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

Susanne Gahl[1] and R. Harald Baayen[2]

1: Department of Linguistics, University of California at Berkeley
2: Seminar für Sprachwissenschaft, Eberhard-Karls University of Tübingen

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

Abstract

Effects of word frequency on spoken word duration are well-documented and have long informed theories of the mental lexicon. In this study, we discuss the two theoretical constructs 'frequency' and 'word' that are implicated by the notion of lexical frequency, in light of recent models of the lexicon which do not contain stable, discrete lexical representations, and in which lexical frequency therefore does not have any place. We compare two approaches (localist spreading-activation vs. discriminative learning (DL) models integrating distributional semantics) by assessing regression models of spoken word duration of English homophones grounded in each. We further show that the relationship between a homophone's form and its semantics is predictive of its duration, consistent with predictions of the DL-based model.*

*The authors would like to thank three anonymous reviewers and the associate editor for their exceptionally thoughtful comments.

Keywords frequency effects, spoken word duration, homophones, mental lexicon, discriminative lexicon model, form-meaning isomorphy, distributional semantics

1. INTRODUCTION. Effects of lexical frequency have for decades had the status of consensus findings in language processing, variation, and historical change (see e.g. Jurafsky 2003; Bybee 2003 for overviews). One such effect concerns spoken word duration: Many researchers have observed that frequent words have shorter durations than infrequent ones (see e.g. Bybee, 2003, 1999, 2002b; Pierrehumbert, 2002; Jurafsky *et al.*, 2001; Bell *et al.*, 2003; Warner, 2011, for discussion). The shortening of frequent words appears to have gained the status of a pre-theoretical observation that competing theories attempt to explain. The ubiquity and robustness of word frequency effects of all kinds may in fact create the impression that lexical frequency itself is a pre-theoretical concept. But the very notion of lexical frequency as a property of words implicates two theoretical constructs: First, the mental lexicon as a 'store of words in long-term memory' (Jackendoff p.130 2002, cited in Elman 2009), and second, frequency as a property of items in such a lexicon.

Conceiving of the mental lexicon as a store of words is not a matter of consensus, however. For example, Elman (2009) argues, on the basis of numerous observations in the psycholinguistic literature, that a lexicon as a 'store of words' is neither feasible, nor necessary: Lexical knowledge is inextricably linked to rich, context-dependent, and ever-changing information, rendering lexical 'entries' infeasible as models of lexical knowledge. Elman (2009) goes on to describe a family of models in which "there is no data structure that corresponds to a lexicon. There are no lexical entries." (Elman, 2009, p.556). In what follows, we refer to models that imply a mental lexicon with words as identifiable, stored, stable units as 'localist', and models without such units as 'distributed'.

It is true that, as Elman (2009) puts it "[e]liminating the lexicon is (...) radical surgery, and it is an operation that at this point many will not agree to". But models entailing that operation have been gaining ground. In a fairly recent class of models, henceforth DISCRIMINATIVE LEARNING (DL) models, such as the Discriminative Lexicon Model (DLM, Baayen *et al.*, 2019; Chuang and Baayen, 2021; Heitmeier *et al.*, 2021), words do not have an existence as representational units within the lexicon. In these models, there are no lexical entries. The DL model makes use of distributed representations for both form and meaning, using word embeddings (see, e.g. Landauer and Dumais, 1997; Mikolov *et al.*, 2013) for the latter. The model makes use of artificial neural networks to predict meanings from forms, and forms from meanings. These networks are constantly updated with experience and, collectively, constitute a body of lexical memory that evolves over time. However, the distributed representations for individual forms and meanings are ephemeral. These short-lived representations come into fleeting existence either as a result of external input (e.g. forms as the starting point for comprehension) or from internal conceptualization (intended meanings that drive production). Neither forms

nor meanings constitute stable units or collections of units in the model.

Lexical frequency effects would seem to pose a challenge for distributed models, because frequency effects are often considered *prima facie* evidence for stored representations. For example, Stemberger and MacWhinney (1986), commenting on a processing advantage for high-frequency vs. low-frequency word forms, conclude that "[t]hese data are best explained by assuming that high frequency inflected forms are stored as separate entries in the lexicon." Along similar lines, frequency effects have been considered evidence for holistic storage of such elusive linguistic units as multiword expressions (Bybee, 2002a; Tremblay *et al.*, 2011; Arnon and Snider, 2010) and syllables. For example, Cholin (2011, p.225) notes that "Because only stored units are expected to exhibit frequency effects, effects of syllable frequency provide strong evidence for the assumption that syllables are (separately) stored units." If frequency effects serve as a diagnostic of what is stored in memory, then the ubiquity of lexical frequency effects would appear to doom distributed models of the lexicon from the start. One goal of the current study is to demonstrate that distributed models can in fact account for effects usually thought to necessitate stored lexical representations.

1.1. LEXICAL FREQUENCY EFFECTS WITHOUT LEXICAL ITEMS. At this point, it is worth asking what it is that measures of lexical frequency actually capture. The success of frequency estimates as predictors of lexical processing invites the conclusion that frequency – the number of times a word has been encountered – is the essence of such estimates. But from a different point of view, lexical frequency may reflect a combination of many facts about a language and its speakers, rather than a property of individual words. Viewed in this way, frequency effects are composite effects somewhat analogous to white light as the combination of colors on the visible spectrum. We are not suggesting that frequency — or indeed white light — is epiphenomenal. Rather, separable components of frequency may be predictive of distinct patterns. The seemingly pre-theoretical term 'frequency' does not reflect a tally of usage events, but rather a complex set of distributional facts in the spectrum of language use.

In the current study, we highlight two such facts. The first concerns practice over the life of a talker, or 'learning'. The second concerns the degree of 'contextual independence' of word meanings in utterances, i.e. the degree to which word meanings are stable across the utterances they occur in. Unlike frequency as a property of lexical items, learning and contextual independence are integral to the conceptual fabric of distributed models. Therefore, if learning and contextual independence can account for patterns usually attributed to frequency, then such patterns do not entail lexical storage and consequently do not pose a challenge to distributed models of the lexicon.

Interpretations of frequency effects on spoken word duration have explicitly appealed to 'practice' as a source of frequency effects. For example, Bybee (Bybee 2002a, 2003 *et alibi* and Newmeyer 2003) point to articulatory routinization as a source of shortening of highly practiced sequences. An additional consequence of practice may be a decrease in the variability of on motor movements (Tomaschek *et al.*, 2021a; Gahl and Baayen, 2019), a point we return to in section 1.3 below.

Prior work has also established beyond doubt the role of utterance context in word duration, as we discuss in greater detail below. "Contextual independence", in the sense of the term that we are proposing in the current study, has not previously been considered as a predictor of word duration. Measures taking the interplay of meaning and variability of contexts into account have, however, been found to be predictive of lexical decision times (see e.g. McDonald and Shillcock, 2001). The reference to contextual information may bring to mind a line of research investigating spoken word duration (and other aspects of linguistic form) using concepts such as surprisal, information entropy, redundancy, and other tools of information theory (Shannon, 1948; see e.g. Levy, 2008; Hale, 2003; Fenk and Fenk, 1980; Van Son and Pols, 2003; Van Son and Van Santen, 2005; Aylett and Turk, 2004). As we explain below, contextual independence differs in important respects from surprisal and is not intended to approximate or replace that concept. Tools of information theory are nevertheless relevant here, in that several proposals have explicitly linked the effect of lexical frequency to a broader pattern of predictability whereby words are phonetically reduced if their occurrence in a given context is highly predictable. For example, (p.229 Jurafsky *et al.*, 2001) propose that "word forms are reduced when they have a higher probability. (. . .) This proposal thus generalizes over earlier models, which refer only to word frequency". Similarly, Aylett and Turk (2004), in an analysis of syllable duration, treat frequency as one of several measures of 'redundancy', i.e. predictability in a given string of speech (see also Bell *et al.*, 2003; Pluymaekers *et al.*, 2005 *et alibi*).

While informativity and frequency both in some way relate to predictability, their effects are separable. Several studies have teased apart effects of lexical frequency and information load. For example, Seyfarth (2014) found that low informativity was associated with shorter word durations even when frequency and local contextual predictability were controlled for. Similarly, Griffin and Bock (1998) found that the effect of lexical frequency on spoken word production was attenuated when target words were predictable, given prior sentence context. Griffin and Bock (1998)'s observations pertained to naming latencies, not word durations; but the interpretation of their findings implicated two distinct aspects of lexical production (lexical selection and phonological encoding) that have in turn been found to be predictive of word duration. Therefore, the findings speak to the separability of different facets of predictability and, potentially, of word frequency. They also highlight a theoretical model of lexical processing

as a component of language and speech production. Such models are the topic to which we now turn.

1.2. CONNECTING MODELS OF THE MENTAL LEXICON TO SPOKEN WORD DURATION. A long line of research has examined spoken word duration in connection with psycholinguistic models of lexical access and retrieval (see e.g. Balota and Chumbley 1985; Shields and Balota 1991; Kahn and Arnold 2012). Typically, the connection between such models and spoken word duration is indirect. And yet, the connection should come as no surprise. There is a long research tradition using behavioral data such as response latencies or eye movements to draw inferences about the speed of mental processes and test predictions of models of the lexicon without attempting to model hand movements, saccades, or speech initiation (see Levelt 2013). The logic of this line of research is that models of the lexicon yield predictions about target accessibility, which in turn is thought to be reflected in reaction times, gaze direction, finger movement, and so on. The same logic applies to research relating models of the lexicon to spoken word duration and many other aspects of pronunciation. Statistical models of lexical decision times, naming latencies, and other behavioral measures have long been used as a means to test competing models of the lexicon. The current study continues that line of research, fitting statistical models of spoken word duration using variables grounded in localist models and in a distributed one. We consider a variable to be 'grounded' in a model if it is inherent in the architecture of the model or at least readily accommodated in its implementations. We return to this point in section 2.1 below. Before we do so, we turn to the empirical domain modeled here.

1.3. DO 'HOMOPHONES' SOUND IDENTICAL?. English homophones have served as a natural experiment for theories of lexical access and retrieval (Ferreira and Griffin, 2003), as well as for theoretical accounts of phonetic detail in pronunciation. With regard to the latter, explanations for the shortening of frequent words have appealed to "late" stages of lexical production, such as phonological encoding and/or articulatory routinization (as hinted in the preceding discussion), as well as 'early' stages of language production that precede articulation, such as lexical retrieval (see e.g. Bybee, 2006; Bell *et al.*, 2009). Many of these proposals are mutually compatible – spoken word duration undoubtedly reflects multiple factors – but their predictions about homophones partially diverge. Gahl (2008) argued that, if the shortening of frequent words solely reflected "late" stages of lexical production, such as articulatory routinization, then a low-frequency word such as *thyme* should have the same duration as a high-frequency homophone twin *time*, other things being equal. If, on the other hand, word duration also reflected "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)

found that spoken word durations of homophones in the Switchboard corpus of telephone conversations differed 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*). Similar results, consistent with the idea that homophone twins can differ in pronunciation as a function of the frequency of each member of the pair, have been reported and discussed in Lohmann (2018a); Phillips (2020); Luef and Sun (2020); Conwell (2018).

The model in Gahl (2008) (henceforth "G2008", and a follow-up analysis, Gahl 2009, henceforth "G2009") did not go unchallenged. Lohmann (2018b) argued that neither G2008 nor G2009 provided direct evidence of any effect of the target-specific frequency (e.g. of *time* being shorter than *thyme*). Instead, according to Lohmann (2018b), the models in G2008 and G2009 only reflected frequency differences across word forms, e.g. the difference between the frequency of *time* vs. *sage*. This characterization is, we believe, misleading. Briefly, the characterization neglects the role played by homophone twins in the statistical models: The model in G2008 asserts that target frequency is predictive of duration when controlling for the duration of a target homophone. Thus, the frequencies of *sage* and *time* are only relevant for predicting the duration of *time* vs. a hypothetical homophone of *time* that had the frequency of *sage*. We concur with Lohmann (2018b) that G2008 and 2009 had serious methodological flaws, however. For example, as pointed out in Lohmann (2018b), the fairly weak correlation between the duration of low-frequency and high-frequency homophones need not indicate lemma-specific lexical characteristics, but may simply reflect uncertainty of estimates based on small numbers of tokens: The reliability of average duration as an estimate of a word's 'true' duration decreases with word frequency. Another issue is that several assumptions underlying the regression models in Gahl (2008, 2009) (and Lohmann 2018b) were violated. One is the assumption of a linear relationship between the predictors and the outcome in the model; another is the homoskedasticity assumption, i.e. the assumption that the variability in the model residuals is equal across the range of each of the predictors. That assumption is inevitably violated, due to the relationship between frequency and variability just mentioned: the smaller the sample size, the higher the variance. Sample sizes are (naturally) smaller for low-frequency words than for high frequency words. Therefore, the model residuals are bound to be smaller with increasing frequency. As a result, the model estimates do not support the conclusions, or they only do so in a hypothetical world in which the modeling assumptions are met.

Complicating this picture is the fact that variability in articulatory movements, and hence in the phonetic realization of words, may in fact be shaped by 'practice', i.e. an aspect of learning. (Tomaschek *et al.*, 2021a, 2018a), for example, have argued that high usage frequency

leads to reduced variability in motor movements. If this is correct, then high variability in token duration of low frequency words may reflect high variability in motor execution, in addition to prediction uncertainty due to small sample size. Dependencies between predictors and variability pose a problem for any model assuming constant variance of model residuals, such as those in G2008, G2009, and Lohmann (2018b). In the current study, we reanalyze the data analyzed previously in G2008, G2009, and Lohmann (2018b), but now making use of Gaussian Location-Scale Generalized Additive Mixed Models (GAMM, see Wood, 2017). These models do better justice to non-linear relations between predictors and outcome, and they are able to model the variance along with the mean, rather than just the mean of the outcome.

1.4. THE ROLE OF SEMANTICS IN SPOKEN HOMOPHONE DURATION. Another shortcoming of the studies of homophone duration just discussed is that they make no mention of semantics, beyond saying that homophones 'differ in meaning'. This broad-brush treatment of word meaning runs counter to facts and intuitions about homophone pairs. For example, there is evidence that emotional tone of voice (whether a word is spoken in a cheerful, neutral, or sad manner) affects listeners' interpretations of homophones differing in emotional valence, such as *bridal* and *bridle* (Nygaard *et al.*, 2009; Nygaard and Queen, 2008). Another factor that is not considered in the models concerns the semantic similarity of homophones. Some homophones seem to be semantically similar, as reflected in spelling uncertainty (e.g. for pairs like *principle, principal* and *drier, dryer*) that might reflect lexical mergers in individual speakers' understanding of these items; others are clearly dissimilar, such as *time, thyme; paws, pause*, or *hoarse, horse*, sometimes giving rise to deliberate puns. A prediction following from DL, which we introduce in greater detail below, is that the degree of semantic similarity should make itself felt in the way (near-)homophones are pronounced: Semantically similar homophones should be more similar in duration, and be longer in duration, than dissimilar ones, other things being equal. More broadly, as we discuss below, DL-based models entail that meaning and form are linked in a manner resulting in a degree of form-meaning isomorphy.

1.5. THE CURRENT STUDY. To summarize: In the current study, we examine the ability of a distributed model (i.e. a 'lexicon without words') to capture the relationship between lexical frequency and spoken word duration. We compare statistical models of spoken word duration using variables grounded in localist (Dell, 1986; Levelt *et al.*, 1999; Schwartz *et al.*, 2006) and distributed models (the discriminative learning (DL) model, Baayen *et al.*, 2019; Chuang and Baayen, 2021; Heitmeier *et al.*, 2021)).The general prediction is that the DL-based statistical model recovers semantic effects on spoken word duration, as well as multiple, separable facets of what are commonly thought of as 'word frequency effects'.

The seemingly simple dichotomy 'localist' vs. 'distributed' belies a complex landscape

of models of the lexicon. To understand the specific questions asked and comparisons made in the current study, we must provide background on models of the lexicon and on variables grounded in the specific models we compare. We do so in section 2, before introducing the theoretical innovations and specific predictions (section 3), and methodological choices (section 4) of the current study. The empirical results and discussion of their implications form sections 5 and 6, respectively.

2. BACKGROUND.

2.1. OVERVIEW. Models of the lexicon as models of linguistic structure are as varied as those of grammar, ranging from lists of lexical idiosyncrasies (e.g. Chomsky 1995) to analyses in which lexicon and grammar are of a piece (e.g. Fillmore *et al.* 2003). Models of the lexicon as models of lexical processing are similarly richly varied. Some models conceive of the lexicon as a single, central repository of words that is tapped by different modalities (such as speaking, auditory, visual, and tactile comprehension, reading, writing, and so on). Alternatively, there may be multiple lexicons that are specific to certain modalities or tasks. Models of the lexicon are not simply models of collections of words, but also of how words relate to one another, to linguistic structure generally, and to other aspects of cognition.

The aspect of architecture of computational models of the lexicon that is critical to the current study concerns the distinction we are drawing between 'localist' vs. 'distributed' models. Questions like 'What word is this?' or 'Where is the word *cat*?' or 'What is the frequency of this word?' can only be meaningfully put to models in which there are 'loci' for individual words and their properties. The presence or absence of such loci leaves room for different model architectures, however. Evaluating our claims about a mental lexicon without words requires an understanding of theoretical models of the lexicon more broadly. We therefore start by laying out the landscape of such models.

2.2. THE LANDSCAPE OF MODELS OF THE MENTAL LEXICON.

BASIC ARCHITECTURE. The landscape of psycholinguistic models can be roughly divided into two kinds: lexical network models on the one hand and artificial neural network models on the other; the latter class can in turn be divided into connectionist and vector-space models. Most of the models we discuss here, including the DL-model, concern the 'computational' level of description (Marr, 1982): They are models of computational tasks that cognition is solving, rather than of its algorithms or implementation.

Lexical network models conceive of the lexicon as sets of interconnected nodes (see Diessel 2017; Siew *et al.* 2019 for overviews). Each node represents a word (or, in morpheme-based models, a morpheme); the connections between nodes represent semantic, morphological, phonological, or co-occurrence relationships between pairs of words. Fodor (1983), for example, describes

the lexicon as "a sort of connected graph, with lexical items at the nodes and with paths from each item to several others." In one class of network models ('multiplex networks'), any given word may appear in multiple networks, connected to other words within and across networks. Lexical network models are 'localist' by definition: 'Where is the word *cat*?' or 'Which word does this node represent?' are felicitous questions as applied to such models.

Connectionist artificial neural network models also involve interconnected nodes. In one subset of connectionist models, all nodes represent units of linguistic structure, such as semantic/syntactic units (known as 'lemmas'), phonological word forms ('lexemes'), syllables, or phonological segments. Examples of such models include McClelland and Rumelhart (1981); McClelland and Elman (1986); Levelt *et al.* (1999); Dell (1986); Norris (1994); Harm and Seidenberg (1999) and Norris and McQueen (2008). In a second subset of connectionist models, there exist 'hidden' layers whose nodes need not correspond to elements that would be discoverable through linguistic analysis. Many connectionist models (e.g. Dell 1986) are 'localist', in that there exist units representing word meanings, forms, and perhaps internal elements such as morphemes or syllables (see e.g. Goldrick, 2006 for discussion). Other connectionist models (e.g. Plaut 1997 and Dell *et al.* 1993) do not envision model-internal lexical representations and are thus non-localist, i.e. 'distributed' (see appendix A.4 for discussion of differences between these "subsymbolic" connectionist models and the model used in the current study).

In vector-space models, finally, information about words emerges from the properties of vectors in multidimensional spaces. The Discriminative Learning (DL) model used here involves several matrices of vectors. None of the vectors, vector elements, or matrices of vectors constitute words. Every encounter with a word form, or (in the case of pseudowords or 'nonce forms') with a potential word form, results in an update of the matrices, not only for the encountered form, but throughout the model. Words, i.e. form-meaning pairings, do not have stable model-internal counterparts.

In the DL model implementation used here, form vectors are binary vectors coding the presence and absence of triphones. There is no claim that either the triphones or the vectors coding their presence in a given signal are 'items' that are stored in human memory. Rather, the triphones represent contrasts in the speech signal to be comprehended or produced; their only function is to establish relationships between sounds and other aspects of thought or experience, which are encoded in the semantic vectors. In fact, the model does not distinguish between actual words and 'pseudo-words' ('non-words' or 'nonce forms'). This (possibly counterintuitive) property of the model correctly predicts that production and perception of pseudo-words engage linguistic knowledge, as we demonstrate in a different set of papers (Chuang *et al.*, 2021b; Cassani *et al.*, 2020; Heitmeier *et al.*, 2023).

There are similarities across lexical networks and some connectionist models, as well

as some connectionist and vector space models. Network models and localist connectionist models have in common that connections among words are used to model strength of connectedness, i.e. the degree to which accessing any one word also causes other words to be accessed. Distributed connectionist and vector-space models have in common that they characterize word learning as incremental association and differentiation processes shaping a mapping between input and output 'layers'. One of these layers consists of elements of forms (e.g. segments, triphones, or – in localist connectionist models – word forms), and one holds conceptual information (e.g. semantic features, word meanings, or distributed semantics). Both classes of models generate predictions about the degree of 'activation' or 'strength' of a given outcome pattern (form) based on a given input pattern (meaning). From this perspective, implementations of distributed connectionist models and the DL model can be quite similar. However, there are also substantial differences between distributed connectionist models and vector space models. We comment on these differences in appendix A.4. Here, we only point out that the implementation of the DL model we use here assumes that semantic vectors and form vectors are located in the same vector space (hence the term 'vector space model') and can be related to each other with linear mappings. Differences and similarities across models are often both more subtle and more far-reaching than they might initially appear.

MODELING FREQUENCY EFFECTS.

In localist models, frequency effects have been modeled by means of several different mechanisms. These include properties of (a) nodes in networks (e.g. resting activation levels, either specified by the researcher or 'learned', i.e. acquired by the model itself), (b) connections between nodes (e.g. connection weights between form and meaning), and (c) criteria deciding which nodes and connections result in items being retrieved, such as activation thresholds or beam width (see e.g.Levelt *et al.* 1999 and Dell 1986 for models of language production, and Jurafsky 1996 for an overview). Any of these mechanisms can capture processing advantages of high-frequency items.

In distributed models, frequency cannot be represented as an item-specific property of words, as a consequence of the absence of word units. There are several approaches to modeling effects of frequency (or effects typically attributed to frequency) in such models. One approach is to attribute 'frequency' effects to lexical properties that are correlated with frequency. (see for example Baayen *et al.*, 2016). Another approach, the one we take here, is based on the idea that frequency estimates may be composite measures reflecting several separable measures. We describe our implementation of the argument we make in section 3 below. Before we can do so, we must describe the components of DL.

2.3. MODELING PRODUCTION WITH THE DL MODEL-ARCHITECTURE. We have described the basic idea behind vector-space models and situated them in the landscape of models. We now provide a more concrete description of the vectors and matrices that make up the DL model. The variables for the statistical models in the empirical portion of the present study are calculated on the basis of these matrices.

Up to this point, we have referred to 'layers' of form and meaning, and to high-dimensional vectors for form and meaning, and hinted that the links between them are 'learned' without saying what the links are or how they are quantified. Here, we describe this process for a toy 'world' in which the learner has encountered the words *time, thyme*, and *lime* and is connecting the triphones occurring in them to semantic vectors. In computational linguistics, semantic vectors are generally referred to as word embeddings. Embeddings encode lexical distributional and collocational properties and are calculated from large corpora. For a general introduction to vector semantics, the reader is referred to Jurafsky and Martin (2019). In a more realistic scenario, a model capable of relating forms to conceptual information contains many more semantic dimensions (empirical embeddings typically have several hundred dimensions) and many more triphones.

For the purposes of this example, the semantic vectors are populated with arbitrarily chosen numbers. Because *time* and *thyme* have different meanings, they have been assigned different semantic vectors. The semantic vector for *time* is boldfaced. The semantic vectors are brought together as the row vectors of a 3×2 semantic matrix $\boldsymbol{S}$, as follows:

$$\boldsymbol{S} = \begin{array}{c} \\ \text{TIME} \\ \text{LIME} \\ \text{THYME} \end{array} \begin{array}{c} \begin{array}{cc} S1 & S2 \end{array} \\ \begin{pmatrix} \mathbf{0.1} & \mathbf{0.3} \\ 0.6 & 0.2 \\ 1.1 & 0.6 \end{pmatrix} \end{array}.$$

The vectors serving as representations for forms specify which triphones are present in a form, using 1 to denote presence and 0 to denote absence. (For expository reasons, the diphthong aɪ is represented here as a sequence of two phones; however, in the actual model that we constructed for the empirical data, diphthongs were represented as single phones.) The form vectors that are present in *time* and *thyme* are boldfaced. As *time* and *thyme* share the same triphones, their form vectors are identical. There are in all 6 different triphones in the three words of our lexicon, and the form matrix $\boldsymbol{C}$ therefore is a 3×6 matrix $\boldsymbol{C}$:

$$\boldsymbol{C} = \begin{array}{c} \\ \textit{time} \\ \textit{lime} \\ \textit{thyme} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{\#ta} & \text{taɪ} & \text{aɪm} & \text{ɪm\#} & \text{\#la} & \text{laɪ} \end{array} \\ \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \end{pmatrix} \end{array}.$$

Readers familiar with earlier models using triphone-derived representations, such as the past-tense model of Rumelhart and McClelland (1986), may wonder if the DL model is able to avoid problems with such models pointed out in Pinker and Prince (1991). Briefly, the algorithms in implementations of the DL model do avoid these problems, as evidenced by the model's predictions about speech errors (see appendix A.6 for further clarification). In the present study, we restrict our attention to an early stage of the production process. The estimated form vectors reflect how well triphones are supported by a word's meaning and, therefore, which triphones are most likely to be needed, but they remain silent about the ordering of triphones. A subsequent algorithm places triphones in the order required for pronunciation, as input to articulation (for details and implementations, see Baayen *et al.*, 2019).

The 'link' between semantic vectors and form vectors is a mapping: We are interested in a mapping $\boldsymbol{G}$ that transforms, as precisely as possible, the semantic vectors in $\boldsymbol{S}$ into the form vectors in $\boldsymbol{C}$. A way to obtain such a mapping is to solve, using matrix algebra:

$$\boldsymbol{SG} = \boldsymbol{C}. \tag{1}$$

Figure 1 spells out equation 1 for our example matrices, and clarifies how the values in $\boldsymbol{C}$ relate to the values in the semantic matrix $\boldsymbol{S}$ and the mapping matrix $\boldsymbol{G}$. Appendix A.1 provides technical details on the way to solve the equation for $\boldsymbol{G}$. For the present data, the resulting mapping matrix $\boldsymbol{G}$ is

$$\boldsymbol{G} = \begin{array}{c} \\ \textit{S1} \\ \textit{S2} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{\#ta} & \text{taɪ} & \text{aɪm} & \text{ɪm\#} & \text{\#la} & \text{laɪ} \end{array} \\ \begin{pmatrix} -1.19 & -1.19 & -0.08 & -0.08 & 1.12 & 1.12 \\ 3.81 & 3.81 & 2.37 & 2.37 & -1.44 & -1.44 \end{pmatrix} \end{array}.$$

PLACE FIGURE 1 APPROXIMATELY HERE

This 'toy' example illustrates the vectors and matrices that together make up the DL model. Additional information on the mapping algorithms can be found in Appendix A.1. Technical details of the model and computational implementation at scale can be found in Heitmeier

et al. (2021, 2024). For software facilitating model implementation, see `https://juliapackages.com/p/judiling`.

WORD FREQUENCY EFFECTS IN THE DL MODEL. With this understanding of the DL model in place, we can return to the question of how word frequency plays out in discriminative approaches to language. Recall that we argued that effects ordinarily attributed to word frequency may result from a combination of many aspects of usage events and their consequences for learning. Considering frequency effects as composite effects does not render it irrelevant how often meanings or forms have been encountered. In fact, individual usage events are key to the workings of the DL model: No two usage events are exactly alike.

We have so far been setting aside the immense variability in the way words are articulated, as well as the context-dependency of meaning – the very rationale for calling into question the lexicon as a store of lexical entries laid out in Elman (2009). We must now turn to the question of how the DL model is shaped by individual usage events.

It is possible to take variability in form and meaning into account fairly directly, by setting up vector space models trained on token-specific forms derived from the speech signal, combined with context-specific word embeddings (see for example Shafaei-Bajestan *et al.*, 2021 for an DL-based comprehension model that takes actual audio tokens as input and Chuang *et al.*, 2024 for a model using context-sensitive embeddings (Devlin *et al.*, 2018)). In the present study, however, modeling and statistical analyses are conducted at the level of word types. Therefore, two things now seemingly stand in the way of the model's accounting for frequency effects as consequences of usage experience: The absence of a 'locus' of word frequency, as well as the type-based nature of the model.

To elaborate on the way the current model is based on types: The procedure for estimating the mapping matrix $\boldsymbol{G}$ is fixed, as a direct consequence of every row vector of the form matrix $\boldsymbol{C}$ and of the semantic matrix $\boldsymbol{S}$) being fixed and unique. As shown in detail by Heitmeier *et al.* (2021, 2024), mappings estimated in this way are blind to the effects of frequency of use on lexical processing. As we discuss below, these types of mappings can be useful for certain purposes. However, for other purposes, they are far from optimal: The model so far effectively assumes that practice is irrelevant. However, the mapping $\boldsymbol{G}$ can also be estimated without making that simplifying assumption, all within the framework of the discriminative lexicon model.

How then does frequency of use come into play in a type-based implementation of the DL model? Given the ephemeral nature of the form and meaning vectors, these vectors themselves cannot be the loci for frequency effects. Any effects of frequency must play out in the mapping. Fortunately, there are two ways in which the mapping $\mathbf{G}$ between semantic and form vectors can be made sensitive to frequency of use. One option is to make use of the learning rule of

Widrow and Hoff (1960) and carry out 'incremental' learning. We comment on that option in appendix A.1. This method presupposes that there is some specific order in which word tokens are encountered during learning. When there is no intrinsic order information available, frequencies of use can be taken into account in a principled way using a recently developed method, "frequency informed learning", proposed in Heitmeier *et al.* (2024) Appendix A.1 provides a formal specification of how to calculate a frequency-informed mapping; for a mathematical proof, see Heitmeier *et al.* (2024).

Incremental learning and frequency-informed learning are ways to make the mapping between semantic and form vectors sensitive to frequency construed as the number of times a given combination of form and meaning has been encountered. However, above we made the point that effects usually attributed to frequency might actually arise from a combination of multiple, separable factors. Frequency-informed learning (henceforth FIL) makes it possible to efficiently model one such factor. In section 3, we discuss a second, independent component of "frequency" and propose a way to model that component, which is also motivated by the general framework of discriminative learning and vector space modeling, applied to the utterance level instead to the word level. Furthermore, we introduce a measure, derived from a vector space model using FIL, that captures the 'paradigmatic' support that a form receives from its semantics.

The idea that the relationship between meaning and form might make itself felt in spoken word duration presupposes an understanding of known predictors of word duration. Therefore, we next turn to prior research on spoken word duration, before introducing ways in which the DL model can serve as the basis for predictions about word duration.

2.4. PREDICTORS OF SPOKEN WORD DURATION. Having introduced the theoretical background for the present study, we now turn to prior research on its empirical focus. Numerous factors have been shown to influence spoken word duration in English (see e.g. Aylett and Turk, 2004; Warner, 2011; Jurafsky, 2003; Fink and Goldrick, 2015; Balota and Chumbley, 1985; Seyfarth, 2014; Gahl *et al.*, 2012). Broadly, these factors fall into three categories: indexical information about the talker, such as age and sex; linguistic context, such as overall speaking rate, proximity to prosodic boundaries, and various conditional probabilities, such as the probability of a target word given the preceding and following context; and lexical information, such as frequency, phonological neighborhood density, and orthographic factors.

With respect to sex and age of talkers, English word duration has been found to be shorter in male talkers compared to female ones (Bell *et al.*, 2009), and to increase with talker age (Bell *et al.* 2009; Horton *et al.* 2010, but see Gahl and Baayen 2019). Word tokens (or, possibly more accurately, syllables) are lengthened in phrase-final, utterance-final, and pre-pausal position (Klatt, 1976; Turk and Shattuck-Hufnagel, 2007; Umeda, 1975; Crystal and House, 1988;

Wightman *et al.*, 1992). Duration averaged over all occurrences of a given word type are therefore bound to be higher for words that often appear immediately before pauses or in phrase-final or utterance-final position. In English, nouns are more often phrase-final than verbs. It stands to reason that they undergo final lengthening more often than verbs do. Gahl (2008) therefore included, as a proxy for prosodic information, an estimate of the proportion of each target word form that represented nouns, as a way of taking into account syntactic category ambiguity of forms like *stake*. As expected, this 'noun bias' measure was associated with longer predicted duration. Longer duration of nouns compared to other parts of speech has also been found in studies directly targeting effects of syntactic category (e.g. Lohmann, 2018a; Sorensen *et al.*, 1978; Lohmann and Conwell, 2020); category-specific duration patterns reflect the syntactic categories of the tokens over which type-based values are averaged; in addition they may represent cumulative effects of how often a given word form undergoes phrase-final lengthening, even when the form is not in fact phrase-final: Along similar lines, Seyfarth 2014; Sóskuthy and Hay 2017 argue that tokens of words that are often highly predictable are shortened even when unpredictable or when other factors associated with shortening are absent in a given local context.

The shortening of frequent words, i.e. the observation at the heart of the current study, has been subsumed under a more general pattern relating shortening to (various measures of) predictability: Frequent words shorten, and so do words that are highly predictable based on the words preceding or following them: single-word frequency is simply one of several measures of the probability of encountering the word (**?**Kilbourn-Ceron *et al.*, 2020; Jurafsky *et al.*, 2001; Aylett and Turk, 2004; Seyfarth, 2014). Theoretical proposals as to why frequency and contextual probability should pattern in this way include proposals refining information-theoretical measures (see e.g. Hale, 2003; Levy, 2008). The basic idea behind these proposals is that words that are unpredictable in context carry a relatively high information load; by spending more time producing unpredictable, highly informative words and less time on predictable, less informative ones, speakers tend to maintain a constant rate of information transfer (Aylett and Turk, 2004; Pluymaekers *et al.*, 2005; Jaeger and Buz, 2016; Fenk and Fenk, 1980; Fenk-Oczlon, 2001).

The probability of a word, conditioned on the surrounding context, e.g. the words preceding or following the target, is a type of 'syntagmatic' probability. 'Paradigmatic probabilities', by contrast, take into account the probabilities of competing candidates that might occur in place of the target. There is evidence suggesting that paradigmatic and syntagmatic probabilities may have opposite effects on spoken duration: High paradigmatic probabilities (based on the relative frequency of words in morphological paradigms) have been found to be associated with phonetic strengthening and lengthening, rather than phonetic reduction or shortening

(Kuperman *et al.*, 2007; Cohen, 2014).

Predictors of spoken word duration in English are subject to continuing debate. One enduring issue concerns how to take into account the phonological segments that a word contains. Speech sounds differ both in their 'inherent duration' and the degree to which their duration varies across contexts, for example as a function of their position within syllables and words, and word length in syllables. Studies of spoken word duration have sought to take these facts into account by using various measures of 'baseline durations' of words based on the segments they contain (see e.g. Seyfarth, 2014 for a thorough discussion of alternative measures of baseline duration). Among the more elusive or controversial factors are morphological complexity (Caselli *et al.*, 2016; Seyfarth *et al.*, 2017; Strycharczuk, 2019), orthography (Warner *et al.*, 2004), and phonological neighborhood density (PND). Increasing PND has been associated with longer (Buz and Jaeger, 2016), as well as shorter (Gahl *et al.*, 2012; Caselli *et al.*, 2016) whole-word duration. A number of studies have identified effects of PND generally and of specific phonological neighbors (e.g. minimal pairs differing in voicing) on aspects of pronunciation other than word duration (such as voice onset time and vowel formants) (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; Baese-Berk and Goldrick, 2009; Nelson and Wedel, 2017; Wedel *et al.*, 2018; Fox *et al.*, 2015). It is conceivable that these variables are also predictors of whole-word duration, i.e. the outcome variable in the current study.

3. THEORY DEVELOPMENT AND SPECIFIC PREDICTIONS BASED ON THE DL MODEL. Earlier, we compared frequency to white light, i.e. to a combination of multiple, separable factors. We pointed out two such factors: practice and 'contextual independence' of word meanings, and we hypothesized that each of these should be predictive of spoken word duration. We also mentioned another prediction based on the logic of the DL model: Semantic similarity of homophones should be predictive of word duration. Each of these predictions is a manifestation of a very general prediction of the DL-model: The model predicts a degree of form-meaning isomophy, via the mapping between meaning and form. Applied to spoken word production, this idea entails that meaning should be predictive of spoken word duration.

More specifically, the prediction is that spoken word duration should be longer the more strongly the relevant triphones are predicted by word meanings; the predicted strength depends, among other things, on how well the mapping is learned. The strategy of basing predictions about speech production on the strength with which triphones are predicted follows the same logic connecting models of the mental lexicon to word production mentioned in section 1.2 above. But why should greater predicted strength be associated with longer (as opposed to shorter) durations? Informally, that question can be answered as follows: Forms that are entirely

ill-suited to expressing the intended meaning should have predicted 'durations' of zero milliseconds. As a mapping comes to be learned better for a form-meaning pair, the predicted duration should increase, other things being equal. Put differently: Going from meaning to form, the correct triphones would ideally be predicted to be present with complete certainty, and all irrelevant triphones should be predicted to be absent: If a triphone needs to be produced, it receives maximum support ('strength'). For empirical evidence consistent with this hypothesis, the reader is referred to Baayen *et al.* (2019); Chuang *et al.* (2021b), and Tomaschek *et al.* (2021b).

Against the backdrop of these general predictions, we can now formulate specific hypotheses. We begin by describing ways to gauge three aspects of the spectrum of distributional facts often bundled under the heading of 'frequency' (and 'predictability', of which frequency may be a simple estimate): First, the practice aspect of frequency (section 3.1), second, contextual independence, for which we introduce a new vector-space based measure (section 3.2), and third the strength of the relationship between meaning and form, termed 'semantic support for form' (section 3.3). Finally, we describe how semantic similarity of homophones can be estimated (3.4).

3.1. FREQUENCY AS 'PRACTICE' SHAPING THE FORM-MEANING MAPPING. As mentioned in section 2.3 above, the effect of frequency in DL must play out in the mapping between form and meaning. The simplest way of estimating the mapping $\mathbf{G}$ is the method described in Appendix A.1, which we refer to as the 'endstate' method. That method has the advantage of simplicity and provides insight into the system, as evidenced by its successful application in previous studies (see, e.g. Chuang and Baayen, 2021, for a review). But it also has an obvious downside, particularly in the context of a model that is, after all, based on theories of learning: it is blind to the role of usage experience in learning. We mentioned frequency-informed learning (FIL) as the alternative method used here. We describe FIL in appendix A.1 for readers wishing to know details of that method. Here, we restrict ourselves to a few high-level comments on the measure.

FIL has limitations. When FIL is used to estimate $\mathbf{G}$, higher-frequency words are learned best, and the lowest-frequency words are not learned well at all, unsurprisingly. That limitation mirrors those of human learners. Another limitation stems from the fact that the frequencies from corpora such as the BNC (Consortium, 2007) or COCA (Davies, 2010) reflect usage in large speech communities. Individual speakers' experiences are much more limited. Speakers may not know many specialist words in use in communities of experts to which they do not belong (see also Heitmeier *et al.*, 2024); conversely, individual speakers may use words frequently that are rare in the corpus overall, depending on their circumstances, interests, and expertise. Since the DL model is a cognitive model of speakers, rather than of communities of speakers, the usefulness of FIL for understanding lexical cognition depends on the validity of the frequencies

as a measure of individual experience. In the present study, we implemented FIL using the British National Corpus (Consortium, 2007; Aston and Burnard, 2020). The results therefore characterize an artificial-intelligence-like super-individual speaker.

Despite the limitations of the training data, FIL is important for the present study. FIL makes it possible to assess, at least to some extent, the consequences of experience for learning the mapping from meaning to form. Comparing the results obtained with FIL with those obtained with the endstate of learning enables us to tease apart the predictions of the system set up by the word types and their properties, from the predictions of that system when it is 'in use', i.e. predicted consequences of learning through many usage events.

3.2. CONTEXTUAL INDEPENDENCE: CIND. In this section, we describe our vector-space based measure of contextual independence. Part of the conceptual motivation of that measure, ultimately, is the observation that word meanings are context-dependent and ever changing (Elman, 2009). But not all word meanings are context-dependent to the same extent or in similar ways. Some appear in many different semantic contexts, each expressible in countless ways, while others are tied to specific scenarios or even specific collocates. Compare, for example, the rich variety of contexts of a word like *side* to the distribution of the word *adjourn*, which collocates almost exclusively with words describing formal gatherings or proceedings, such as meetings or trials, or *inclement*, which usually pertains to weather and in fact collocates almost exclusively with the word form *weather*. We say that words, or more accurately, word meanings, like the meaning of *side* have high contextual independence, and those like *adjourn* and *inclement* have low contextual independence. We wished to quantify the degree of context-dependency of word meanings, in order to test the DL model's general prediction that meaning-based properties of words should make themselves felt in spoken word duration.

The property we are getting at is similar to a measure ('Contextual Distinctiveness') proposed in McDonald and Shillcock (2001), which reflects frequency distributions of lexical contexts. As McDonald and Shillcock (2001) point out, high-frequency words tend to occur in many different contexts, whereas many low-frequency words keep fairly specific lexical company. McDonald and Shillcock (2001) argue that certain supposed effects of lexical frequency are perhaps better understood as effects of this general relationship between contextual diversity and lexical frequency. The general idea of tracing supposed effects of lexical frequency to broader distributional patterns and to lexical semantics is similar to the goals of the current study, although the measure we are proposing differs from that in McDonald and Shillcock (2001), at many levels: We are not asking how many contexts a word occurs in, or how frequent those contexts are. The measure we are proposing gauges the extent to which a word meaning can be inferred from other words in utterance contexts. High values of contextual independence indicate that the semantic vector as constructed based on a corpus cannot be very precise.

With a slight change in perspective, one can read contextual independence as reflecting how much information about the word's meaning can only be learned by considering many different contexts: A learner who understood the expression *inclement weather* had learned pretty much everything there is to learn about the meaning of *inclement* (with the exception of its collocational restrictedness and resulting stylistic connotations). A learner whose only encounters with the word *side* were the sentences *This side up* and *I would like coleslaw on the side*, by contrast, still had a long way to go.

Importantly, what we are describing here is a property of types, rather than of specific tokens of words. Just as it does not make sense to ask 'Does this token have a frequency of 17 per million?', it does not make sense to ask about the contextual independence of a specific token. The measure we are proposing can only be estimated based on an entire corpus, as a property of word types in linguistic usage. The type-based nature of contextual independence should already make it clear that this measure is fundamentally different from measures of the predictability of a word's occurring in a given context, such as surprisal. Nevertheless, the reference to context-dependency almost inevitably brings to mind information-theoretic measures, which are well-established as predictors of linguistic variability generally and spoken word duration in particular (Seyfarth, 2014; Bell *et al.*, 2003, 2009). Readers may wonder if a failure to include surprisal in a statistical model of word duration (for example, as the average surprisal value of all tokens of a word) does not deprive the model of important information. To alleviate this concern, anticipating our empirical findings somewhat: We did explore two estimates of a word's average probability of occurrence, calculated from word-based bigram probabilities (conditioned on the words preceding or following the target). These average measures did not improve model fit in our (type-based) models, leading us not to pursue average surprisal as a covariate. We believe surprisal to be a valuable predictor of token duration; however, not all properties of word tokens yield good predictions when averaged at the level of word types.

In what follows, we develop our measure of contextual independence, which we term `Cind`, and walk through a small set of utterances by way of example. Readers not wishing to engage with the specifics of how we calculated the measure are invited to skip ahead to section 3.3.

ESTIMATING CONTEXTUAL INDEPENDENCE (`CIND`). Following up on earlier work Baayen *et al.* (2019), a simple network (with no hidden layers) was trained incrementally to predict all words in an utterance from the very same words in that utterance, using naive discriminative learning (Baayen *et al.*, 2011), which makes use of the learning rule of Rescorla and Wagner (1972) rather than the learning rule of Widrow and Hoff (1960). Naive discrimination models are also vector space models, but the input and output vectors of these models are constrained to consist of only zeroes and ones. We illustrate the type of network that we used here for a

toy example with just five utterances, `my time is short`, `my good time`, `my fragrant thyme`, `my lime is bad`, and `my lime is good`. In what follows, we use 'words' to refer to word meanings, which for computational simplicity we assume to be all uniquely distinct.

We first construct a matrix $\boldsymbol{U}$ that specifies for each utterance which words are present, and for each word which utterances it appears in: The first row, for example, indicates that the utterance *my time is short* contains the words shown in the first four columns and does not contain the words shown in columns 5–9. The first column indicates that the word *my* appears in all five utterances, i.e. all five rows. Each sentence is multiple-hot encoded for the word meanings that occur in it.

$$\boldsymbol{U} = \begin{array}{c} \\ \texttt{my time is short} \\ \texttt{my good time} \\ \texttt{my fragrant thyme} \\ \texttt{my lime is bad} \\ \texttt{my lime is good} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \texttt{my} & \texttt{time} & \texttt{is} & \texttt{short} & \texttt{good} & \texttt{fragrant} & \texttt{thyme} & \texttt{lime} & \texttt{bad} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Next, the model needs a means of connecting information about words across utterances. To that end, we estimate a word-to-word mapping matrix $\boldsymbol{W}$ such that $\boldsymbol{UW} = \boldsymbol{U}$. (For technical details and further discussion of this equation, see Appendix A.5.) The resulting matrix, rounded to two decimal digits, is this:

$$\boldsymbol{W} = \begin{array}{c} \\ \texttt{my} \\ \texttt{time} \\ \texttt{is} \\ \texttt{short} \\ \texttt{good} \\ \texttt{fragrant} \\ \texttt{thyme} \\ \texttt{lime} \\ \texttt{bad} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \texttt{my} & \texttt{time} & \texttt{is} & \texttt{short} & \texttt{good} & \texttt{fragrant} & \texttt{thyme} & \texttt{lime} & \texttt{bad} \end{array} \\ \begin{pmatrix} 0.58 & 0.25 & 0.14 & 0.03 & 0.17 & 0.21 & 0.21 & 0.11 & 0.17 \\ 0.25 & 0.65 & -0.08 & 0.18 & 0.10 & -0.13 & -0.13 & -0.27 & 0.10 \\ 0.14 & -0.08 & 0.62 & 0.32 & -0.06 & -0.07 & -0.07 & 0.30 & -0.06 \\ 0.03 & 0.18 & 0.32 & 0.46 & -0.21 & -0.01 & -0.01 & -0.14 & -0.21 \\ 0.17 & 0.10 & -0.06 & -0.21 & 0.73 & -0.08 & -0.08 & 0.15 & -0.27 \\ 0.21 & -0.13 & -0.07 & -0.01 & -0.08 & 0.39 & 0.39 & -0.06 & -0.08 \\ 0.21 & -0.13 & -0.07 & -0.01 & -0.08 & 0.39 & 0.39 & -0.06 & -0.08 \\ 0.11 & -0.27 & 0.30 & -0.14 & 0.15 & -0.06 & -0.06 & 0.44 & 0.15 \\ 0.17 & 0.10 & -0.06 & -0.21 & -0.27 & -0.08 & -0.08 & 0.15 & 0.73 \end{pmatrix} \end{array}$$

The product $\boldsymbol{UW}$ is not shown here because, for this simple example, it is identical to $\boldsymbol{U}$ up

to 15 decimal digits.

We have discussed the word-to-word mapping matrix $\boldsymbol{W}$ elsewhere: Baayen *et al.* (2019) shows that, when the diagonal elements of $\boldsymbol{W}$ are set to zero, its row vectors can be used as word embeddings that perform on a par with word embeddings based on latent semantic analysis (Landauer and Dumais, 1997). For our current purposes, however, the diagonal elements of $\boldsymbol{W}$ (highlighted in grey) are not set to zero; in fact, they provide estimates of words' contextual independence.

Recall that the task of the model is to predict each word meaning in an utterance from all the word meanings in that utterance. We can now use the $\mathbf{W}$ matrix to do just that: The column vectors of the word-to-word mapping matrix $\boldsymbol{W}$ gauge the collocation-strengths of the words on the rows of $\boldsymbol{W}$ with the words on the columns of $\boldsymbol{W}$. The degree to which a given word is predicted to occur in a given utterance is the sum of the products of the elements of the row ('utterance') vector in $\boldsymbol{U}$, and the column ('word') vector in $\boldsymbol{W}$. We refer to this quantity as the word's 'prediction strength'. For example, the prediction strength for *time* in the sentence *my time is short* is the sum of the first four elements of the second column in $\boldsymbol{W}$. This is because, in the product $\boldsymbol{UW}$ (identical to $\boldsymbol{U}$), the first four elements in the row for the utterance *my time is short* are 1, and the remaining elements are 0. Multiplying the non-zero values by the first four elements of the column for *time* in $\boldsymbol{W}$ gives us

$$1 \times 0.25 + 1 \times 0.65 + 1 \times -0.08 + 1 \times 0.18 = 1.$$

To take another example, the prediction strength for *time* in the phrase *my good time* is the sum of the first, second, and fifth element of the column for *time* in $\boldsymbol{W}$, because the utterance *my good time* contains the words in the first, second, and fifth columns of $\boldsymbol{U}$. The prediction strength for *time* in *my good time* is thus

$$1 \times 0.25 + 1 \times 0.65 + 1 \times 0.10 = 1.$$

Notice that, in each of these sums, the term with the largest value is 1×0.65, which is the diagonal element of *time* in $\boldsymbol{W}$. That term will be present in the sum representing the prediction strength for *time* in any utterance in which *time* appears: For any utterance containing *time*, the total prediction strength for *time* always depends for 35% on its collocates, and for 65% on *time* itself, according to this simple model.

More generally, the diagonal elements in $\boldsymbol{W}$ indicate how strongly each word predicts itself in any utterance in which it appears. In our example, the diagonal elements are the largest values in their rows and columns, and this state of affairs is typical for empirical matrices as well. A high diagonal value indicates that a word appears in many, and diverse, contexts. As

a consequence of cue competition during learning, such a word loses its associations with the words that it co-occurs with. As a consequence, a high diagonal value also indicates that a context-independent embedding (such as used in the present study) cannot be very precise, the reason being that that highly variegated semantics is being collapsed into a single embedding.

Importantly, given the general goal of understanding the complexity of word frequency, the diagonal values based on realistically-sized datasets are positively correlated with word frequency, as illustrated in the left panel of Figure 2. They can in fact be understood as frequencies that (1) are rescaled to proportions in the $(0, 1)$ interval and that (2) measure the extent to which a word meaning can be predicted without the help of other word meanings instantiated in an utterance.

As the distribution of d_i is highly skewed with a long right tail, when using this construct as a predictor in a statistical model, a transformation is called for. A logarithmic transformation alleviates the skew, but does not eliminate it. We therefore applied an additional power transformation, resulting in the following definition of the contextual independence measure `Cind`:

$$\textsc{Cind}(w_i) = \left[\log\left(\frac{1}{d_i}\right)\right]^{0.25}.$$

The (negative) correlation of this transformed measure with log lemma frequency is illustrated in the right panel of Figure 2. For further discussion and independent validation of this measure, see Appendix A.5.

Crucially, the diagonal elements of $\boldsymbol{W}$ are informed by all sentences, but specify fixed proportions of self-prediction strength that are independent of these sentences. The fact that the diagonal values are informed by context, but are themselves context-independent, makes `Cind` attractive as a predictor in a regression model for our type-based data.

PLACE FIGURE 2 APPROXIMATELY HERE

3.3. SEMANTIC SUPPORT FOR FORM. Having introduced a way in which the 'practice' and 'contextual independence' aspects of frequency of use can be taken into account in the DL model, we now turn to another part of the complex set of distributional facts implicated by the concept of 'frequency' (and predictability), one that captures a type of paradigmatic probability of word forms. We developed a new measure based on the DL model, termed Semantic Support for Form. Semantic Support for Form (henceforth SSF) gauges the extent to which a given form vector receives support from a meaning vector, taking into account the probabilities of other form vectors that might occur in place of the target. Quantifying the degree to which a

meaning predicts a form's triphones, via a mapping, opens up an additional way to test the DL model's general prediction that meaning-based properties of words should make themselves felt in production. The measure has since been successfully used in a model of tongue tip height in particular German vowels (Saito *et al.*, 2023; Saito, 2023), with greater support predicting "hyperacticulation".

The procedure for estimating SSF can informally be described as follows: Recall that a mapping $\boldsymbol{G}$ takes a matrix of semantic vectors and projects it onto a matrix of form vectors. The values of each of the resulting (i.e. 'predicted') form vectors reflect, for each triphone, how strongly a semantic vector predicts that triphone. In the terminology of the DL model, we say that the triphones are "supported by" the semantic vector to varying degrees. The sum of the predicted values for a word's triphones represents their total semantic support. Readers not wishing to engage with the procedure for calculating these sums for the time being are invited to skip ahead to section 3.5.

ESTIMATING SEMANTIC SUPPORT FOR FORM. The measure that we propose is the total support that a word's triphones jointly receive from its semantic vector. Given a mapping $\boldsymbol{G}$ and the semantic vector $\boldsymbol{s}_i$ of word ω_i, the predicted form vector $\hat{\boldsymbol{c}}_i$,

$$\hat{\boldsymbol{c}}_i = \boldsymbol{s}_i \boldsymbol{G}, \tag{2}$$

specifies, for all triphones known to the model, how much support they receive from $\boldsymbol{s}_i$. For our toy example, the matrix with predicted form vectors $\hat{\boldsymbol{C}}$ (shown in Figure 1 and partly repeated here for convenience with the relevant triphones highlighted), specifies supports for six triphones, only four of which occur in *time*.

	#ta	taɪ	aɪm	ɪm#	#la	laɪ
time	1.024	1.024	0.704	0.704	−0.320	−0.320
lime	0.046	0.046	0.428	0.428	0.383	0.383
thyme	0.973	0.973	1.339	1.339	0.366	0.363

The two triphones that are irrelevant for *time* have small negative weights. For *lime*, the first two triphones are irrelevant, and have small weights, whereas the last four triphones are present in this word and have larger weights. Ideally, the supports for those triphones that actually occur in a given word should be large, whereas supports for the remaining triphones should be close to zero or negative. The total support that a word's relevant triphones receive is a measure of how well the word's semantic vector predicts the word. In the present example, the total support for *time* is 3.456, for *lime* it is 1.714, and for *thyme*, it is 4.624.

The predicted form vector is the input for the second stage of the production process briefly

mentioned above, which places the triphones in the required order for articulation. We assume that the support for a triphone from the semantics is still available once triphones have been placed in order, and hence can inform articulation. The many triphones that do not occur in a particular word, and hence receive little or no support from its meaning, are predicted to be produced with zero (or close to zero) duration. Conversely, the more evidence there is that a triphone should be pronounced, the longer its duration will be. Therefore, we predict that greater semantic support should be associated with longer spoken word duration. This prediction is further motivated by prior findings relating paradigmatic probability to longer duration (cf. section 2.4 above).

In the current study, we make use of two ways of estimating semantic support for form, one based on a model that uses 'type-based' or 'endstate' learning' (using equation 1 in the Appendix), and one based on a model that makes use of 'frequency-informed learning' (using equation 5). Table 1 illustrates that the two methods can make very different predictions, using the toy example data set (see Appendix A.2 for further details).

The endstate-based measure ($\text{SSF}_{\text{ENDSTATE}}$) provides a window on the relation between meaning and form that is largely independent of frequency of use. By contrast, the frequency-informed measure (SSF_{FIL}) does take individual usage events into account. As a consequence, the measure is correlated with word frequency ($r = 0.61$), which may give rise to problems of collinearity when fitting regression models for spoken word duration.

It is an empirical question which measure is the superior predictor for spoken word durations. The analyses we report below use the measure estimated with frequency-informed learning, because it is better motivated theoretically. However, irrespective of which measure of semantic support for form is used, the theory predicts that greater support should be associated with longer spoken word duration.

PLACE TABLE 1 APPROXIMATELY HERE

3.4. QUANTIFYING THE SEMANTIC SIMILARITY OF HOMOPHONE PAIRS. In this section, we describe our measure of semantic similarity of homophone pairs, using the tools introduced in sections 2.3 and 3.1. We estimated semantic similarity as the Pearson correlation between the semantic vectors of a homophone pair. A few comments are in order on what that correlation reflects: The greater the semantic similarity between homophones, the more consistently the mapping from the respective target meanings to the identical form vectors will be learned. Put differently: Mapping very dissimilar meanings onto the same form goes against the grain of associating cues with outcomes with a linear mapping. Forcing a model that is designed

to map distinct semantic vectors onto distinct form vectors to instead map distinct semantic vectors onto identical form vectors induces frailty in the mapping (see Chuang *et al.*, 2021a, 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. Further detail is provided in Appendix A.3. Therefore, we predicted semantically similar homophones to be longer, as well as more similar to each other in duration.

3.5. SUMMARY OF DL-BASED METHODS AND PREDICTIONS. In this section, we described a method for estimating the mapping from meaning to form in a manner that implements the 'practice' aspect of frequency. We then introduced three measures grounded in the DL: contextual independence (Cind), Semantic Support for Form (SSF, implemented in two ways), and semantic similarity of homophone pairs. We use these measures to test the DL model's general prediction that meaning-based properties of words should make themselves felt in spoken word duration.

Ultimately, all of the ideas in this section - the process of learning a mapping between meaning and form, the degree of contextual independence of word meanings, and the strength of the relationship between meaning and form – touch on facets of the 'frequency' spectrum. Quantifying these facets enables us to test predictions of the DL. We do so by means of statistical models grounded in the DL-based measures proposed here vs. localist models of the lexicon, i.e. models in which words do have stable, model-internal representations. That is the empirical part of the current study, to which we now turn.

4. METHODS.

4.1. THE DATA SET. We analyzed the same data set that was used in Gahl (2008) and Lohmann (2018b). The word list for that data set initially contained all English word forms homophonous with at least one other word form that differed in spelling, according to the transcriptions in the CELEX database (Baayen *et al.*, 1995). The data set consists of the spoken word duration of these items, 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. As in previous analyses, 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. For three words (the names *Phil, Marx, Thais*), DL-based measures were not available. These words were therefore excluded from all analyses. The resulting list contained 409 target words.

4.2. GAMs. We made use of the Gaussian Location Scale Additive Model, using the packages **mgcv** (Wood, 2011; **?**, 2017) and **itsadug** (van Rij *et al.*, 2020) in R (R Core Team, 2022). 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 nonlinearity. If a predictor is truly linear, a GAM will detect this and not report artifactual non-linearity.

Like other types of regression models, a GAM estimates the relationship between a set of predictor variables and an outcome variable. In GAM, the shape of the relationship between predictor and outcome is modeled as the ("additive") combination of two sets of functions: parametric and non-parametric. The parametric functions resemble linear predictors in linear mixed effects regression (LMER). The non-parametric smoothing terms, by contrast, estimate the shape of the relationship between predictors and outcome without that shape being specified by the researcher ahead of time. The shape is modeled as the sum of successively more complex (more "wiggly") basis functions. The number of these functions and their coefficients are determined by a procedure balancing model fit and parsimony. GAMs can include Gaussian random effects, analogous to the random effects in LMER. GAMs (and GAMMs) can handle interactions among continuous variables far more flexibly than linear mixed-effects regression models. Such interactions yield model estimates of surfaces and are fitted with either thin plate regression splines (if the variables are on the same scale, justifying identical smoothing parameters and penalties) or with tensor product smooths (if the variables are on different scales, necessitating separate smoothing parameters).

Importantly for the present study, Gaussian Location Scale GAMs (Wood *et al.*, 2016) relax another assumption of the linear regression model, namely, the assumption of equal variance. Gaussian Location Scale GAMs allow the variance, which is assumed to be Gaussian, to change with the predicted mean, if necessary in a non-linear way. This allows us to address the question of whether lower frequency words have more variable durations directly, by modeling duration variance as a function of word frequency (in the localist model) or its component variables (in the DL-based model).

4.3. VARIABLES IN THE STATISTICAL MODELS.

PREDICTORS COMMON TO BOTH MODELS. The outcome variable in all of our statistical models was the log-transformed duration of each target word, averaged over its tokens in Switchboard.

Spoken word durations reflect many factors that are independent of localist and DL-specific assumptions about the lexicon. We included variables indexing such factors in both sets of statistical models, as follows:

Baseline duration Estimated as the log-transformed sum of the average duration of the target's segments in the Buckeye corpus (Pitt *et al.*, 2007). Homophones have identical baseline duration estimates, as they contain identical segments.

Morphological Complexity A binary variable distinguishing morphological simple vs. complex targets, e.g. *lax* vs. *lacks*.

Noun bias 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. The rationale for this variable concerned syntactic preferences for phrase-final positions (Gahl, 2008), a factor equally beyond the scope of localist and DL models of the lexicon. Therefore, this is a control variable for both localist and DL model statistical models.

Orthographic regularity A measure of orthographic regularity (Berndt *et al.*, 1987), called "m-score" in G2008.

Pause quotient The proportion of tokens of a given target that immediately preceded pauses. Like noun bias, this is a control variable for both localist and DL models, treating pauses as determined by forces operating outside the lexicon.

PREDICTORS SPECIFIC TO MODELS GROUNDED IN LOCALIST APPROACHES.

Frequency The frequency of each target (e.g. *time* vs. *thyme*), estimated as the target's (log-transformed) frequency in the CELEX database (Baayen *et al.*, 1995).

Relative Frequency A further frequency measure, estimated as the lemma frequency of each target, 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). The more frequent a word is, the more it can exceed its homophone twin in frequency. As a consequence, the effect of relative frequency may vary with target frequency. Therefore, we included an interaction between `Lemma Frequency` and `Relative Frequency`. An advantage of using target frequency in combination with `Relative Frequency`, rather than using the frequencies of the target and its homophone twin, is that this move somewhat reduces collinearity.

Phonological Neighborhood Density (PND) 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).

PREDICTORS SPECIFIC TO THE MODELS GROUNDED IN THE DISCRIMINATIVE LEARNING APPROACH. Three predictors were specific to the DLM: `Homophone Semantic Similarity`, `Semantic Support For Form`, and `Cind`.

Homophone Semantic Similarity Estimated as the Pearson correlation between the semantic vectors of a homophone pair, using tweet-based `fasttext` word embeddings of dimension 200 (Cieliebak *et al.*, 2017), cf. section 3.4. This measure informs us about how similar in meaning the words of a homophone pair are.

Semantic Support For Form Calculated as described in section 3.3 and appendix A.2. We predicted that duration should be longer for higher values of `Semantic Support For Form`, for the reason laid out in section 3 above.

Cind The measure of contextual independence. Details on how that measure was calculated for the present dataset are given in Appendix A.5, along with a validation study of this measure against visual lexical decision times. We predicted that duration should be longer for higher values of `Cind`.

4.4. STATISTICAL MODELING STRATEGY.

The full sets of variables in our 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), and the same holds for non-linear regression. Various corrective methods for addressing collinearity are available (see, e.g. Tomaschek *et al.*, 2018b, 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 predictors mentioned. After removing predictors that failed to receive evidence for their relevance, 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 variables, independently for both sets of predictors, and used the residuals, henceforth `Residual Baseline Duration`, as a predictor. Details on the residualization can be found in the supplementary materials. 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 $\alpha = 0.0001$.

4.5. SUMMARY OF PREDICTED OUTCOMES. Table 2 summarizes the predicted effects of all the variables.

PLACE TABLE 2 APPROXIMATELY HERE

5. RESULTS AND DISCUSSION. In what follows, we first report analyses with predictors grounded in the localist model (section 5.1), before moving to models with predictors grounded in the DL model (section 5.2). Analyses and comments following up on specific modeling decisions are discussed along the way. Patterns pertaining to the main empirical and theoretical implications of the study are taken up in the General Discussion, in section 6 below.

5.1. GAM ANALYSIS WITH LOCALIST PREDICTORS. The generalized additive model using localist variables is summarized in Table 3. 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.

`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 ($p = 0.0004$; recall that we set $\alpha = 0.0001$), so it is doubtful that this effect will replicate consistently.

PLACE TABLE 3 APPROXIMATELY HERE

PLACE FIGURE 3 APPROXIMATELY HERE

Section B of the table summarizes the smooth terms in the model, visualized in Figure 3. The effect of `Orthographic Regularity` is fully linear (edf=1.0000), in the expected direction, i.e. with higher orthographic regularity being associated with shorter predicted duration. The effect of this variable was small, and its p-value far above our preset α-level, so it remains doubtful that this variable is predictive of spoken word duration of homophones when controlling for frequency. Predicted duration increased in a nearly linear fashion with proportion of prepausal tokens (however, there were very few words in the upper range of this variable, as reflected in the wide confidence region in that range), and linearly with residualized baseline duration. The correlation of residualized baseline duration and the original baseline duration is substantial (.75). As a consequence, since length in milliseconds is predicted from length in phones, the linear relation between residualized baseline duration and spoken word duration is as expected. Increasing phonological neighborhood density was associated with shorter duration, consistent with the findings in (Gahl *et al.*, 2012).

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 3. Frequency shows the expected effect of durational shortening: the general gradient in the regression surface is negative (i.e. going from lighter shades of gray to darker ones). There is also an effect of frequency ratio, but only for targets with log frequencies below about 4. In that range, words with very low frequency ratios (values below -4) are predicted to have shorter durations than words with similar frequency, but higher frequency ratios. That is to say, very low-frequency words with very high frequency homophone twins have shorter duration than would be predicted based on their own frequency. That pattern is consistent with proposals under which duration is expected to vary with form frequency, either as an effect of "frequency inheritance" on lexical retrieval (Dell, 1990; Jescheniak and Levelt, 1994) or as an effect of articulatory practice on speech production (Bybee and Hopper, 2001). Given its theoretical interest, we explored this pattern further, asking whether form frequency provides a superior explanation for the effect of lemma frequency: If that is the case, then cumulative form frequency, i.e. the summed frequency of each pair of homophones, should improve model fit. That was not the case: Replacing lemma frequency with cumulative form frequency resulted in a model with a higher AIC (by 42 units), indicating substantially poorer model fit.

Finally, the lower right panel of Figure 3 clarifies that increased lemma frequency is associated with decreased variance in duration (as well as shorter duration); however, uncertainty for the variance was high near the extremes of the frequency distribution, where data are sparse.

5.2. GAM ANALYSIS WITH DL PREDICTORS. We now turn to the analysis using discrimination-based variables. The Gaussian Location-Scale GAM for this set of predictors is summarized in Table 4 and visualized in Figure 4.

PLACE TABLE 4 APPROXIMATELY HERE

The partial effects of the control variables are very similar to what was found in the model with localist predictors: Predicted duration increased with `Proportion with Following Pauses` and with `Residual Baseline Duration`. The effect of `Residual Baseline Duration` (correlation with Baseline Duration of .91) was completely linear. Of the DL-specific variables, increasing `Cind` was associated with longer predicted duration and higher variance, consistent with our prediction. The fact that both of these patterns were in the opposite direction of the effect of frequency in the localist model was to be expected, given the negative correlation of frequency and `Cind`. Spoken word duration increased with `Homophone Semantic Similarity`, also in line with our expectations. The partial effect of this predictor was linear. The partial effect of `Semantic Support for Form` was nonlinear: For the scatter of lowest values of this predictor, where data were sparse, confidence intervals were wide with no clear trend for mean duration. However, for values above about -2.5, mean duration increased with increasing `Semantic Support for Form`, in line with our expectation.

PLACE FIGURE 4 APPROXIMATELY HERE

5.3. ALTERNATIVE MODELS WITH DL-BASED PREDICTORS. Computational modeling entails choosing between alternative approaches. The model reported in Table 4 and Figure 4 reflects three such choices. First, `Semantic Support for Form` was calculated with frequency-informed learning, rather than endstate-of-learning. Secondly, we made use of contextual independence rather than frequency, given that the DL-model does not contain words as stable representations of which frequency could be a property. Third, we did not include `Orthographic Regularity` as a control variable. Table 5 summarizes the effects of these modeling decisions. In what follows, we address these choices in turn, to gain insight into the extent to which the results depend on them.

PLACE TABLE 5 APPROXIMATELY HERE

First consider the consequences of estimating `Semantic Support for Form` with frequency-informed learning, or with the endstate of learning: Table 5 clarifies that the advantage of using frequency-informed learning over using the endstate of learning is modest, hovering around 2 AIC units. We believe that this is due to the correlations of the two variants of `Semantic Support for Form` with `Cind` and (log-transformed) frequency of occurrence, shown side-by-side in Table 6: When frequency-informed learning (left-hand column) is used instead of the endstate of learning (right-hand column), the magnitude of the correlation almost doubles. As a consequence, `Semantic Support for Form` contributes less independent information about the semantic support for a word's form when it is estimated using frequency-informed learning. When `Semantic Support for Form` is estimated by means of frequency-informed learning, it unavoidably becomes entangled with frequency. As a consequence, it cannot be ruled out that the model estimates are affected by enhancement or suppression, due to collinearity (Friedman and Wall, 2005). Closer inspection suggests that this may indeed be the case: `Semantic Support for Form` is negatively correlated with duration ($r - 0.10, t(407) = -2.099, p = 0.0364$), but in combination with the `Cind` measure in the regression model, greater support predicts longer durations, not shorter ones. By contrast, when semantic support is calculated using the endstate of learning, it enters into a stronger and, importantly, positive correlation with duration $r = 0.36, t(407) = 7.84, p < 0.0001$). The estimate based on the endstate of learning, because of its weaker inherent entanglement with frequency, allows the correlation between semantic support for form and duration to capture the expected relationship between these two variables.

PLACE TABLE 6 APPROXIMATELY HERE

Next, consider the consequence of using `Lemma Frequency` vs `Cind` as a predictor. As can be seen in Table 5, GAM models with `Cind` invariably outperform GAMs with `Lemma Frequency`. `Cind` may simply be the more informative predictor of spoken word duration. Appendix A.5 provides some independent evidence that supports this possibility. The evidence ratios in Table 5 clarify that inclusion of `Cind` rather than `Lemma Frequency` as a predictor results in a better fit to the data. This conclusion holds for GAMs using DL-based predictors compared to a GAM with localist predictors, for different variants of GAMs with DL-based predictors compared to one another, and for GAMs using localist predictors: Given the consistent effect of `Cind`, and given the theoretical interest of frequency measures, we also fitted a GAM

replacing Lemma Frequency with Cind in the 'localist' model discussed in section 5.1. This also resulted in model improvement (of 5 AIC units, from -231.69 to -236.97), as shown in the top two rows of Table 5. When Cind is replaced by Lemma Frequency, fits are invariably worse.

These conclusions are independent of whether Orthographic Regularity is added as a covariate to DL-based models. This variable was not included in the DL-based model reported in Table 4 above because we were not attempting to model the consequences of orthography-phonology consistency within the DLM, although in principle this is possible (see, e.g. Baayen *et al.*, 2019; Chuang *et al.*, 2021b). When Orthographic Regularity was added as a control covariate, model fit improved by around 5 AIC units (see Table 5), irrespective of the other modeling choices.

5.4. VARIABLE IMPORTANCE ANALYSIS. The comparisons of localist vs. DL-based models invite the question what the relative value is of the predictors that are specific to each type of model. Recall that Phonological Neighborhood Density and Frequency Ratio only appear in the localist model, whereas Homophone Semantic Similarity and Semantic Support for Form appear only in the DL-based model. We opted for an assessment using a random forest, a non-parametric regression technique from machine learning (Strobl *et al.*, 2008). Random forests are useful for assessing which variables are more effective as predictors, conditional on the set of predictors available to the model, without necessarily providing meaningful insight into how these predictors interact. We used this method, rather than combining all predictors in one large regression model, for two reasons. First, a regression model with many correlated variables runs the risk of becoming uninterpretable due to high collinearity. Second, the regression models grounded either in localist or in the DL model were motivated by theories of the lexicon; a regression model combining all variables lacks theoretical justification.

PLACE FIGURE 5 APPROXIMATELY HERE

Figure 5 presents the variable importances of the total set of predictors across models. The most important predictor is Semantic Support for Form estimated with the endstate of learning, followed by Phonological Neighborhood Density. The least important predictors are Homophone Semantic Similarity, Frequency Ratio, and Semantic Support for Form estimated with frequency-informed learning. In between these extremes, we have Orthographic Regularity, Lemma Frequency, Noun Bias, Proportion with Following Pauses, and Cind.

The solid variable importance of lemma frequency according to the random forest is unsurprising given the many strong correlations that frequency has with other variables: This outcome is a consequence of the logic of random forests. For seven out of nine predictors (other than lemma frequency itself), frequency is highly correlated, with rank 3, 2, or 1 (and all correlations significant minimally at 0.025). When a recursive partitioning tree has access to frequency, but not to a strongly correlated variable such as Semantic Support estimated with FIL ($r = 0.61$), prediction accuracy hardly suffers. As a consequence, the latter inevitably receives a low variable importance. Conversely, since Semantic Support estimated with endstate learning is less strongly correlated with frequency ($r = -0.38$), its predictiveness depends less on whether frequency is available to the recursive partitioning tree. Furthermore, when frequency is available to the model, it can take over for other variables correlated with frequency that are withheld. As a a consequence, frequency is a highly effective predictor that indirectly thrives on collinearity.

The strong support provided by the random forest for endstate-of-learning based `Semantic Support for Form` provides further evidence for the potential relevance of the precision of the mapping between meaning and form for spoken word duration. This estimate of semantic support is 'uncontaminated' by frequency of use. It is therefore, from a statistical perspective, ideal for gauging the importance of semantic support. The cognitively more plausible measure, semantic support estimated with frequency-informed learning, which blends frequency with semantic support (as it should), is much less useful for statistical modeling in the presence of frequency as highly correlated covariate. When the goal is to optimize prediction using machine learning, endstate-based semantic support is preferable, but when the goal is to understand human lexical learning, FIL-based semantic support is the superior measure.

Among the models evaluated in Table 5, those that have access to the semantic support measure based on frequency-informed learning are superior to those with the measure based on endstate support, as evidenced by the lower AIC values. The differences are small — smaller than 2, the minimum value that is usually considered to indicate a substantial difference — but consistent across pairs of models. We leave it to the reader to decide whether these differences are meaningful. We believe that the model based on frequency-informed learning is more realistic from a cognitive perspective. However, for understanding the contributions of frequency on the one hand, and semantic support on the other, the measure based on a model using endstate learning provides greater analytical clarity: The endstate-of-learning measures bring out the effects of analytical constructs in models of lexical processing, at the expense of revealing effects of usage.

The large variable importances for `Semantic Support for Form` estimated with the endstate of learning and `Phonological Neighborhood Density` suggest the possibility

that these measures are probing the same underlying causal factor. The two measures are negatively correlated ($r = -0.381$, $t(407) = -8.31$, p < 0.0001). From a discriminative point of view, this negative correlation makes sense. Words with more phonological neighbors are more difficult to discriminate between when mapping from meaning to form. As a consequence, words with more neighbors will receive less support from their semantics compared to words with very different forms. In earlier work (Gahl *et al.*, 2012), we found increasing PND to be associated with shortening; the negative correlation between PND and semantic support for form, and the positive correlation of that measure with spoken word duration are consistent with that pattern. Our analyses thus suggest a different interpretation of what phonological neighborhood density is capturing. From a discriminative perspective, phonological neighborhood density reflects (or, also reflects) the precision with which meaning can be realized in form, rather than lexical competition or support.

We conclude with a note on the behavior of `Log Relative Frequency`. We interpret the extremely low variable importance of that variable as a consequence of the 'composite' nature of frequency estimates. `Log Relative Frequency` is confounded with several separable components of such estimates. Some of these, including `Cind`, `Semantic Support FIL`, and `Semantic Support Endstate` are available to the random forest analysis. Paired t-tests on the pairs of homophones in our dataset offer another glimpse of the relationship between `Log Relative Frequency` and other variables in the random forest analysis: Homophones with a positive frequency ratio, i.e those that are more frequent than their twins, have lower values for endstate-based estimates of semantic support compared to their twins ($t(202) = -3.5, p = 0.0005$), higher values for FIL based estimates of semantic support ($t(202) = 9.7, p < 0.0001$), and lower values for `Cind` ($t(202) = -16.2, p < 0.0001$). This suggests differences between homophones with respect to both contextual independence and practice: higher-frequency homophones tend to be less contextually independent and tend to have lower endstate semantic support. The extremely low variable importance of `Log Frequency Ratio` is likely due to these confounds. This result is consistent with the idea that supposed effects of lexical frequency, far from being basic observations, reflect the interplay of multiple components of a complex set of distributional facts.

6. GENERAL DISCUSSION. We began this study by observing that the relationship between word frequency and spoken word duration, a much cited consensus finding, appeared to pose a challenge for models in which words did not have stable representations that could be the bearer of frequency information. We asked whether such models were therefore doomed to fail. To answer that question, we fitted statistical models of spoken word duration using variables grounded in two different approaches to the lexicon: models in which words are represented as model-internal units ('localist' models) and models based on DISCRIMINATIVE LEARNING in

which words do not have stable, model-internal representations. The results of our statistical models confirm earlier findings about the phonetic realization of homophones: The duration of supposed homophones such as *time* and *thyme* reflects variables specific to each homophone twin. Homophones differ in spoken duration – and presumably in other aspects of pronunciation. These observations remain a challenge for any account associating effects of frequency exclusively with form representations or articulatory practice (see Gahl 2008 for discussion). Here, we discuss broader goals and theoretical implications of our analyses, specifically (1) the comparison of word-based (localist) and DL-based models (section 6.1), (2) limitations of the DLM in its current form (section 6.2); (3) the role of meaning in phonetic realization (section 6.3, and, (4) the interpretation of the pre-theoretical concept of lexical frequency (section 6.4.

6.1. COMPARING WORD-BASED (LOCALIST) AND DL-BASED MODELS. Despite the absence of 'words' in the DL-based model, models grounded in DL were capable of modeling variation in spoken word duration, recovering effects usually attributed to word frequency. In fact, we found that the statistical models grounded in DL outperformed those grounded in localist models (Dell, 1986; Levelt *et al.*, 1999; Schwartz *et al.*, 2006). The performance of the particular models we presented does not, of course, imply that models grounded in DL should always outperform those grounded in localist approaches. As a reviewer points out, the comparison is between specific statistical models grounded in the two approaches, rather than between all localist vs. distributed models. That is of course correct. All of the models presented here could almost certainly be improved (or weakened), for example by adding or dropping variables or by using different estimates of the existing variables. After many modeling decisions, one might find the optimal model in each approach.

Finding optimal models was not our goal, however. At one level, we wished to demonstrate that a DL-based model was capable of capturing findings usually attributed to frequency. At a different level, our aims were broader than the demonstration that a model without frequency as a predictor could perform on a par with other models. The construct of lexical frequency, does not clearly have any place in a DL-model, i.e. a lexicon without lexical representations. This makes the DL-based statistical model an important proof of concept. Models taking into account lexical frequency have a long track record of success in linguistic and psycholinguistic research. Each of these successes has further entrenched the general acceptance of usage frequency as a property of words. Showing that a model without lexical frequency can be similarly successful paves the way for alternative approaches.

Relatedly, frequency effects have been taken as diagnostic of representational status, as mentioned in the Introduction: If the frequency of a particular unit of structure (such as words, but also syllables, multiword expressions, and syntactic patterns) affects the way it is processed, the argument goes, then the unit in question corresponds to a 'stored' mental representation.

The success of a model grounded in an approach in which words cannot be bearers of frequency suggests that frequency effects do not necessarily entail 'storage'.

6.2. LIMITATIONS AND THE ROAD AHEAD FOR THE DL MODEL. The DL model makes a number of simplifying assumptions. Speaker's true mental lexicons undoubtedly are much more intricate and effective, and must in some way connect to or encode 'non-linguistic' information, as argued in discussions of exemplar-based models of the lexicon (see e.g. Johnson 2006; Walker and Hay 2011). Restricting the network architecture to linear mappings is undoubtedly 'wrong', — but, we hope to have shown, results in models that are both interpretable and 'useful' (cf. Box, 1976).

Another limitation of the present study is that our endstate-of-learning implementation assumes invariant semantic and form vectors for each word type. However, word meanings vary with context, and no two audio tokens of the same word are identical. The DL model can use the Widrow-Hoff rule for incrementally learning mappings on the basis of token-specific pairs of spoken forms and context-specific embeddings. In principle, then, the DL model is poised to model token-level variability of both meaning and form.

A third limitation of the DL model as presented here concerns the form vectors using triphones as a simple representation of the targets of continuous articulatory planning. We posited identical form vectors for homophones in order to model the consequences of different amounts of semantic support for homophonous forms — despite prior evidence (which the subsequent analyses confirmed) that homophones do not necessarily sound identical. The use of identical vectors of triphones was workable: Predicted form vectors, unlike vectors coding the presence or absence of triphones categorically, showed homophones' triphones to be supported to different degrees. We were able to show, furthermore, that the amount of semantic support received was predictive of spoken word duration and resulted in different predictions for homophone pairs: the model predicted longer duration for whichever member of a pair had stronger semantic support for form. Thus, the simple representation using triphones proved useful. Nevertheless, more refined form vectors than the triphone-based forms are certainly desirable and the target of ongoing research.

Perhaps the most major fundamental limitation of the DL model as a tool for investigating properties of continuous speech is that it is a model for single words. The variable we proposed (`Cind`) for taking into account the role of utterance context, albeit at the level of word types rather than word tokens, is not derived within a worked-out theory of utterance processing.

However, we wish to emphasize an implication of our analyses that would still hold even if an utterance-level (or discourse-level) implementation of the DL model were available: The `Cind` measure may actually capture usage effects on words' meanings. Within the framework of the DL model, the effect of `Cind` plays out during the conceptualization process that precedes

the mapping of meanings onto forms. Ultimately, the claim is that the richness of usage experience shapes not just the realization of word forms, but also word meanings.

6.3. THE ROLE OF MEANING IN PHONETIC REALIZATION. The fact that two semantic predictors, homophone semantic similarity, and semantic support for form, are predictive for homophones' spoken word durations, shows that it is indeed useful to take semantics into account when studying spoken word duration.

Currently the most detailed semantic representations available are word embeddings, capable of nuance far beyond either semantic features of words or specific collocates. Embeddings are widely used in present-day natural language processing and artificial intelligence. Beyond their practical usefulness, they are also capable of mirroring speakers' intuitions about meaning (see e.g. Landauer and Dumais, 1997; Mikolov *et al.*, 2013; Wang *et al.*, 2019; Günther *et al.*, 2019; Boleda, 2020; Shahmohammadi *et al.*, 2021). The architecture of the DL model provides a way of integrating embeddings into a model of the mental lexicon: Absent embeddings, some other vector representation of semantics could of course be used, but embeddings are currently the most richly nuanced semantic representations available to us.

The fact that the simple linear mappings used by the DL model predict meaning and form suggests considerable isomorphism between the form space and the semantic space, beyond specific lexical substrates of onomatopoeia, ideophones, or phonaesthemes. Effects of semantics on phonological form and phonetic realization have long been recognized. For example, Goldrick (2006) reviews reaction time data and speech errors indicating that word-level syntactic/semantic information affects the availability of word forms. Goldrick (2006) further reviews evidence suggesting some limits on the interaction of phonological, lexical, and conceptual information: There is evidence for feedback from phonological to word-level syntactic/semantic representations; by contrast, feedback from word-level representations to semantic features (i.e. conceptual knowledge) is limited or entirely absent. The semantic representations in the models discussed in Goldrick (2006) consist of features ('furry', 'pet', and 'feline' as a means of distinguishing *cat*, *dog*, and *rat* from one another and from (non-furry) *hat* and *cab*). These could in principle be replaced by some other, more nuanced, representation. Such a move invites several questions. One is whether a localist model with rich semantic representations of semantics can (a) match the successes of existing 'sparse' localist models and (b) outperform models without word units as bearers of semantic information. A related question for DL-based models is whether such models can recover limits on the interaction of lexical and conceptual knowledge along the lines discussed in Goldrick (2006). We leave these questions for future research. However, we believe that the main issue standing in the way of any model assuming stable semantic word units, is the context-dependent nature of lexical meaning discussed in Elman (2009).

Correspondences between sound and meaning beyond sound symbolism in the usual sense

have long been pointed out (see e.g. Nuckolls 1999 for discussion). For example, Monaghan *et al.* (2014) observed a small but significant positive correlation between the similarities between words' forms (gauged with an edit-distance measure) and word's meanings (approximated by embeddings). This observation was recently replicated for form similarities based on acoustic signals (Shafaei-Bajestan *et al.*, 2022).

Our findings also dovetail with prior literature demonstrating consequences of correspondences between sound and meaning for lexical processing. For instance, Nygaard *et al.* (2009) demonstrated that sound to meaning correspondences facilitated word learning, in a study of adult English speakers' learning of Japanese word forms paired with their English translations, their English antonyms, or semantically unrelated English words. The actual meanings were learned better than the control words — and, importantly, so were antonyms. This last fact differs from the usual finding of 'iconicity' or sound symbolism and suggests a relationship not just between forms and particular meanings, but with entire semantic fields.

6.4. THE CONSTRUCT OF FREQUENCY. Our aims were broader than the demonstration that a model without frequency as a predictor could perform on a par with other models. We wished to better understand the nature of frequency effects, by treating such effects as arising through the combination of multiple, separable distributional patterns. The finding that variables tracking these patterns outperformed frequency as a predictor of word duration suggests to us that the distributional variables can help shed light on the interpretation of effects usually attributed to frequency. We are emphatically not proposing that the DL-based variables we employed should take the place of frequency in discussions of lexical representations. In fact, we would caution against reifying any particular variable on the basis of its predictiveness in a statistical model. That caveat holds for DL-based variables just as much as for frequency or any other variable.

For instance, `Semantic Support for Form` is one of many quantities one can derive from matrices and mappings that constitute DL models. This measure captures an aspect of an intermediate, dynamically generated and ephemeral state in the production process. The predictiveness of this semantic variable for spoken word duration on the one hand and its correlations with PND and lexical frequency on the other suggests a place for rich and dynamic semantic information in speech production; it does not, however, entail any privileged *representational* status.

7. CONCLUSION. We began the current study by observing that the apparent consensus finding of frequent words shortening relied on two complex constructs: the mental lexicon as a 'store of words', and frequency as a property of items in such a lexicon. We hope to have shown that the seemingly pre-theoretical term 'frequency' as a simple tally of usage events fails to

do justice to the combined, and sometimes antagonistic, effects of distributional patterns. We believe that these distributional facts pertain to meaning as much as to form. Taking the 'arbitrariness of the sign' for granted may obscure areas of isomorphy of meaning and form.

More generally, we hope that our discussion illustrates the value of scrutinizing familiar measures, variables, and analytical constructs. The theoretical constructs proposed in the current study are no exception: We do not wish to reify contextual independence, homophone semantic similarity, or any other measure we used in our attempts to understand the interplay of meaning and form. Observations about any of these variables, just like long-standing consensus findings, are probably best thought of not as 'effects' so much, but as patterns to be explained.

A. APPENDIX.

A.1. MAPPINGS IN THE DISCRIMINATIVE LEXICON MODEL.

ENDSTATE LEARNING (EL). To obtain the mapping matrix $\boldsymbol{G}$ from the equality

$$\boldsymbol{SG} = \boldsymbol{C}, \tag{3}$$

Endstate Learning makes use of the normal equations for regression (see, e.g. Faraway, 2005):

$$\begin{aligned} \mathbf{SG} &= \mathbf{C} \\ \mathbf{S}^T\mathbf{SG} &= \mathbf{S}^T\mathbf{C} \\ (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{SG} &= (\mathbf{C}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{C} \\ \mathbf{IG} &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{C} \\ \mathbf{G} &= (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{C}. \end{aligned} \tag{4}$$

We use Cholesky-decomposition to calculate $(\mathbf{C}^T\mathbf{C})^{-1}$. Matrix $\mathbf{G}$ can be thought of as the result of infinite learning experience with the word types in a given data set. For detailed discussion of the balance between memorization and generalization in these models, see Heitmeier *et al.* (2021).

INCREMENTAL LEARNING. Incremental learning works through a given ordered list of word tokens. Each time a next token is encountered, the mapping $\mathbf{G}$ is updated. This method is used by Heitmeier *et al.* (2023) to document trial-to-trial learning in a large lexical decision experiment (Keuleers *et al.*, 2012). When this incremental learning algorithm is taken through an ordered list of word tokens repeatedly, it will eventually converge towards the mapping obtained by solving $\mathbf{SG} = \mathbf{C}$ (see also Shafaei-Bajestan *et al.*, 2021). This is why we refer to the mapping obtained by solving $\mathbf{SG} = \mathbf{C}$ as representing the 'endstate' of learning, and refer to learning with the Widrow-Hoff rule as 'incremental' learning. Learning a mapping by incremental regression has two disadvantages. A technical disadvantage is that the token-by-token updating of the mapping is computationally demanding, and prohibitively so for data sets with many millions of tokens, even when using a numerically optimized language such as julia (Heitmeier *et al.*, 2024). A practical disadvantage is that dense training data with valid temporally ordered tokens are extremely rare. Corpora, for instance, bring together texts from various registers, written by a variety of different authors at different points in time. Although an order can be imposed on the texts in a corpus, and although within a given text, words do have a natural order, the resulting sequence of tokens is very different from the actual experience of any individual language user. However, when order information is available, incremental

learning can be very effective.

FREQUENCY-INFORMED LEARNING (FIL). For unordered data, regression with "Frequency-Informed Learning" (FIL) offers an alternative and highly efficient method for taking frequencies of use into account. Suppose that the words *time, lime,* and *thyme* have frequencies 100, 10, and 1. We place these frequencies on the main diagonal of a matrix $\boldsymbol{Q}$,

$$\boldsymbol{Q} = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and then solve

$$(\sqrt{\boldsymbol{Q}}\boldsymbol{S})\ \boldsymbol{G} = (\sqrt{\boldsymbol{Q}}\boldsymbol{C}), \tag{5}$$

using the normal equations of regression. We now have a mapping $\boldsymbol{G}$ that takes into account the token frequencies with which the words in our lexicon occur. Instead of working with form and meaning matrices that have 100 repeated entries of *time*, 10 repeated entries of *lime*, and one entry for *thyme*, we multiply both the form matrix and the semantic matrix with the square root of the diagonal matrix $\boldsymbol{Q}$, and solve the normal equations with just a single pair of vectors for each word type. For mathematical details, the reader is referred to Heitmeier *et al.* (2024). This is the method we used to estimate the semantic support for form used in the GAM model reported in Table 4. When using FIL, it is important not to log-transform frequencies, as doing so distorts learning compared to incremental learning, leading to overlearning of low frequency words, and underlearning of high frequency words (see Heitmeier *et al.*, 2024, for detailed discussion).

Both incremental learning and FIL are less accurate than EL when evaluated on types: low-frequency words are learned less well. However, when evaluated on the number of tokens with correct predictions, FIL and incremental learning outperform EL (Heitmeier *et al.*, 2024).

A.2. SEMANTIC SUPPORT FOR FORM. In what follows, we describe more formally how we estimate `Semantic Support for Form`, given a mapping $\boldsymbol{G}$. A matrix $\boldsymbol{T}$ tabulating the total support that word forms receive from semantic vectors, not only their own, but also those of other words, is obtained as follows:

$$\boldsymbol{T} = \hat{\boldsymbol{C}}\boldsymbol{C}^T, \tag{6}$$

where $\boldsymbol{C}^T$ is the transpose of $\boldsymbol{C}$, i.e., the matrix obtained from $\boldsymbol{C}$ by flipping rows and columns (for details, see Figure A.1). Using (1) to calculate $\hat{\boldsymbol{C}}$ for the endstate of learning, for our

running example, the following matrix is obtained:

$$\boldsymbol{T} = \begin{array}{c} \\ \text{TIME} \\ \text{LIME} \\ \text{THYME} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{time} & \textit{lime} & \textit{thyme} \end{array} \\ \begin{pmatrix} \mathbf{3.455} & 0.767 & 3.455 \\ 0.948 & \mathbf{1.622} & 0.948 \\ 4.623 & 3.409 & \mathbf{4.623} \end{pmatrix} \end{array}.$$

The bolded values on the main diagonal highlight the amount of support that words' forms receive from their own meanings. Because *time* and *thyme* have exactly the same triphones, their forms receive the same amount of support from the semantic vector of *time*. Similar observations can be made for the support provided by LIME for the forms of *time* and *thyme*, and for the support provided by THYME for the forms of *time* and *thyme*. In this example, the meaning of *thyme* provides stronger support for the form of *thyme* (4.623) compared to the amount of support that the form of *time* receives from its meaning (3.455).

When frequency-informed learning is used to estimate $\boldsymbol{T}$, very different support values are obtained:

$$\boldsymbol{T}_{\text{FIL}} = \begin{array}{c} \\ \text{TIME} \\ \text{LIME} \\ \text{THYME} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{time} & \textit{lime} & \textit{thyme} \end{array} \\ \begin{pmatrix} \mathbf{39.805} & 19.560 & 39.805 \\ 5.137 & \mathbf{9.965} & 5.137 \\ 6.225 & 7.029 & \mathbf{6.225} \end{pmatrix} \end{array}.$$

Because *time* is now much more frequent than *thyme*, the semantic support for the form of *time* given TIME (39.805) is much greater than the semantic support for the form of *thyme* given THYME (6.225). Importantly, the correlation between the bolded diagonal values and the token frequencies is close to 1 (0.9998). For this small lexicon, the semantic support for form is therefore overwhelmingly determined by frequency of use.

We calculated the mappings $\boldsymbol{G}$ for both type-based and frequency-informed learning on the basis of a dataset of 10,636 words. These words were taken from the dataset studied in Baayen *et al.* (2019), 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 that are the row vectors of $\boldsymbol{C}$. As there are 5,600 different triphones in this dataset, the dimensionality of $\boldsymbol{C}$ is $10,636 \times 5,600$. It will be observed that the number of triphones is far smaller than the number of all possible triphones. Only those triphones are used that are required for the data under consideration.

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. $\boldsymbol{S}$ is therefore a $10,636 \times 200$ matrix. The mapping matrix $\boldsymbol{G}$ is a 200 $\times$ 5,600 matrix, irrespective of whether it is estimated based on types, or with frequency-informed learning. For frequency-informed learning, we used word frequency counts from the written part of the British National Corpus (Consortium, 2007; Aston and Burnard, 2020).

PLACE FIGURE A.1 APPROXIMATELY HERE

A reviewer raises the question whether SFF introduces circularity into the DL model as a production model; the concern raised was that the model already needed to know which triphones were needed. To clarify: The DL model passes on the complete $\hat{\boldsymbol{c}}$ vector to later stages of the production system (which include an ordering algorithm) as opposed to only the triphones of a given word. The semantic support for the triphones in the word selected continues to be available to guide articulation at those later stages. Hence, there is no circularity.

A.3. SYNONYMS AND HOMOPHONES. When a matrix $\mathbf{G}$ is square and not singular, the mapping that it defines is one-to-one. However, for our data, $\mathbf{G}$ is a 200 $\times$ 5,600 matrix, and although the mapping $\boldsymbol{G}$ is optimal in the least-squares sense, it is not one-to-one.

PLACE FIGURE A.2 APPROXIMATELY HERE

To clarify how the mapping deals with synonyms and homophones, which violate the one meaning - one form principle, consider Figure A.2, which illustrates the issue using univariate regression. Synonyms are similar to datapoints with the same x-coordinates, as exemplified by the black squares. Their predicted value strikes a balance between the two observed values, while also taking into account the other observations. As a consequence, only one y-value is predicted for these two datapoints, which is located on the regression line. For the $\boldsymbol{G}$ mapping, the vectors of synonyms are likewise predicted to have exactly the same form. Since true synonyms do not exist (see, e.g. Clark, 1993, for discussion), the semantic vectors of near-synonyms are not indentical, and hence are mapped onto different forms.

With respect to homophones, the black triangles in Figure A.2 illustrate the analogous situation for the univariate case. We now have two observations with different x-coordinates, that are assumed to have exactly the same y-value. This is possible only when the regression line is a horizontal line. For all other regression lines, different y-values that are located on the regression line are predicted. In other words, the high-dimensional mapping $\mathbf{G}$ predicts that homophones will not have exactly the same forms: the amount of support their triphones receive from their semantics will differ. As shown in the present study, these differences in semantic support are predictive for the differences is spoken word duration observed by Gahl (2008).

A.4. COMPARISON WITH DEEP LEARNING MODELS. Vector-space models assume that high-dimensional representations of forms and their corresponding meanings are situated in the same vector space, and that as a consequence, a form vector can be mapped onto its meaning vector, and a meaning vector onto its form vector, by simple linear transformations. Linear transformations can be implemented with neural networks with just an input layer and an output layer, without any hidden layers nor any special activation functions.

By contrast, deep learning models, make use of one or more hidden layers (e.g. Plaut 1997 and Dell *et al.* 1993) and implement special functions that modulate how the information reaching a node is processed. Widely used functions are the sigmoid function, which squashes the activation of nodes between 0 and 1, and the rectified linear unit function, which sets negative activations to zero, and leaves positive values unchanged. These activation functions enable deep artificial neural networks to handle problems that cannot be solved linearly, i.e., by simply rotation and or stretching (or shrinking) points in a vector space.

A second difference between vector space models and deep learning models is that the latter rely on the backpropagation of error algorithm to drive learning. Vector space models, because they don't have hidden layers, can use much simpler and faster learning algorithms.

A third difference is that deep learning networks may employ recurrent connections (see, for production, e.g. Dell *et al.*, 1993) to model incremental processing. Vector space models set up much simpler mappings between form and meaning. At first sight, this would suggest that incremental processing is out of reach of this class of models. This is not the case, however, but it is accomplished in a different way, see, e.g. Shafaei-Bajestan *et al.* (2021) for auditory comprehension.

Within the general framework of the discriminative lexicon model, linear mappings can in principle be replaced by non-linear mappings, using deep neural networks; code for doing so is available in the **JudiLing** package.

A.5. CONTEXTUAL INDEPENDENCE. The ideal solution for the equation

$$\mathbf{UW} = \mathbf{U}, \tag{7}$$

where $\boldsymbol{U}$ is a binary matrix using multiple-hot encoding to indicate which words (columns) are present in utterances (rows), is the identity matrix $\mathbf{I}$:

$$\begin{aligned} \mathbf{UW} &= \mathbf{U} \\ \mathbf{U}^{-1}\mathbf{UW} &= \mathbf{U}^{-1}\mathbf{U} \\ \mathbf{W} &= \mathbf{I}. \end{aligned} \tag{8}$$

The elements on the diagonal of the $\mathbf{W}$ matrix typically have the largest values in their rows and columns. However, because $\mathbf{U}$ is not a square matrix, there is no unique inverse and we have to use the pseudoinverse, or, alternatively, incremental learning with the Rescorla-Wagner rule. As a consequence, the matrix $\mathbf{W}$ is an approximation of the optimal matrix $\mathbf{I}$. From a learning perspective, especially when using incremental learning (which we applied to the BNC), this makes sense: learners do not have all the information about how words are used available at the same time. Instead, they learn step by step, from utterance to utterance. As a consequence of incomplete information, learners cannot arrive at the perfect solution, but instead find a solution that is co-determined by their experiences with words' collocational preferences.

The empirical estimates that we use in our regression analysis are based on a network $\boldsymbol{W}$ that, following Baayen *et al.* (2019), 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 Consortium, 2007). 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. In other words, the matrix $\boldsymbol{U}$ is a 87,906,894 $\times$ 23,572 matrix.

However, rather than using the normal equations for regression to estimate $\boldsymbol{W}$ (which we used for the toy model in the main text), we use the Rescorla-Wagner equations to incrementally build a network with connection weights characterized by a 23,562 $\times$ 23,562 weight matrix $\mathbf{W}$, with a learning rate η (the product of the α and β parameters of the Rescorla-Wagner equations, with $\lambda = 1.0$) of 0.001.

In this incremental learning process, cue competition drives associations between context words to a target word towards zero, the more these context words occur in utterances without the target word. As a consequence, more polysemous target words that occur in richly varying contexts will have larger diagonal values. These larger values are indicative of decreasing semantic integration, and are an index for lower-quality of context-free embeddings.

VALIDATION. To obtain independent evidence that the diagonal elements of $\mathbf{W}$ capture an important aspect of lexical knowledge, we extracted the 10,224 words for which we have available both this measure and the reaction time (averaged over subjects, and inverse transformed to -1000/rt) in the British Lexicon Project (Keuleers *et al.*, 2012). We compared the AIC of Gaussian location-scale GAMs with as predictor either log frequency in the British National Corpus or `Cind`. In addition, we also considered the untransformed, raw diagonal values of $\boldsymbol{W}$. Furthermore, we also fitted a model with the log of log frequency (backing off from zero by adding 2 to the original counts). Results are summarized in Table A.2. The models with measures based on d_i

outperform the models fitted with log frequency by 2 and 24 AIC units.

PLACE TABLE A.1 APPROXIMATELY HERE

THEORETICAL CONSIDERATION. A theoretical consideration concerns the question of how to bring together the Cind measure and measures derived from the DL model within an integrated theoretical framework. The Cind measure, which is an utterance-level measure, can be integrated with the DL model as follows. Recall that according to the DL model, the starting point for production of a word ω_i is a semantic vector $\boldsymbol{s}_i$. Thus far, we have assumed that a predicted form vector $\hat{\boldsymbol{c}}$ is obtained by multiplying $\boldsymbol{s}$ with $\boldsymbol{G}$: $\hat{c} = \boldsymbol{sG}$. However, the strength of a semantic vector can vary depending on context. In the simplest case, this strength is simply a non-negative scalar α. To take into account a word's contextual independence, for a given word ω_i, α_i can be set to be proportional to $\texttt{Cind}_i$. As a consequence, the triphones of words with a higher value of Cind, i.e., words with lower contextual independence, will receive more support from the semantics than words with greater contextual independence. In this case, their spoken word durations will be longer. Effects of surprisal on spoken word duration at the token level can be accounted for along similar lines.

METHODOLOGICAL CONSIDERATIONS. The learning rate η is a free parameter. When η is high, many words will have diagonal values very close to 1, and d becomes useless as predictor. We found setting $\eta = 0.001$ to provide reasonably well-differentiated diagonal values, for the utterances in the British National Corpus. Training $\boldsymbol{W}$ with many epochs through the same set of utterances with a large learning rate will cause $\boldsymbol{W}$ to converge to the identity matrix.

A further caveat is in order. As observed by Hollis (2020) for word frequency and contextual diversity, one measure may appear to outperform another measure due to minor changes in probability distributions and correlational structure. As a consequence, uncertainty remains about the predictivity of Cind for spoken word duration (or reaction times in lexical decision) as compared to word frequency. This uncertainty is aggravated by the choice of transformation. For the present data, changing the transformation to a double log transformation results in a decrease in AIC by some 10 units, whereas using the untransformed d_i values leads to a worse fit, with an increase of some 20 AIC units.

There is one more aspect of Cind that should be mentioned here. We have taken embeddings as a-priori givens, but in real life, embeddings also have to be learned. Embeddings, however, are engineered to be of high quality, without being strongly dependent on frequency of occurrence. For the present dataset, there is a modest correlation between vector length (estimated with L1

norm) and log lemma frequency ($r = -0.44$), as well as with Cind ($r = 0.48$). Although the length of an embedding is a crude way of assessing the consequences of frequency of use for an embedding, the correlation of Cind with embedding length suggests that Cind can be conceptualized as a means of assessing the consequences of frequency of use for the 'availability' of an embedding. As the L1 norms of embeddings are far less effective at predicting spoken word durations than Cind, the L1 norms were not considered further in this study.

A.6. PRODUCTION IN THE DL MODEL. The mapping from embeddings to triphone vectors, from which we calculated the semantic support for form measure, is the first step in the production algorithm that has been studied most intensively in the DL framework. This first step, which estimates the amount of support triphones receive from a word's embedding, is complemented by a second step in which the triphones are ordered for articulation.

The algorithm for this second step that is implemented in the **JudiLing** package combines positional learning with a beam-search like procedure. The algorithm implemented in Baayen *et al.* (2019) makes use of procedures from graph theory. In contrast to the speech errors generated by the binding algorithm used by Rumelhart and McClelland (1986), both of the abovementioned algorithms produce phonotactically legal speech errors that typically involve morphological or semantic errors, see, e.g. Chuang *et al.* (2020) for a discussion of speech errors made by the model for Estonian nouns. We also note here that Rumelhart and McClelland (1986) generated forms from forms, whereas the modeling approach taken in the present study generates forms from meanings.

A.7. SOFTWARE. A julia package, **JudiLing** (https://juliapackages.com/p/judiling), facilitates setting up the form and meaning matrices and calculating the mappings between them. The **pyndl** package for python Sering *et al.* (2022) includes functionality for efficiently calculating Cind. The supplementary materials provide the relevant code.

REFERENCES.

ARNON, INBAL, and NEAL SNIDER. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62.67–82.

ASTON, GUY, and LOU BURNARD. 2020. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh University Press.

AYLETT, MATTHEW, and ALICE TURK. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47.31–56.

BAAYEN, R. HARALD, YU-YING CHUANG, ELNAZ SHAFAEI-BAJESTAN, and JAMES P. BLEVINS. 2019. 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.

BAAYEN, R. HARALD, PETAR MILIN, DUSICA FILIPOVIĆ ĐURDEVIĆ, PETER HENDRIX, and MARCO MARELLI. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118.438–482.

BAAYEN, R. HARALD, PETAR MILIN, and MICHAEL RAMSCAR. 2016. Frequency in lexical processing. *Aphasiology* 30.1174–1220.

BAAYEN, R HARALD, RICHARD PIEPENBROCK, and LEON GULIKERS. 1995. The CELEX lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania* .

BAESE-BERK, MELISSA, and MATTHEW GOLDRICK. 2009. Mechanisms of interaction in speech production. *Language and cognitive processes* 24.527–554.

BALOTA, DAVID A., and JAMES I. CHUMBLEY. 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, DAVID A., MELVIN J. YAP, KEITH A. HUTCHISON, MICHAEL J. CORTESE, BRETT KESSLER, BJORN LOFTIS, JAMES H. NEELY, DOUGLAS L. NELSON, GREG B. SIMPSON, and REBECCA TREIMAN. 2007. The English Lexicon Project. *Behavior Research Methods* 39.445–459.

BELL, ALAN, JASON M. BRENIER, MICHELLE GREGORY, CYNTHIA GIRAND, and DAN JURAFSKY. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60.92–111.

BELL, ALAN, DANIEL JURAFSKY, ERIC FOSLER-LUSSIER, CYNTHIA GIRAND, MICHELLE GREGORY, and DANIEL GILDEA. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113.1001–1024.

BELSLEY, DAVID A., EDWIN KUH, and ROY E. WELSCH. 1980. *Regression Diagnostics. Identifying Influential Data and sources of Collinearity*. New York: Wiley.

BERNDT, RITA SLOAN, JAMES A. REGGIA, and CHARLOTTE C. MITCHUM. 1987. Empirically derived probabilities for grapheme-to-phoneme correspondences in English. *Behavior Research Methods, Instruments, & Computers* 19.1–9.

BOJANOWSKI, PIOTR, EDOUARD GRAVE, ARMAND JOULIN, and TOMAS MIKOLOV. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5.135–146.

BOLEDA, GEMMA. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6.213–234.

BOX, GEORGE E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71.791–799.

BUZ, ESTEBAN, and T. FLORIAN JAEGER. 2016. The (in)dependence of articulation and lexical planning during isolated word production. *Language, Cognition and Neuroscience* 31.404–424.

BYBEE, JOAN. 1999. Usage-based phonology. *Functionalism and formalism in linguistics* 1.211–242.

BYBEE, JOAN. 2002a. Phonological evidence for exemplar storage of multiword sequences. *Studies in second language acquisition* 24.215–221.

BYBEE, JOAN. 2002b. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change* 14.261–290.

BYBEE, JOAN. 2003. *Phonology and language use*. Cambridge University Press.

BYBEE, JOAN. 2006. From usage to grammar: The mind's response to repetition. *Language* 82.711–733.

BYBEE, JOAN, and PAUL HOPPER (eds.) 2001. *Frequency and the emergence of linguistic structure*. Typological studies in language 45. John Benjamins Publishing.

CASELLI, NAOMI K., MICHAEL K. CASELLI, and ARIEL M. COHEN-GOLDBERG. 2016. Inflected words in production: Evidence for a morphologically rich lexicon. *Quarterly Journal of Experimental Psychology* 69.432–454.

CASSANI, GIOVANNI, YU-YING CHUANG, and R. HARALD BAAYEN. 2020. On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46.621–637.

CHOLIN, JOANA. 2011. Do syllables exist? Psycholinguistic evidence for the retrieval of syllabic units in speech production. In *Handbook of the Syllable*, 225–253. Brill.

CHOMSKY, NOAM. 1995. *The minimalist program*. Cambridge, MA: MIT Press.

CHUANG, YU-YING, and R. HARALD BAAYEN. 2021. Discriminative learning and the lexicon: NDL and LDL. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

CHUANG, YU-YING, MELANIE J. BELL, ISABELLE BANKE, and R. HARALD BAAYEN. 2021a. Bilingual and multilingual mental lexicon: a modeling study with Linear Discriminative Learning. *Language Learning* 71.219–292.

CHUANG, YU-YING, MELANIE J. BELL, YU-HSIANG TSENG, and R. HARALD BAAYEN, 2024. Word-specific tonal realizations in Mandarin. Manuscript, National Taiwan Normal University, Anglia Ruskin University, Cambridge, UK, and the university of Tübingen, Germany.

CHUANG, YU-YING, KAIDI LÕO, JAMES P. BLEVINS, and R. HARALD BAAYEN. 2020. Estonian case inflection made simple. A case study in Word and Paradigm morphology with Linear Discriminative Learning. In *Advances in Morphology*, ed. by L. Körtvélyessy and P. Štekauer, 119–141. Cambridge University Press.

CHUANG, YU-YING, MARIE LENKA VOLLMER, ELNAZ SHAFAEI-BAJESTAN, SUSANNE GAHL, PETER HENDRIX, and R. HARALD BAAYEN. 2021b. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior research methods* 53.945–976.

CIELIEBAK, M., JAN DERTU, F. UZDILLI, and D. EGGER. 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.

CLARK, EVE V. 1993. *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

CLOPPER, CYNTHIA G, and RORY TURNBULL. 2018. Exploring variation in phonetic reduction: Linguistic, social, and cognitive factors. In *Rethinking reduction*, ed. by Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler, and Margaret Zellers, 25–72. De Gruyter Mouton.

COHEN, CLARA. 2014. Probabilistic reduction and probabilistic enhancement. *Morphology* 24.291–323.

CONSORTIUM, BNC. 2007. British national corpus, xml edition. *Oxford Text Archive Core Collection* .

CONWELL, ERIN. 2018. Token frequency effects in homophone production: An elicitation study. *Language and Speech* 61.466–479.

CRYSTAL, THOMAS H., and ARTHUR S. HOUSE. 1988. The duration of American-English vowels: An overview. *Journal of Phonetics* 16.263–284.

DAVIES, MARK. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing* 25.447–464.

DELL, GARY S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93.283.

DELL, GARY S. 1990. Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes* 5.313–349.

DELL, GARY S., CORNELL JULIANO, and ANITA GOVINDJEE. 1993. Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science* 17.149–195.

DESHMUKH, NEERAJ, ARAVIND GANAPATHIRAJU, ANDI GLEESON, JONATHAN HAMAKER, and JOSEPH PICONE. 1998. Resegmentation of SWITCHBOARD. In *Fifth international conference on spoken language processing*.

DEVLIN, JACOB, MING-WEI CHANG, KENTON LEE, and KRISTINA TOUTANOVA. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

DIESSEL, HOLGER. 2017. Usage-based linguistics. In *Oxford research encyclopedia of linguistics*. Oxford University Press.

ELMAN, JEFFREY L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science* 33.547–582.

FARAWAY, JULIAN J. 2005. *Linear Models with R*. Boca Raton, FL: Chapman & Hall/CRC.

FENK, AUGUST, and GERTRAUD FENK. 1980. Konstanz im Kurzzeitgedächtnis –Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie* 27.400–414.

FENK-OCZLON, GERTRAUD. 2001. Familiarity, information flow, and linguistic form. In *Frequency and the emergence of linguistic structure*, ed. by J. Bybee and P. Hopper, 431–448. Amsterdam: John Benjamins.

FERREIRA, VICTOR S., and ZENZI M. GRIFFIN. 2003. Phonological influences on lexical (mis) selection. *Psychological Science* 14.86–90.

FILLMORE, CHARLES, PAUL KAY, LAURA A. MICHAELIS, and IVAN SAG. 2003. *Construction grammar*. Stanford: CSLI Publications.

FINK, ANGELA, and MATTHEW GOLDRICK. 2015. The influence of word retrieval and planning on phonetic variation: Implications for exemplar models. *Linguistics Vanguard* 1.215–225.

FODOR, JERRY A. 1983. *The modularity of mind*. Cambridge, MA: MIT Press.

FOX, NEAL P., MEGAN REILLY, and SHEILA E. BLUMSTEIN. 2015. Phonological neighborhood competition affects spoken word production irrespective of sentential context. *Journal of memory and language* 83.97–117.

FRICKE, MELINDA, MELISSA M. BAESE-BERK, and MATTHEW GOLDRICK. 2016. Dimensions of similarity in the mental lexicon. *Language, cognition and neuroscience* 31.639–645.

FRIEDMAN, LYNN, and MELANIE WALL. 2005. Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician* 59.127–136.

GAHL, SUSANNE. 2008. 'Time' and 'thyme' are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84.474–496.

GAHL, SUSANNE. 2009. Homophone duration in spontaneous speech: A mixed-effects model. *UC Berkeley Phonology Lab Technical Report* .

GAHL, SUSANNE. 2015. Lexical competition in vowel articulation revisited: Vowel dispersion in the easy/hard database. *Journal of Phonetics* 49.96–116.

GAHL, SUSANNE, and R HARALD BAAYEN. 2019. Twenty-eight years of vowels: Tracking phonetic variation through young to middle age adulthood. *Journal of Phonetics* 74.42–54.

GAHL, SUSANNE, YAO YAO, and KEITH JOHNSON. 2012. Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66.789–806.

GODFREY, JOHN J., EDWARD C. HOLLIMAN, and JANE MCDANIEL. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, 517–520. IEEE Computer Society.

GOLDRICK, MATTHEW. 2006. Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence. *Language and Cognitive Processes* 21.817–855.

GOLDRICK, MATTHEW, CHARLOTTE VAUGHN, and AMANDA MURPHY. 2013. The effects of lexical neighbors on stop consonant articulation. *The Journal of the Acoustical Society of America* 134.EL172–EL177.

GRIFFIN, ZENZI M., and KATHRYN BOCK. 1998. Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language* 38.313–338.

GÜNTHER, FRITZ, LUCA RINALDI, and MARCO MARELLI. 2019. Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science* 14.1006–1033.

HALE, JOHN. 2003. The information conveyed by words in sentences. *Journal of psycholinguistic research* 32.101–123.

HARM, MICHAEL W, and MARK S SEIDENBERG. 1999. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review* 106.491.

HEITMEIER, MARIA, YU-YING CHUANG, SETH AXEN, and R. HARALD BAAYEN. 2024. Frequency effects in linear discriminative learning. *Frontiers in Human Neuroscience, Sec. Speech and Language* 17.1242720.

HEITMEIER, MARIA, YU-YING CHUANG, and R HARALD BAAYEN. 2021. Modeling morphology with linear discriminative learning: Considerations and design choices. *Frontiers in psychology* 12.720713.

HEITMEIER, MARIA, YU-YING CHUANG, and R. HARALD BAAYEN. 2023. How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology* 146.101598.

HOLLIS, GEOFF. 2020. Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language* 114.104146.

HORTON, WILLIAM S., DANIEL H. SPIELER, and ELIZABETH SHRIBERG. 2010. A corpus analysis of patterns of age-related change in conversational speech. *Psychology and Aging* 25.708.

JACKENDOFF, RAY S. 2002. *Foundations of language*. Oxford: Oxford University Press.

JAEGER, T. FLORIAN, and ESTEBAN BUZ. 2016. Signal reduction and linguistic encoding. *Handbook of psycholinguistics. Wiley-Blackwell* .

JESCHENIAK, JÖRG D., and WILLEM J. M. LEVELT. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20.824–843.

JOHNSON, KEITH. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of phonetics* 34.485–499.

JURAFSKY, DAN. 2003. Probabilistic modeling in Psycholinguistics: Linguistic comprehension and production. In *Probabilistic linguistics*, ed. by Rens Bod, Jennifer Hay, and Stefanie Jannedy, 39–95. MIT Press.

JURAFSKY, DANIEL. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive science* 20.137–194.

JURAFSKY, DANIEL, ALAN BELL, MICHELLE GREGORY, and WILLIAM D. RAYMOND. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the emergence of linguistic structure*, ed. by J. Bybee and P. Hopper, 229–254. Amsterdam: John Benjamins.

JURAFSKY, DANIEL, and JAMES H MARTIN. 2019. *Vector semantics and embeddings*, 270–85. Prentice Hall.

KAHN, JASON M., and JENNIFER E. ARNOLD. 2012. A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language* 67.311–325.

KEULEERS, EMMANUEL, PAULA LACEY, KATHLEEN RASTLE, and MARC BRYSBAERT. 2012. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods* 44.287–304.

KILBOURN-CERON, ORIANA, MEGHAN CLAYARDS, and MICHAEL WAGNER. 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.

KLATT, DENNIS H. 1976. Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59.1208–1221.

KUPERMAN, VICTOR, MARK PLUYMAEKERS, MIRJAM ERNESTUS, and HARALD BAAYEN. 2007. Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America* 121.2261–2271.

LANDAUER, THOMAS K., and SUSAN T. DUMAIS. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104.211–240.

LEVELT, WILLEM J. M., ARDI ROELOFS, and ANTJE S. MEYER. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22.1–38.

LEVELT, WILLEM J.M. 2013. *A history of psycholinguistics: The pre-Chomskyan era*. Oxford University Press.

LEVY, ROGER. 2008. Expectation-based syntactic comprehension. *Cognition* 106.1126–1177.

LOHMANN, ARNE. 2018a. Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics* 54.753–777.

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

LOHMANN, ARNE, and ERIN CONWELL. 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.

LUEF, EVA MARIA, and JONG-SEUNG SUN. 2020. Wordform-specific frequency effects cause acoustic variation in zero-inflected homophones. *Poznań Studies in Contemporary Linguistics* 56.711–739.

MARR, DAVID. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. WH Freeman.

MCCLELLAND, JAMES L., and JEFFREY L. ELMAN. 1986. The TRACE model of speech perception. *Cognitive psychology* 18.1–86.

MCCLELLAND, JAMES L., and DAVID E. RUMELHART. 1981. An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review* 88.375–407.

MCDONALD, SCOTT A., and RICHARD C. SHILLCOCK. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech* 44.295–323.

MIKOLOV, TOMAS, KAI CHEN, GREG CORRADO, and JEFFREY DEAN. 2013. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* 1–12.

MONAGHAN, PADRAIC., RICHARD C. SHILLCOCK, MORTEN H. CHRISTIANSEN, and SIMON KIRBY. 2014. How arbitrary is language. *Philosophical Transactions of the Royal Society B.* 369.20130299.

NELSON, NOAH RICHARD, and ANDREW WEDEL. 2017. The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English. *Journal of Phonetics* 64.51–70.

NEWMEYER, FREDERICK J. 2003. Grammar is grammar and usage is usage. *Language* 79.682–707.

NORRIS, DENNIS, and JAMES M MCQUEEN. 2008. Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review* 115.357.

NORRIS, DENNIS G. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52.189–234.

NUCKOLLS, JANIS B. 1999. The case for sound symbolism. *Annual Review of Anthropology* 28.225–252.

NYGAARD, LYNNE C., DEBORA S. HEROLD, and LAURA L. NAMY. 2009. The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive science* 33.127–146.

NYGAARD, LYNNE C., and JENNIFER S. QUEEN. 2008. Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance* 34.1017.

PHILLIPS, BETTY S. 2020. Spread across the lexicon: Frequency, borrowing, analogy, and homophones. *The Handbook of Historical Linguistics* 2.343–356.

PIERREHUMBERT, JANET. 2002. Word-specific phonetics. *Laboratory Phonology* 7.101–139.

PINKER, STEPHEN, and ALAN PRINCE. 1991. Regular and irregular morphology and the psychological status of rules of grammar. In *Proceedings of the 1991 meeting of the Berkeley Linguistics Society*.

PITT, MARK A., LAURA DILLEY, KEITH JOHNSON, SCOTT KIESLING, WILLIAM RAYMOND, ELIZABETH HUME, and ERIC FOSLER-LUSSIER, 2007. Buckeye Corpus of Conversational Speech (2nd release).

PLAUT, DAVID C. 1997. Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and cognitive processes* 12.765–806.

PLUYMAEKERS, MARK, MIRJAM ERNESTUS, and R. HARALD BAAYEN. 2005. Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America* 118.2561–2569.

R CORE TEAM, 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RESCORLA, ROBERT A., and ALLAN R. WAGNER. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current research and theory*, ed. by A. H. Black and W. F. Prokasy, 64–99. New York: Appleton Century Crofts.

RUMELHART, DAVID E., and JAMES L. MCCLELLAND. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing. Explorations in the Microstructure of*

Cognition. Vol. 2: Psychological and Biological Models, ed. by J. L. McClelland and D. E. Rumelhart, 216–271. Cambridge, Mass.: MIT Press.

SAITO, MOTOKI, 2023. *Enhancement effects of frequency: An explanation from the perspective of Discriminative Learning*. University of Tübingen dissertation.

SAITO, MOTOKI, FABIAN TOMASCHEK, and R. HARALD BAAYEN. 2023. Articulatory effects of frequency modulated by inflectional meanings. In *Interfaces of Phonetics*, ed. by Marcel Schlechtweg, 125–155. De Gruyter.

SCARBOROUGH, REBECCA. 2013. Neighborhood-conditioned patterns in phonetic detail: Relating coarticulation and hyperarticulation. *Journal of Phonetics* 41.491–508.

SCHWARTZ, MYRNA F., GARY S. DELL, NADINE MARTIN, SUSANNE GAHL, and PAULA SOBEL. 2006. A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language* 54.228–264.

SERING, KONSTANTIN, MARC WEITZ, ELNAZ SHAFAEI-BAJESTAN, and DAVID-ELIAS KÜNSTLE. 2022. pyndl: Naïve discriminative learning in python. *Journal of Open Source Software* 7.4515.

SEYFARTH, SCOTT. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133.140–155.

SEYFARTH, SCOTT, MARC GARELLEK, GWENDOLYN GILLINGHAM, FARRELL ACKERMAN, and ROBERT MALOUF. 2017. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 1–18.

SHAFAEI-BAJESTAN, ELNAZ, MASOUMEH MORADIPOUR-TARI, PETER UHRIG, and R. HARALD BAAYEN. 2021. LDL-AURIS: Error-driven learning in modeling spoken word recognition. *Language, Cognition and Neuroscience* .

SHAFAEI-BAJESTAN, ELNAZ, PETER UHRIG, and R. HARALD BAAYEN. 2022. Making sense of spoken plurals. *The Mental Lexicon* 17.337–367.

SHAHMOHAMMADI, HASSAN, HENDRIK LENSCH, and R HARALD BAAYEN. 2021. Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. 158–170. arXiv preprint arXiv:2104.07500.

SHANNON, CLAUDE E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423.

SHIELDS, LYNNE W., and DAVID A. BALOTA. 1991. Repetition and associative context effects in speech production. *Language and Speech* 34.47–55.

SIEW, CYNTHIA S.Q., DIRK U. WULFF, NICOLE M. BECKAGE, and YOED N. KENETT. 2019. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity* 2019.

SORENSEN, JOHN M., WILLIAM E. COOPER, and JEANNE M. PACCIA. 1978. Speech timing of grammatical categories. *Cognition* 6.135–153.

SÓSKUTHY, MÁRTON, and JENNIFER HAY. 2017. Changing word usage predicts changing word durations in new zealand english. *Cognition* 166.298–313.

STEMBERGER, JOSEPH P., and BRIAN MACWHINNEY. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition* 14.17–26.

STROBL, CAROLIN, ANNE-LAURE BOULESTEIX, THOMAS KNEIB, THOMAS AUGUSTIN, and ACHIM ZEILEIS. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9.

STRYCHARCZUK, PATRYCJA. 2019. Phonetic detail and phonetic gradience in morphological processes. *Oxford Research Encyclopedia of Linguistics* .

TOMASCHEK, FABIAN, DENIS ARNOLD, FRANZISKA BRÖKER, and R. HARALD BAAYEN. 2018a. Lexical frequency co-determines the speed-curvature relation in articulation. *Journal of Phonetics* 68.103–116.

TOMASCHEK, FABIAN, DENIS ARNOLD, KONSTANTIN SERING, BENJAMIN V. TUCKER, JACOLINE VAN RIJ, and MICHAEL RAMSCAR. 2021a. Articulatory variability is reduced by repetition and predictability. *Language and speech* 64.654–680.

TOMASCHEK, FABIAN, PETER HENDRIX, and R. HARALD BAAYEN. 2018b. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* 71.249–267.

TOMASCHEK, FABIAN, INGO PLAG, MIRJAM ERNESTUS, and R. HARALD BAAYEN. 2021b. Phonetic effects of morphology and context: Modeling the duration of word-final s in English with naïve discriminative learning. *Journal of Linguistics* 57.123–161.

TREMBLAY, ANTOINE, BRUCE DERWING, GARY LIBBEN, and CHRIS WESTBURY. 2011. Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning* 61.569–613.

TURK, ALICE E., and STEFANIE SHATTUCK-HUFNAGEL. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35.445–472.

UMEDA, NORIKO. 1975. Vowel duration in American English. *The Journal of the Acoustical Society of America* 58.434–445.

VAN RIJ, JACOLIEN, MARTIJN WIELING, R. HARALD BAAYEN, and HEDDERIK VAN RIJN, 2020. itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4.

VAN SON, ROB J. J. H., and LOUIS C. W. POLS. 2003. Information Structure and Efficiency in Speech Production. In *Proceedings of Eurospeech-2003*, 769–772, Geneva, Switzerland.

VAN SON, ROB J. J. H., and JAN P. H. VAN SANTEN. 2005. Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication* 47.100–123.

WALKER, ABBY, and JEN HAY. 2011. Congruence between 'word age'and 'voice age'facilitates lexical access. *Laboratory Phonology* .

WANG, BIN, ANGELA WANG, FENXIAO CHEN, YUNCHENG WANG, and C. C. JAY KUO. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing* 8.e19.

WARNER, NATASHA. 2011. Reduction. *The Blackwell companion to phonology* 1–26.

WARNER, NATASHA, ALLARD JONGMAN, JOAN SERENO, and RACHEL KEMPS. 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 251–276.

WEDEL, ANDREW, NOAH NELSON, and REBECCA SHARP. 2018. The phonetic specificity of contrastive hyperarticulation in natural speech. *Journal of Memory and Language* 100.61–88.

WIDROW, BERNARD, and MARCIAN E. HOFF. 1960. Adaptive switching circuits. *1960 WESCON Convention Record Part IV* 96–104.

WIGHTMAN, COLIN W., STEFANIE SHATTUCK-HUFNAGEL, MARI OSTENDORF, and PATTI J. PRICE. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* 91.1707–1717.

WOOD, SIMON N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73.3–36.

WOOD, SIMON N. 2017. *Generalized Additive Models*. New York: Chapman & Hall/CRC.

WOOD, SIMON N, NATALYA PYA, and BENJAMIN SÄFKEN. 2016. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association: Theory and Methods* 111.1548–1575.

WRIGHT, RICHARD. 2004. Factors of lexical competition in vowel articulation. *Papers in Laboratory Phonology VI* 75–87.

FIGURES

$$\boldsymbol{S} \qquad\qquad \boldsymbol{G} \qquad\qquad = \qquad\qquad \hat{\boldsymbol{C}}$$

$$\begin{array}{l} \\ \textsc{time} \\ \textsc{lime} \\ \textsc{thyme} \end{array} \begin{pmatrix} S1 & S2 \\ 0.1 & 0.3 \\ 0.6 & 0.2 \\ 1.1 & 0.6 \end{pmatrix} \quad \begin{array}{l} \\ \textit{S1} \\ \textit{S2} \end{array} \begin{pmatrix} \text{\#ta} & \text{ta\textscr} & \text{a\textscm} & \text{\textscm\#} & \text{\#la} & \text{la\textscr} \\ -1.19 & -1.19 & -0.08 & -0.08 & 1.12 & 1.12 \\ 3.81 & 3.81 & 2.37 & 2.37 & -1.44 & -1.44 \end{pmatrix} = \begin{array}{l} \\ \textit{time} \\ \textit{lime} \\ \textit{thyme} \end{array} \begin{pmatrix} \text{\#ta} & \text{ta\textscr} & \text{a\textscm} & \text{\textscm\#} & \text{\#la} & \text{la\textscr} \\ 1.024 & 1.024 & 0.704 & 0.704 & -0.320 & -0.320 \\ 0.046 & 0.0046 & 0.428 & 0.428 & 0.383 & 0.383 \\ 0.973 & 0.973 & 1.339 & 1.339 & 0.366 & 0.363 \end{pmatrix}$$

Figure 1. The value of the triphone ta\textscr in the predicted form vector of *time* (the first row vector of $\boldsymbol{S}$) is obtained by pairwise multiplication of the values of this semantic vector (highlighted in gray) and the values in the column vector of ta\textscr in the transformation matrix $\boldsymbol{G}$ (also highlighted in gray): $1.024 = 0.1 \times -1.19 + 0.3 \times 3.81$. Borrowing notation from statistics, we write $\hat{\boldsymbol{C}}$ instead of $\boldsymbol{C}$ as the elements of the form matrix are predicted values that differ from the values in $\boldsymbol{C}$, which are either 1 or 0. The more accurate the mapping $\boldsymbol{G}$ is, the more similar (albeit not identical) the row vectors of $\hat{\boldsymbol{C}}$ will be to the row vectors of $\boldsymbol{C}$.

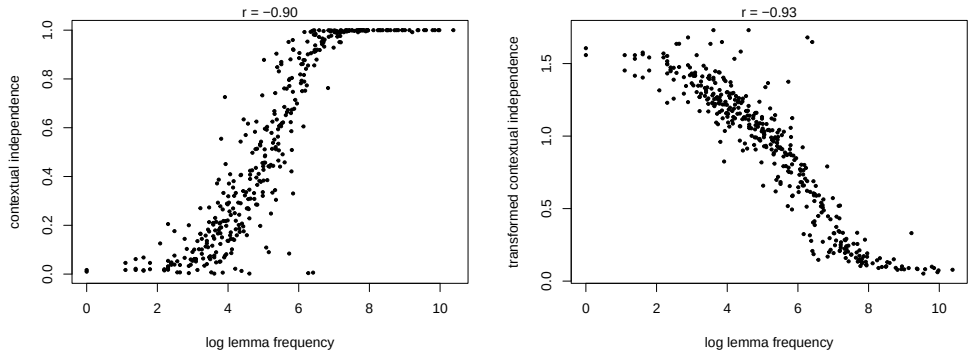

Figure 2. Scatterplots for the correlation of log lemma frequency with untransformed (left) and transformed (right) contextual independence.

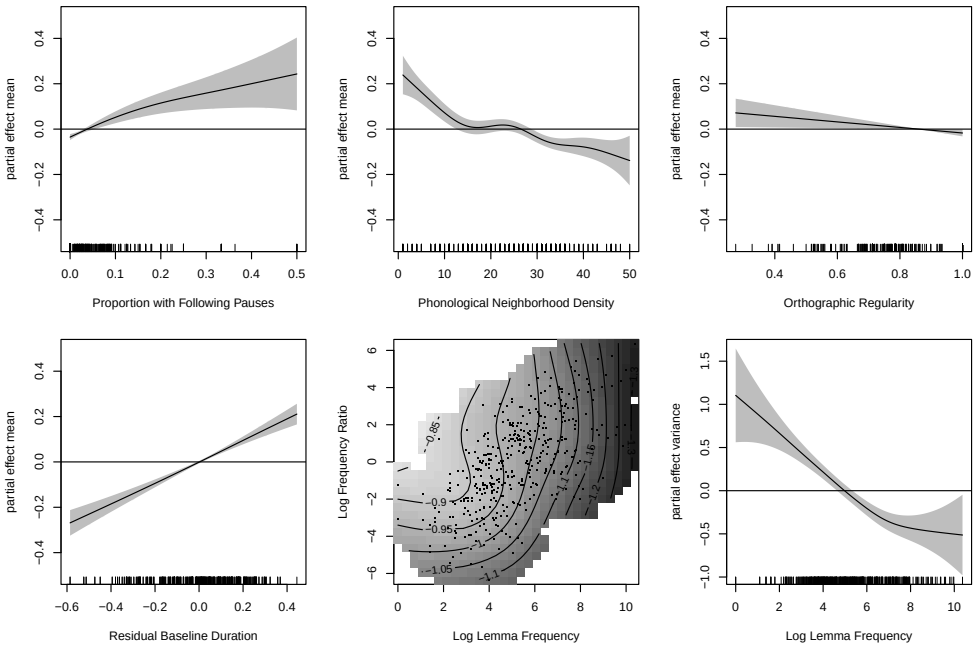

Figure 3. Partial effects according to a Gaussian Location-Scale GAM fitted to the average duration of the homophone word types, using control variables and localist predictors. Rugs indicate the unique values of predictors. Y-axis scales are fixed across panels for the partial effects on the mean, to facilitate comparison of effect sizes. In the contour plot, lighter shades of gray indicate longer spoken word duration. AIC: -231.7; -REML = -94.98.

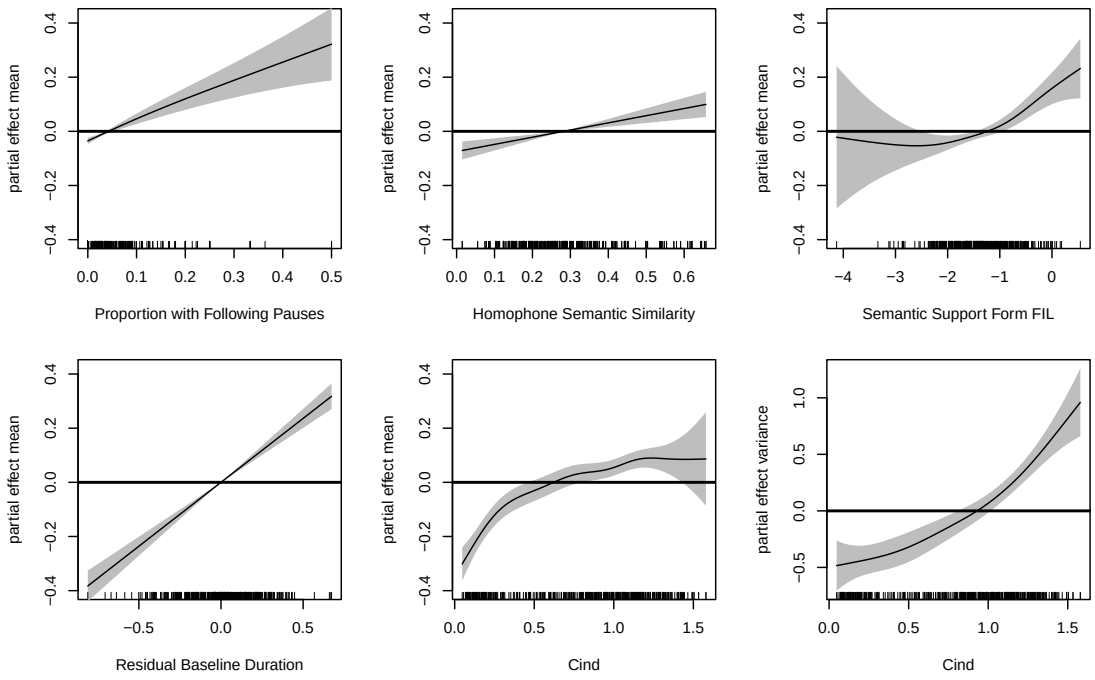

Figure 4. Partial effects according to the Gaussian Location-Scale GAM using control and DL-based predictors fitted to the average log-transformed duration of the homophones in Table 4.

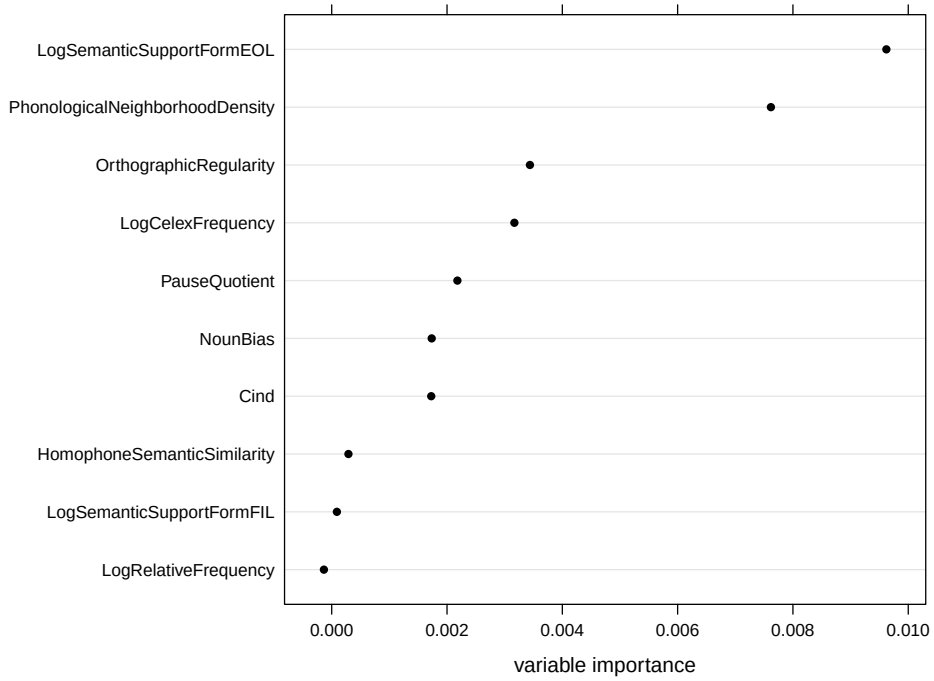

Figure 5. Variable importance of all predictors according to a random forest.

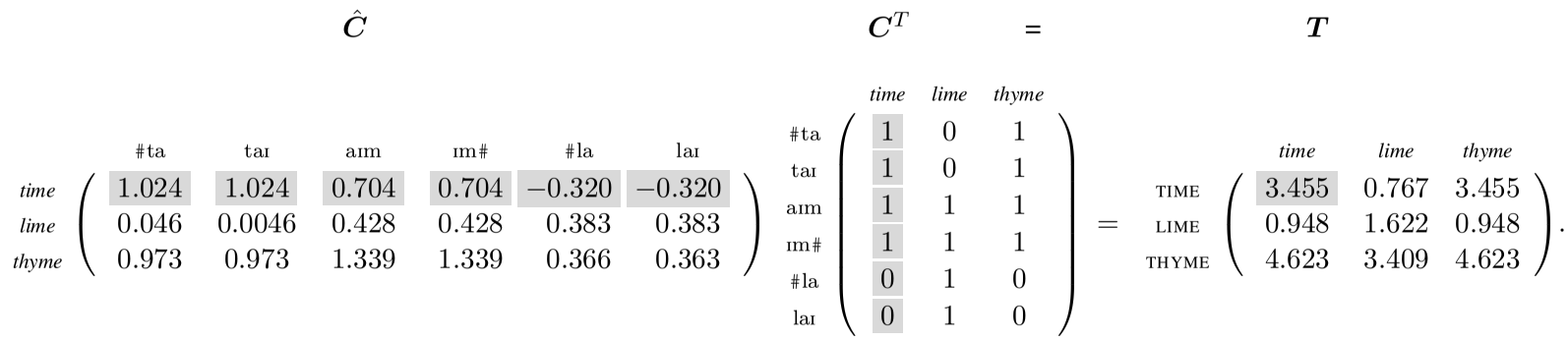

Figure A.1. The total semantic support for the form of *time* (highlighted in $\boldsymbol{T}$) is obtained by pairwise multiplication of the support values in its row vector of $\hat{\boldsymbol{C}}$ (highlighted) and the 1/0 values in this word's column vector in $\boldsymbol{C}^T$ (also highlighted), followed by summation. The support of TIME for *time* therefore is $1.024 \times 1 + 1.024 \times 1 + 0.704 \times 1 + 0.704 \times 1 - 0.320 \times 0 - 0.320 \times 0 = 3.455$.

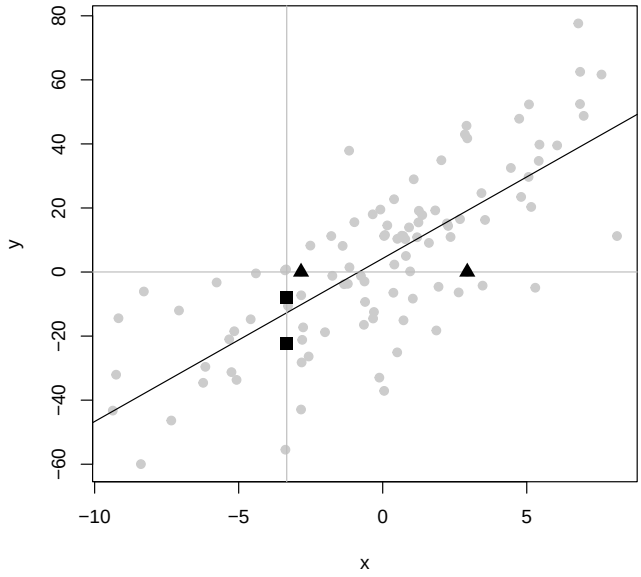

Figure A.2. Analogons of synonyms (squares) and homophones (triangles) in univariate regression.

TABLES

Table 1. Semantic support for form estimated for the toy example using endstate learning and FIL learning.

word	frequency	endstate learning	FIL learning
time	100	3.455	39.805
lime	10	1.622	9.965
thyme	1	4.623	6.225

Table 2. Predictions about the relationship between variables considered and spoken word duration. Unless otherwise indicated, 'Predicted effect' (rightmost column) refers to the predicted change in mean duration for increasing values of the variable.

Variable	Predicted change in mean duration as values of continuous variables increase
A. Variables common to localist and DL-based models	
`Baseline duration`	Longer
`Orthographic regularity`	Shorter
`Noun-bias`	Longer for noun-biased targets
`Morphological complexity`	Longer for complex targets
`Pause quotient`	Longer
B. Variables specific to localist-based models	
`Phonological Neighborhood Density`	Shorter
`Lemma frequency`	Shorter; increased variance
`Relative Frequency`	Shorter
C. Variables specific to DL-based models	
`Homophone Semantic Similarity`	Longer
`Semantic Support For Form`	Longer
`Cind`	Longer; increased variance

Table 3. Model summary for a Gaussian Location-Scale GAM fitted to the mean log-transformed durations of the homophone word types using control and localist predictors. AIC: -230.8

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 4. Model summary for a Gaussian Location-Scale GAM fitted to the mean log-transformed durations of the homophone word types using control variables and variables grounded in the DL model. AIC: -239.13, -REML = -96.58. Semantic support for form was estimated using frequency-informed learning of the mapping between meaning and form.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
`Intercept` [mean]	-1.0448	0.0139	-75.1163	< 0.0001
`Noun Bias` =yes	0.0666	0.0167	3.9957	0.0001
`Intercept` [variance]	-1.8207	0.0376	-48.4458	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(`Proportion with following pauses`)	1.3296	1.5866	38.1509	< 0.0001
s(`Homophone Semantic Similarity`)	1.0000	1.0001	18.3399	< 0.0001
s(`Semantic Support Form`)	2.7544	3.5395	28.8927	< 0.0001
s(`Residual Baseline Duration`)	1.0000	1.0001	181.1099	< 0.0001
s(`Cind`) [mean]	4.4438	5.4597	117.3170	< 0.0001
s(`Cind`) [variance]	2.6485	3.2862	92.3508	< 0.0001

Table 5. Overview of GAM models based on different selections of predictors and modeling choices, for localist and DL models, cross-tabulating AIC for `Semantic Support for Form` estimated with frequency-informed learning (FIL) vs. endstate-of-learning; `Contextual Independence` (Cind) vs. `Frequency`; and excluding vs. including a measure of orthographic consistency. All relevant predictors are significant across all models. Evidence ratios are calculated with respect to the localist model using `frequency`.

Lexicon	Method for estimating Sem. Sup.	Usage measure	Orth. consist.	AIC	Evidence ratio
Localist		frequency	+	-231.69	
		Cind	+	-236.97	14.01
DL	endstate	Cind	−	-237.52	18.45
	endstate	Cind	+	-243.87	441.42
	endstate	frequency	−	-219.97	0.0029
	endstate	frequency	+	-226.01	0.0584
	frequency-informed	Cind	−	-239.13	41.26
	frequency-informed	Cind	+	-244.77	692.29
	frequency-informed	frequency	−	-221.28	0.0055
	frequency-informed	frequency	+	-226.64	0.0801

Table 6. Correlations of two frequency measures (Cind and Frequency) and Semantic support for form as estimated with frequency-informed learning vs. endstate of learning.

	Semantic Support for Form	
	Frequency-informed learning	Endstate of learning
Cind	-0.640	0.315
Frequency	0.612	-0.378

Table A.1. AIC for four models fitted to the inverse-transformed visual lexical decision times in the British Lexicon Project, using untransformed and transformed values of contextual independence and frequency. Frequency counts are based on the BNC, for the same sentences on which Cind is calculated.

AIC	measure	transformation
-13369	Cind	$[\log(1/d_i)]^{0.25}$
-13345	Cind	none (d_i)
-13343	frequency	$\log(f+1)$
-13341	frequency	$\log(\log(f+2))$