatively different effect on vowel duration. The Fisher/Buckeye probability effect was the most robust of the three measures, revealing longer vowel duration for relatively more predictable words, albeit primarily for higher-frequency, first mention words in plain speech. The cloze probability effect was only observable in a five-way interaction, and the Google 1T probability effect was only observable in an interaction with phonotactic probability, discourse mention, and style. Moreover, unlike the cloze and Fisher/Buckeye probability effects, Google 1T probability did not interact with frequency. The qualitative differences in significance patterns across models are difficult to quantify, although they are clearly visible in Figures 2–4, and the model fits for the three analyses do not provide an obvious choice for selecting one model over the others: the BIC is smallest for the Fisher/Buckeye probability analysis, but the model pseudo-r^2s are indistinguishable (all $= 0.84$). The higher-order interactions also require independent replication to ensure that they are not simply noise (see also Foulkes et al. 2018).

The qualitative differences between the three analyses are nevertheless surprising given that the three measures of contextual probability are significantly positively correlated and are typically assumed to reflect the same underlying phenomenon. However, the correlations are only moderate (see Figure 1), providing evidence that these three measures may tap different aspects of local contextual predictability (Smith and Levy 2011; Tremblay and Tucker 2011). That is, each of these measures is one estimate of local contextual predictability, but none of them may perfectly align with talkers' assessment of predictability in the speech production context because they are related to the read stories in different ways. The cloze probability measure was based on isolated sentences extracted from the stories, whereas the talkers had access to the entire story and therefore may have had a different context for interpreting the target words, especially during the clear speech reading, when they had already read the story before. The Google 1T and Fisher/Buckeye probabilities were based on trigram models and therefore reflect an even more local context than the cloze probabilities. Furthermore, the Google 1T corpus is written, whereas the Fisher/Buckeye corpora are spoken. None of these three corpora may adequately reflect the genre of read stories that comprised the current data, and this mismatch between the measures and the materials may explain the measures' varied impact on the production of these stories.

Together, these analyses suggest that researchers should exercise caution both in selecting measures of predictability, particularly contextual probability, and in interpreting predictability effects, especially at the level of contextual or discourse effects. If we had considered only one of our three contextual probability measures, we may have confidently interpreted our results in the context of one of the three qualitatively different patterns we observed. Instead, the variation in our three analyses points to the need to carefully consider how our measures of predictability align with the speech material we are analyzing and whether our results should be generalized beyond a particular measure based on a particular corpus to a more general phenomenon. Careful attention to the specific estimates of predictability that lead to significant effects within and across studies will allow us to identify how predictability at different levels of linguistic structure shapes speech production. We have also focused exclusively on vowel duration in this study and may therefore have missed critically different effects of predictability in other acoustic dimensions, including vowel quality, f0, and/or intensity, or in the domain of prosodic structure, which may also affect acoustic variability in all of these dimensions (Lam and Watson 2010; Burdin and Clopper 2015; Turnbull 2015; Turnbull 2017). Together, these results underscore the need for better statistical and quantitative standards for assessing predictability effects in speech research as well as a more nuanced and complex theoretical understanding of predictability effects and their interactions.

Funding: National Science Foundation, Division of Behavioral and Cognitive Sciences, Funder Id: 10.13039/100000169, Grant Number: BCS-1056409.

Appendix

Table A1: Summary of the fixed effects in the model with cloze probability.

Factor	Estimate	t-value
Intercept	0.024	0.90
Frequency	−0.031	−1.66
Phonotactic probability	**−0.058**	**−3.08**
Cloze probability	0.003	0.21
Mention	−0.001	−0.14
Style	**0.058**	**4.71**
Frequency × Phonotactic probability	−0.013	−0.65
Frequency × Cloze probability	−0.005	−0.43
Frequency × Mention	−0.006	−0.77
Frequency × Style	**−0.008**	**−3.25**
Phonotactic probability × Cloze probability	−0.007	−0.60
Phonotactic probability × Mention	−0.001	−0.19
Phonotactic probability × Style	0.002	0.63
Cloze probability × Mention	0.001	0.11
Cloze probability × Style	0.001	0.60
Mention × Style	0.002	0.98
Frequency × Phonotactic probability × Cloze probability	0.000	−0.01
Frequency × Phonotactic probability × Mention	0.005	0.66
Frequency × Phonotactic probability × Style	−0.002	−0.60
Frequency × Cloze probability × Mention	**0.020**	**2.60**
Frequency × Cloze probability × Style	−0.003	−1.20
Frequency × Mention × Style	0.001	0.75
Phonotactic probability × Cloze probability × Mention	−0.008	−0.88
Phonotactic probability × Cloze probability × Style	**0.005**	**2.29**
Phonotactic probability × Mention × Style	0.000	0.27
Cloze probability × Mention × Style	0.002	1.40
Frequency × Phonotactic probability × Cloze probability × Mention	0.009	1.15
Frequency × Phonotactic probability × Cloze probability × Style	−0.004	−1.71
Frequency × Phonotactic probability × Mention × Style	0.002	1.37
Frequency × Cloze probability × Mention × Style	0.000	−0.01
Phonotactic probability × Cloze probability × Mention × Style	−0.003	−1.63
Frequency × Phonotactic probability × Cloze probability × Mention × Style	**−0.004**	**−2.31**

Significant effects ($|t| > 2$) are shown in bold.

Table A2: Summary of the fixed effects in the model with Google 1T probability.

Factor	Estimate	t-value
Intercept	0.027	1.01
Frequency	−0.031	−1.56
Phonotactic probability	**−0.055**	**−2.83**
Google 1T probability	0.001	0.10
Mention	−0.005	−0.65
Style	**0.059**	**4.81**
Frequency × Phonotactic probability	−0.009	−0.44
Frequency × Google 1T probability	−0.011	−0.88
Frequency × Mention	−0.005	−0.64
Frequency × Style	**−0.010**	**−3.68**
Phonotactic probability × Google 1T probability	0.001	0.10
Phonotactic probability × Mention	−0.003	−0.41
Phonotactic probability × Style	0.002	0.63
Google 1T probability × Mention	0.000	0.05
Google 1T probability × Style	0.004	1.84
Mention × Style	0.001	0.46
Frequency × Phonotactic probability × Google 1T probability	−0.010	−0.69
Frequency × Phonotactic probability × Mention	0.005	0.56
Frequency × Phonotactic probability × Style	−0.002	−0.73
Frequency × Google 1T probability × Mention	0.014	1.77
Frequency × Google 1T probability × Style	−0.004	−1.81
Frequency × Mention × Style	0.002	0.94
Phonotactic probability × Google 1T probability × Mention	−0.009	−0.86
Phonotactic probability × Google 1T probability × Style	0.001	0.60
Phonotactic probability × Mention × Style	0.000	−0.13
Google 1T probability × Mention × Style	0.001	0.35
Frequency × Phonotactic probability × Google 1T probability × Mention	0.006	0.67
Frequency × Phonotactic probability × Google 1T probability × Style	−0.003	−1.18
Frequency × Phonotactic probability × Mention × Style	**0.005**	**2.50**
Frequency × Google 1T probability × Mention × Style	0.002	1.00
Phonotactic probability × Google 1T probability × Mention × Style	**−0.007**	**−2.96**
Frequency × Phonotactic probability × Google 1T probability × Mention × Style	0.001	0.44

Significant effects ($|t| > 2$) are shown in bold.

Table A3: Summary of the fixed effects in the model with Fisher/Buckeye probability.

Factor	Estimate	t-value
Intercept	0.008	0.28
Frequency	−0.027	−1.28
Phonotactic probability	**−0.081**	**−3.80**
Fisher/Buckeye probability	0.001	0.07
Mention	−0.006	−0.83
Style	**0.061**	**4.92**
Frequency × Phonotactic probability	−0.025	−1.10
Frequency × Fisher/Buckeye probability	**0.030**	**2.07**
Frequency × Mention	−0.016	−1.75
Frequency × Style	**−0.008**	**−2.76**
Phonotactic probability × Fisher/Buckeye probability	0.016	0.96
Phonotactic probability × Mention	0.004	0.51
Phonotactic probability × Style	0.003	0.90
Fisher/Buckeye probability × Mention	0.015	1.64
Fisher/Buckeye probability × Style	0.000	0.17
Mention × Style	0.001	0.67
Frequency × Phonotactic probability × Fisher/Buckeye probability	**0.037**	**2.31**
Frequency × Phonotactic probability × Mention	−0.004	−0.37

Table A3 (continued)

Factor	Estimate	t-value
Frequency × Phonotactic probability × Style	−0.001	−0.27
Frequency × Fisher/Buckeye probability × Mention	**0.015**	**2.30**
Frequency × Fisher/Buckeye probability × Style	**−0.005**	**−2.42**
Frequency × Mention × Style	0.000	0.08
Phonotactic probability × Fisher/Buckeye probability × Mention	0.005	0.51
Phonotactic probability × Fisher/Buckeye probability × Style	0.001	0.22
Phonotactic probability × Mention × Style	−0.001	−0.30
Fisher/Buckeye probability × Mention × Style	0.002	0.95
Frequency × Phonotactic probability × Fisher/Buckeye probability × Mention	0.000	−0.07
Frequency × Phonotactic probability × Fisher/Buckeye probability × Style	−0.004	−1.66
Frequency × Phonotactic probability × Mention × Style	0.004	1.76
Frequency × Fisher/Buckeye probability × Mention × Style	0.000	0.02
Phonotactic probability × Fisher/Buckeye probability × Mention × Style	−0.003	−1.27
Frequency × Phonotactic probability × Fisher/Buckeye probability × Mention × Style	0.001	0.33

Significant effects ($|t| > 2$) are shown in bold.

References

Arnold, J. E. 2008. Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes* 23. 495–527.

Aylett, M. & A. 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. 31–56.

Baayen, R. H., D. J. Davidson & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.

Baese-Berk, M. & M. Goldrick. 2009. Mechanisms of interaction in speech production. *Language and Cognitive Processes* 24. 527–554.

Baker, R. E. & A. R. Bradlow. 2009. Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech* 52. 391–413.

Bard, E. G., A. H. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon & A. Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42. 1–22.

Bard, E. G., A. J. Lowe & G. T. M. Altmann. 1989. The effect of repetition on words in recorded dictations. *Proceedings of Eurospeech* 89. 573–576.

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

Brants, T. & A. Franz. 2006. Web 1T 5-gram version 1. Philadelphia, PA: Linguistic Data Consortium.

Burdin, R. S. & C. G. Clopper. 2015. Phonetic reduction, vowel duration, and prosodic structure. *Proceedings of the 18th International Congress of Phonetic Sciences*. 378.

Burdin, R. S., R. Turnbull & C. G. Clopper. 2015. Interactions among lexical and discourse characteristics in vowel production. *Proceedings of Meetings on Acoustics* 22. 060005.

Cieri, C., D. Graff, O. Kimball, D. Miller & K. Walker. 2005. *Fisher English Training Part 2*. Philadelphia, PA: Linguistic Data Consortium.

Clopper, C. G. & R. Turnbull. 2018. Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler & M. Zellers (eds.), *Rethinking reduction*, 25–72. Berlin: de Gruyter Mouton.

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

Cohen Priva, U. & F. Jaeger. 2018. The interdependence of frequency, predictability, and informativity. *Linguistics Vanguard* 4(S2).

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

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).

Fowler, C. A. 1988. Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech* 31. 307–319.

Fowler, C. A. & J. Housum. 1987. Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26. 489–504.

Fowler, C. A., E. T. Levy & J. M. Brown. 1997. Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language* 37. 24–40.
Hawker, T., M. Gardiner & A. Bennetts. 2007. Practical queries of a massive n-gram database. *Proceedings of the Australasian Language Technology Workshop*. 40–48.
Lam, T. Q. & D. G. Watson. 2010. Repetition is easy: Why repeated referents have reduced prominence. *Memory and Cognition* 38. 1137–1146.
Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer.
Nusbaum, H. C., D. B. Pisoni & C. K. Davis. 1984. Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception Progress Report no. 10*, 357–376. Bloomington, IN: Speech Research Laboratory, Indiana University.
Pitt, M. A., L. C. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume & E. Fosler-Lussier. 2007. *Buckeye corpus of conversational speech*. Columbus, OH: Department of Psychology, Ohio State University.
Seyfarth, S. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133. 140–155.
Shaw, J. A. & S. Kawahara. in press. Effects of surprisal and entropy on vowel duration in Japanese. *Language and Speech*.
Smith, N. J. & R. Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. 1637–1642.
Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. *Proceedings of Interspeech*. 901–904.
Stolcke, A., J. Zheng, W. Wang & V. Abrash. 2011. SRILM at sixteen: Update and outlook. *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.
Tomaschek, F., M. Wieling, D. Arnold & H. Baayen. 2013. Word frequency, vowel length and vowel quality in speech production: An EMA study of the importance of experience. *Proceedings of Interspeech*. 1302–1306.
Tomaschek, F., B. Tucker, M. Fasiolo & H. Baayen. 2018. Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard* 4(S2).
Tremblay, A. & B. V. Tucker. 2011. The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon* 6. 302–324.
Turnbull, R. J. 2015. *Assessing the listener-oriented account of predictability-based phonetic reduction.* Columbus, OH: Ohio State University doctoral dissertation.
Turnbull, R. 2017. The role of predictability in intonational variability. *Language and Speech* 60. 123–153.
Turnbull, R. 2018. Patterns of probabilistic segment deletion/reduction in English and Japanese. *Linguistics Vanguard* 4(S2).
Turnbull, R., R. S. Burdin, C. G. Clopper & J. Tonhauser. 2015. Contextual predictability and the prosodic realisation of focus: A cross-linguistic comparison. *Language, Cognition and Neuroscience* 30. 1061–1076.
Vitevitch, M. S. & P. A. Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40. 374–408.
Vitevitch, M. S. & P. A. Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers* 36. 481–487.
Watson, D. G. 2010. The many roads to prominence: Understanding emphasis in conversation. *Psychology of Learning and Motivation* 52. 163–183.
Yuan, J. & M. Liberman. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics ’08*. 5687–5690.