# *Time* and *thyme* again: Connecting spoken word duration to models of the mental lexicon

Susanne Gahl, Harald Baayen

January 2022

Effects of lexical frequency have long informed theories of the mental lexicon. This study compares two approaches (localist spreading-activation vs. discriminative learning models integrating distributional semantics) by assessing regression models of spoken word duration of English homophones grounded in each. The regression models avoid assumption violations of earlier analyses of the same data set. We point out a major methodological flaw besetting corpus-based research on pronunciation. We also show that the relationship between a homophone's form and its semantics is predictive of its duration. Implications for theories of the mental lexicon are discussed.

**keywords** homophones, spoken word duration, mental lexicon, localist models, discriminative learning, distributional semantics, Gaussian location-scale generalized additive models

# 1 Introduction

Word frequency effects have for decades had the status of consensus findings (see e.g.Jurafsky 2003; Baayen et al. 2016 for overviews). One aspect of lexical frequency that has played a particularly important role in theories of language processing, variation, and historical change is its effect on spoken word duration (see e.g. Bybee, 2001, 1999, 2002; Pierrehumbert, 2002; Jurafsky et al., 2001; Bell et al., 2003; Warner, 2011; Jurafsky, 2003). The apparent ubiquity of word frequency effects may have created the impression that frequency is a pre-theoretical concept. But the very notion of lexical frequency as a property of words implicates a theoretical construct — the word as a representational unit — that is very far from being a matter of consensus. The current study puts the question of lexical frequency in the context of two fundamentally different types of models of the mental lexicon: A well-established class of LOCALIST MODELS of lexical access and retrieval (Dell, 1986b; Levelt et al., 1999; Schwartz et al., 2006) vs. a more recent class of models, henceforth DISCRIMINATIVE LEARNING (DL) models, in which words do not have an existence of their own as representational units within the lexicon itself (Baayen et al., 2019a; Chuang and Baayen, 2021; Heitmeier et al., 2021).

The consensus about the existence of word frequency effects may also have distracted somewhat from areas of disagreement. The relationship between frequency, predictability, and spoken word duration, for example, has received two seemingly opposite interpretations (see e.g. Seyfarth, 2014; Fox et al., 2015; Lohmann and Conwell, 2020): The phonetic reduction of frequent forms has been interpreted as part of a broader pattern of reduction due to high predictability (Jurafsky, 2003; Aylett and Turk, 2004). But high predictability of word forms (in morphological paradigms) has also been argued to be associated with phonetic strengthening and lengthening (Kuperman et al., 2007; Baese-Berk and Goldrick, 2009; Cohen, 2014).

Explanations for the shortening of frequent forms (e.g. Bybee, 2006; Bell et al., 2009) have appealed to many mechanisms at different steps in the processes that take place before and during word production. Many of these proposals are mutually compatible – spoken word duration undoubtedly reflects multiple factors – but they occasionally entail different predictions. The pronunciation of homophones is a case in point. Gahl (2008) argued that, if the shortening of frequent words solely reflected "late" stages of lexical production, such as phonological encoding and/or articulatory processes, a low-frequency word such as *thyme* should have the same duration as a high-frequency homophone twin *time* with identical phonological form, other things being equal. If, on the other hand, word duration also reflected the "earlier" steps of access and retrieval of word meanings, then duration should reflect each twin's specific frequency, again other things being equal. Consistent with the latter possibility, Gahl (2008) observed that spoken word durations of homophones in the Switchboard corpus of telephone conversations differed even when other factors were brought under statistical control, such that the more frequent member of a pair of homophones (e.g. *time*) tended to be shorter than the less frequent member of the pair (e.g. *thyme*).

The argument in Gahl (2008) was rooted in a localist model of lexical access and retrieval (Dell, 1986b). In a DL lexicon without words, there are no 'items' of which lexical frequency could be a property. How, then, can such a lexicon account for effects of word frequency?

We propose that the answer to that question holds the key to reconciling the seemingly contradictory effects of high predictability leading to shortening, as well as lengthening. Here, we argue that the DL framework offers a principled explanation for both effects: One effect reflects the strengthening that comes from better learning and can be estimated as the predictability of a form, given its meaning. The other effect arises when one steps outside of the lexicon and asks about the predictability of a target word given all words in an utterance.

A particularly appealing property of DL models is that they allow us to extend the investigation to effects of word meanings, beyond saying that homophones "differ in meaning". DL models integrate distributional semantics (Landauer and Dumais, 1997; Mitchell and Lapata, 2008; Mikolov et al., 2013). To preview our argument: We show that a measure quantifying the extent to which a homophone's form is supported by its meaning is a surprisingly strong predictor of its spoken duration. Before introducing the technical and conceptual properties of that measure, we must turn again to Gahl (2008)'s model of homophone duration.

The model in Gahl (2008) (henceforth "G2008", and a follow-up analysis, Gahl 2009, henceforth "G2009") did not go unchallenged. Lohmann (2018b) asserted that neither G2008 nor G2009 provided direct evidence of homophone pairs differing in duration depending on the target-specific frequency (e.g. of *time* being shorter than *thyme*). Instead, according to Lohmann (2018b), the models in G2008 and G2009 only showed that *time* was shorter than *sage*, not that *time* was shorter than *thyme*. This characterization is, we believe, misleading.[1] We do however concur with Lohmann (2018b) that G2008 and 2009 had many methodological shortcomings. For example, as pointed out in Lohmann (2018b), the fairly weak correlation between the duration of low-frequency and high-frequency homophones in the corpus need not indicate lemma-specific lexical characteristics, but may simply reflect uncertainty about duration estimates based on small numbers of tokens. Another issue with the regression models in Gahl (2008, 2009) (and Lohmann 2018b) is that they violated two modeling assumptions: One is the assumption of a linear relationship between predictors and spoken word duration. The second assumption concerns the relationship between frequency and variability in duration. As Lohmann (2018b) points out, the reliability of average duration as an estimate of a word's 'true' duration decreases with word frequency: The smaller the sample size, the higher the variance.

Complicating this picture is the fact that the variability in motor execution may in fact decrease as frequency increases (Tomaschek et al., 2020). If this is correct, then high variability in token duration of low frequency words may reflect high variability in motor execution, rather than (or in addition to) uncertainty about a true mean due to small sample size. The possible connection between frequency and duration variability poses a problem for any model assuming constant variance of model residuals, which includes those in G2008, G2009, and Lohmann (2018b). Here, we reanalyze the data analyzed previously in G2008, G2009, and Lohmann (2018b), but now using statistical models that do better justice to non-linear relations between predictors and acoustic duration, when present, and that are

[1] Briefly, Lohmann 2018b omits mentioning the role played by homophone twins in the models: G2008 asserts that the frequency of *time* is predictive of duration when controlling for the expected duration, given the target phonemes. The model in G2008 predicts the duration of *sage* depending on the frequency of *sage*, but only when the expected duration of a homophone (e.g. the first name *Saige*) is taken into account.

also better able to model uncertainty about the mean acoustic duration, by making use of Gaussian Location-Scale Generalized Additive Models (Wood, 2017).

A methodological goal of the current study is to point out an issue that has crept into research on pronunciation variation. Such studies have generally taken one of two approaches: The first aggregates information over word types. This is the approach taken in Gahl (2008), as well as in numerous analyses of controlled, balanced data from experiments (e.g. Mousikou and Rastle, 2015) and Wright (2004). The second approach considers word tokens. This approach is the tool of choice in most corpus-based analyses, e.g. Bell et al. (2003), Caselli et al. (2016), Tanner et al. (2017), Bell et al. (2009), Aylett and Turk (2004), Dilts (2013), Seyfarth (2014), Hay et al. (2015), and Kilbourn-Ceron et al. (2020). The appeal of that second approach is the richness of the information that can be considered, such as indexical information about the talker, or the words preceding or following the target in an utterance. Indeed, the rise of token-based analyses has gone hand in hand with developments in, and awareness of, statistical models that make such analyses feasible, possibly creating the impression that token-based models represent the gold standard, with type-based ones a less sophisticated or less informative alternative. We argue that token-based models introduce methodological pitfalls and biases that render them potentially problematic when used to draw inferences about lexical processing.

The main goal of the current study is broader, however. We compare the predictions for spoken word duration of two different theoretical approaches to the lexicon, by examining the ability of variables grounded in localist vs. DL models to predict the spoken word duration of homophones in the Switchboard corpus. The wide-spread consensus around the relationship between word frequency and spoken word duration makes this empirical domain a prime testing ground for asking how, and how well, a DL-model can account for effects attributed to word frequency in localist models. We begin by outlining the two conceptions of the mental lexicon under discussion, before turning to the choice of variables in our regression models of homophone duration, the specific implementation of the DL model, our data, and methods.

# 2 Background

## 2.1 Localist models

Most models of the lexicon, unsurprisingly, involve representations of words. The lexicon, in these models, constitutes a repository of discrete pieces of meaningful linguistic information, such as words and morphemes, connected to one another and to phonological information. Access, retrieval, encoding, and recognition are modeled by means of algorithms that produce predictions of degrees of accessibility, such as interactive activation (Dell, 1986b), spreading activation (Levelt et al., 1999), or Bayesian inference (e.g. Norris and McQueen, 2008). These models are localist, in the sense that they work with discrete linguistic units for phonemes, syllables, morphemes, words, and concepts.

Localist models have yielded numerous testable predictions. Most relevantly to the cur-

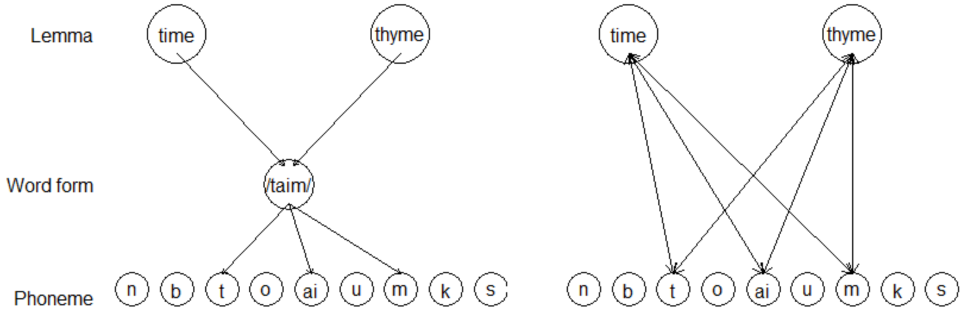

Figure 1: Two examples of localist models, based on Levelt et al., (1999) (left panel) and Dell (1990) (right panel)

.

rent discussion, the model depicted in the left-hand panel of Figure 1 leads to the prediction that homophones, by virtue of having identical word forms (or 'lexemes'), should show identical effects of word form frequency. Indeed, there is empirical support for that position: Low-frequency homophones of high-frequency words behave in some respects as though their own frequency were also high, an effect known as "frequency inheritance" (Jescheniak and Levelt 1994; Dell 1990). The model depicted in the right-hand panel of Figure 1 likewise predicts frequency inheritance (see Dell, 1990). In addition, as argued in Dell and Gordon, that model leads to the prediction that phonological neighbors should facilitate target retrieval in production, via feedback from phonemes shared by the target and its neighbors (see Chen and Mirman 2012; Middleton et al. 2015 for further discussion). The predicted effect of PND follows from the model architecture. As Dell (1988) points out, other model predictions follow from parameters that are specified by the researcher. Effects of lexical frequency, for example, can be modeled by means of differences in resting activation levels, with increasing frequency being modeled by setting higher resting activation levels. Numerous studies of neurotypical speakers and individuals with language disorders attest to the success and theoretical interest of such models (see e.g. Schwartz et al. 2006; Dell et al. 1999; Levelt et al. 1999; Foygel and Dell 2000).

## 2.2 Discriminative learning models

Spreading-activation models fall in a broad category of 'exposure-based' models, in which the mind is seen as an adaptive system whose behavior is shaped by experience. However, they are not learning models, but rather depict the result of a presumed process of learning and development, although an interest in learning was at the heart of many early connectionist approaches to the lexicon (see, e.g., Rumelhart and McClelland, 1986; MacWhinney and

Leinbach, 1991; Harm and Seidenberg, 2004). More recently, deep learning networks have been harnessed for the learning of morphology (see, e.g., Kirov and Cotterell, 2018; McCurdy et al., 2020).

The modeling framework that we make use of in this study, introduced in Baayen et al. (2019a) also makes use of error-driven learning, but uses a mathematically much simpler algorithm for learning mappings between form and meaning. Their model, to which we henceforth refer as the Discriminative Learning Model, or DLM, sets up several simple linear mappings (depending on modality) between modality-specific form representations and semantic representations. The model's semantic representations are inspired by distributional semantics: Words' meanings are represented by real-valued vectors, referred to as "embeddings" in the natural language processing literature. We will refer to these vectors as 'semantic vectors'. Semantic vectors can be derived from corpora, or, in the absence of resources, they can be simulated (see, e.g., Heitmeier et al., 2021; Chuang and Baayen, 2021, for detailed discussion). The model's form representations can be set up in many ways, for a systematic overview, see Heitmeier et al. (2021). Form representations can be derived from the acoustic signal (Shafaei-Bajestan et al., 2021), from visual bitmaps (Linke et al., 2017), or they can be set up at a more abstract level as binary vectors specifying which n-phones, n-grams, or n-syllables jointly define a word's form.

The mappings from form to meaning (for comprehension) and from meaning to form (for production) are implemented with simple input-to-output networks, without hidden layers, the connection weights of which are mathematically equivalent to the beta weights of multivariate multiple regression. Although the mappings are linear and therefore less powerful than the mappings that deep learning makes available, they work surprisingly well for modeling lexical processing. Mappings can also be derived incrementally, by training the networks word by word, using the update rules of Rescorla and Wagner (1972) and Widrow and Hoff (1960), which implement incremental linear regression (Shafaei-Bajestan et al., 2021). Measures derived from the model's networks have been found to be good predictors for lexical measures ranging from reaction times in unprimed (Baayen et al., 2019a) and primed lexical decision (Baayen and Smolka, 2020; Chuang and Baayen, 2021) to spoken word durations (Chuang and Baayen, 2021; Chuang et al., 2021). Of central importance for the present study is that the DLM integrates distributional semantics within a model that includes a production mapping from words' semantic vectors to their constituent segments. This makes it possible to examine to what extent words' meanings co-determine the spoken word durations of homophones. Below, we introduce several measures for assessing the learning of homophones' forms given their meanings, derived from the mappings of the DLM.

In the context of the current discussion, the most conspicuous feature of the DLM is the fact that it does not imply or make use of words as 'items'. Model predictions follow from principles of learning (Rescorla and Wagner, 1972) and the distributional properties of the data, rather than item-specific specifications designed by the analyst. Because in the DLM forms and meanings are generated dynamically, on the fly, during production and comprehension respectively, there are no representations internal to the DLM that a frequency measure can be associated with. There are two ways in which frequency of occurrence plays out within a discrimination-learning based approach to language. First, during learning, higher-frequency words are encountered more often than lower-frequency words, and this will differentially affect learning. Second, there is more than the 'lexicon': words are used in sentences. We can therefore ask how words are learned in contexts where they appear together with other words. For reasons detailed below, in this study we propose a measure that follows the second approach: this measure gauges frequency of occurrence tempered by cue-competition at the sentence level.

Before introducing the specifics of our implementation of the DLM variables, we turn to prior work on spoken word duration.

## 2.3 Predictors of spoken word duration

Numerous factors have been shown to influence spoken word duration in English (for overviews, see e.g. Aylett and Turk, 2004; Warner, 2011; Jurafsky, 2003; Fink and Goldrick, 2015; Tucker et al., 2019; Balota and Chumbley, 1985). Broadly, these factors fall into three categories: Lexical information, such as frequency, phonological neighborhood density (PND), part of speech, and orthography; indexical information about the talker, such as age and sex; and linguistic context, such as speaking rate, prosodic boundaries, or the probability of a target word given the preceding and following context.

The effects of some of these factors are relatively uncontroversial, at least for English and Dutch, the languages in which they have been studied the most. In English, word duration has been found to be shorter in male talkers compared to female ones (Bell et al., 2009), and in nouns compared to other parts of speech (Lohmann, 2018a; Sorensen et al., 1978). Word duration tends to increase with increasing talker age (Bell et al. 2009; Horton et al. 2010, but see Gahl and Baayen 2019) and with decreasing frequency and contextual predictability (Lieberman, 1963; Kilbourn-Ceron et al., 2020; Jurafsky et al., 2001; Aylett and Turk, 2004).

Other effects have not been investigated extensively or are subject to continuing debate. Among these are morphological complexity (Caselli et al., 2016) and, especially, PND. Increasing PND has been associated with longer (Buz and Jaeger, 2016) or shorter (Gahl et al., 2012; Caselli et al., 2016) whole-word duration. Effects of PND on other aspects of pronunciation have similarly yielded mixed results (see e.g. Scarborough, 2013; Wright, 2004; Gahl et al., 2012; Goldrick et al., 2013; Fink and Goldrick, 2015; Buz and Jaeger, 2016; Clopper and Turnbull, 2018; Gahl, 2015; Fricke et al., 2016; Caselli et al., 2016). Indeed, just as lexical frequency has come to be treated as a consensus variable, PND may be the most polarizing of variables.

We believe that some of the differences across studies are the result of differences in the unit of analysis: Some of the literature on PND effects uses data aggregated over word types. Other studies analyze word tokens. Assessing the consequences for interpretation of analyzing token-level vs. type-level information is difficult in part because any given study only reports token-level or type-level results, but not both. In the current study, we report both type-level and token-level models, in order to draw attention to the consequences of analyzing tokens vs. types.

# 3 Methods

## 3.1 The data set

We analyzed the same data set that was used in Gahl (2008) and Gahl (2009). The initial word list contained all English lexemes with a word form that is homophonous with the word form of another lexeme, according to the transcriptions in the CELEX database (Baayen et al., 1995), and that differed in spelling from their homophone twins. Spoken word duration of these items were extracted from the time-aligned orthographic transcript (Deshmukh et al., 1998) of the Switchboard corpus (Godfrey et al., 1992), a corpus of 240 hours of telephone conversations between strangers. Word forms with identical spelling were pooled: For example, the plural noun and the third-person singular verb *laps* were treated as a single item. In cases of three or more homophonous spellings (e.g. *praise, prays, preys*), only the two highest-frequency forms were included in the analysis. Several classes of items were removed from the word list: (1) spellings associated with more than one phonemic representation, e.g. *tear* (homophonous with *tier* and *tare*); (2) Pairs involving function words, such as *in, inn* and *or, ore* and interjections, such as *whoa, woe*; (3) pairs such as *source, sauce* that are homophones in the CELEX transcriptions, which are based on British English Received Pronunciation, but that were unlikely to be homophones in the (American English) Switchboard corpus; (4) items containing transcription errors in CELEX; and (5) names of letters in the alphabet. The resulting list contained 409 homophones.

For the analysis of token duration, disfluent tokens, defined as tokens immediately preceding hesitation sounds or periods of silence of 0.5 seconds or longer, were excluded from analysis. We initially retained such tokens in our models and allowed the Fluency factor to interact with talker age and with the relative frequency of target and homophone, to rule out spurious effect of lexical frequency variable due to fluency. However, the disfluent tokens were too unevenly distributed to allow meaningful models of the effects of lexical predictors on these tokens. Excluding the disfluent tokens left 56,024 observations for analysis.

## 3.2 Measures based on discriminative learning

We now turn to the measures grounded in the DLM, beginning with an alternative to frequency of occurrence and then turning to measures taking semantics into account.

### 3.2.1 A DLM-based alternative to frequency of occurrence

As frequency of occurrence is a robust predictor of spoken word duration, we first consider how frequency plays out in discriminative approaches to language. As already mentioned above, when mappings between form and meaning are estimated using incremental, word-by-word, learning (see, e.g., Shafaei-Bajestan et al., 2021; Heitmeier et al., 2021), then the mappings will be more precise for words that have been encountered more often (but see Ramscar et al., 2013, for how frequency imbalances influence learning). However, modeling the mapping between meaning and form using incremental learning is beyond the scope of this study, for two reasons. Firstly, incremental learning using the Widrow-Hoff learning rule (see also Milin et al., 2020) is computationally highly demanding. Secondly, during actual word learning, words' semantic vectors should also be learned incrementally, reflecting speakers' increasing semantic and textual knowledge. What we have available to us are semantic vectors, trained on large volumes of text that surpass the experience of any individual speaker, that represent well-discriminated meanings expected with near-infinite language experience.

Therefore, instead of seeking to implement incremental learning to capture frequency effects, we consider instead the question of the learnability of words in sentential context. A core insight of discrimination learning (see, e.g., Rescorla, 1988; Marsolek, 2008; Ramscar et al., 2010) is that input features (henceforth, adopting the terminology of (Danks, 2003), cues) compete for predicting output classes (henceforth, outcomes). Since words normally do not occur in isolation but in utterances, we examined how well word outcomes can be learned when there are competing word cues.

More specifically, following Baayen et al. (2019a), we constructed a simple network (with no hidden layers) that was trained incrementally to predict all words in an utterance from the very same words in that utterance, using NAIVE DISCRIMINATIVE LEARNING (NDL Baayen et al., 2011; Milin et al., 2017). The network was trained, using the update rule of Rescorla and Wagner (1972), on 6,020,399 sentences from the written part of the British National Corpus (BNC Burnard, 1995). Words with a frequency less than or equal to 200 were not included, unless they were among the homophones in our dataset. The total number of word tokens taken into account during training was 87,906,894; the number of different word types was 23,562. For each sentence, the input to the network specified which words were present in the sentence, using one-hot encoding for each word. Thus, for a sentence with 10 unique word types, 10 bits in the input were on, and all other bits were off. The network was given the task to predict which words occurred in the very same sentence. Thus, the output vector presented to the model as target was identical to the input vector. Training the model on the British National Corpus resulted in a network characterized by a 23,562 × 23,562 weight matrix $\mathbf{W}$. The row vectors of $\mathbf{W}$ (with diagonal elements set to zero) perform on a par with word embeddings based on latent semantic analysis (Landauer and Dumais, 1997), see Baayen et al. (2019a) for detailed discussion.

The extent to which words support themselves is captured by the diagonal elements of the weight matrix. Let $d_i$ denote the diagonal element of word $\omega_i$. Values of $d_i$ range between 0 and 1, and hence can be interpreted as probabilities. The amount of information carried by $\omega_i$, henceforth $\mathrm{Id}_i$, is therefore (see, e.g., Shannon and Weaver, 1949; Moscoso del Prado Martín et al., 2004; Levy, 2008) given by

$$\mathrm{Id}_i = \log_2 \left( \frac{1}{d_i} \right).$$

This measure of information differs from the standard estimate of information based on frequency only in the way that the probabilities are defined. Instead of using raw word frequencies of occurrence to derive probabilities (Aylett and Turk, 2004), this measure makes

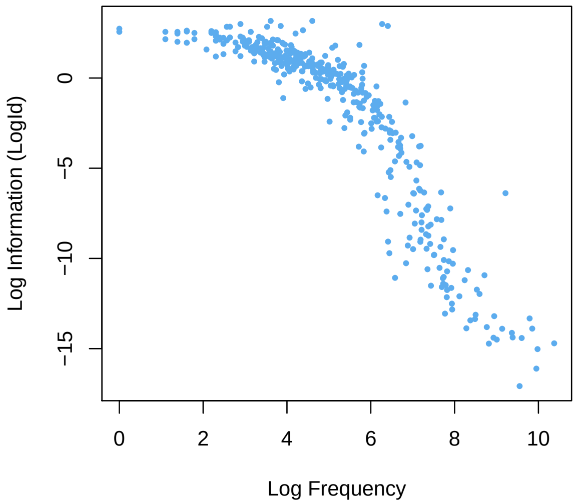

Figure 2: Log Information (`LogId`) as a function of log frequency of occurrence for the homophones in the present study.

use of frequency estimates that are modulated by cue-competition with other words in sentences.

For the present set of homophones, $\text{Id}_i$ has a distribution with a long right tail. In order to avoid outlier effects in the regression, we applied a log transformation to $\text{Id}_i$, resulting in a measure (`LogId`) that is correlated with, but not identical to, frequency of use (see Figure 2). As predictor for by-word mean reaction times in the British Lexicon Project (Keuleers et al., 2012), the `LogId` measure outperforms log frequency of occurrence in the British National Corpus by 563 AIC units (see the supplementary materials for further details on the single predictor Gaussian location-scale models that we fitted). We therefore will consider the `Id` and `LogId` measures as predictors for spoken word duration, as a learning-informed counterpart to standard frequency counts.

### 3.2.2 Measures incorporating semantics

In order to understand the measures that we have found useful for predicting homophone duration from homophones' semantic vectors, we need to provide a brief introduction to core concepts of the mappings used by the DLM. Consider the following example. The starting point is a pair of matrices with information about form and meaning, respectively, and mappings between the two. A form matrix $\boldsymbol{C}$ and a semantic matrix $\boldsymbol{S}$ for a simple two-word lexicon are shown in Figure 3. The two words in the example are *time* and *lime*. Importantly, these labels do not exist as "entries" in a lexicon, or as representational units to which information about form or meaning is linked. The only representations in the model are the high-dimensional numeric vectors, which are understood to be generated on the fly from auditory or visual input in the case of comprehension, and from conceptualization in production. Thus the row names *time* and *lime* are included in Figure 3 purely as a matter of convenience.

$$\boldsymbol{C} = \begin{array}{ccccccc} & \text{ti} & \text{tim} & \text{ime} & \text{me} & \text{li} & \text{lim} \\ \textit{time} & 1 & 1 & 1 & 1 & 0 & 0 \\ \textit{lime} & 0 & 0 & 1 & 1 & 1 & 1 \end{array} \qquad \boldsymbol{S} = \begin{array}{ccccccc} & S1 & S2 & S3 & S4 & S5 & S6 \\ \textit{time} & 0.1 & 0.3 & 0.1 & 0.7 & 0.4 & 1.2 \\ \textit{lime} & 0.6 & 0.2 & 0.9 & -0.1 & 1.4 & 0.4 \end{array}$$

Figure 3: The form matrix $\boldsymbol{C}$ brings together words' form vectors, the semantic matrix $\boldsymbol{S}$ brings together the corresponding semantic vectors.

The row vectors of $\boldsymbol{C}$ specify which triphones are present in a word's phonological form. The triphones are contextualized phones that reflect co-articulation and that also do justice to the dependence of the identity of phones such as stops on the formant transitions in adjacent vowels. The row vectors of $\boldsymbol{S}$ are semantic vectors (word embeddings). It is worth reiterating that form and meaning representations do not have an existence of their own within the framework of the discriminative lexicon, unlike in models in which processes act on units such as morphemes or words. Given external form input, a semantic representation is created dynamically. Similarly, given an internal conceptualization, a form representation is created on the fly. The "discriminative lexicon" is a lexicon without words.

The DLM sets up mappings between the $\boldsymbol{C}$ and $\boldsymbol{S}$ matrices by using the mathematics of multivariate multiple regression to solve the equation $\boldsymbol{CF} = \boldsymbol{S}$ for comprehension mappings, and the equation $\boldsymbol{SG} = \boldsymbol{C}$ for production mappings. The resulting transformation matrix $\boldsymbol{F}$ for the matrices in Figure 3 is a tabulation of beta coefficients, i.e. regression weights.

$$\boldsymbol{F} = \begin{array}{ccccccc} & \text{S1} & \text{S2} & \text{S3} & \text{S4} & \text{S5} & \text{S6} \\ \text{\#ti} & -0.067 & 0.067 & -0.117 & 0.25 & -0.1 & 0.333 \\ \text{tim} & -0.067 & 0.067 & -0.117 & 0.25 & -0.1 & 0.333 \\ \text{ime} & 0.117 & 0.083 & 0.167 & 0.10 & 0.3 & 0.267 \\ \text{me\#} & 0.117 & 0.083 & 0.167 & 0.10 & 0.3 & 0.267 \\ \text{\#li} & 0.183 & 0.017 & 0.283 & -0.15 & 0.4 & -0.067 \\ \text{lim} & 0.183 & 0.017 & 0.283 & -0.15 & 0.4 & -0.067 \end{array}$$

This matrix of regression weights is mathematically equivalent to the matrix of weights on the connections in a network linking input dimensions (for comprehension, the columns of $\boldsymbol{C}$) to output dimensions (for comprehension, the columns of $\boldsymbol{S}$). Each beta coefficient can thus be re-interpreted as the association strength between an input and an output unit. The number of rows of the network's weight matrix $\boldsymbol{F}$ is equal to the number of different trigrams. The number of its columns equals the dimensionality of the word embeddings. In this example, this dimensionality is set to 6, but common dimensions for empirical word embeddings are 200 or 300. However, this dimensionality can be much larger as in the model of Baayen et al. (2019a). When the weights of the network are estimated using the multivariate multiple regression equations, they represent the best weights that can be learned with infinite experience. This way of estimating weights is referred to as LINEAR DISCRIMINATIVE LEARNING (LDL). As mentioned above, the weights of the network can also be learned incrementally. When using incremental learning with the learning rules of Rescorla and Wagner (1972) or Widrow and Hoff (1960), cycling through the training data many times results in estimates that come ever closer to those shown for $\boldsymbol{F}$ (see Chuang and Baayen, 2021; Shafaei-Bajestan et al., 2021, for further details). In this study, we make use of LDL, as modeling with incremental learning is computationally intensive and requires precise data on lexical acquisition over time that is not available to us.

A production matrix $\boldsymbol{G}$ mapping semantic vectors (embeddings) onto form vectors was calculated for a dataset of 10,636 words. These words were taken from the dataset studied in Baayen et al. (2019a), augmented with the homophones studied in G2008. For each of these words, the constituent triphones were calculated from their DISC phoneme representations in the CELEX lexical database (Baayen et al., 1995), and used to create the binary form vectors. For each word, a semantic vector was extracted from Cieliebak et al. (2017), which provides 200-dimensional embeddings obtained with `fasttext` (Bojanowski et al., 2017) applied to tweets. Thus, $\boldsymbol{G}$ is a 5,600 × 200 matrix mapping the row vectors of a 10,636 × 200 semantic matrix $\boldsymbol{S}$ onto the row vectors of a 10,636 × 5,600 form matrix $\hat{\boldsymbol{C}}$. Here, we borrow notation from statistics, as the predicted form vectors $\hat{\boldsymbol{C}}$ are not identical to words' actual 'gold standard' form vectors in $\boldsymbol{C}$. However, the more accurate the mapping $\boldsymbol{G}$ is, the more similar the row vectors of $\hat{\boldsymbol{C}}$ will be to the row vectors of $\boldsymbol{C}$.

Two related measures quantifying the support for a homophone's form provided by its meaning can now be introduced. Both measures are derived from homophones' predicted form vectors $\hat{\boldsymbol{c}}$, which are row vectors of $\hat{\boldsymbol{C}}$. The first measure, `Semantics To Form Mapping Precision`, quantifies how close the predicted vector $\hat{\boldsymbol{c}}$ is to the 'gold standard' vector $\boldsymbol{c}$ that specifies which triphones are present in a word's form. This measure is defined as the Pearson correlation of $\boldsymbol{c}$ and $\hat{\boldsymbol{c}}$. The greater this correlation is, the more accurate the mapping from meaning to form is.

The second measure quantifies the total support that a word's triphones receive from its semantics. We refer to it as `Semantic Support For Form`. Let $\mathcal{C}_i$ denote the set of indices of the triphone cues of word $i$ in the columns of the form matrix $\boldsymbol{C}$, and let $\hat{\boldsymbol{c}}$ denote its estimated form vector. A word's `Semantic Support For Form` can now be defined formally as $\sum_{j \in \mathcal{C}_i} \hat{\boldsymbol{c}}_j$. Unsurprisingly, these two measures are strongly correlated ($r = 0.896$), and it is an empirical question which of the two will be the more precise predictor. We note that it will not make sense to include both predictors in the same regression model.

## 3.3 Predictors and predictions

### 3.3.1 Predictors common to both types of models

There are several predictor variables for spoken word duration, listed in Table 1, that are independent of localist and DL-specific assumptions about the lexicon. These general variables include age and sex of the speaker (cf. Horton et al. 2010; Yuan and Liberman 2008), and the conditional probability of a target word, given the word(s) preceding and following it (e.g. Bell et al. (2009); Kilbourn-Ceron et al. (2020)). Another control variable that cuts across models of the lexicon is the duration that can be expected, given a word's phonemes. We estimated the 'baseline' duration of a word as the log-transformed sum of the average duration of the target segments in the Buckeye corpus (Pitt et al., 2007). Homophones have identical baseline duration estimates, as they have identical segments.

Variables capturing properties of specific tokens entered the type-based regression models in the form of aggregated information. One such variable is `Pause quotient`. Words that occur preceding a pause tend to be longer than pre-pausal words. Furthermore, tokens of words that frequently occur phrase-finally have been found to be longer, compared to their baseline, even when not in final position (Sóskuthy and Hay, 2017). In a token-based analysis, a binary variable stating whether a token is pre-pausal is straightforward to include, but for a type-based analysis, we included as predictor the proportion of tokens of a given target word that preceded pauses. Similarly, local speaking rate entered the type-based analyses in aggregated form, as the average of the local speaking rate of the tokens.

Our token-based models did not include by-word random effects. In general, when word items have very different frequencies, it is advisable to not specify a random effect for word as the posterior modes may then diverge considerably from normality (Douglas Bates, p.c.). Furthermore, when item-bound predictors are of interest, as in the present regression study, the by-item adjustments (posterior modes) are confounded with the item predictors. This confound is complete for predictors that have non-repeating values: in this case, there is a one-to-one relation between predictor values and posterior modes. In GAMs, this gives rise to concurvity values near 1 when smoothing splines are used. When predictors have values that are instantiated across multiple tokens of the same type, whether a predictor will survive inclusion of a random effect will depend, in unpredictable ways, on which values of the predictor are repeated by these tokens.

### 3.3.2 Variables specific to localist models

We now turn to predictors that are specific to localist theories of the mental lexicon. These variables were also included in G2008, with the exception of PND. A key localist variable, given the theoretical interest of homophone frequency effects, was lexical frequency, i.e. the frequency of each target lemma (e.g. *time* vs. *thyme*). In models attributing frequency effects to mechanisms that affect homophone twins equally, such as retrieval of a word form (lexeme) or articulatory fluency, the frequency variable of interest would be the `form frequency`, i.e. the cumulative frequency of both homophones. We expected spoken word duration to reflect lemma-specific frequency and therefore included `Lemma Frequency` as a predictor, defined as the log-transformed frequency of the lemma provided by the CELEX database. We expected spoken word duration to decrease with increasing frequency of each target lemma. We further expected lemma frequency to be a stronger predictor than form frequency. We tested that prediction in a comparison of models using lemma frequency vs. form frequency.

| Variable name | Description |
|---|---|
| `Age` | The talker's age. [Tok] |
| `Baseline duration` | The (log-transformed) sum of the average duration of the target's segments, based on the Buckeye corpus (Pitt et al., 2007). |
| `Bigram probability` | The word-based bigram probability of the target, given the word following it. [Avg] |
| `Sex` | The talker's sex (male or female, as coded in (Godfrey et al., 1992)). [Tok] |
| `Biphone probability` | The average of the target's position-specific biphone probabilities, based on Vitevitch and Luce (2004). |
| `Orthographic regularity` | A measure of orthographic regularity (Berndt et al., 1987). This variable was called "m-score" in G2008. |
| `Pause quotient` | The proportion of target tokens immediately followed by a pause. [Typ] |
| `Phonological form` | A factor coding the phonemic content of the target and its homophones, e.g. /taɪm/ in the case of *time* and *thyme*. |
| `Speaking rate` | The local speaking rate, in syllables per second, in the stretch of speech from the end of the target to the end of the utterance. [Avg] |
| `Noun quotient` | A binary variable coding whether the estimated proportion of nouns among the tokens of a given form was above vs. below 0.5, based on the syntactic category-specific frequency counts in CELEX. |

Table 1: Predictor Variables considered in 'benchmark' models, as well as in the LDL models. Avg = raw value included in token-models; average included in type-models; Tok = included in token-models only; Typ = included in type-models only (see text).

We included a further frequency measure, `Relative Frequency`, estimated as the (log-transformed) frequency of each target homophone in the CELEX database (Baayen et al., 1995), divided by the frequency of its homophone twin. The relative frequency is thus greater than zero for the higher-frequency member of the pair, and smaller than zero for the lower-frequency member of the pair. The same variable was used in Lohmann (2018b). This variable makes it possible to take into account the fact that the more frequent a word is, the more it can exceed its homophone twin in frequency, by including an interaction between `Lemma Frequency` and `Relative Frequency`. If the frequency of the homophone is indeed a co-determinant of its duration, as argued by G2008, then homophones with higher relative frequency should have shorter durations.

A binary variable `Morphological Complexity` was included to distinguishing morphological simple vs. complex target words, e.g. *lax* vs. *lacks*. Finally, `Phonological Neighborhood Density (PND)` was included as predictor. Neighborhood density was estimated as the number of words differing from the target word by addition, deletion, or substitution, based on the English Lexicon Project (Balota et al., 2007). We expect spoken word duration to decrease as PND increases (Gahl et al., 2012).

### 3.3.3 Variables specific to the DLM

Four predictors are specific to the DLM. Three of these have already been introduced: `Semantics To Form Mapping Precision`, `Semantic Support For Form`, and `LogId`. The fourth measure is the Pearson correlation between the semantic vectors of a homophone pair, again using the tweet-based `fasttext` word embeddings of dimension 200 (Cieliebak et al., 2017), henceforth `Homophone Semantic Similarity`. This measure informs us about how similar in meaning the words of a homophone pair are.

With respect to the predictions for the learning-based measures, we note the following. The theory of the discriminative lexicon predicts that duration should be longer for higher values of `Semantics to Form Mapping Precision` and for `Semantic Support For Form`. The reason is straightforward: forms that are inappropriate given the semantics to be expressed should be realized with zero length. As a mapping is learned better for a form-meaning pair, a word will be produced with greater duration, other things (such as contextual predictability) being equal. For empirical evidence, the reader is referred to Baayen et al. (2019a); Chuang et al. (2020c); Tomaschek et al. (2021a).

With respect to `Homophone Semantic Similarity`, our hypothesis is that greater similarity will afford longer duration, the reason being that greater similarity reduces the stress in the mapping from meaning to form, and thus will provide stronger support for words' forms. The mapping from semantic space to form space has to associate different meanings with different forms for a large part of the vocabulary. Mapping different meanings onto the same form is less often required, and because it is more exceptional, learning will be more tentative. Importantly, forcing a model, designed to map distinct semantic vectors onto distinct form vectors, to map onto identical form vectors induces frailty in the mapping (see Chuang et al., 2020a, for detailed simulation studies on homophone-induced frailty). Conversely, the more similar in meaning homophones are, the more they approximate non-homophones, and hence the better the corresponding form vectors are expected to be learned.

Finally, for the information measures LogId and its untransformed counterpart Id, higher values are predicted to give rise to longer mean spoken word durations. With respect to the variance in spoken word durations, we expected greater amounts of information to give rise to greater variance, as the network on which the information measures are based has had reduced opportunities for learning words with higher information load. In addition, since the actual context in which words were spoken in the Switchboard corpus is more likely to be different from the context of learning in the British National Corpus, the estimates based on the British National Corpus are expected to be more off for the Switchboard Corpus especially for lower-frequency, high-information words, leading to greater variance in context-specific spoken word durations.

Classical lexical-distributional predictors that have no theoretical motivation within the DLM framework, are phonological neighborhood density, biphone probability, lemma frequency, and relative frequency. We also did not include the orthography-phonology consistency measure, because developing a learning-based consistency measure within the model as laid out in Baayen et al. (2019a) is beyond the scope of the present study.

It is debatable whether Baseline Duration should be included in a statistical model grounded in discriminative learning, as the model does not incorporate phones as theoretically motivated units. However, as the model represents words' forms with indicators of words' triphones, the number of bits that are on for a row vector $\boldsymbol{c}$ of $\boldsymbol{C}$ is the same as the number of a words' phonemes (except for words with repeated phones, as $\boldsymbol{c}$ vectors only register the presence of triphones and not their location or number). We therefore included Baseline Duration as a control variable, which also facilitated comparing localist and DL models.

For three words (the names *Phil, Marx, Thais*), values for LogId and Id were unavailable. The analyses of the predictors grounded in the DLM are therefore based on 403 of the 406 homophones in the analyses with localist predictors.

## 3.4 Statistical modeling strategy

We made use of the Gaussian Location Scale Additive Model, using the packages **mgcv** (Wood, 2011, 2004, 2017) and **itsadug** (van Rij et al., 2020) in R (R Core Team, 2017). The Generalized Additive Model (GAM) is a regression model that relaxes the assumption that the effects of numeric predictors are linear. GAMs make use of smoothing splines that are set up such that an optimal balance is reached between staying faithful to the data and keeping model complexity down (by penalizing for nonlinearity). If a predictor is truly linear, a GAM will detect this and not report artifactual non-linearity. Applications to linguistic data and tutorials can be found in Baayen et al. (2010); Wieling et al. (2011); Wieling (2018); Wieling et al. (2014); Lee et al. (2016); Baayen et al. (2017); Sóskuthy (2017); Baayen and Linke (2020); Chuang et al. (2020b). GAMs are also much more flexible than the standard linear regression model for modeling interactions between numeric predictors.

Gaussian Location Scale GAMs relax yet another assumption of the linear regression model, namely, that the variance is not supposed to change with the mean. Instead of requiring residuals to be homoskedastic, Gaussian Location Scale GAMs allow the variance, which is assumed to be Gaussian, to change with the mean, if necessary in a non-linear way. This allows us to address the question of whether lower frequency words have more variable durations.

The full sets of variables in the type-based analyses show high collinearity for both the localist and the DL models: $\kappa = 41.7$ for the former, and 40.8 for the latter, using the collinearity index of Belsley et al. (1980). Without corrective measures, magnitude and sign of coefficients in linear regression may become theoretically uninterpretable due to suppression or enhancement (Friedman and Wall, 2005). In non-linear regression, main trends can likewise reverse. Various corrective methods for addressing collinearity are available (see, e.g., Tomaschek et al., 2018a, for an overview), but these are not straightforward to apply when the goal is to study both mean and variance of spoken word duration. We therefore proceeded as follows. In a first step, we used all relevant predictors. After removing predictors that failed to receive evidence for their relevance, as well as the `Semantics to Form Mapping Precision` measure, which is highly correlated with `Semantic Support for Form` ($r = 0.90$) and which showed evidence of suppression, the model was refit. This reduced collinearity to 15.2 and 16.3 respectively. `Baseline Duration` still induced suppression. As it is a control variable in our model, we regressed it on the other remaining variables, for both sets of predictors, and used the residuals, henceforth `Residual Baseline Duration`, as predictor. This further reduced collinearity down to 9.8 and 9.9 respectively. According to Belsley et al. (1980), this low level of collinearity is unlikely to lead to distorted estimates of regression coefficients.

For our analyses, we restricted the number of basis functions for the smoothing splines, in order to bring out main trends in the data. We note, however, that increased numbers of basis functions, resulting in far more wiggly partial effects, are well-supported statistically. In order to facilitate interpretation, we have avoided these more complex smooths.

Given that the dataset under investigation has been studied several times, and in light of our exploratory approach to statistical modeling, we set $\alpha = 0.0001$.

# 4 Results

In what follows, we first report analyses with predictors grounded in the localist model, before moving to models with predictors grounded in the DLM.

## 4.1 Localist model

The type-based model using localist variables is summarized in Table 2. The upper part of this table lists the effects for factorial predictors, as well as for the intercept. Because mean and variance are modeled jointly, there are two intercepts, one for the mean and one for the variance. The only measure that was predictive for the variance was word frequency. We discuss its effect below.

A `Noun-bias` was associated with longer mean duration, consistent with the idea that nouns occur phrase-finally more often than verbs do, and are hence more likely to undergo phrase-final lengthening (Gahl, 2008; Sóskuthy and Hay, 2017). However, this effect was associated with a relatively high p-value ($p = 0.0004$), rendering it doubtful to us (recall that we set $\alpha = 0.0001$) that this effect will replicate consistently .

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| `Intercept` [mean] | -1.0387 | 0.0139 | -74.7519 | < 0.0001 |
| `Noun Bias` =yes | 0.0598 | 0.0169 | 3.5387 | 0.0004 |
| `Intercept` [variance] | -1.8293 | 0.0378 | -48.4416 | < 0.0001 |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(`Proportion with Following Pauses`) | 1.8335 | 2.2863 | 32.2159 | < 0.0001 |
| s(`Phonological Neighborhood Density`) | 4.7174 | 5.7485 | 71.3616 | < 0.0001 |
| s(`Orthographic regularity`) | 1.0000 | 1.0001 | 5.2430 | 0.0220 |
| te(`Lemma Freq., Rel. Freq.`) | 7.2717 | 9.4159 | 108.3451 | < 0.0001 |
| s(`Residual Baseline Duration`) | 1.1915 | 1.3567 | 122.3611 | < 0.0001 |
| s(`Lemma Frequency`) [variance] | 2.6120 | 3.2966 | 75.3180 | < 0.0001 |

Table 2: Gaussian Location-Scale GAM fitted to the average durations of the homophone word types, using control and localist predictors. AIC: -230.8

The second part of the table summarizes the smooth terms in the model. When the effective degrees of freedom (edf) are close to 1, the effect of a predictor is close to linear. Figure 4 visualizes the smooth terms; by fixing the scales across panels for the mean, the effect sizes of the predictors emerge straightforwardly.

The effect of `Orthographic Regularity` is fully linear (edf=1.0000). Predicted duration increased in a nearly linear fashion with proportion of prepausal tokens, and linearly with residualized baseline duration. Increasing phonological neighborhood density and increasing orthographic regularity were each associated with shorter duration. The linear effect of orthographic regularity is small, and its p-value is far above our preset $\alpha$-level, so it remains doubtful that this predictor is relevant for spoken word duration of homophones.

The interaction of word frequency and frequency ratio is visualized by means of a contour plot in the center panel of the second row of Figure 4. Warmer colors indicate longer spoken word durations. Frequency shows the expected effect of durational shortening: the general gradient in the regression surface is negative. There is also an effect of frequency ratio, but only for targets with log frequencies below about 4. In that range, targets with very low frequency ratios (values below -4) are predicted to have shorter durations, compared to targets with similar frequency, but higher frequency ratios: very low frequency targets with very high frequency homophone twins have shorter durations than would be predicted based on their lemma frequency. That pattern is consistent with proposals under which duration is expected to vary with form frequency. Such proposals include "frequency inheritance" (Dell, 1990; Jescheniak and Levelt, 1994), as well as proposals attributing frequency-dependent duration to articulatory practice. That being the case, we explored this pattern further, asking whether this possible inheritance effect could provide a superior explanation for the effect of lemma frequency. If so, then form frequency, i.e. the summed frequency of each

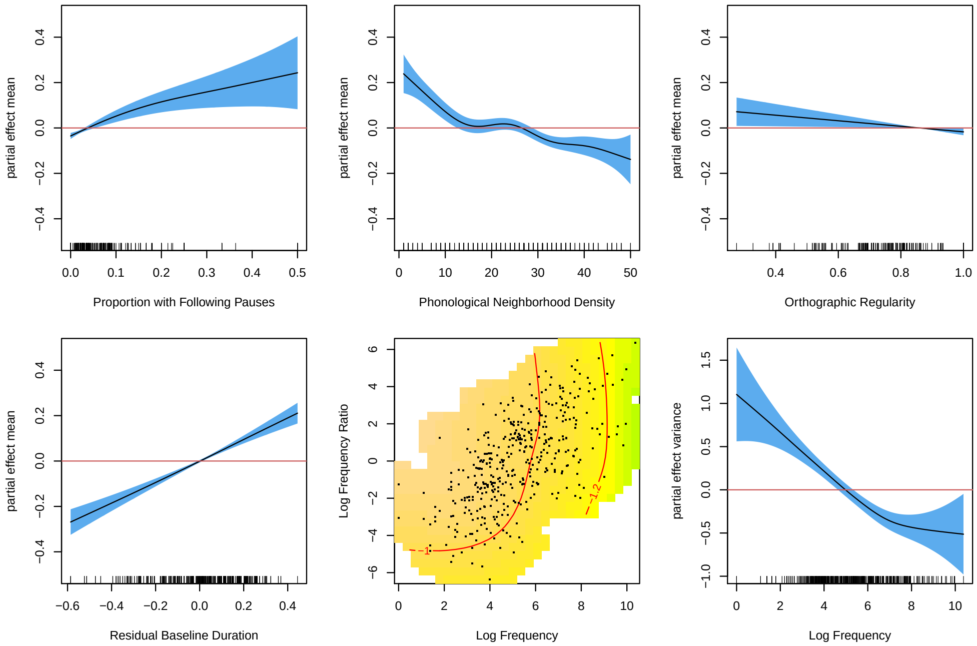

Figure 4: Partial effects according to a Gaussian Location-Scale GAM fitted to the average duration of the homophone word types, using control variables and localist predictors. Rugs indicate the unique values of predictors. In the contour plot, warmer colors indicate longer spoken word durations. AIC: -231.7; -REML = -94.98.

pair of homophones, should improve model fit. We found that replacing lemma frequency with the summed frequency of the homophones resulted in a model with a higher AIC (by 42 units), indicating substantially poorer model fit. Finally, the lower right panel clarifies that increased lemma frequency is associated with decreased variance in duration (as well as decreased duration); however, uncertainty was high near the extremes of the frequency distribution, where data are sparse.

The analysis we just presented was type-based. We now turn to the token-based analysis. As mentioned in the Introduction, token-based models are used in most corpus-based analyses of pronunciation duration (e.g. Bell et al. 2003; Seyfarth 2014; Gahl et al. 2012, and Caselli et al. 2016). Examining the token-based models is therefore an important step enabling comparisons to previous work. We present these models only briefly, however, because we perceive serious methodological problems raised by token-based models generally.

Figure 5 presents the partial effects of the smooths in the localist GAM based on tokens; the model summary is available in the supplementary materials. Even though we severely restricted all splines for the mean (by setting a low upper bound to the number of basis functions), several predictors show sine-like curves that are theoretically not straightforward to interpret, and that in all likelihood are artefactual. For instance, it is entirely unclear why the effect of baseline duration would be non-linear, with an initial reduction in spoken word duration as baseline duration is increased. The extreme wiggliness in smooths that results when the default number of basis functions is used is illustrated by the lower right panel, which presents the partial effect of frequency for the variance. When more basis functions are used also for predicting the mean, model fit improves substantially, at the cost of further reductions in interpretability.

In the type-based analysis, the interaction of Lemma Frequency and Frequency Ratio suggested an effect of Frequency Ratio for the lowest values of Lemma Frequency. A similar interaction emerges from the token-based analysis, as shown in contour plot in Figure 5. The regression surface is substantially more wiggly, but as before, spoken word duration tends to decrease for items with very low frequency ratio, in increments varying substantially with frequency.

## 4.2 DLM analysis

We now turn to the analyses using discrimination-based variables. The type-based Gaussian Location-Scale GAM for this set of predictors is summarized in Table 3 and visualized in Figure 6.

Most of the predictors have a linear or near-linear effect, and upon inspection, a Gaussian Location Scale GAM with only linear predictors turned out to have an equivalent AIC. The effects of the control variables (`Proportion with Following Pauses` and `Residual Baseline Duration`) show similar effects as in the analysis with localist-grounded predictors. As anticipated, spoken word duration increased with `Homophone Semantic Similarity` as well as with `Semantic Support Form`. Furthermore, in the mean, words with a higher information load (`LogId`) were realized with longer durations. The variance also increased with information load, but here, the `Id` measure turned out to be the superior predictor.

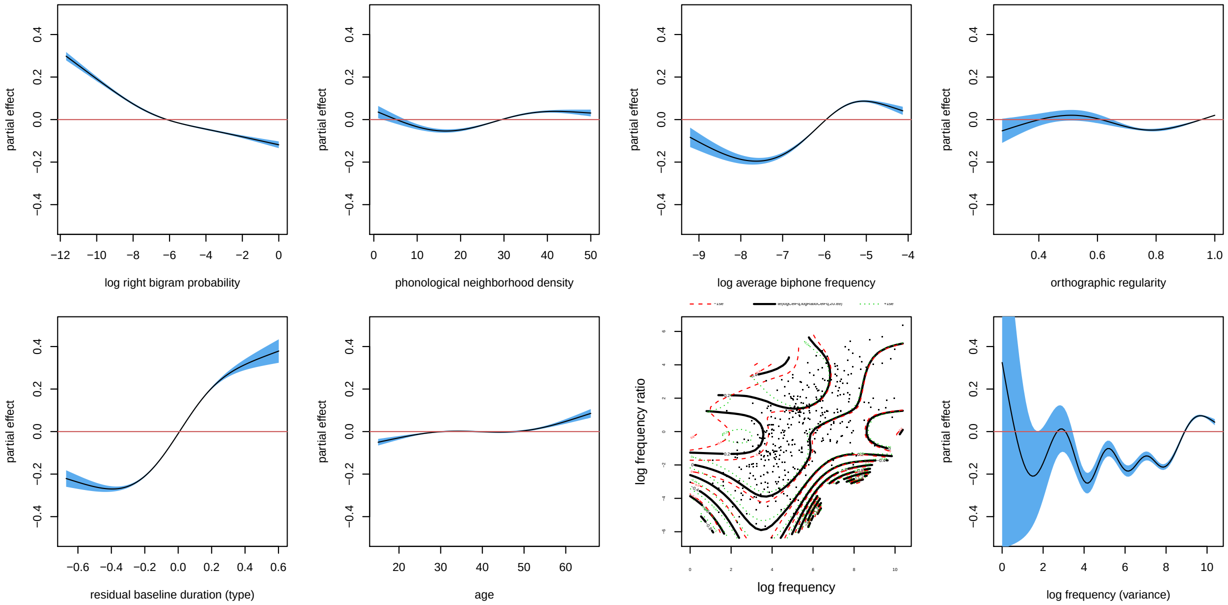

Figure 5: Partial effects according to a Gaussian Location-Scale GAM fitted to spoken word duration, for all word tokens, using predictors grounded in localist theory. AIC: 72144; -REML = 36208. In the contour plot, the direction of the gradient is indicated by dashed green lines (upper one-standard-error confidence contour) and dotted red lines (lower one-standard-error confidence contour). For many contour lines, confidence regions are so narrow that they are hardly visible. We note that in general the gradient along the vertical axis is upwards.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| `Intercept` [mean] | -1.0511 | 0.0131 | -80.3967 | $< 0.0001$ |
| `Noun Bias` = yes | 0.0774 | 0.0166 | 4.6483 | $< 0.0001$ |
| `Intercept` [variance] | -1.8224 | 0.0373 | -48.8418 | $< 0.0001$ |
| B. smooth terms | edf | Ref.df | F-value | p-value |
| s(`Proportion with Following Pauses`) | 1.4334 | 1.7438 | 36.8387 | $< 0.0001$ |
| s(`Homophone Semantic Similarity`) | 1.0000 | 1.0000 | 17.3556 | $< 0.0001$ |
| s(`Semantic Support Form`) | 1.9174 | 2.4631 | 43.0121 | $< 0.0001$ |
| s(`Residual Baseline Duration`) | 1.0000 | 1.0001 | 134.6418 | $< 0.0001$ |
| s(`LogId`) [mean] | 1.0002 | 1.0004 | 48.1098 | $< 0.0001$ |
| s(`Id`) [variance] | 1.0001 | 1.0002 | 77.3014 | $< 0.0001$ |

Table 3: Model summary for a Gaussian Location-Scale GAM fitted to the mean spoken word durations, using control variables and variables grounded in the DLM. AIC: -256.32

Replacing either `LogId` by `Id` or vice versa resulted in inferior fits with more nonlinear effects.

The model based on localist predictors has an AIC score equal to -231.7, whereas the model based on DLM predictors had an AIC score equal to -256.3. The evidence ratio, 219,696, thus is massively in favor of the DLM-based model, indicating that this model is more likely to minimize the information loss than the localist model. It is noteworthy that in the latter model, the effect sizes of Log Information (`LogId`) and `Semantic Support Form` are similar in size. Clearly, the semantic measure makes a solid independent contribution to the model fit.

A token-based model for spoken word durations using DL predictors is available in the supplementary materials. Although the collinearity for the predictors in this model was less severe ($\kappa = 21$), it was still high. Unsurprisingly, partial effects became nonlinear, and showed effects of suppression and enhancement. The partial effect of baseline duration in this model is completely enigmatic, as it levels off two-thirds into the range of predictor values.

## 4.3 Variable importance analysis

Instead of asking about the shape of the relationship between the predictors and word duration in token-based analysis, which is far from straightforward in the presence of high collinearity, we next asked a different question, namely, which predictors made a solid contribution to the model fit.

As our point of departure, we took the token-based model with localist predictors. We constrained all effects to be linear, in order to bring out main trends and avoid issues with uninterpretable smooths. We first left one predictor at a time out of the regression equation, and observed the increase in AIC. The greater the increase in AIC, the more important a predictor is. The upper panel of Figure 7 presents the resulting variable importances. Residual baseline duration is the most important predictor, as expected, as it predicts word duration from (average) phone duration. The next most important predictor is the tensor product of frequency by frequency ratio, which is closely followed by frequency by itself.

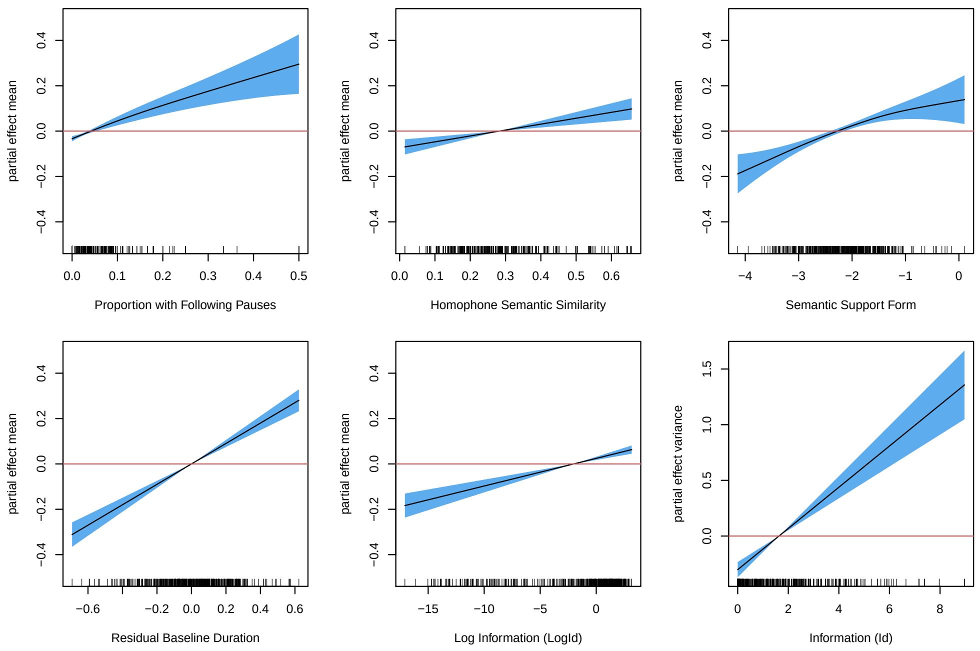

Figure 6: Partial effects according to a Gaussian Location-Scale GAM fitted to the average duration of the homophone word types, using control and DLM-based predictors. AIC: -256.3; -REML = -106.63.

Frequency ratio also makes a clear contribution, whereas the contributions of remaining variables are modest, though not negligible (AIC increments range from 30 to 11,338).

Next, we considered the question of whether predictors based on the DLM help improve model fit beyond what is possible to achieve with localist variables. To address this question, we added DL predictors to the model, one at a time. If the AIC decreases substantially, we can conclude that the predictor is enhancing prediction accuracy. The lower panel of Figure 7 presents the reduction in AIC when DL based predictors are added to the token-based model. The decrease in AIC ranges from 26 to 695. Clearly, taking semantics into account helps improve the token-based localist model fit further. From these results, we conclude that the semantics of homophones have to be taken into account in discussions about the spoken word durations of homophones. This conclusion is independent of whether analyses are based on types or on tokens.

## 5 General Discussion

We modeled word duration averaged over word types, as well as word token duration of homophones in a corpus of spontaneous speech, using predictors grounded in two conceptions of the mental lexicon: localist models, i.e. models containing words as units of representation, and recent models based on discriminative learning (DL). Our results contribute to ongoing research comparing discrimination-based measures of lexical processing with classical lexical-

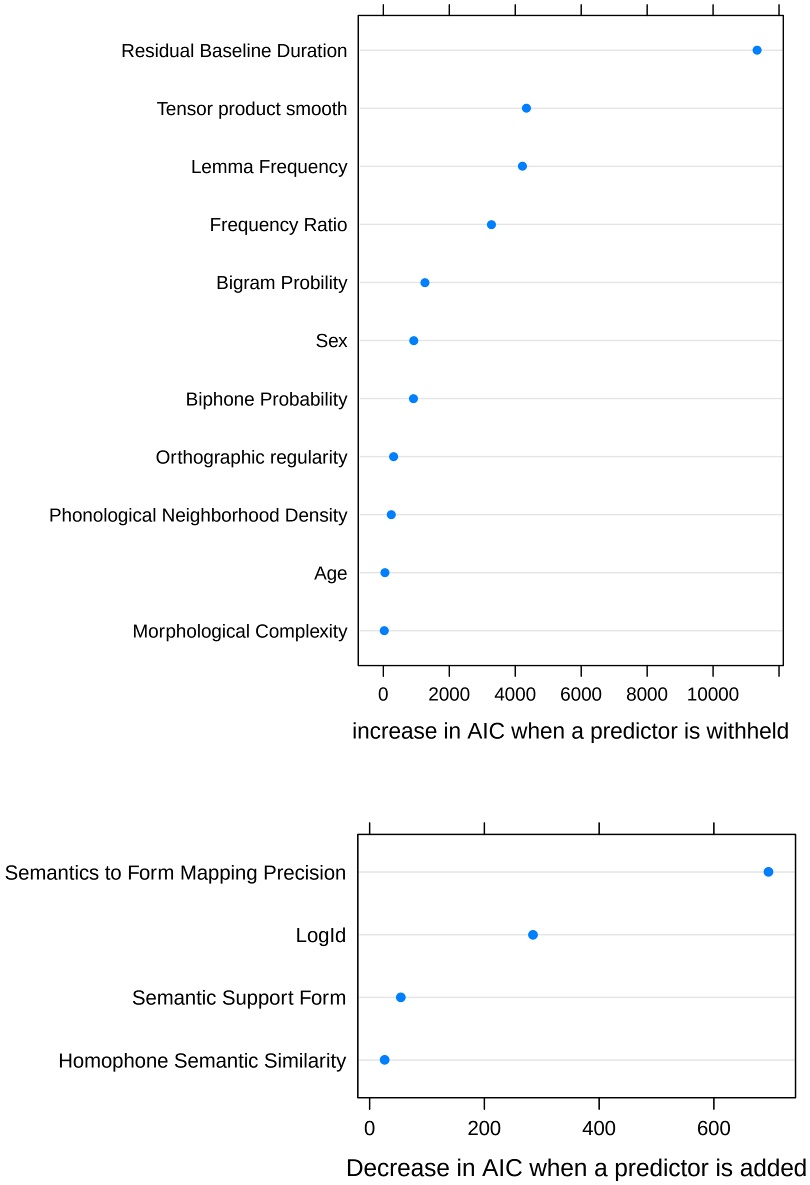

Figure 7: Variable importance in a token-based analysis of spoken word duration. Upper panel: increase in AIC when a predictor is **withheld** from the localist model; lower panel: decrease in AIC when an DL predictor is **added** to the model.

distributional predictors (Milin et al., 2017;Chuang and Baayen, in press).

We discuss our results in the context of three broad sets of issues: The first concerns the choice between the type-based vs. the token-based analyses. We argue that token-based models result in major methodological problems for corpus-based research about lexical effects on fine phonetic detail. We argue that these problems may be responsible for at least some of the divergent findings on effects of Phonological Neighborhood Density (PND) on pronunciation, the variable we called the most polarizing figure among lexical variables. The second broad issue concerns the two conceptions of the mental lexicon, and in particular, the role played in each by frequency of occurrence, the variable seemingly enjoying the greatest general consensus. The third issue, related to the second, concerns the role of words' meanings in shaping word production. We discuss these three broad issues in turn.

## 5.1 Is more always better? The consequences of analyzing types vs. tokens

We begin by considering some of the differences between our 'localist' type-level and token-level models. Increasing Phonological Neighborhood Density (PND) was nearly linearly associated with shorter duration in the type-level model. In the token-based model, the same relationship followed a U-shaped pattern. Increasing homophone semantic similarity was associated with longer duration at the type level, but followed an inverse U-curve at the token level. Analogous differences in type-level vs. token-level models arise with other variables, both localist and DL-based. What explains the differences, and which sets of models should one trust?

At first glance, token-based models might seem superior: These models are able to take into account utterance-specific or indeed token-specific information, such as local speaking rate, or the words preceding and following the target. Also, token-based investigations boast impressively large data sets that can support nuanced statistical models. For example, Bell et al. (2009) analyze 6,938 tokens of function and content words; Dilts (2013) analyzes 137,319 content word tokens; and Aylett and Turk (2004) analyze 169,464 syllable tokens. In the current study, that *prima facie* strength is evident in the narrow confidence intervals of the token-based estimates, compared to the high uncertainty of some of the type-based ones. These advantages would seem to make token-based analyses the instrument of choice.

However, the data sets for token-based analyses of unbalanced corpora are dominated by high-frequency items. As a consequence, lexical frequency enters token-based models of such corpora twice: as a predictor variable and as the count of tokens. It might be objected that frequency estimates based on sources external to the corpus need not correlate with corpus frequencies: Word frequency estimates differ across corpora (Kilgarriff, 2001). However, sometimes they do correlate. For our target words, for instance, the (Pearson) correlation between Switchboard and CELEX frequency was 0.83. Because deviations of model predictions are punished for every single observation, token-based models achieve excellent fit if and only if they work well for high-frequency words.

The overwhelming influence of high-frequency words is particularly pernicious in the context of research on lexical processing because frequency is correlated with several other lexical variables, such as orthographic regularity, length in letters and syllables, and PND. Moreover, as demonstrated in Yap and Balota (2009), the effects of some variables on lexical processing decrease with increasing frequency. As a consequence, models driven by high-frequency words are liable to miss effects of such variables.

Yet another issue becomes apparent when one considers the pairwise correlations among variables, shown in Table 4. We realize that the bivariate relationships are nonlinear and hence not well captured by the correlation coefficients; however, the problems we wish to point out persist regardless of how the relationships are quantified.

The first problem to note is that correlations at the token level are inevitably significant. This is because tokens of any given word have identical values for any given lexical property. For an extreme case of this, consider the variables Baseline duration and Relative frequency, i.e. the (log-transformed) frequency of each target divided by that of its homophone twin: Homophones have identical baseline duration, by virtue of containing identical phones; they also have identical absolute values of relative frequency, but with opposite sign. As a consequence, the type-level correlation between Baseline duration and Relative frequency is zero. At the token level, however, the higher-frequency homophone contributes a greater number of tokens than its twin, each with a positive value of relative frequency. This results in a positive correlation of baseline duration with relative frequency. Our models used a residualized measure of baseline duration, rather than raw baseline duration; we mention this example by way of illustrating that token-level correlations are significant even for orthogonal variables.

| | Baseline | Target Fq. | Rel.Fq. | Orthogr. | PND | Biphone |
|---|---|---|---|---|---|---|
| Baseline dur. | | **-0.22** | 0 | **-0.18** | **-0.6** | **0.27** |
| Target Frequency (log) | **-0.37** | | **0.65** | **0.11** | **0.18** | 0.03 |
| Relative Freq. (log) | **0.24** | **0.43** | | 0.01 | 0.01 | -0.01 |
| Orthographic regularity. | **-0.27** | **0.35** | **0.02** | | **0.16** | **0.13** |
| PND | **-0.38** | **0.29** | **-0.15** | **0.36** | | **-0.16** |
| Avg. biphone prob. (log) | **0.33** | **-0.13** | **-0.21** | **-0.16** | **-0.14** | |

Table 4: Token-level (lower triangular) and type-level (upper triangular) Pearson correlations among localist variables; correlations with p-values below .05 are shown in boldface.

These inevitable correlations have two immediate consequences: They produce high collinearity, and they explain the apparent predictive power of some of the variables in the token-based models, which arise because tokens of high-frequency words move *en bloc*. To make matters worse, token-level correlations, although inevitable, are sometimes weaker than type-level ones. This seemingly paradoxical pattern is due to the distribution of lexical characteristics in the lexicon. For example, Baseline duration and PND are more strongly correlated at the type level vs. the token level (-.60 vs. -.39). Evidently, words with long baseline duration tend to have few phonological neighbors — and these words tend to be infrequent. This pattern highlights the fact that unbalanced token-based data sets can mask patterns that are present in the lexicon.

The influence of high-frequency words on token-based models has pervasive consequences. One such consequence concerns the variable we called baseline duration, estimated here as the sum of the mean duration of the target phones in the Buckeye corpus Pitt et al. (2005). We residualized baseline duration on other predictors (PND, form frequency, and biphone probability), using tokens as the unit of analysis. That procedure was justified, and arguably necessary, given that the goal of residualization was to reduce concurvity, which it did. From a different perspective, another rationale for taking baseline duration into account is the need to control for the distribution of segments in the lexicon: For example, if high-frequency words contain phones that are inherently shorter than those appearing in low-frequency words, then a failure to control for inherent phone duration at the type level may result in a spurious effect of frequency: high frequency words will be shorter than low-frequency ones due to their inherent phone duration, rather than due to usage frequency *per se*. To see whether frequency affects duration when controlling for inherent phone duration, one might therefore residualize baseline duration based on word types, rather than tokens, and enter the 'type-based' residualized baseline duration in a token-level model. In order to check the effects of this choice, we fitted such a model (available in the supplementary material). The resulting model leads to a completely different characterization of the relationship between PND and duration: The partial effect of PND is no longer U-shaped. Instead, predicted duration decreases with increasing PND.

In fact, even seemingly negligible differences in word lists can produce large differences between models. In our data, for example, the eight highest-frequency word types account for over 50% of the tokens, and the 30 highest-frequency types account for over 80%. One might object that replacing high-frequency words with other high frequency items might produce similar results, given the correlation of frequency with other variables. That is not necessarily the case, as the comparison of the effects of PND in the current study vs. the study of Gahl et al. (2012) demonstrates.

Type-based analyses avoid many of these issues - and they need not forego information about contextual variables altogether. In fact, when used in combination with token-level information, they can shed light on the consequences of the accumulation of utterance-specific properties for lexical processing. For example, Seyfarth (2014) found that "Words that usually appear in predictable contexts are reduced in all contexts, even those in which they are unpredictable." Similar cumulative effects, e.g. of the positions a word most often occurs in, on its production even in other contexts have been also been reported in Sóskuthy and Hay (2017); Brown et al. (2021).

The choice between analyses based on tokens and analyses based on types requires careful reflection on the goal of the analysis. For example, if the goal is to predict word duration, then token-based models taking into account pre-pausal lengthening and local speech rate may be perfectly satisfactory for many purposes. But these predictions will be blind to processes that may nevertheless be operative in human language production. Token-based models can also help address some research questionss about learning and development: For such questions, one may actually want to embrace that high-frequency words are represented in the data proportionally to their frequency: During learning, word types are encountered proportional to their frequencies; as a consequence, the cognitive system receives its fine-tuning predominantly from the higher-frequency words. On the other hand, if the goal is to evaluate models of lexical retrieval, then type-based models are often called for. Type-based models can be based on aggregated data, as in our analyses, or they might be based on sampling a fixed number of tokens for each type, so as to keep the data from being overwhelmed by high-frequency words (see, for an example, Pluymaekers et al., 2005). It is beyond the scope of the present study to evaluate these specific strategies. Rather, our goal is to point out that the ability to model large sets of observations has led to a proliferation of models that may actually run counter to the goals of psycholinguistic research.

## 5.2 Comparing discrimination-based measures with classical lexical-distributional predictors

We now turn to our third set of issues, the comparison of the statistical models based on localist predictors and DL predictors, and the consequences of this model comparison for linguistic theory. Given the issues with token-based models we just discussed, the discussion is based on the type-based models.

We first note that the model with DL predictors provides a fit with a substantially better AIC (-231.7 and -256.3, evidence ratio 219,696). Moreover, the effects of the predictors in the DL model are linear, whereas predictors show more complex non-linear effects and a non-linear interaction in the model with localist predictors. Thus, the DL framework provides not only a better fit, but also a simpler explanation of spoken word duration. This indicates that there are demonstrable advantages to bringing learning and distributional semantics together in models of the (mental) lexicon. It is also clear that form is pervasively influenced by meaning, a finding that fits well with other recent work on acoustic duration (Tomaschek et al., 2019; Chuang et al., 2021). In what follows, we first reflect on the similarities and differences between localist and DL explanations of spoken word duration. We then offer some thoughts on how to adjucate between the two theories, given the present findings.

A challenge for localist models is how to understand the effect of the Semantic Support for words' forms. Although semantic effects can be incorporated in localist theories, Oppenheim et al. (as demonstrated in 2010), it seems to us that the precision offered by recent advances in distributional semantics may be beyond the reach of models building on symbolic representations of word meanings or semantic features. The challenge faced by the DLM, on the other hand, is how to understand predictors such as frequency of occurrence, phonological neighborhood density and biphone probability.

Starting with frequency of occurrence, four points are worth noting. First, a frequency measure that takes into account cue competition with other words, and that in this sense is learnable, makes basically the same kind of predictions as localist frequency of occurrence: mean spoken word duration increases with information load, and the same holds for the variance of spoken word duration. Second, our data suggest that, possibly, the relation between duration and frequency (or information) becomes simpler once cue competition during word learning is taken into account: the often-observed non-linear effect of (log) frequency of occurrence is exchanged for a linear effect of (log) information in the DLM. If this finding replicates, then it may offer a window on understanding why the effect of localist

Table 5: Triphones of two phonological neighbors.

| | | | | | |
|---|---|---|---|---|---|
| hands | #ha | han | and | nds | ds# |
| lands | #la | lan | and | nds | ds# |

frequency of occurrence is non-linear, and levels off for higher frequencies: high-frequency words co-occur with more other words, hence suffer more cue-competition. This results in a reduced information load, and hence shorter spoken word duration.

Third, in principle, frequency also plays a role during the learning of the mappings between form and meaning, and meaning and form. Words that are encountered more often are learned better. For technical reasons (computational complexity and lack of adequate training data), we have made use of mappings estimated with the regression equations of LDL, which are not sensitive to frequency of use (see Chuang and Baayen, 2021; Heitmeier et al., 2021, for discussion). Importantly, replacing learning in the limit of experience with incremental learning is expected to generate stronger semantic support for form for higher frequency words, and hence to longer spoken word durations (see also Kuperman et al., 2006), as well as more distinct articulation (see Tomaschek et al., 2018b), for higher-frequency words. Thus, we have two opposing effects of frequency on spoken word duration: a higher 'syntactic' frequency (tempered by cue competition with other words in the sentence) corresponds with a lower LogId, which in turn predicts shorter durations, whereas a higher 'lexical' frequency (tempered by cue competition with other triphones within the word, all competing for meaning) predicts longer spoken word durations (see also Tomaschek et al., 2021b).

The final issue related to frequency of occurrence concerns the frequency inheritance effect in localist models (Jescheniak and Levelt, 1994; Dell, 1990). Our regression model with localist predictors led us to conclude that homophones have their own frequency signatures, consistent with earlier findings (Gahl, 2008; Lohmann, 2018b). This conclusion is further strengthened by the fact that when lemma frequency is replaced with the summed lemma frequencies of the homophone twins, the AIC of the GAM increases (from -231.7 to -188.8, for the type-based model). However, the same model also suggests that words of low frequency with very high frequency homophones were shorter than their own frequency would lead one to expect. Under the assumption that greater frequency gives rise to shorter duration, this would imply that low-frequency homophones inherit, albeit incompletely, and not consistently across the full range of lemma frequencies, part of the frequency of their homophone twins. If this is correct, then our results are consistent with localist proposals suggesting multiple loci of "lexical" frequency: lemmas, i.e. holistic representations of word meaning, as well as phonological forms (Kittredge et al., 2008; Antón-Méndez et al., 2012). In the DLM, the issue of frequency inheritance does not arise. However, there is an effect that could be described as "semantic inheritance", a kind of mirror image of frequency inheritance: The greater the semantic similarity of homophones, the stronger the support for their forms, leading to longer spoken word durations. Below, we discuss this effect in further detail.

Next consider phonological neighborhood density (pnd). Within the dlm, neighborhood effects play out primarily during vocabulary learning. Consider the phonological neighbors *hands* and *lands* and their triphones, as listed in Table 5. The three triphones shared by the two words cannot discriminate between their meanings in comprehension. The burden of doing so is carried by the first two triphones. As these triphones also occur in other words, their discriminatory potential is shaped also by these other words. The comprehension mapping finds weights from triphones to semantic dimensions that take all partial similarities between words into account in a principled way. However, the more similar words are, the more frail their mappings become. The same holds for production. This is illustrated for the present data by the effect of the *semantic* similarity of the two homophones: more similar homophones have longer durations, a straightforward consequence of a reduction in the 'stress' of the mapping from meaning to form. Further discussion of this frailty in the mappings can be found in the simulation studies of Chuang et al. (2020a) on multilingual learning. Importantly, since phonological neighbors are similar in form (they are, in a sense, near-homophones), they receive reduced support from the semantics. That prediction is borne out for our data: Words with many phonological neighbors are characterized by reduced semantic support for their forms ($r = -0.38, t(407) = -8.3, p < 0.0001$), and accordingly, their spoken word durations are predicted by the dlm to be shorter. Thus, both localist models and the dlm predict that increased pnd should be associated with shorter spoken word duration, though for different reasons.

Adjucating between the localist and learning theories is not straightforward. Localist theories offer theoretical transparency, but building on localist units has its price, especially when it comes to semantics. The discriminative lexicon model (dlm) is an attempt to work out in mathematical detail how fine-grained high-dimensional semantic representations based on large corpora and numeric representations of words' forms relate to each other. The dlm seeks to keep the mathematics of the mappings between meaning and form as simple as possible, by using the core ideas behind multivariate multiple regression, in an attempt to maintain clarity of understanding model performance. What we have shown is that once learning is taken into account, well-known effects can be considered in a new light, and novel effects (such as Semantic Support for Form) emerge. We therefore believe that the dlm model provides a useful tool for understanding the lexicon and lexical processing specifically from a learning perspective.

There are many open questions that require further research in connection with the dl model. In this study, we made use of semantic vectors that are based entirely on written corpora. These vectors are problematic in several ways. First, they do not incorporate information that is not available in text, such as information about shape, size, and color. Improved vectors can be obtained by merging in information from vision (see, e.g., Shahmohammadi et al., 2021). Second, the semantic vectors of orthographic homophones are identical. Context-sensitive word embeddings (Devlin et al., 2018) may help alleviate this problem. Finally, current embeddings are trained on corpora of billions of words, that exceed by far the amount of experience with language that a given speaker can ever encounter. In short, there is considerable room for improvement for the semantic representations. There is also room for improvement of the form representations. In this study, we built form vectors around triphones, but form vectors can also be constructed for diphones or syllables (see Heitmeier et al., 2021, for detailed discussion). Sequenced triphones can be brought together into syllables, just as in the model of Dell (1986a) phones are assigned to syllabic templates. However, the real challenge is to find a mapping from semantic vectors to the time series of articulatory parameters that drive articulation.

## 5.3 Conclusion

This study contrasted localist theories with theories grounded in discriminative learning. The discriminative lexicon model (Baayen et al., 2019b; Chuang and Baayen, 2021) provides a natural framework for integrating distributional semantics into theories of the mental lexicon. A regression analysis using predictors building on discriminative predictors provided a tighter fit to the duration data, while providing a simpler model. The amount of support that a homophone's form receives from its semantics emerged as a strong predictor for its spoken word duration.

We have also sounded a cautionary note. Large naturalistic corpora have held enormous appeal in light of their ecological validity. The availability of statistical modeling tools capable of handling large, unbalanced data sets has fueled analyses using ever larger sets of observations. Larger is not always better, however. We have argued that data sets dominated by tokens of high-frequency words are in fact liable to masking effects of lexical diversity in some cases and creating spurious effects in others.

The *prima facie* correlation of lexical frequency with word duration caused us to reflect on how frequency of occurrence might play out within the discriminative learning framework underlying LD models. Based on those reflections, we proposed a novel measure that takes into account cue-competition between words in utterances. This 'utterance-based frequency' measure was predictive of duration in the same way as frequency of occurrence in localist models. Importantly, the semantic measure predicted lengthening with increasing 'lexical' semantic support: the integration of distributional semantics within a theory of the lexicon offers new opportunities for studying the relation between form and meaning.

# References

Antón-Méndez, I., Schütze, C. T., Champion, M. K., and Gollan, T. H. (2012). What the tip-of-the-tongue (tot) says about homophone frequency inheritance. *Memory & cognition*, 40(5):802–811.

Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Baayen, R. and Linke, M. (2020). An introduction to the generalized additive model. In

Gries, S. and Paquot, M., editors, *Practical Handbook of Corpus Linguistics*, page in press. Springer.

Baayen, R. H., Chuang, Y., Shafaei-Bajestan, E., and Blevins, J. (2019a). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, 2019:39.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019b). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*.

Baayen, R. H., Kuperman, V., and Bertram, R. (2010). Frequency effects in compound processing. In Scalise, S. and Vogel, I., editors, *Compounding*. Benjamins, Amsterdam/Philadelphia.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482.

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania.*

Baayen, R. H. and Smolka, E. (2020). Modelling morphological priming in German with naive discriminative learning. *Frontiers in Communication, section Language Sciences.* preprint on PsyArXiv, doi:10.31234/osf.io/nj39v.

Baayen, R. H., Vasishth, S., Bates, D., and Kliegl, R. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Baese-Berk, M. and Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and cognitive processes*, 24(4):527–554.

Balota, D. A. and Chumbley, J. I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language*, 24:89–106.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

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

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and sources of Collinearity.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Berndt, R. S., Reggia, J. A., and Mitchum, C. C. (1987). Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers*, 19(1):1–9.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brown, E. L., Raymond, W. D., Brown, E. K., and File-Muriel, R. J. (2021). Lexically specific accumulation in memory of word and segment speech rates. *Corpus Linguistics and Linguistic Theory*.

Burnard, L. (1995). *Users reference guide for the British National Corpus.* Oxford University Computing Services.

Buz, E. and Jaeger, T. F. (2016). The (in) dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience*, 31(3):404–424.

Bybee, J. (1999). Usage-based phonology. *Functionalism and formalism in linguistics*, 1:211–242.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change*, 14(3):261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4):711–733.

Bybee, J. L. (2001). *Phonology and language use.* Cambridge University Press, Cambridge.

Caselli, N. K., Caselli, M. K., and Cohen-Goldberg, A. M. (2016). Inflected words in production: Evidence for a morphologically rich lexicon. *Quarterly Journal of Experimental Psychology*, 69(3):432–454.

Chen, Q. and Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, 119(2):417.

Chuang, Y. Y. and Baayen, R. H. (in press, 2021). Discriminative learning and the lexicon: NDL and LDL. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Chuang, Y.-Y., Bell, M., Banke, I., and Baayen, R. H. (2020a). Bilingual and multilingual mental lexicon: a modeling study with Linear Discriminative Learning. *Language Learning*, page in press.

Chuang, Y. Y., Fon, J., Papakyritsis, I., and Baayen, R. H. (2020b). Analyzing phonetic data with generalized additive mixed models. In Ball, M. J., editor, *Handbook of Clinical Phonetics*, page in press. Routledge.

Chuang, Y. Y., Kang, M., Luo, X. F., and Baayen, R. H. (2021). Vector space morphology with linear discriminative learning. In Crepaldi, D., editor, *Linguistic morphology in the mind and brain*. Routledge.

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2020c). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*.

Cieliebak, M., Dertu, J., Uzdilli, F., and Egger, D. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media (SocialNLP 2017)*, Valencia, Spain.

Clopper, C. G. and Turnbull, R. (2018). Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In Francesco Cangemi, Meghan Clayards, O. N. B. S. and Zellers, M., editors, *Rethinking reduction*, pages 25–72. De Gruyter Mouton.

Cohen, C. (2014). Probabilistic reduction and probabilistic enhancement. *Morphology*, 24(4):291–323.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

Dell, G. (1986a). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, 93:283–321.

Dell, G., Chang, F., and Griffin, Z. (1999). Connectionist Models of Language Production: Lexical Access and Grammatical Encoding. *Cognitive Science*, 23:517–542.

Dell, G. S. (1986b). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3):283.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, 27(2):124–142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes*, 5(4):313–349.

Dell, G. S. and Gordon, J. K. Neighbors in the lexicon: Friends or foes? In Schiller, Niels O. and Meyer, Antje S., editor, *Phonetics and phonology in language comprehension and production: Differences and similarities*.

Deshmukh, N., Ganapathiraju, A., Gleeson, A., Hamaker, J., and Picone, J. (1998). Resegmentation of SWITCHBOARD. In *Fifth international conference on spoken language processing*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dilts, P. C. (2013). *Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression*. PhD thesis, University of Alberta.

Fink, A. and Goldrick, M. (2015). The influence of word retrieval and planning on phonetic variation: Implications for exemplar models. *Linguistics Vanguard*, 1(1):215–225.

Fox, N. P., Reilly, M., and Blumstein, S. E. (2015). Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of Memory and Language*, 83:97–117.

Foygel, D. and Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of memory and language*, 43(2):182–216.

Fricke, M., Baese-Berk, M. M., and Goldrick, M. (2016). Dimensions of similarity in the mental lexicon. *Language, cognition and neuroscience*, 31(5):639–645.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59:127–136.

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

Gahl, S. (2009). Homophone duration in spontaneous speech: A mixed-effects model. *UC Berkeley Phonology Lab Technical Report*.

Gahl, S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the easy/hard database. *Journal of Phonetics*, 49:96–116.

Gahl, S. and Baayen, R. H. (2019). Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. *Journal of Phonetics*, 74:42–54.

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4):789–806.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Goldrick, M., Vaughn, C., and Murphy, A. (2013). The effects of lexical neighbors on stop consonant articulation. *The Journal of the Acoustical Society of America*, 134(2):EL172–EL177.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111:662–720.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., and LaShell, P. (2015). Tracking word frequency effects through 130years of sound change. *Cognition*, 139:83–91.

Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: considerations and design choices. *Frontiers in Psychology: Language Sciences*.

Horton, W. S., Spieler, D. H., and Shriberg, E. (2010). A corpus analysis of patterns of age-related change in conversational speech. *Psychology and Aging*, 25(3):708.

Jescheniak, J. D. and Levelt, W. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(4):824–843.

Jurafsky, D. (2003). Probabilistic modeling in Psycholinguistics: Linguistic comprehension and production. In Bod, Rens, H. J. and Jannedy, S., editors, *Probabilistic linguistics*, pages 39–95.

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

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, 44(1):287–304.

Kilbourn-Ceron, O., Clayards, M., and Wagner, M. (2020). Predictability modulates pronunciation variants through speech planning effects: A case study on coronal stop realizations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11(1).

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Kittredge, A. K., Dell, G. S., Verkuilen, J., and Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *CognitiveNeuropsychology*, 25(4):463–492.

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

Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2006). Morphological predictability and acoustic salience of interfixes in Dutch compounds. *JASA*, 122:2018–2024.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lee, S. J., Cho, Y., Song, J. Y., Lee, D., Kim, Y., and Kim, H. (2016). Aging effect on Korean female voice: Acoustic and perceptual examinations of breathiness. *Folia Phoniatrica et Logopaedica:International Journal of Phoniatrics, Speech Therapy and Communication Pathology*, 67(6):300–307.

Levelt, W., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–38.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6:172–187.

Linke, M., Broeker, F., Ramscar, M., and Baayen, R. H. (2017). Are baboons learning "orthographic" representations? Probably not. *PLOS-ONE*, 12(8):e0183876.

Lohmann, A. (2018a). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics*, 54(4):753–777.

Lohmann, A. (2018b). Time and thyme are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language*, 94(2):e180–e190.

Lohmann, A. and Conwell, E. (2020). Phonetic effects of grammatical category: How category-specific prosodic phrasing and lexical frequency impact the duration of nouns and verbs. *Journal of Phonetics*, 78:100939.

MacWhinney, B. and Leinbach, J. (1991). Implementations are not conceptualizations: revising the verb learning model. *Cognition*, 40:121–157.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Science*, 12(5):176–181.

McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. *arXiv preprint arXiv:2005.08826*.

Middleton, E. L., Chen, Q., and Verkuilen, J. (2015). Friends and foes in the lexicon: homophone naming in aphasia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(1):77.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. Discrimination in lexical decision. *PLoS ONE*, 12(2).

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS-one*, 12(2):e0171935.

Milin, P., Madabushi, H. T., Croucher, M., and Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. PsyArXiv.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *ACL*, pages 236–244.

Moscoso del Prado Martín, F., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94:1–18.

Mousikou, P. and Rastle, K. (2015). Lexical frequency effects on articulation: a comparison of picture naming and reading aloud. *Frontiers in psychology*, 6.

Norris, D. and McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357.

Oppenheim, G. M., Dell, G. S., and Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114(2):227–252.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, 7:101–139.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release).

Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, 118:2561–2569.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4):760–793.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *American Psychologist*, 43(3):151–160.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64–99. Appleton Century Crofts, New York.

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L. and Rumelhart, D. E., editors, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, pages 216–271. The MIT Press, Cambridge, Mass.

Scarborough, R. (2013). Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics*, 41(6):491–508.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., and Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, 54(2):228–264.

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

Shafaei-Bajestan, E., Tari, M. M., P., U., and Baayen, R. H. (2021). LDL-AURIS: Error-driven learning in modeling spoken word recognition. *Language, Cognition and Neuroscience.*

Shahmohammadi, H., Lensch, H., and Baayen, R. H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *CoNLL 2021*. arXiv preprint arXiv:2104.07500.

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.

Sorensen, J. M., Cooper, W. E., and Paccia, J. M. (1978). Speech timing of grammatical categories. *Cognition*, 6(2):135–153.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*.

Sóskuthy, M. and Hay, J. (2017). Changing word usage predicts changing word durations in new zealand english. *Cognition*, 166:298–313.

Tanner, J., Sonderegger, M., and Wagner, M. (2017). Production planning and coronal stop deletion in spontaneous speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8(1).

Tomaschek, F., Arnold, D., Sering, K., Tucker, B. V., van Rij, J., and Ramscar, M. (2020). Articulatory variability is reduced by repetition and predictability. *Language and speech*, page 0023830920948552.

Tomaschek, F., Hendrix, P., and Baayen, R. H. (2018a). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics*, 71:249–267.

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2019). Modeling the duration of word-final s in english with naive discriminative learning. *Journal of Linguistics*. https://psyarxiv.com/4bmwg, doi = 10.31234/osf.io/4bmwg.

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2021a). Phonetic effects of morphology and context: Modeling the duration of word-final s in english with naïve discriminative learning. *Journal of Linguistics*, 57(1):123–161.

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, R. H. (2018b). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, 4(s2).

Tomaschek, F., Tucker, B. V., Ramscar, M., and Baayen, R. H. (2021b). Paradigmatic enhancement of stem vowels in regular english inflected verb forms. *Morphology*, 31(2):171–199.

Tucker, B. V., Sims, M., and Baayen, R. H. (2019). Opposing forces on acoustic duration. *PsyArXiv*.

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2020). itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.4.

Vitevitch, M. S. and Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

Warner, N. (2011). Reduction. *The Blackwell companion to phonology*, pages 1–26.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, pages 96–104.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, 70:86 –116.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman & Hall/CRC, New York, 2nd edition.

Wright, R. (2004). Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology VI*, pages 75–87.

Yap, M. J. and Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, 60(4):502–529.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of the Eighth International Conference on Acoustics, Speech, and Signal Processing*, pages 5687–5690.