Speech Communication 171 (2025) 103239

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# The role of informativity and frequency in shaping word durations in English and in Polish

Kamil Kaźmierski

*Adam Mickiewicz University, Poznań, Poland*

ARTICLE INFO

*Keywords:*
Informativity
Frequency
Temporal reduction
Speech corpus

ABSTRACT

Overall lexical frequency has long been known to play a role in sound change. Specifically, lexical frequency is negatively correlated with phonetic duration, and as such can be seen as a driver of diachronic reduction processes. However, recent findings suggest that it is the frequency of occurrence in a phonetic environment that favors a particular type of sound change, rather than overall lexical frequency, that shapes phonetic forms. For temporal reduction, Seyfarth (2014) shows that words that have a high frequency of occurrence in predictable contexts (*low informativity* words) are more temporally reduced than words that have a lower frequency of occurrence in predictable contexts (*high informativity* words). In this paper, I replicate Seyfarth's (2014) finding using another corpus of unscripted English — the Nationwide Speech Project corpus (Clopper and Pisoni, 2006), as well as using a corpus of another language, Polish — the Greater Poland Spoken Corpus (Kaźmierski et al., 2019; Kul et al., 2019). In both cases, informativity is included as a predictor of theoretical interest in mixed-effects linear regression models of word durations. Informativity, i. e. the frequency of occurrence in low-predictability contexts is shown to have a statistically significant effect on word durations in both English and Polish. Extending the analysis beyond a replication of Seyfarth (2014), a comparison of the effect of informativity and overall lexical frequency shows that the effect of informativity is somewhat weaker in Polish than in English, lending some support to the notion that morphologically rich languages are less sensitive to contextual predictability.

## 1. Introduction

The negative correlation between lexical frequency and word length has been known at least since Zipf (1949). Arguably, the shorter length of frequent words results from a phonologization of phonetic reduction of high-frequency words. The underlying mechanism could be speaker-driven, with repeated use resulting in articulatory reduction (Bybee, 2001). Alternatively, or perhaps complementarily, the mechanism could be driven by listeners: highly frequent forms are recovered more easily by listeners, and so do not need to be pronounced with a great deal of phonetic detail to be decoded (Lindblom, 1990). In either case, online adjustments in word durations might result in diachronic change through the reinterpretation of reduced forms as intended pronunciations.

Recently, effects of the frequency of occurrence in an environment that favors a particular phonetic realization, in contrast to overall lexical frequency, have come to light. Eddington and Channer (2010) show that English words ending in /t/ which are often followed by consonant-initial words have higher rates of glottaling than words which are followed by consonant-initial words less often, and that this effect holds even in prevocalic environment. Brown and Raymond (2012) show that the diachronic path that accounts for the present-day distribution of *f*-[f] vs. *h*-[∅] words in Spanish is best modeled when taking into account the frequency of occurrence of a given word after words beginning in non-high vowels. Similarly, Raymond and Brown (2012) show that for /s/ initial words in Spanish spoken in New Mexico, the probability of reduction of the word-initial /s/ to [h] or to ∅ is driven by the frequency of occurrence of the [s] initial word after words beginning with non-high vowels. Raymond et al. (2016) show that the probability of word-final t/d deletion in English is positively correlated with the frequency of occurrence of a word before consonant-initial words. Forrest (2017) found that for (ING) in English the effect of frequency of occurrence in a particular environment interacts with overall frequency. Frequent occurrence in environments favoring *-in* boosts the frequency effect, whereas frequent occurrence in environments favoring *-ing* dampens it. Bybee (2017) gives an overview of sound changes previously accounted for with regard to grammatical and lexical factors, and argues that they too can be captured with regard to phonetic context

*E-mail address:* kamil.kazmierski@amu.edu.pl.

https://doi.org/10.1016/j.specom.2025.103239
Received 22 May 2023; Received in revised form 19 March 2025; Accepted 4 April 2025
Available online 5 April 2025

only, when the notion of phonetic context is expanded to include a word's frequency of occurrence in an environment conducive to change. An effect of an environment conducive to change that is revisited in this paper is one found by Seyfarth (2014), who found that words which frequently occur in predictable contexts (low *informativity* words, term coined by Cohen Priva and Jurafsky (2008)) show shorter durations than words which frequently occur in unpredictable contexts (high *informativity* words). Seyfarth's study is an extension of Piantadosi et al.'s (2011) finding that informativity is correlated with word length in a sample of 11 languages, and that this correlation is stronger than a correlation of word length with frequency. As Piantadosi et al. (2011, p. 3526) posit, "the amount of information conveyed by a word should be linearly related to the amount of time it takes to produce". This, in turn, builds on Aylett's (1999) finding that speakers speak more slowly when their messages are more information rich.

The effects of frequency of occurrence in a particular phonetic context are challenging to models of speech production which assume abstract phonological representations and feed-forward speech production architecture, such as Levelt et al.'s (1999). Such models do incorporate overall lexical frequency effects. Temporal reduction can be construed as an online effect driven by higher activation levels of high-frequency lexical items. However, information regarding the typical environment in which a given word occurs is not available during online processing. A class of speech production models with access to such information are those based on rich storage, where multiple phonetically specified tokens of a word are available. The mechanism for the influence of occurrence in predictable contexts on phonetic duration could be as follows. The tokens of productions in high predictability contexts are shorter, and tokens of productions in low predictability contexts are longer due to online temporal reduction effects. During production, when all available tokens are drawn upon for a production plan, the frequency of occurrence in high versus low predictability contexts will play a role in shaping the duration of the output form. That is one possible way of accounting for the effects of frequency of occurrence in a phonetic environment favoring temporal reduction. Regardless of whether it is the right mechanism, a major modification of abstractionist feed-forward models is the least that would be called for, given that these effects are real. Testing the generalizability of the frequency in favoring environment effects is therefore warranted. The effect found by Seyfarth (2014) is tested in this paper. First, the effect is replicated on the same language, i. e. English, but on the basis of a different data set than in the source article. Consequently, the effect is replicated on another language, namely on Polish. Extending research findings beyond English, and beyond the Germanic language family is worthwhile. Polish has a much richer morphology than English, and Polish words are generally longer than English words. Koplenig et al. (2022) failed to find an effect of informativity on phonetic durations in Polish, which can be due to typological differences between English and Polish (Levshina, 2022). It is therefore conceivable that the effect of informativity will not hold for Polish. Finally, the relative weight of informativity versus lexical frequency as predictors of word durations is compared for the two languages.

## 2. Predictability

One simple and useful operationalization of the predictability of a given word in context is its transitional probability (Bush, 2001). The predictability of a word $w_i$ given word $w_{i-1}$ can be expressed as the ratio of the joined probability of $w_{i-1}$ and $w_i$ (bigram probability) to the probability of $w_{i-1}$, as summarized in Eq. (1). Frequencies of occurrence in a reference corpus are typically taken as estimates of probabilities.

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{1}$$

A limitation of this conditional probability measure is that it yields zero for any bigram not encountered in the reference corpus, thus

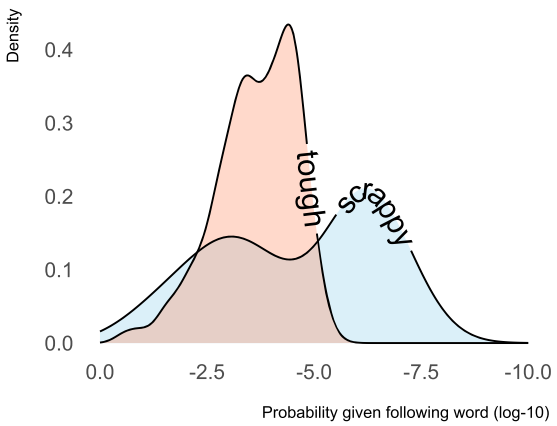


**Fig. 1.** A comparison of right-context predictability of scrappy and tough across all their contexts of occurrence given the language model constructed on the SUBTLEX-US (Brysbaert and New, 2009) corpus. scrappy has a bimodal distribution, with the higher of the two density peaks in the lower predictability region compared to the central tendency for tough.

treating all occurrences of novel bigrams as equally improbable. Additionally, conditional probability cannot be computed at all for any bigram for which $w_{i-1}$ does not occur in the reference corpus. A technique for alleviating these issues employed in Seyfarth (2014) is Modified Kneser-Ney smoothing (Chen and Goodman, 1998), described by Chen and Goodman (1998), and implemented in the SRILM Toolkit (Stolcke, 2002; 2011). Thanks to smoothing, non-zero probabilities are computed even in cases where one of the elements of the bigram is missing. In the present paper, the same method is used, albeit implemented with different tools. The *cmscu* package (Davis and Vinson, 2016) was used for training a smoothing function. While Seyfarth (2014) used the Fisher corpus for constructing the language model, here, large and freely available corpora were used instead: SUBTLEX-US (Brysbaert and New, 2009) for English and SUBTLEX-PL (Mandera et al., 2015) for Polish.

To illustrate the relationship between predictability and phonetic duration, compare the predictability of *home* in *nice home* and in *fortress-like home*. The predictability of *home* given *nice* in *nice home* is very low ($< 0.001$), since many other lexical items appear routinely after *nice* in English speech. Given this relatively low predictability, relatively little temporal reduction of *home* in *nice home* is expected. Conversely, the predictability of *home* given *fortress-like* in *fortress-like home* is much higher (0.412). Consequently, more temporal reduction of *home* in *fortress-like home* is predicted. Both these examples consider left-context predictability, i. e. the predictability of a word given the word that precedes it. In addition to left-context predictability, right-context predictability has also been shown to exert its influence on temporal reduction (Lohmann, 2017; Seyfarth, 2014). To illustrate, the predictability of *home* given *course* in *home course* is relatively low ($< 0.001$), whereas the predictability of *home* given *furnishings* in *home furnishings* is relatively high (0.256). Consequently, relatively little reduction of *home* in *home course* is predicted, and more reduction of *home* in *home furnishings* is predicted. Right-context predictability has been shown to exert a stronger influence on phonetic duration than left-context predictability (Lohmann, 2017; Seyfarth, 2014).

## 3. Informativity

There is variation across lexical items as to how often they appear in predictable contexts. For quantifying how *un*predictable a word is overall, Seyfarth (2014) invokes *informativity*. It is a metric proposed by Piantadosi et al. (2011) as 'average information content', and proposed to be an improvement on the role of lexical frequency put forward by Zipf. Using a large scale Google corpus as their data set, Piantadosi et al.

(2011) show that for 11 languages that they studied, informativity is more strongly correlated with orthographic word length than lexical frequency is.

Informativity is expressed by Eq. (2).

$$\frac{-1}{N}\sum_{i=1}^{N} logP(W = w \mid C = c_i) \tag{2}$$

A word's informativity is the negative weighted mean of smoothed log probabilities across all contexts. Informally, the more often a word occurs in low-predictability contexts, the higher its informativity. Conversely, the more often a word occurs in high-predictability contexts, the lower its informativity. To illustrate, Fig. 1 shows the distributions of right-context predictability of all occurrences of two words, *scrappy* and *tough* in the language model. The word *scrappy* is often less predictable from its right context than the word *tough*, and so it has higher informativity.

## 4. Informativity vs frequency

The robustness of the primacy of informativity over lexical frequency has since been questioned.

Care is needed to assess the effect of informativity versus frequency on word durations, as informativity and frequency are well correlated. Seyfarth (2014) notes this correlation in his data ($r = -0.58$). Finding a stronger correlation of informativity ($r = 0.52$) than of frequency ($r = -0.28$) with duration, he sees frequency as a "suppressor variable" (Seyfarth, 2014: 147), and excludes frequency from the final models he reports. Including both frequency and informativity in the same regression model would lead to paradoxical result of lexical frequency being positively correlated with word durations. Such paradoxes may happen when two strongly correlated predictors, one of which is more strongly correlated with the response than the other, are included in the same regression model (Seyfarth, 2014: 147; cf. Friedman and Wall, 2005).

Meylan and Griffiths (2021) found that Piantadosi et al.'s (2011) result, with informativity trumping frequency in all 11 languages studied, fades when different - arguably, better - methodological decisions are introduced in the analysis. Informativity remains better correlated with word length in English, but for other languages there is no difference between frequency and informativity. For Polish, the effect is reversed: frequency beats informativity. As a possible reason for the across-languages difference, they note morphology, including case-marking: while English has two cases (Quirk, Greenbaum, Leech, and Svartvik, 1985), Polish has 6+ cases (Bielec, 2012). With regard to case-marking, Polish is the polar opposite of English in the sample, making it an interesting testing ground for the role of informativity vs frequency in word durations, in addition to word length.

Following up on Meylan and Griffiths (2021) finding and recommendations, Koplenig et al. (2022) used a large corpus of German to train 3-gram models to test the role of frequency vs. informativity on word length for lists of most frequent word-form types of increasing sizes (from 12.5k to 800k most frequent types). They find only partial support for Piantadosi et al.'s finding: whether frequency or informativity is more strongly correlated with word length depends on the size of the list of types selected. Additionally, Koplenig et al. (2022) computed Spearman correlations with average phonetic durations (obtained from a separate speech corpus): consistently, frequency is correlated with phonetic word durations, while informativity is not: this is in contrast to Seyfarth (2014).

Levshina (2022) hypothesized that there are "'frequency-sensitive' languages and 'informativity-sensitive' languages', and that belonging to one group or the other is patterned, belonging in large part on the morphological richness of a given language. She studied the relationship between informativity (left-context and right-context) and lexical frequency on the one hand and orthographic word length on the other in seven typologically diverse languages. She argues that higher type-token ratio in morphologically rich languages causes data sparseness, which makes neighboring words less reliable predictors. Indeed, she found that the languages in her sample with highest type-token ratios (Finnish and Hungarian) showed a dominant role of frequency compared to the language with lowest type-token ratio (Indonesian), which showed a dominant role of informativity. Levshina also hypothesized that relatively free word order, which is often a feature of morphologically rich languages, makes words less predictable from their context. This hypothesis, however, was not supported in her data set. For Russian, typologically the closest language of her sample to Polish, Levshina (2022) found left-context informativity to have a stronger effect than right-context informativity.

In light of findings and arguments of Meylan and Griffiths' (2021), Koplenig et al. (2022), Levshina (2022), Polish, in contrast to English, is expected to show a dominant role of frequency rather than of informativity in driving temporal reduction of words.

## 5. Goals

In the target paper, Seyfarth (2014) investigated the influence of informativity on temporal reduction in two unscripted speech corpora: Buckeye (Pitt et al., 2007) and Switchboard-1 Release 2 (Calhoun et al., 2009; Godfrey and Holliman, 1997). Predictability and informativity were estimated based on the Fisher Part 2 corpus (Cieri et al., 2005). The findings were that predictability, both left-context and right-context, lead to more reduction, in both corpora. For informativity, higher right-context informativity was correlated with less reduction in both corpora, and higher left-context informativity was correlated with less reduction in the Switchboard corpus. To test the robustness of this finding, I attempted to replicate Seyfarth's (2014) finding on the same language (English) using different data sources than the target article: Nationwide Speech Project corpus as a source of phonetic duration data and SUBTLEX-US as a reference corpus for computing predictability and informativity metrics. Given contrasting predictions for Polish, I have extended the same design to that language as well, using the Greater Poland Spoken Corpus (Kaźmierski et al., 2019; Kul et al., 2019) as source of phonetic data, and SUBTLEX-PL as a reference corpus for computing predictability and informativity metrics.

## 6. Study 1: English

### 6.1. Data

For English, the Nationwide Speech Project (Clopper and Pisoni, 2006) corpus was used.[1] It contains speech of 60 speakers, 10 from each of the 6 dialect regions of American English (Labov et al., 2006), 5 female and 5 male speakers for each region. A number of speech styles were elicited from each speaker. For the present investigation, only unscripted speech is relevant. Therefore, only one speech style, 5-minute-long interviews, were used. These include 30,775 word-form tokens, and 3052 word-form types. The transcripts provided with the corpus were manually aligned on the breath-group level in Praat (Boersma, 2018), and force-aligned with the audio on the word and phoneme levels with FAVE-align (Rosenfelder et al., 2014). Further annotation and querying was done in LaBB-CAT (Fromont and Hay, 2012). All data transformation was done using the *dplyr* package in R (R Core Team, 2022).

A Modified Kneser-Ney smoothing function was trained on the SUBTLEX-US corpus using the *cmscu* package (Davis and Vinson, 2016) in R. Consequently, a complete list of bigrams from SUBTLEX-US was annotated with both left-context and right-context smoothed

[1] Data sets and scripts for fitting the models reported on here are available in a public OSF repository at https://osf.io/w52cd/

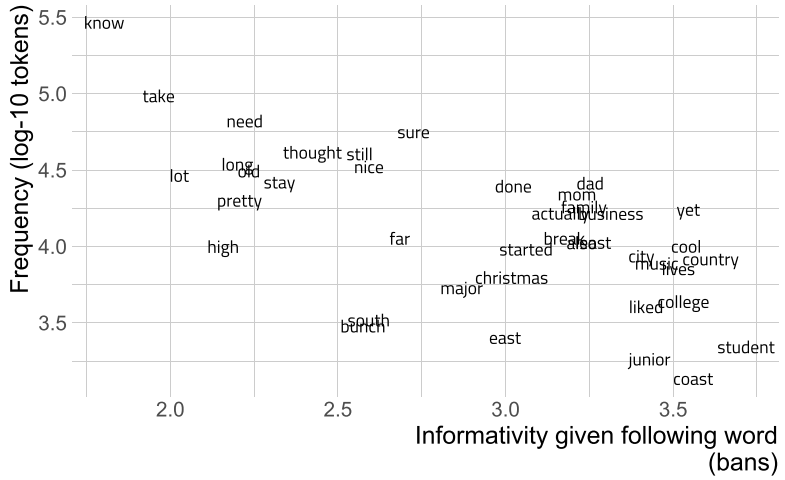


**Fig. 2.** The correlation of right-context informativity (in bans) and log10 lexical frequency for 40 randomly sampled words with token count above 10 $r = -0.64$.

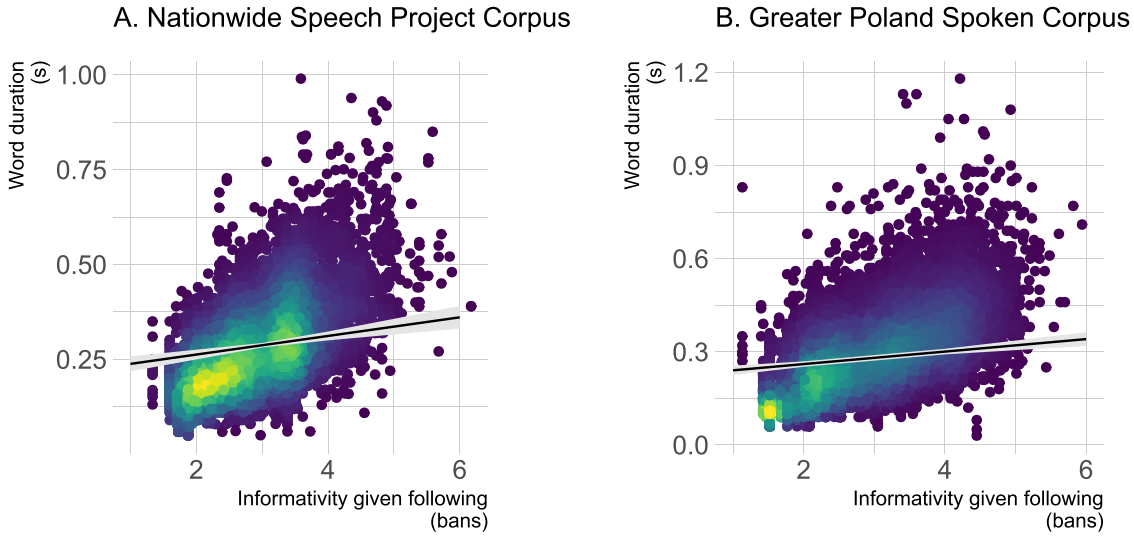


**Fig. 3.** Partial-effect plots for the significant informativity predictors in the two corpora: right-hand informativity in each case. Points show individual data points, with densities indicated by color: brighter spots show higher concentrations of data points.

conditional probabilities. This list was then used to compute left-context and right-context informativity for each word using the formula from Eq. (2). Its implementation was made feasible thanks to the *data.table* R package (Dowle and Srinivasan, 2022).

Each word retrieved from the NSP corpus was annotated with its left-context and right context smoothed conditional probability (=local predictability), as well as with its left-context and right-context informativity (=global *un*predictability). To account for other factors influencing word durations, the following annotations were also added to serve as control variables:

**Part of speech**: Words were tagged with dominant part of speech tags from SUBTLEX-US (Brysbaert and New, 2009). Only words tagged as Nouns, Adjectives, Adverbs and Verbs were included. Part of speech was a treatment-coded categorical predictor, with *Noun* as the reference level.

**Orthographic length**: Number of letters in the word, a numerical predictor, log-transformed, centered and standardized.

**Number of syllables**: Number of syllables in the word, retrieved from CELEX (Baayen et al., 1995), a numerical predictor, log-transformed and standardized.

**Dialect area**: It has been suggested that different dialects of American English are associated with different speech rates. It is therefore conceivable that there are differences across the dialects with regard to temporal reduction as well. To control for this, dialect area, specifying one of the six dialect areas: Mid-Atlantic, Midland, New England, North, South, West, was entered as a sum-coded categorical predictor.

**Average per speaker articulation rate**: Average number of syllables per second for each speaker, log-transformed, centered and standardized.

**Per utterance per speaker speech rate deviation**: Reduction patterns may be linked to local deviations from a speaker's average speaking rate. This was captured by the rate deviation numerical predictor expressed as a difference, in syllables per second, between the articulation rate of the utterance containing the target bigram and that speaker's average articulation rate (cf. Tanner et al., 2017); log-transformed, centered and standardized.

Durations of all words in the corpus were extracted. The following words were discarded: all words which do not belong to one of the content-word categories, i.e. noun, verb, adjective, adverb; as well as all words with missing left-hand and/or right-hand context. In the end, 7152 word tokens (853 word types) were kept for analysis.

Both left-context informativity (that is informativity of $w_i$ given $w_{i-1}$) and right-context informativity (that is informativity of $w_{i-1}$ given $w_i$) of the words present in the final data set are negatively correlated with log-

**Table 1**
Summary of mixed-effects linear regression model of word durations in English (the Nationwide Speech Project corpus).

| Term | Est. | CI | t | p-value |
|---|---|---|---|---|
| (Intercept) | 0.256 | 0.227–0.286 | 17.131 | **<0.001** |
| Local predictability given previous | −0.001 | −0.004–0.001 | −0.952 | 0.341 |
| Informativity given previous | 0.002 | −0.008 – 0.012 | 0.444 | 0.657 |
| Local predictability given following | −0.029 | −0.031 – −0.026 | −20.253 | **<0.001** |
| Informativity given following | 0.025 | 0.015 – 0.034 | 5.004 | **<0.001** |
| POS [Adjective] | −0.013 | −0.029 – 0.002 | −1.745 | 0.081 |
| POS [Adverb] | −0.036 | −0.051 – −0.021 | −4.683 | **<0.001** |
| POS [Verb] | −0.031 | −0.042 – −0.020 | −5.486 | **<0.001** |
| Orthographic length | 0.035 | 0.028 – 0.042 | 9.489 | **<0.001** |
| Dialect [Mid-Atlantic] | −0.012 | −0.019 – −0.004 | −3.103 | **0.002** |
| Dialect [Midland] | 0.007 | −0.000 – 0.015 | 1.835 | 0.067 |
| Dialect [New England] | −0.003 | −0.010 – 0.004 | −0.805 | 0.421 |
| Dialect [North] | 0.005 | −0.002 – 0.012 | 1.351 | 0.177 |
| Dialect [South] | 0.004 | −0.003 – 0.011 | 1.201 | 0.230 |
| Average articulation rate | −0.015 | −0.018 – −0.011 | −8.259 | **<0.001** |
| Articulation rate deviation | −0.026 | −0.028 – −0.024 | −28.619 | **<0.001** |
| Number of syllables | 0.027 | 0.021 – 0.034 | 8.827 | **<0.001** |

10 lexical frequency ($r = -0.9, p < 0.001; r = -0.81, p < 0.001$ respectively).

The relationship between right-hand informativity and lexical frequency for a random subset of 40 words is illustrated in Fig. 2.

A mixed-effects linear regression model was fit to the data with word duration as the response variable, and the factors listed in the preceding paragraph as predictor variables. Additionally, to account for word-level and speaker-level idiosyncrasies in word durations, by-word and by-speaker random intercepts were included. Further, to account for differing effects of the predictor variables of greatest theoretical interest, i.e. right-context and left-context informativity, by-speaker random slopes for these two variables were also included. The model was fit with the *lme4* package (Bates et al., 2015) in R, with *p*-values being calculated with the Satterthwaite method, as implemented in the *lmerTest* package (Kuznetsova et al., 2017).

### 6.2. Results

Due to initial convergence issues, the default optimizer was replaced with "optimx", from the R package of the same name (Nash and Varadhan, 2011). The model converged with a Marginal $R^2$ of 0.564 and Conditional $R^2$ of 0.704. Of the predictors of theoretical interest, left-context local predictability ($\widehat{\beta} = -0.001$) and left-context informativity ($\widehat{\beta} = 0.002$) did not turn out to be significant ($p = 0.341$ and $p = 0.657$, respectively). Conversely, right-context local predictability ($\widehat{\beta} = -0.029$, $p < 0.001$) and right-context informativity ($\widehat{\beta} = 0.025$, $p < 0.001$) did turn out to be significant.[2] The effect of the significant right-context informativity predictor is visualized in Fig. 3, Panel A.

For part of speech, the first of the control variables in the model, the difference between adjectives and nouns was not significant ($\widehat{\beta} = -0.013$, $p = 0.081$), while the differences between adverbs and nouns ($\widehat{\beta} = -0.036$, $p < 0.001$), as well as between verbs and nouns ($\widehat{\beta} = -0.031, p < 0.001$), were. Additionally, an omnibus test revealed part of speech to be significant overall ($\chi^2(3) = 37.74$, $p < 0.001$). Continuing with control variables, orthographic length ($\widehat{\beta} = 0.035$), average articulation rate ($\widehat{\beta} = -0.015$), articulation rate deviation ($\widehat{\beta} = -0.026$), and number of syllables ($\widehat{\beta} = 0.027$) all turned out to be statistically significant ($p < 0.001$). The sum-coded dialect predictor showed only one significant difference from the mean, for the Mid-Atlantic area ($b = -0.015$, $p = 0.002$), while an omnibus test did show dialect to be a significant predictor overall ($\chi^2(5) = 20.19, p = 0.001$). All model terms are summarized in Table 1.

### 6.3. Study 2: Polish

Levshina hypothesizes that in the higher type-token ratio of morphologically rich languages can result in the sparsity of data for contextual predictability. To verify the difference in type-token ratios between English and Polish, for comparability with Levshina, the following computations were performed. From each corpus, 1000 samples of 1000,000 tokens each were randomly drawn. Using these samples, 1000 type-token ratios per corpus (SUBTLEX-US for English and SUBTLEX-PL for Polish) were computed. Word forms with at least three occurrences in the corpus were included. For English, the thus calculated mean type-token ratio is 0.0245 (95 % of means between 0.0243 and 0.0246). For Polish, the mean type-token ratio is 0.0736 (95 % of means between 0.0733 and 0.0739). The higher type-token ratio in Polish compared to English supports the notion that data for computing contextual predictability is sparser in Polish than in English.

### 6.4. Data

For Polish, the Greater Poland Spoken Corpus (Kaźmierski et al., 2019; Kul et al., 2019) was used. The corpus contains phonemically annotated interview speech of 64 speakers (51 self-identified as female,

[2] A Reviewer, self-identified as Vsevolod Kapatsinski, warns that there is a risk of the effect of informativity being an artifact of partial pooling, given that the local predictability estimates are noisy and might be better approximated, particularly in rare contexts, by the addition of informativity estimates (i.e., by making use of predictability in other contexts). The Reviewer suggests modeling a series of simulated datasets, as outlined in Kapatsinski (2024) to verify this possibility. The datasets are generated such that word durations are not affected by informativity at all. I performed 1,000 simulations for this model, as well as for the model reported on for Study 2. As the estimated effects of informativity in the models fit to the simulated data sets are smaller than those in the original models, there is reason to believe that the effects of informativity reported here are not artifacts. Given that none of the 1,000 estimates of the models fit to the simulated datasets are as large as the actual estimates (for either model), the probability of obtaining such large effects of informativity is < .001. It is likely, though, that the effect sizes reported here may be overestimated: a little more than a half of the effect size might be due to partial pooling. The implementation and results of the simulations are available as scripts and plots in the OSF repository.

13 as male). Two corpus creators acted as interviewers in all interviews, and some interviewees were interviewed on their own, and some in pairs. Both read speech and interview speech was elicited from each speaker. For the present investigation, only unscripted speech is relevant. Therefore, only interview speech was used. 15 min long stretches of selected interviews are accompanied by orthographic transcriptions. These have been force-aligned in Labb-CAT (Fromont and Hay, 2012), by training HTK models (Young et al., 2006) for each speaker separately, with the help of a pronunciation dictionary compiled by the author. It is these 15 min-long stretches that lend themselves to querying and measurement, so they formed the basis for the present analysis. They contain 68,326 word-form tokens and 8378 types.

In most respects, the study on Polish paralleled the study on English described above. One variable present in the analysis of English that is missing from the analysis of Polish is dialect. While the English speakers of the NSP corpus come from six dialect areas of the United States, all speakers of the GPSC come from the same dialect area, Greater Poland.

A Modified Kneser-Ney smoothing function was trained on the SUBTLEX-PL corpus using the *cmscu* package (Davis and Vinson, 2016) in R. Since full sentences are not available for the SUBTLEX-PL corpus, the training was conducted using a pre-compiled list of bigrams. Consistent with Meylan and Griffith's (2021) recommendation, UTF-8 encoding was used to maintain a full range of orthograpic distinctions made in Polish, e.g. *widzą* 'they see' vs. *widza* 'spectator's'. Consequently, a complete list of bigrams from SUBTLEX-PL was annotated with both left-context and right-context smoothed conditional probabilities. This list was then used to compute left-context and right-context informativity for each word using the formula from Eq. (2), with the help of the *data.table* R package (Dowle and Srinivasan, 2022).

Each word retrieved from the GPSC was annotated with its left-context and right context smoothed conditional probability (=local predictability), as well as with its left-context and right-context informativity (=global *un*predictability). To account for other factors influencing word durations, the following annotations were also added to serve as control variables:

**Part of speech**: Words were tagged with dominant part of speech tags from SUBTLEX-PL (Mandera et al., 2015). Only words tagged as Nouns, Adjectives, Adverbs and Verbs were included. Part of speech was a treatment-coded categorical predictor, with *Noun* as the reference level.

**Orthographic length**: Number of letters in the word, a numerical predictor, log-transformed, centered and standardized.

**Number of syllables**: Number of syllables in the word was derived automatically from phonemic transcriptions prepared for the pronunciation dictionary compiled by the author for the purposes of force-alignment of the GPSC. It is a numerical predictor, log-transformed and standardized.

**Average per speaker articulation rate**: Average number of syllables per second for each speaker, log-transformed, centered and standardized.

**Per utterance per speaker speech rate deviation**: A numerical predictor expressed as a difference, in syllables per second, between the articulation rate of the utterance containing the target bigram and that speaker's average articulation rate; log-transformed, centered and standardized.

Durations of all words in the corpus were extracted. The following words were discarded: all words which do not belong to one of the content-word categories, i.e. noun, verb, adjective, adverb; as well as all words with missing left-hand and/or right-hand context. In the end, 10,589 word-form tokens (1712 word-form types) were kept for analysis.

A mixed-effects linear regression model was fit to the data with word duration as the response variable, and the factors listed in the preceding paragraph as predictor variables. Additionally, to account for word-level and speaker-level idiosyncracies in word durations, by-word and by-speaker random intercepts were included. Further, to account for differing effects of the predictor variables of greatest theoretical interest, i.e. right-context and left-context informativity, by-speaker random slopes for these two variables were also included. The model was fit with the *lme4* package (Bates et al., 2015) in R, with *p*-values being calculated with the Satterthwaite method, as implemented in the *lmerTest* package (Kuznetsova et al., 2017).

**Table 2**
Summary of mixed-effects linear regression model of word durations in Polish (the Greater Poland Spoken Corpus).

| Term | Est. | CI | t | p-value |
|---|---|---|---|---|
| (Intercept) | 0.251 | 0.227 – 0.275 | 20.612 | **<0.001** |
| Local predictability given previous | −0.007 | −0.009 – −0.004 | −5.299 | **<0.001** |
| Informativity given previous | −0.007 | −0.015 – 0.000 | −1.881 | 0.060 |
| Local predictability given following | −0.016 | −0.018 – −0.013 | −12.281 | **<0.001** |
| Informativity given following | 0.020 | 0.013 – 0.027 | 5.507 | **<0.001** |
| POS [Adjective] | −0.001 | −0.012 – 0.011 | −0.112 | 0.911 |
| POS [Adverb] | −0.001 | −0.015 – 0.014 | −0.132 | 0.895 |
| POS [Verb] | −0.020 | −0.028 – −0.011 | −4.349 | **<0.001** |
| Orthographic length | 0.074 | 0.066 – 0.082 | 18.011 | **<0.001** |
| Average articulation rate | −0.019 | −0.021 – −0.017 | −20.167 | **<0.001** |
| Articulation rate deviation | −0.023 | −0.024 – −0.021 | −30.546 | **<0.001** |
| Number of syllables | 0.040 | 0.033 – 0.046 | 12.097 | **<0.001** |

### 6.5. Results

Similarly to the first model, no convergence could be obtained with the default optimizer. In this case, the "bobyqa" optimizer allowed the model to converge. Marginal $R^2$ was 0.703, and conditional $R^2$ was 0.804. Of the predictors of theoretical interest, left context local predictability ($\widehat{\beta} = -0.007$), right context local predictability ($\widehat{\beta} = -0.016$) and right-context informativity ($\widehat{\beta} = 0.02$) turned out to be significant ($p < 0.001$). Only left-context informativity did not come out as significant ($\hat{\beta} = -0.007$, $p = 0.06$). The effect of the significant right-context informativity predictor is visualized in Fig. 3, Panel B.

The first of the control variables, part of speech, showed a significant difference between verbs and nouns ($\widehat{\beta} = -0.02$, $p < 0.001$), but did not show a significant difference between adjectives and nouns ($\widehat{\beta} = -0.001$, $p = 0.911$), or adverbs and nouns ($\widehat{\beta} = -0.001$, $p = 0.895$). An additional omnibus test showed the part of speech predictor to be significant overall ($\chi^2(3) = 27.1$, $p < 0.001$). All of the remaining control predictors, that is orthographic length ($\widehat{\beta} = 0.074$), average articulation rate ($\widehat{\beta} = -0.019$), articulation rate deviation ($\widehat{\beta} = -0.023$), and number of syllables ($\widehat{\beta} = 0.04$) were significant ($p < 0.001$). All model terms are summarized in Table 2.

## 7. Discussion

The results of Seyfarth (2014) have been largely replicated on an additional English data set. Seyfarth (2014) found a significant effect of right-context informativity in both the Buckeye corpus and the Switchboard corpus, and additionally an effect of left-context informativity (though with a smaller effect size) in the Switchboard corpus. The present replication found an effect of right-context informativity only. In this study, just like in the target study, higher informativity (i.e. higher overall **un**predictability) is associated with longer word durations.

This effect, going in the same direction, has additionally been extended to another language — Polish. This extension may be seen as

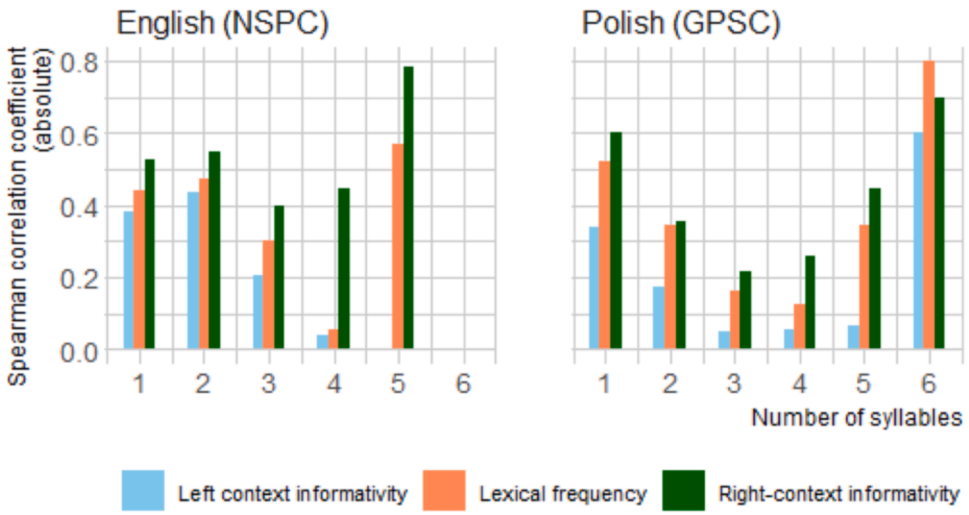


**Fig. 4.** Spearman correlation coefficients (absolute values) of left-context informativity, lexical frequency, and right-context informativity with mean word durations in English (the Nationwide Speech Project corpus) and Polish (the Greater Poland Spoken Corpus) broken down by number of syllables.

surprising, given Koplenig et al.'s (2022) failure to find an effect of informativity on phonetic durations in Polish, hypothesized to stem from typological differences between Polish and English (Levshina, 2022).

Since informativity can be seen as a case of frequency of occurrence in an environment favoring particular phonetic shapes (here: frequency of occurrence in low predictability contexts, favoring little temporal reduction), this article has provides support for such effects more broadly. The effect of right-hand informativity for word durations in English, and, to an extent, the same effect in Polish, poses a challenge for models of phonological storage and processing which assume abstract representations devoid of phonetic detail. Since phonological models able to account for these effects must assuming access to stored phonetically-rich representations, the present article indirectly furnishes support for rich-storage models.

## 8. Informativity vs. lexical frequency

Given strong correlations of the informativity measures and lexical frequency, they could not be entered in the same regression models. Especially for English, the same rationale as that applied by Seyfarth (2014) holds: given that both informativity and lexical frequency are strongly correlated with word durations, lexical frequency can be seen as a suppressor variable.

To try to compare the relative influence of the two, one could compare the correlation coefficients of each of the three metrics: left-context informativity, right-context informativity, and lexical frequency, with mean word durations.

In English, Spearman's correlation coefficients show that the association between right-context informativity and word durations ($r = 0.6$) is the strongest, followed by lexical frequency ($r = -0.58$) and left-context informativity ($r = 0.54$) .[3] The difference between $r = 0.6$ and $r = 0.58$ is not statistically significant, the difference between $r = 0.58$ and $r = 0.54$ is. In Polish, Spearman's correlation coefficients show that lexical frequency ($r = -0.46$) and right-context informativity (0.46) are correlated with mean word durations to similar degrees. Left-context informativity shows a weaker correlation ($r = 0.32$). The difference between $r = -0.46$ and $r = 0.46$ is not statistically significant, the difference between $r = 0.46$ and $r = 0.32$ is.

While left-context informativity (i.e. the average unpredictability of a word given the word that precedes it) shows weaker correlations with word durations, in both English and in Polish, than lexical frequency, the correlation of **right**-context informativity (i.e. the average unpredictability of a word given the word that *follows* it) with word durations is comparable to lexical frequency. This crucial difference has to be borne in mind when assessing the role of informativity versus lexical frequency: the choice of right-hand versus left-hand informativity will have an impact on the outcome of the comparison. Additionally, this curious finding that phonetic durations seem to be more affected by following words than be preceding words (cf. also Lohmann, 2017) demands further attention.

A serious limitation of the correlations computed above is that they disregard all the other variables affecting word durations which are accounted for in the regression models reported on above. Crucially, words of different lengths (measured in number of syllables, or in number letters) are bound to have different durations. One could therefore break down the correlations by a length measure. Fig. 4 shows the correlations of each of the three metrics with mean word durations broken down by the number of syllables in a word. For English, the order of the strengths of the correlations is the same regardless of number of syllables. In Polish, while right-context informativity has the strongest correlation with mean word durations for most syllable counts, its effect is comparable to that of lexical frequency for disyllabic words, and it trumps it for six-syllable words. Taken together, this comparison does point to a somewhat less decisive role of informativity for Polish word durations compared to English.

## 9. Orthographic word length vs. phonetic word durations

For English, informativity appears to shape phonetic duration (Seyfarth 2014, present replication) and as well as orthographic word length (Meylan and Griffiths, 2021; Piantadosi et al., 2011). For Polish, informativity may play a role in shaping phonetic durations (as indicated in the present study) but not word length (Meylan and Griffiths, 2021). The less consistent primacy of informativity over lexical frequency in Polish compared to English in speech production could be due to lower reliability of contextual predictability measures in languages with richer morphology, as stipulated by Levshina (2022). Thus, the present study provides some support for Levshina's postulated division into frequency-sensitive and informativity-sensitive languages.

[3] Statistical significance of the differences between correlations was tested by means of the *cocor* (Diedenhofen and Musch, 2015) R package, at an alpha level of 0.05.

## Funding

Funding: This work was supported by the Polish National Science Center [grant number UMO-2017/26/D/HS2/00027]

## CRediT authorship contribution statement

**Kamil Kaźmierski:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kamil Kazmierski reports financial support was provided by National Science Centre Poland.

## Acknowledgments

I would like to thank Jason K. Davis, the creator of the *cmscu* package, for making the code publicly available and for answering my queries regarding its implementation. I'd also like to thank Scott Seyfarth for answering my queries, and for his help with diagnosing an error in my computations of informativity in Polish specifically. Finally, I am grateful to the Reviewer, Vsevolod Kapatsinski for helping me diagnose my models for artifactual effects. Naturally, all remaining errors are my own.

## Data availability

A link to data and scripts is shared within the manuscript.

## References

Aylett, M., 1999. Stochastic suprasegmentals: relationships betweenredundancy, prosodic structure and syllabic duration. In: Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D., Bailey, A.C. (Eds.), Proceedings of the 14th International Congress of Phonetic Sciences. ICPhS Archive. Retrieved from. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0289.pdf.

Baayen, R.H., Piepenbrock, R., Gulikers, L., 1995. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, Philadelphia.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bielec, D., 2012. Polish an Essential Grammar: an Essential Grammar. Routledge.

Boersma, D., Paul & Weenink. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.41, retrieved 15 Aug 2018 from https://www.fon.hum.uva.nl/praat/manual/FAQ_How_to_cite_Praat.html.

Brown, E.L., Raymond, W.D., 2012. How discourse context shapes the lexicon: explaining the distribution of Spanish f- /h- words. Diachronica 29 (2), 139–161.

Brysbaert, M., New, B., 2009. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav. Res. Methods 41 (4), 977–990. https://doi.org/10.3758/brm.41.4.977. Retrieved from.

Bush, N., 2001. Frequency effects and word-boundary palatalization in English. In: Bybee, J.L., Hopper, P. (Eds.), Frequency and the Emergence of Linguistic Structure. John Benjamins, pp. 255–280.

Bybee, J., 2001. Phonology and Language Use. Cambridge University Press, Cambridge. https://doi.org/10.1017/cbo9780511612886.

Bybee, J., 2017. Grammatical and lexical factors in sound change: a usage-based approach. Lang. Var. Change 29 (03), 273–300. https://doi.org/10.1017/s0954394517000199.

Calhoun, S., Carletta, J., Jurafsky, D., Nissim, M., Ostendorf, M., & Zaenen, A., (2009). NXT switchboard annotations (Tech. rep.). Philadelphia: Linguistic DataConsortium.

Chen, S., & Goodman, J., (1998). An empirical study of smoothing techniques for language modeling. Tech. Rep. TR-10-98, Harvard University.

Cieri, C., Graff, D., Kimball, O., Miller, D., Walker, K., 2005. Fisher English training part 2. Linguistic Data Consortium. Linguistic Data Consortium. Philadelphia.

Clopper, C.G., Pisoni, D.B., 2006. The Nationwide Speech Project: a new corpus of American English dialects. Speech Commun. 48 (6), 633–644. https://doi.org/10.1016/j.specom.2005.09.010.

Cohen Priva, U., & Jurafsky, D., (2008). Phone information content influences phone duration. A poster Presented At Experimental and Theoretical Advances in Prosody; Cornell University: Ithaca, NY, USA. Retrieved from http://www.prosodylab.org/~chael/etap/abstracts/posters/cohen-priva.pdf.

Davis, J.K., & Vinson, D.W., (2016). Cmscu: a count-min-sketch with conservative update implementation for r. Retrieved from https://github.com/jasonkdavis/r-cmscu.

Diedenhofen, B., Musch, J., 2015. cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10 (4), e0121945. https://doi.org/10.1371/journal.pone.0121945.

Dowle, M., & Srinivasan, A., (2022). Data.table: extension of 'data.frame'. Retrieved from https://CRAN.R-project.org/package=data.table.

Eddington, D., Channer, C., 2010. American English has go? a lo? of glottal stops: social diffusion and linguistic motivation. Am. Speech 85 (3), 338–351. https://doi.org/10.1215/00031283-2010-019.

Forrest, J., 2017. The dynamic interaction between lexical and contextual frequency: a case study of (ING). Lang. Var. Change 29 (02), 129–156. https://doi.org/10.1017/s0954394517000072.

Friedman, L., Wall, M., 2005. Graphical views of suppression and multicollinearity in multiple linear regression. Am. Stat. 59 (2), 127–136. https://doi.org/10.1198/000313005x41337.

Fromont, R., Hay, J., 2012. LaBB-CAT: an annotation store. In: Proceedings of the Australasian Language Technology Association Workshop. Dunedin, New Zealand, pp. 113–117.

Godfrey, J.J., Holliman, E., 1997. Switchboard-1 release-2 (Tech. rep.). Philadelphia: Linguistic Data Consortium.

Kaźmierski, K., Kul, M., Zydorowicz, P., 2019. Educated Poznań speech 30 years later. Stud. Linguist. Univ. Iagell. Crac. 136 (4), 245–264. https://doi.org/10.4467/20834624SL.19.021.11314.

Kapatsinski, V., 2024. Informativity effects can be probability effects in disguise. In: Proceedings of the Paper Presented at the Corpus Phonetics Workshop, LabPhon 19. Seoul, South Korea. https://labphon.org/sites/default/files/labphon19/Papers/corpusphon/kapatsinski.pdf.

Koplenig, A., Kupietz, M., Wolfer, S., 2022. Testing the relationship between word length, frequency, and predictability based on the German reference corpus. Cogn. Sci. (6), 46. https://doi.org/10.1111/cogs.13090.

Kul, M., Zydorowicz, P., Kaźmierski, K., 2019. The Greater Poland spoken corpus: data collection, structure and application. In: Wrembel, M., Kiełkiewicz-Janowiak, A., Gąsiorowski, P. (Eds.), Approaches to the Study of Sound Structure and Speech. Interdisciplinary work in Honour of Katarzyna Dziubalska-Kołaczyk. Taylor & Francis, New York, pp. 198–212. https://doi.org/10.4324/9780429321757-15.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Labov, W., Ash, S., Boberg, C., 2006. The Atlas of North American English. Mouton de Gruyter. https://doi.org/10.1515/9783110167467.

Levelt, W.J.M., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Behav. Brain Sci. 22, 1–75. https://doi.org/10.1017/s0140525x99001776.

Levshina, N., 2022. Frequency, informativity and word length: insights from typologically diverse corpora. Entropy 24 (2), 280. https://doi.org/10.3390/e24020280.

Lindblom, B., 1990. In: Hardcastle, W., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer, Dordrecht, pp. 403–439.

Lohmann, A., 2017. Cut (n) and cut (v) are not homophones: lemma frequency affects the duration of nounverb conversion pairs. J. Linguist. 54 (4), 753–777. https://doi.org/10.1017/s0022226717000378.

Mandera, P., Keuleers, E., Wodniecka, Z., Brysbaert, M., 2015. SUBTLEX-PL: subtitle-based word frequency estimates for Polish. Behav. Res. Methods 47 (2), 471–483.

Meylan, S.C., Griffiths, T.L., 2021. The challenges of large-scale, web-based language datasets: word length and predictability revisited. Cogn. Sci. (6), 45. https://doi.org/10.1111/cogs.12983.

Nash, J.C., Varadhan, R., 2011. Unifying optimization algorithms to aid software system users: optimx for R. J. Stat. Softw. 43 (9), 1–14. Retrieved from. http://www.jstatsoft.org/v43/i09/.

Piantadosi, S.T., Tily, H., Gibson, E., 2011. Word lengths are optimized for efficient communication. Proc. Natl. Acad. Sci. 108 (9), 3526–3529. https://doi.org/10.1073/pnas.1012551108.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W.D., Hume, E., & Fosler-Lussier, E., (2007). Buckeye corpus of conversational speech. Columbus, OH. Retrieved from www.buckeyecorpus.osu.edu.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., 1985. A Comprehensive Grammar of the English language. Longman, p. 1779.

R Core Team. (2022). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

S. G. Raymond, W.D., Brown, E.L., 2012. Learning and processing. In: Divjak, D. (Ed.), Frequency Effects in Language Learning and Processing, Frequency Effects in Language Learning and Processing, 2. Mouton de Gruyter, Berlin, pp. 35–52.

Raymond, W.D., Brown, E.L., Healy, A.F., 2016. Cumulative context effects and variant lexical representations: word use and English final t/d deletion. Lang. Var. Change 28 (02), 175–202. https://doi.org/10.1017/s0954394516000041.

I. Rosenfelder, J. Fruehwald, K. Evanini, S. Seyfarth, K. Gorman, H. Prichard, & J. Yuan (2014). Fave 1.2.2. 10.5281/zenodo.9846.

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

Stolcke, A., 2002. SRILM — An extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing. Denver, CO.

Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011. SRILM at sixteen: update and outlook. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. Waikoloa, HI.

Tanner, J., Sonderegger, M., Wagner, M., 2017. Production planning and coronal stop deletion in spontaneous speech. Lab. Phonol. J. Assoc. Lab. Phonol. 8 (1), 1–39. https://doi.org/10.5334/labphon.96.

Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2006. The HTK Book Version 3.4. Cambridge University Press.

Zipf, G.K., 1949. Human Behavior and the Principle of Least effort: an Introduction to Human Ecology. Addison-Wesley, Cambridge, Mass.