Opposing forces on acoustic duration

Benjamin V. Tucker[a,*], Michelle Sims[a], R. Harald Baayen[b]

[a]*Department of Linguistics, University of Alberta, 4-32 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada*
[b]*Seminar für Sprachwissenschaft, Universitat Tübingen, Wilhelmstrasse 19, Tübingen, Germany*

Abstract

The present paper investigates the influence of opposing lexical forces on speech production using the duration of the stem vowel of regular and irregular verbs as attested in the Buckeye corpus of conversational North-American English. We compared two sets of predictors, reflecting two different approaches to speech production, one based on competition between word forms, the other based on principles of discrimination learning. Classical measures in word form competition theories such as word frequency, lexical density, and gang size (types of vocalic alternation) were predictive of stem vowel duration. However, more precise predictions were obtained using measures derived from a two-layer network model trained on the Buckeye corpus. Measures representing strong bottom-up support predicted longer vowel durations. Conversely, measures reflecting uncertainty predicted shorter vowel durations, including a measure of the verb's semantic density. The learning-based model also suggests that it is not a verb's frequency as such that gives rise to shorter vowel duration, but rather a verb's collocational diversity. Results are discussed with reference to the Smooth Signal Redundancy Hypothesis and the Paradigmatic Signal Enhancement Hypothesis.

Keywords: vowel duration, Smooth Signal Redundancy Hypothesis, Paradigmatic Signal Enhancement Hypothesis, naive discriminative learning, semantic density, neighborhood density

1. Introduction

Many forces influence the acoustic realization of words. These forces, which range from frequency of use to audience design, shape the durations with which words and segments are produced, as well as the details of their articulation and pitch contours. The present paper investigates the role of some of these forces in the production of irregular and regular verbs from a corpus of spontaneous speech.

Frequency of use is one such force that has been studied intensively: Words that are used more often tend to have shorter realizations (e.g., Zipf, 1929; Jurafsky et al., 2001; Bell et al., 2003). Homophones such as *thyme* and *time* have been shown to differ in duration as a function of their frequency of occurrence: *time* is more frequent and has a shorter duration in unscripted speech than *thyme* (Gahl, 2008). Moreover,

[*] Corresponding author

Email addresses: bvtucker@ualberta.ca (Benjamin V. Tucker), mnsims@ualberta.ca (Michelle Sims), harald.baayen@uni-tuebingen.de (R. Harald Baayen)

Pluymaekers et al. (2005) reported that the duration of Dutch affixes are co-determined by the frequency of the complex words in which they occurred.

The Probabilistic Reduction Hypothesis of Jurafsky et al. (2001) proposes that probability is the causal factor driving acoustic reduction of higher-frequency forms. The relevant probabilities can be highly context-sensitive, and depend on the preceding and following words and even the probabilities of word n-grams (Bell et al., 2003; Pluymaekers et al., 2005; Tremblay and Tucker, 2011). Aylett and Turk (2004, 2006) build on the Probabilistic Reduction Hypothesis with their Smooth Signal Redundancy Hypothesis. They argue that high-frequency forms have a lower degree of communicative information, which is argued to render them more redundant, and hence more prone to reduction, i.e., to shortening and to less distinct articulation. By contrast, the high information load of low-frequency words is hypothesized to require lengthening and more distinct articulation. As a consequence, the speech signal is argued to be smoother than it otherwise would be, in the sense that the amount of information transmitted per unit of time is more constrained and less variable. Evidence for the Smooth Signal Redundancy Hypothesis is not restricted to vowel durations, but extends to their spectral characteristics, with more centralization for high-frequency words (Aylett and Turk, 2006).

A second force argued to shape the acoustic characteristics of words is lexical neighborhood structure. Neighbors are words that are phonologically similar, for example *lad*, *read*, and *leak* are all neighbors of *lead*. Research on phonological neighborhood density in speech production has shown both inhibitory and facilitatory effects on speech production when the neighborhood is large. Inhibitory effects for large neighborhoods have been posited on the basis of positive correlations of neighborhood size with vowel dispersion (Wright, 2004; Munson and Solomon, 2004; Munson, 2007), vowel duration (Munson, 2007), nasal coarticulation (Scarborough, 2004, 2010, 2013), and stop voice-onset-times (Baese-Berk and Goldrick, 2009; Fox et al., 2015; Goldrick et al., 2013).

Investigations by Flemming (2010) and Gahl (2015) point out some shortcomings of this previous research and indicate that some of these early findings may have been due to limitations in the design, the kind of speech analysed, or the way the data were analysed. Gahl et al. (2012) studied the effects of frequency and phonological neighborhood on acoustic duration in the Buckeye Corpus of Conversational Speech (Pitt et al., 2007), instead of words produced in isolation in the lab, and observed that vowels were more centralized and shorter in words with large neighborhoods when frequency is statistically controlled. Gahl and Strand (2016) also demonstrate, using the same spontaneous speech corpus, that words with large neighborhoods have shorter durations than words with small neighborhoods. Gahl and colleagues interpret these results as indicating that pronunciation is driven by the relations between words in the mental lexicon. The neighborhood density effect could also reflect speakers' accumulated practice with the articulation of sound sequences. Since words with more neighbors share articulatory trajectories with many phonological neighbors (see also Nusbaum, 1985), the greater experience with these partial trajectories would then give rise to shorter acoustic durations, and more centralized vowels, i.e., less distinct articulations, for words with larger neighborhoods.

However, the interpretation of the consequences of frequency and probability for speech production may not be as straightforward as the above studies suggest. Evidence obtained with electromagnetic articulography suggests that the articulatory trajectories for higher-frequency words may, all other things being equal, show signs of more expert articulatory control, with more articulatory differentiation, instead of less distinct articulation. Tomaschek et al. (2013, 2014, 2017) observed for higher-frequency words more extreme articulatory trajectories of tongue sensors and earlier co-articulatory anticipation of upcoming inflectional suffixes. (Note that earlier anticipatory co-articulation is consistent with higher degrees of vowel centralization as reported for vowels in higher-frequency words.) In other words, with increasing experience, speakers master more difficult articulatory challenges, just as someone learning to play the cello will, over time, be able to execute difficult passages both faster (with shorter durations) and more beautifully (with improved motor control). The experience gained over time with the repeated use of words may result in more skillful use of the articulators. A finding pointing in the same direction is that over the lifetime, the vowel space of speakers of English may be expanding (Gahl and Baayen, 2017).

A third force exerting pressure on the acoustic characteristics of words was reported by Kuperman et al. (2007), who studied the duration of interfixes in Dutch compounds (e.g., *oorlog*-**s**-*verklaring* 'announcement of war', and *dier*-**en**-*arts* 'veterinary'). Several previous studies had indicated that the choice of interfixes is determined by the conditional probability of an interfix in the distribution of interfixes following the initial constituent (Krott et al., 2001, 2002). Kuperman and colleagues observed that the acoustic duration of interfixes was positively correlated with this conditional probability, thus, as the conditional probability increased, the duration of the interfix also increased. Their finding contradicts predictions of the Smooth Signal Redundancy Hypothesis, which expects durational shortening for the more probable and hence less informative interfix. Kuperman et al. (2007) proposed the Paradigmatic Signal Enhancement Hypothesis to account for this unexpected finding. Kuperman et al. argued that in the absence of syntagmatic rules providing unambiguous support for a given morphological realization, morphological forms with strong paradigmatic support will be enhanced. In other words, in pockets of grammatical indeterminacy, paradigmatic support is taken to allow speakers to realize morphological forms more confidently, resulting in longer durations. Recent work by Cohen (2014) has found additional evidence supporting the claims of the Paradigmatic Signal Enhancement Hypothesis.

Kuperman et al. (2007) predicted that a similar process of paradigmatic durational strengthening should hold for irregular verbs in English. Specifically, irregular past tense verbs which receive substantial paradigmatic support from other irregular past tense forms sharing the same kind of vowel alternation (e.g., *sing*/*sang* and *ring*/*rang*) should reveal effects of signal enhancement, such as increased acoustic duration and more extreme formant characteristics. For English irregular verbs, the way to code paradigmatic support (Bybee and Slobin, 1982; Bybee and Moder, 1983) would be to count the number of irregular verbs that share the same vocalic alternation, henceforth, a verb's gang size.

One purpose of the present study is to test this prediction, and to clarify, using the Buckeye corpus (Pitt et al., 2007), whether verbs are indeed realized with longer vowel durations when the morphological tense has stronger paradigmatic support, as predicted by the Paradigmatic Signal Enhancement Hypothesis. As shown by Plag et al. (2017) for the acoustic duration of word-final s in English, the morpho-semantic function of a morphological exponent can be a co-determinant of its phonetic realization.

The second purpose of our study is to reflect on this issue from the perspective of discrimination learning. As the above literature review reveals, the general conceptual framework within which acoustic strengthening and weakening is discussed builds on word form units. The effect of frequency is typically assumed to reflect word forms' resting activation levels, whereas the number of neighbors is taken to reflect the extent to which words with similar forms hinder or gang up (compete) during lexical access. In what follows, we refer to this family of models as word form competition (WFC) approaches.

For auditory comprehension and speech production, WFC approaches are highly problematic given the enormous variability of the spoken word. If this variability would only concern Gaussian noise on a canonical form, the problem would be less severe than it actually is. The variation in the spoken word is such that segments and even whole syllables can be missing, a widespread phenomenon described by Johnson (2004) as 'massive reduction' (see also Ernestus, 2000; Dilts, 2013). This property of spoken language raises many issues, such as whether reduced forms have their own form frequency effect, and whether reduced forms are neighbors of other reduced forms of the same word (compare, for instance, a subset of the variants of Dutch *natuurlijk*, 'of course': tyk, tək, ntyk, tylək, tyrlək). A further problem is that words spliced from spontaneous speech tend to be very difficult to understand, with low identification rates in the range of 20–40% (Ernestus et al., 2002; Arnold et al., 2017). If words indeed have their own form representations, as assumed by WFC approaches, then why is it that these form representations fail to give listeners access to their semantics? Computational WFC approaches are typically engineered in such a way that given the input, the correct word is often recognized with $p = 1$, and likely present an overly optimistic perspective on human comprehension.

To avoid these problems with word forms, we therefore also made use of measures derived from the Buckeye corpus by means of discrimination learning to better understand the forces driving signal enhancement and signal reduction. Our computational implementation, inspired by the work of Ramscar and Yarlett (2007); Ramscar et al. (2010), systematically explores the consequences of error-driven learning (Baayen et al., 2011, 2016b; Milin et al., 2017b). This novel theory moves away from the phoneme as a basic unit (see Port and Leary, 2005; Ramscar and Port, 2016; Arnold et al., 2017, for why this is a desideratum for linguistic theory). Given that formant transitions within vowels are crucial for discriminating the place of articulation of flanking consonants, and following Baayen et al. (2011), who used letter pairs to model visual comprehension, the present study explores diphones as sublexical features. Naive discriminative learning offers a way of assessing the functional load of diphones (see Wedel et al., 2013, for an operationalizing of the functional load for phonemes using minimal pairs). Of special interest to us is the 'discriminative load' of the diphones for the transitions into and out of the vowel.

Naive discriminative learning (NDL) departs from standard conceptualizations not only by moving away from phonemes, but also by eschewing word form representations. NDL sets itself the goal to pursue whether it is possible to discriminate between lexical meanings straightforwardly on the basis of sublexical features of form (in the present study, diphones). In this approach, there is no competition between form units, and there are no gangs of similar form units that would co-determine acoustic durations. Because there are no form units, reduced and unreduced variants of a given word are not a problem: We can train the model on what people actually said (using the diphones of a phonetic transcription) to predict what people actually meant. A straightforward prediction follows: If we were to train the model on the canonical dictionary forms, prediction accuracy for the empirical vowel durations should be inferior compared to a model trained on what speakers actually said.

Of course, such a radical departure from the standard conceptualization of language comprehension and speech production raises the question of whether measures grounded in discrimination learning are truly competitive with classical measures based on WFC approaches, such as word frequency, neighborhood density, and gang size. Do discriminative measures improve our understanding of the competing forces for which the diametrically opposed hypotheses of the smooth signal and the enhanced signal have been formulated?

To anticipate the results, statistical models for vowel duration that have access to measures derived from discrimination learning indeed outperform models based on classical predictors, and thus provide a new perspective on the opposing forces shaping acoustic durations in speech production.

In the next section we provide a brief introduction to key concepts of naive discriminative learning. We then provide further detail on the data set which we extracted from the Buckeye Corpus of Conversational Speech (Pitt et al., 2007) and the set of verbs that we examined. The data are analysed using methods from both machine-learning and regression analysis, i.e., with random forests (Strobl et al., 2009) and with generalized additive mixed models (Hastie and Tibshirani, 1990; Wood, 2006; Baayen et al., 2017c). We conclude this study with a discussion of the forces that shape stem vowel durations in regular and irregular verbs and the implications of our findings for understanding speech production.

2. Naive Discriminative Learning

As previously noted, more frequent words tend to have shorter durations (see e.g., Bell et al., 2003). Frequency counts, and probability estimates based on these counts, presuppose that the counted objects or events can be discriminated from each other. Discrimination, in turn, presupposes that objects or events have characteristic features, henceforth *cues*, on the basis of which of these objects or events, henceforth *outcomes*, can be distinguished. Naive discriminative learning (NDL) seeks to quantify the consequences of discrimination for language structure and processing by using a powerful yet simple formalization of error driven learning, the learning rule of Rescorla and Wagner (1972). The Rescorla-Wagner learning rule is closely related to the perceptron (Rosenblatt, 1962) and the adaptive learning rule of Widrow and Hoff (1960). A range of empirical studies indicate that error-driven learning characterizes important aspects of human learning (Ramscar and Yarlett, 2007; Marsolek, 2008; Ramscar et al., 2010) and language processing (Ellis, 2006; Baayen et al., 2011, 2016b). Furthermore, using as cues acoustic features derived from the speech signal, a discrimination-based model for auditory word recognition, trained on 20 hours of German conversational speech containing some 13,000 different words, performed within the range of human accuracy (Arnold et al., 2017).

Previous research has shown that quantitative measures derived from NDL networks trained on large corpora are competitive with lexical measures based on counts of occurrences or counts of lexical neighbors (Baayen et al., 2016a; Hendrix, 2015; Milin et al., 2017a,b). The present study therefore also explores the potential of NDL-based learning measures for understanding acoustic durations and the competing claims of the smooth signal redundancy hypothesis and the paradigmatic signal enhancement hypothesis.

ARPAbet	IPA	lazy	lead	leadership	leads	PAST	PRESENT
ey.z	/eɪz/	0.0171	0.0003	-0.0000	-0.0000	0.0169	0.1964
l.iy	/li/	-0.0071	0.0036	0.0020	0.0017	0.0553	0.0319
d.#	/d#/	0.0008	0.0001	-0.0024	0.0001	0.0553	0.0384
ae.f	/æf/	-0.0023	-0.0002	0.0007	-0.0000	-0.0321	0.0487
iy.dx	/iɾ/	0.0028	0.0010	0.0049	-0.0007	0.0511	0.0717
iy.ih	/iɪ/	0.0018	-0.0023	-0.0017	-0.0011	-0.0522	-0.0600
dx.er	/ɾɹ̩/	0.0017	-0.0011	0.0016	-0.0005	0.1018	-0.0846
f.ih	/fɪ/	-0.0015	0.0010	0.0004	0.0001	-0.0253	0.0380
iy.d	/id/	0.0021	0.0041	0.0038	0.0036	-0.0304	0.0731
#.l	/#l/	0.0015	0.0021	0.0012	0.0011	0.0147	-0.0004

Table 1: Sample weight matrix calculated on the observed phonetic realizations in the Buckeye corpus. Each row represents a diphone, represented with the ARPAbet and IPA transcriptions systems, and its association weights to the lexical outcomes. Lexical outcomes (lexomes) are listed above the columns. The weights for the connections from the cues for *lead* to the lexome LEAD are highlighted in dark gray. The vowel diphones for *lead* and their weights to tense are highlighted in light gray.

An NDL network consists of an input layer of cues and an output layer of outcomes. We trained an NDL network on the Buckeye corpus, using as cues the diphones as available in the phonetic transcriptions in this corpus. Thus, words with reduced forms contributed the diphones for these reduced forms, not those of the canonical phonological form. As outcomes, we used the lemmata of the word forms in the corpus, as well outcomes for present and past tense. In what follows, we refer to these output units as lexical outcomes, or lexomes (see below for further discussion). This resulting network is summarized by a weight matrix $\boldsymbol{W}$ of dimension 2524 diphone cues × 7222 lexical outcomes. Each cell $\boldsymbol{w}_{ij}$ of $\boldsymbol{W}$ specifies the estimated association strength of the i-th cue to the j-th outcome. Part of the weight matrix is shown in Table 1. This example illustrates diphones as cues (rows) and has as outcomes (columns) a sample of the lexical outcomes, including the outcomes for PAST and PRESENT.

In the framework of naive discriminative learning, the lexical outcomes are interpreted not as units of form, but rather as pointers to vectors in a high-dimensional semantic space. The term 'lexome' was introduced to be able to distinguish, terminologically, between a word's form, a word's dictionary entry (lexeme), and a word's semantics, on the one hand, and in the NDL framework, a word's pointer to a location in semantic space on the other hand.

When a word is presented to the network, the diphones of that word are extracted, and one unit of activation is propagated through the network for each of these active diphones. For a given lexical outcome, the activation received from all active cues is summed. We refer to this sum as the activation of the lexome. This sum is identical to the sum of the weights on its afferent connections. Thus, for the outcome LEAD, the activation is the sum of the weights from `#.l, l.iy, iy.d`, and `d.#` to LEAD. In Table 1, these connection weights are highlighted in dark grey.

The `l.iy` cue (highlighted with light gray) supports the lexical outcomes LEAD, LEADERSHIP, and LEADS, but has a negative weight for the outcome LAZY. The `iy.d` diphone (also highlighted with light gray) has positive weights to all content lexomes, with strongest support again for LEAD. The contributions of the edge cues (`d.#` and `#.l`) to LEAD are small, especially in the case of `d.#`: Edge cues tend to be shared by many words, and hence are less discriminative.

Of special interest to us are the weights on the connections from the diphone transitions into and out of the stem vowel of a verb to the lexomes for present and past tense. In Table 1, the weights on the connections from the diphones `l.iy` and `iy.d` to the lexomes for tense are highlighted with light gray. The total support for the two tenses provided by the diphones of LEAD is $0.0553 + 0.0553 - 0.0304 + 0.0147 = 0.0949$ for PAST, and $0.0319 + 0.0384 + 0.0731 - 0.0004 = 0.1430$ for PRESENT. The model thus provides stronger support for `present` than for `past`, as expected given that we are considering the diphones of a present-tense form.

Two further measures, calculated from the NDL weight matrix, have been found to be excellent predictors of lexical processing costs (see Arnold et al., 2017; Milin et al., 2017b,a; Baayen et al., 2016a). These measures are the L1-norms of the row and column vectors of the weight matrix. The L1-norm, widely used in machine learning (Hastie et al., 2001), is a distance measure. It quantifies the distance traveled from one point to another under the restriction that one can move only in parallel with the axes. Thus, while the Euclidean distance (the L2-norm) from the origin to the point (-3,4) equals 5, the L1-norm, also know as city-block distance, is 7. The city-block distance and the Euclidean distance are highly correlated for the present data, but, as also observed by the above mentioned studies, the city-block distance turned out to be the superior predictor for vowel durations.

The L1-norm for a cue i, $\sum_j |\boldsymbol{w}_{ij}|$ is calculated over its row in the weight matrix. Distributions of connection weights are spiky, with most of the probability mass concentrated near zero, as illustrated for the diphone `l.iy` in Figure 1 by the density estimate (blue line). Typically, both strong negative and strong positive weights develop for a small number of lexical outcomes. Figure 1 shows all lexical outcomes for which the absolute connection strength exceeds 0.05. L1-norms are determined primarily by such outliers. Note that the lexical outcomes with large absolute afferent weights need not contain the cue itself (e.g., *culture* in Figure 1 does not contain `l.iy`). Such weights can develop as a consequence of the same cue being used for other lexical outcomes, such as the word *culturally*. (see also Mulder et al., 2014, for a discussion of how this

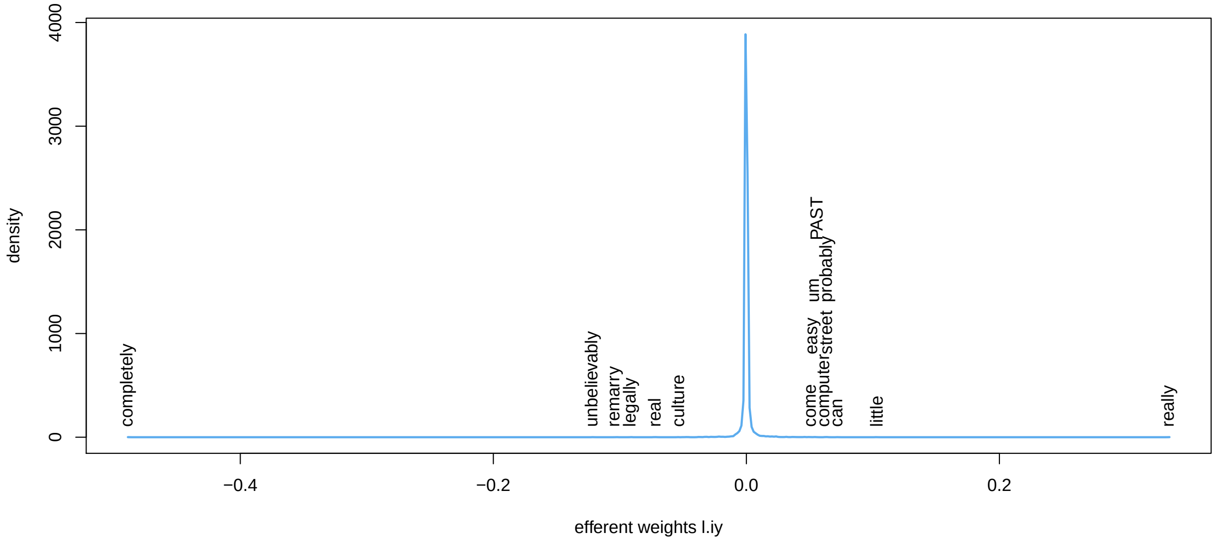

Figure 1: The ‘spiky’ density of the weights on the efferent connections of the vowel diphone cue `l.iy`. Most weights are very close to zero, but larger negative and positive weights exist for a minority of lexomes. Shown are all lexomes for which the absolute weight exceeds 0.05.

co-learning can give rise to the secondary family size effect).

In what follows, we refer to the sum of the L1-norms for the cues representing articulations into and out of the vowel (`l.iy` and `iy.d` for the present tense form *lead*) as the activation diversity (`a-diversity`) of the vowel cues. It is a measure of the extent to which other outcomes compete with the targeted outcome. A small `a-diversity` indicates that the cue does not provide strong support for any lexical outcomes. A large `a-diversity` indicates that the cue provides strong support for at least one, but typically more than one, lexical outcome. In other words, a large `a-diversity` is an index of a lack of discrimination for the target word. At the same time, a large `a-diversity` implies that partial articulations into and out of the vowel are shared with many other lexical outcomes, and perhaps are well-practiced for articulation.

The L1-norm of the column vector in the weight matrix of a trained model (e.g., the column vector for the outcome LEAD, which specifies the weights on the connections to LEAD from all cues known to the model) will be referred to as this outcome’s `prior availability`. Unlike the activation diversity, the prior availability does not depend on the input presented to the network. Instead, it assesses the entrenchment of the outcome in the network. This measure tends to be correlated with the activation measure. The L1-norm for a lexome j, $\sum_i |\boldsymbol{w}_{ij}|$ is independent of the cues that are actually present in a given input, and thus functions as a measure of the lexome’s `prior availability`, similar to the priors of Bayesian theories such as Shortlist-B (Norris and McQueen, 2008). A lexome’s `prior availability` also tends to be strongly correlated with, but not identical to, its frequency of occurrence.

Following Milin et al. (2017b); Baayen et al. (2016a), the semantic space is itself constructed with a second discrimination network. For the present study, this network was trained on the utterances in the Buckeye corpus, now using the words in the utterances both as cues and as outcomes. The rows of the resulting 9636 × 9636 weight matrix $\boldsymbol{V}$ are semantic vectors, in the sense that the correlation of two row vectors (or equivalently, the cosine of the angle between the two vectors) is a measure of the extent to which the corresponding lexical outcomes are similar in meaning. Although measures from classical models for distributional semantics such as LSA, HAL or word2vec (Landauer and Dumais, 1997; Lund and Burgess, 1996; Mikolov et al., 2013) could have been used, we wanted to clarify whether measures based on semantic vectors derived from the Buckeye corpus itself contribute to understanding acoustic durations in spontaneous speech. Thus, lexomes are pointers to, or indices of, semantic vectors in the space defined by $\boldsymbol{V}$.

We derived two semantic measures from $\boldsymbol{V}$. First, we calculated the L1-norms for the row vectors of $\boldsymbol{V}$. In what follows, we refer to this measure as the semantic activation diversity (`semantic a-diversity`). This measure, which is the city-block length of a lexome's semantic vector, reflects the extent to which a given word predicts other words. A word that has hardly any collocational preferences will have a `semantic a-diversity` close to zero. The more a word shows preferences or dispreferences for other words, the higher its `semantic a-diversity` is.

Second, from $\boldsymbol{V}$, we derived the correlation matrix $\boldsymbol{C}$, with entry $c_{i,j}$ specifying the correlation between rows i and j of $\boldsymbol{V}$. In what follows, we refer to the correlation between a given row vector $\boldsymbol{c}_i$ of $\boldsymbol{C}$ and the average of all row vectors of $\boldsymbol{C}$ as the `semantic typicality` of lexome i. The greater the semantic typicality of lexome i is, the more it resembles, semantically, the average lexome. A low typicality indicates that a word's lexical co-occurrence patterns are distributed differently from what one would expect on average.

The discrimination networks were trained incrementally on the full Buckeye corpus, using the order of words in which they appear in the corpus for the first network (weight matrix $\boldsymbol{W}$) and the order of utterances for the second network (weight matrix $\boldsymbol{V}$). For each word, c.q. utterance, the learning rule of Rescorla and Wagner was applied to update the weights on the networks' connections from diphone cues to lexome outcomes.

3. Data

To investigate the forces that shape stem vowel duration, we used data from the Buckeye Corpus of Conversational Speech (Pitt et al., 2007). The Buckeye Corpus comprises spontaneous interview speech with about 350,000 words from 40 speakers, split across gender (20 male, 20 female) and age (20 old, 20 young) from the Columbus, Ohio area. The corpus includes time-aligned orthographic and phonetic transcriptions of the recorded speech. Each transcription contains time stamps denoting the word and segmental boundaries. The corpus also provides two transcriptions of each word: a dictionary entry style transcription and a phonemic/broad phonetic transcription of the word (which is closer to the actual production). We calculated naive discriminative learning measures for both transcriptions. As is documented in more detail below, training the NDL network on the phonetic transcriptions provided superior measures for predicting acoustic durations compared to training on the 'canonical' forms. In the analyses reported below, we therefore only discuss measures derived from the phonetic transcriptions.

We created a list of all verbs in the Buckeye Corpus by using the part of speech tags provided with this corpus. The full list was further restricted to irregular and regular verbs which are monosyllabic in their present tense form, which includes regular verbs that may be disyllabic in the past tense (e.g., *walk/walked*, *code/coded*). The set of regular verbs serves as a control for comparison to the irregular verbs, since regular verbs do not have a vocalic alternation in their past/present tense morphological derivations.

We limited the irregular verbs to those which differ between their past and present forms by a single vowel alternation (e.g. *sing/sang*, *hold/held*, *lead/led*). Thus, verbs such as *weep/wept*, that differ by more than the vowel alternation, were excluded from this study. As a consequence, the vowel of irregular verbs is the segment which carries the majority of the tense information.

A total of 11,061 verbs were extracted from the corpus, of which 6,456 were irregular and 4,605 were regular. For each verb, its duration was extracted from the corpus along with the duration of each of its segments. Of these segment durations, the duration of the stem vowel is the variable of focus.

4. Predictors

We group the predictors in four categories: speech control variables, speaker variables, stem vowel variables, and words carrying these stem vowels variables.

4.1. Speech Control Variables

The duration of a stem vowel largely depends on the local speech rate, as well as on whether the carrier word is in phrase-final position, where words tend to be lengthened (Klatt, 1976; Wightman et al., 1992). The local `speech rate` was estimated by first using the pauses in the corpus to determine phrase boundaries, next, the number of syllables per second were calculated for each phrase (Dilts, 2013). A factor for `Phrase-Final Position` was included, using treatment dummy coding with non-final position as the reference level.

4.2. Speaker

`Speaker` is a random-effect factor distinguishing the 40 speakers of the Buckeye Corpus, with 10 speakers for each combination of `Age` (young/old) and `Sex` (female/male). `Age` and `Sex` are fixed effect factors, entered into the models with treatment coding, with 'old' and 'female' as the reference levels.

4.3. Vowel

A fixed-effect factor, `Vowel Quality` (tense/lax), distinguishes between phonologically tense and lax vowels; acoustically the tense vowels are generally longer than the lax vowels in English. In this data set, diphthongs are combined with the tense vowels as an initial analysis indicated that there was no statistical difference between the group mean durations. We used treatment coding, with 'lax' as the reference level.

Vowel-Tense Activation is defined as the sum of the weights on the connections in the network from a vowel's diphones to the tense lexome (PRESENT or PAST). We expected stronger activations to correspond to longer vowel durations. This correspondence should be especially strong for irregular verbs, as it is here that different vowels are in paradigmatic opposition.

4.4. Word

Word forms were cross-classified by Tense (present/past) and Regularity (irregular/regular). We again used treatment coding for these fixed-effect factors, with 'past' and 'irregular' as reference levels.

Classic word-related predictors are frequency, length, and neighborhood density. More frequent words (Bell et al., 2003), and in a stress-timed language such as English, longer words, tend to have shorter segments (Kemps et al., 2005). For neighborhood density, it has been argued that a greater density implies more shared articulatory gestures and hence affords faster articulation (Gahl et al., 2012; Stemberger, 2004; Vitevitch, 1997, 2002; Vitevitch and Sommers, 2003).

Frequency of occurrence (Frequency) was defined as the number of occurrences of the spoken forms in the Buckeye Corpus. In addition, we included as predictors the frequency of the next word (ForwardFrequency) and the bigram frequency of the target word and the next word (ForwardBigramFreq). These measures allowed us some control over the probability of the upcoming word, as well as the conditional probability of the current word given the next word, $p(w_1|w_2) = p(w_1, w_2)/p(w_2)$.

For Length, we used the number of phones in the corresponding canonical citation form. We evaluated neighborhood density by means of a count of the word forms that differ in one segment. We calculated these neighbors both for the spoken forms in the Buckeye Corpus (NcountBuckeye) and for the citation forms given in the IPHOD database (Vaden et al., 2009), henceforth Ncount.

For each verb, we calculated its Gang Size, the number of verbs sharing the same vowel alternation (see also Bybee and Slobin, 1982; Bybee and Moder, 1983). For instance, *sing-sang* and *ring-rang* are part of the same 'gang'. Regular verbs were grouped into one large 'identity gang'. In the distribution of Gang Size, the size of the regular gang is an outlier. The two theoretical frameworks that we compare in this study, word form competition (WFC) and naive discriminative learning (NDL), make opposite predictions. Given the WFC perspective, a past-tense form with a large gang size has strong paradigmatic support which, given the paradigmatic signal enhancement hypothesis should give rise to longer durations, just as interfixes with stronger paradigmatic support have longer acoustic durations (Kuperman et al., 2007). However, from the perspective of NDL, it is less clear what to expect. On the one hand, a large gang size implies that diphone transitions into and out of the vowel are shared across many verbs. Hence these diphones are less effective discriminators for the verbs' lexomes. With less support from these diphones, acoustic durations are expected to be shorter instead of longer. On the other hand, a large gang size implies that there are many verbs that have diphones into and out of the vowel that support a given present or past tense lexome. With more support for the appropriate tense lexome, acoustic durations may increase. Thus, from an NDL perspective, there are two opposing forces at issue, and only the data can inform us about their joint effect. The vowel predictor `Vowel-Tense Activation`, introduced above, will allow us to explore these effects from the NDL perspective.

At the word level, we considered four learning measures. The first is a lexome's `Prior Availability` in the network (the L1-norm of the lexome's column vector in weight matrix $\boldsymbol{W}$). Our expectation is that, just like frequency of occurrence, with which `Prior Availability` is strongly correlated ($r = 0.91$), greater priors will give rise to shorter durations.

Our second learning measure is the `Activation Diversity`, the sum of the L1 norms of the word's diphone row vectors in $\boldsymbol{W}$. A greater activation diversity implies greater uncertainty: The diphones perturb to a greater extent the state of the lexical system when the activation diversity is higher (Baayen et al., 2017a). Therefore, shorter acoustic durations are predicted (see Arnold et al., 2017, for independent evidence). This prediction is actually in line with the paradigmatic signal enhancement hypothesis, according to which durations increase under low entropy (the case in which one interfix has a much higher probability than the other interfixes in the Kuperman et al. (2006) study), and decrease under high entropy (when interfixes are roughly equiprobable).

The third discrimination measure is `Semantic Activation Diversity` (the L1-norm of a lexome's semantic vector in the semantic space estimated through the network $\boldsymbol{V}$, i.e., the city block length of the semantic vector). If a word enters into more and stronger collocational relations, then it is more semantically confusable. As it does not make sense to durationally amplify a signal that is only creating confusion and cannot contribute to discrimination, we expect the acoustic duration of the vowel to decrease with increasing semantic a-diversity.

The fourth learning measure is `Semantic Typicality`. This measure is defined as the cosine similarity of a lexome to the mean semantic vector obtained by averaging over the row vectors of $\boldsymbol{V}$. Since semantically typical words are unsurprising, the smooth signal redundancy hypothesis predicts that acoustic durations should decrease with increasing semantic typicality. As semantically more typical words are more similar to each other and hence less discriminable, it follows from NDL that words with a high semantic typicality have shorter vowel durations.

Many of the numeric variables in our data have distributions that are suboptimal for regression analysis. In order to avoid artefactual effects due to outliers or other intermittent intervals of data sparseness, we transformed many of the variables. The response variable was log-transformed (as indicated by a Box-Cox analysis (Box and Cox, 1964)), as was frequency of occurrence. Note that a logarithmic transform of the frequencies allows us to rewrite the conditional probability of the current word given the next word as follows: $\log(p(w_1|w_2)) = \log(p(w_1, w_2)) - \log(p(w_2))$. The negative slope for $\log p(w_1|w_2)$ observed by Bell et al. (2003) predicts a negative slope for $\log(p(w_1, w_2))$ and a positive slope for $\log(p(w_2))$, with the two slopes having roughly the same absolute magnitude. The `Prior Availability` was subjected to a power transform with exponent 0.1, which turned out to be optimal for reducing the strong rightward skew of this variable. Furthermore, all numeric predictors were centered and scaled.

	Rate	ForFreq	ForBiFreq	SubtFreq	NCountBuck	Sem A-div	GangSize	Length	FormFreq	NCount	SemTyp	V2TAct	A-Div	Prior
Rate	1.00	0.03	0.16	0.14	0.02	0.14	-0.08	-0.08	0.14	0.04	-0.06	0.07	-0.04	0.13
ForFreq	0.03	1.00	0.48	0.01	-0.04	0.00	0.11	0.01	0.01	-0.01	-0.10	-0.03	-0.08	-0.04
ForBiFreq	0.16	0.48	1.00	0.46	-0.01	0.51	-0.23	-0.26	0.51	0.17	-0.24	0.30	-0.09	0.42
SubtFreq	0.14	0.01	0.46	1.00	-0.10	0.91	-0.62	-0.59	0.92	0.44	-0.40	0.64	-0.21	0.91
NCountBuck	0.02	-0.04	-0.01	-0.10	1.00	-0.07	0.03	0.16	-0.07	-0.31	0.18	0.03	0.02	-0.06
Sem A-div	0.14	0.00	0.51	0.91	-0.07	1.00	-0.56	-0.60	1.00	0.42	-0.38	0.64	-0.15	0.89
GangSize	-0.08	0.11	-0.23	-0.62	0.03	-0.56	1.00	0.51	-0.56	-0.40	0.08	-0.53	0.14	-0.62
Length	-0.08	0.01	-0.26	-0.59	0.16	-0.60	0.51	1.00	-0.60	-0.73	0.43	-0.44	0.50	-0.48
FormFreq	0.14	0.01	0.51	0.92	-0.07	1.00	-0.56	-0.60	1.00	0.43	-0.39	0.64	-0.15	0.89
NCount	0.04	-0.01	0.17	0.44	-0.31	0.42	-0.40	-0.73	0.43	1.00	-0.32	0.21	-0.26	0.35
SemTyp	-0.06	-0.10	-0.24	-0.40	0.18	-0.38	0.08	0.43	-0.39	-0.32	1.00	-0.18	0.25	-0.17
V2TAct	0.07	-0.03	0.30	0.64	0.03	0.64	-0.53	-0.44	0.64	0.21	-0.18	1.00	-0.06	0.69
A-Div	-0.04	-0.08	-0.09	-0.21	0.02	-0.15	0.14	0.50	-0.15	-0.26	0.25	-0.06	1.00	-0.09
Prior	0.13	-0.04	0.42	0.91	-0.06	0.89	-0.62	-0.48	0.89	0.35	-0.17	0.69	-0.09	1.00

Table 2: Pairwise Pearson correlations of the numeric predictors for the acoustic duration of the stem vowel. ForwFreq: Forward Frequency; FwBiFreq: Forward Bigram Frequency; SubtFreq: film subtitle frequency; NcntBuck: neighhbor count based on the Buckeye corpus; SemA-Dev: Semantic Activation Diversity; SemTyp: Semantic Typicality; V2TAct: Vowel to Tense Activation; A-Div: Activation Diversity; Prior: Prior Availability.

Table 2 shows the results of pairwise Pearson correlations of the numeric predictors. Buckeye `Frequency` shows the expected negative correlation with `Length` ($r = -0.51$). It is also negatively correlated with `Gang Size` ($r = -0.52$) and positively correlated with film `Subtitle Frequency` ($r = 0.92$). Strong positive correlations are present for Buckeye `Frequency` with `Prior Availability` ($r = 0.91$) and `Semantic Activation Diversity` ($r = 0.94$), as well as a more modest correlation with `Vowel-Tense Activation` ($r = -0.41$). `Gang Size` shows moderate correlations not only with `Frequency` ($r = -0.52$) and `Length` ($r = -0.58$), but also with `Semantic Activation Diversity` ($r = -0.58$) `Vowel-Tense Activation` ($r = -0.56$) and `Prior Availability` ($r = -0.50$). The learning-based alternative for `Gang Size`, `Vowel-Tense Activation`, was most strongly correlated with `Semantic Activation Diversity` ($r = 0.60$), `Gang Size` ($r = -0.56$), `Form Frequency` ($r = 0.57$), and `Prior Availability` ($r = 0.59$). It is noteworthy that the Buckeye form-based density measure (`NCountBuckeye`) shows a moderate negative correlation with the `NCount` measure ($r = -0.32$).

Figure 2 provides a further illustration of the correlations among the predictors by means of a heatmap. A hierarchical cluster analysis of the correlation matrix $\boldsymbol{C}$, using $\mathbf{1} - \boldsymbol{C}$ as distance matrix, is depicted in the left and upper margins of the heatmap, and rows and columns of the correlation matrix are reordered accordingly. The heatmap suggests two clusters with medium to strong positive correlations internally, and negative correlations with the predictors of the other cluster. The smaller cluster comprises `GangSize, Length, NcountBuckeye, Semantic Typicality`, and `Activation Diversity`, the other group comprises within its largest subcluster the variables `Ncount, Vowel-Tense Activation, Prior Availability, Form Frequency, Subtitle Frequency`, and `Semantic Activation Diversity`.

5. Results

5.1. Random forest analysis

We first analyzed the log-transformed duration of the vowel with a machine learning method, random forest analysis, using the `party` package version 1.0-25 (Strobl et al., 2009) for `R` (R Core Team, 2015). The random forest is a non-parametric machine-learning modeling technique that iteratively implements conditional inference trees by performing binary splits on the predictor values and subsequently selecting the most useful predictors (see Tagliamonte and Baayen, 2012, for a more detailed explanation and application to linguistic data). We use this non-parametric method as a way to investigate the variable importance of our predictors. It also allows us to check for convergent evidence from two different modeling methods: machine-learning and regression. Random forests do not depend on any of the many assumptions of and requirements for regression modeling, and therefore provide a window on the relative importance of the predictors independent of the transformations applied to the response and covariates. Furthermore, random forests are exquisitely sensitive to complex interactions that escape modeling with GAMMs. However, `Word` was not included as predictor, as the large numbers of words in the Buckeye corpus makes the random forest calculations computationally

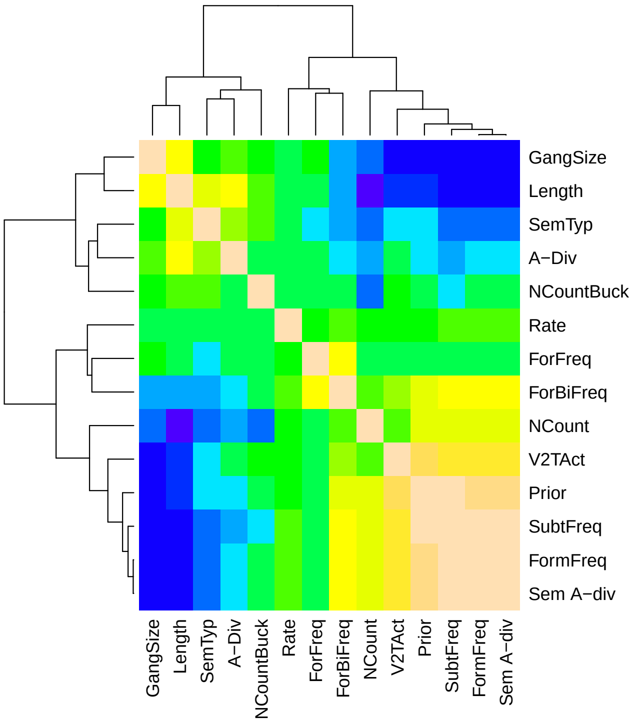

Figure 2: Heatmap for the Spearman correlation matrix of the variables in Table 2. Darker shades of yellow/brown indicate stronger positive correlations, deeper shades of blue represent stronger negative correlations.

intractable. Figure 3 presents variable importance estimates, obtained by averaging over the variable importance of 10 random forests. The average squared correlation of the predictions of this machine-learning technique and the acoustic durations of the stem vowels was 0.625.

The most important predictor in the random forest model is the `Vowel Quality` itself, followed by whether the vowel occurs phrase-finally (`Phrase Final`), which in turn is followed by `Speech Rate` and `Length` of the written form, which are, for the present purposes, all control variables. `Age` and `Sex` of the speaker are the least predictive. Between the subject variables and the control variables we find the predictors of interest to the present study: `Frequency, Regularity, Tense`, and `Gang Size`, as well as the learning-based measures `Prior Availability`, `Activation Diversity`, `Vowel-Tense Activation`, together with the neighborhood density measures `N-Count`, and the count of neighbors based on the observed (often reduced) forms in the Buckeye Corpus (`N-Count Buckeye`). Of these predictors, `Activation Diversity` and `N-Count` along with the control variable `Speaker` have the largest variable importance. The remaining variables of interest, such as `Forward Bigram Freq`, `Gang Size`, `Vowel-Tense Act` and `Forward Freq` fall in the middle region of the variable importance.

5.2. Analyses with generalized additive mixed modeling

To investigate the functional relations between the predictors and the response variable, we proceeded with fitting generalized additive mixed models (Hastie and Tibshirani, 1990; Wood, 2006, 2011; Baayen et al., 2017c,b) to the vowel durations, using the mgcv package Version 1.8.10 by Wood for R (R Core Team, 2015), fitting the model to the data with the method of maximum likelihood (ML). In the absence of a-priori predictions concerning the potentially nonlinear form of the functional relation between vowel duration and a given predictor, the models presented here are the result of an exploratory investigation of the quantitative structure of the data. Factor contrasts, linear slopes, thin plate regression spline smoothers, or tensor product smooths for interactions of numeric variables were retained only when their inclusion in the model afforded a significant increase in goodness of fit, assessed with the help of a chi-squared test on the difference in ML (maximum likelihood) or fREML (for random effects) scores (as assessed by the compareML function from the itsadug package; van Rij et al., 2016, Version 1.0.1). Furthermore, adversarial attacks on novel predictors were carried out, which, as will become clear below, led to further hypotheses and the exploration of additional theoretically motivated predictors.

The wiggliness of the thin plate regression spline smooths were constrained by setting the maximum number of basis functions to 5. The distribution of residuals from the GAMMs showed marked departure from normality, with long tails characteristic of a t-distribution. We therefore refitted models, using the link

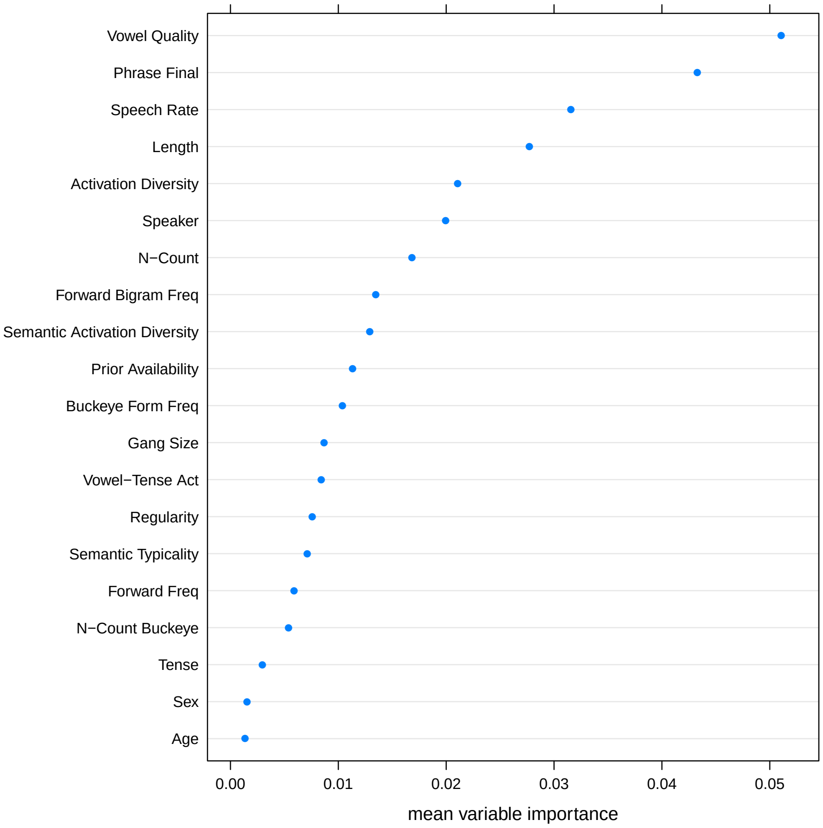

Figure 3: Variable importance averaged over 10 random forest models for the acoustic duration of the stem vowels of English verbs.

function for the scaled t-distribution. That is, given the linear predictor μ, the parameters σ and ν, and the response y, $y - \mu$ scaled by σ is taken to follow a t-distribution with ν degrees of freedom:

$$\frac{y - \mu}{\sigma} \sim \tau_\nu.$$

A quantile-quantile plot (using `qq.gam` from the **mgcv** package) showed that the residuals of models refitted in this way followed the nominal distribution.

Word was not included as a random-effect factor in the model. No less than 205 out of the 505 verbs (41%) occurred once in the corpus. Including word as a predictor, together with other predictors such as word length, vowel quality, tense, and regularity, would result in a heavily over-specified model. We also did not include word length as a predictor, as a cross-tabulation of the predictors: length, vowel quality, age, sex, regularity, and phrase-final position indicated that 13% of the cells had only one observation. After removal of length from this set of predictors, each cell of the design had at least 10 observations. Removal of length also reduced the collinearity of the predictors.

Table 3 reports goodness of fit statistics for five models. The baseline model included only control predictors (phrase rate, vowel quality, age, sex, tense, regularity, phrase-final position, forward frequency, forward bigram frequency, and speaker). The second model had as additional predictors form frequency in the Buckeye, as well as the neighborhood density based on the observed phonetic forms as spoken in the corpus. Thus, this second model includes predictors that are central to word form competition models. The third model replaces the frequency and neighborhood density measures by corresponding counts based on much larger resources (Vaden et al., 2009; Keuleers et al., 2012a). The fourth model replaces counts of occurrences and counts of neighbors with predictors calculated from the Buckeye corpus using naive discriminative learning: `Semantic Activation Diversity`, `Semantic Typicality`, `Vowel-Tense Activation`, and `Activation Diversity`. Prompted by difficulties of interpretation of the effects observed for some of these learning variables, further learning measures were developed, resulting in the fifth model.

In what follows we first discuss the effects of the control variables, which were very similar across models. We then discuss the second and third models with the classic predictors frequency and neighborhood density, and then turn to the models with the NDL predictors. Table 3 indicates that model fits become better as we move down the table. The AIC values listed in this table should be interpreted with caution as the parameters of the scaled t-distribution are not taken into account. Pairwise differences in ML scores are well supported (all $p < 0.0001$) according to the Chi-Square test implemented in `compareML` function available in the **itsadug** package.

5.2.1. GAMM with classical predictors

Table 4 summarizes the GAMM model with classical predictors. This model was fitted with REML, but is otherwise identical to the second model listed in Table 3, which was fitted with ML to allow comparison of models with different fixed-effects structure. Parametric coefficients are listed in the upper subtable (A), and the statistics for the thin plate regression spline smooths, denoted by `s()`, and the random intercepts, denoted

Model	ML Score	Edf	AIC
Baseline model	6303.0	14	12551.9
Classic model, Buckeye-based predictors	6115.8	20	12151.0
Classic model, large resource based predictors	5849.8	20	11598.9
NDL model	5811.6	23	11556.6
Expanded NDL model	5624.2	29	11146.2

Table 3: `ML` score, effective degrees of freedom `Edf`, and `AIC` for five generalized additive mixed models (using the scaled t-distribution) fitted to the duration of the stem vowel.

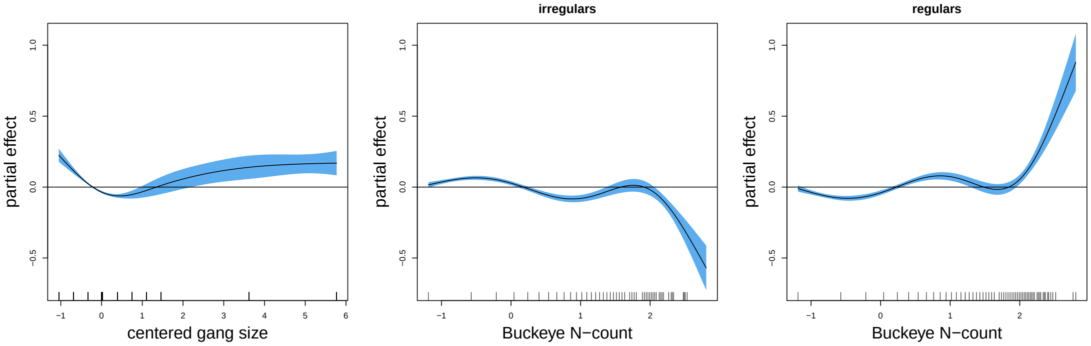

Figure 4: Nonlinear partial effects according to the generalized additive mixed model fitted to the acoustic durations of stem vowels. Left panel: centered gang size. Center and right panels: the interaction of `Buckeye N-count` by `Regularity`. The center panel presents the partial effect of `Buckeye N-count` for irregular verbs, and the right panel its effect for regular verbs.

by `re()`, are listed in the lower subtable (B). The squared correlation of fitted and observed durations was 0.390. Although the random forest ($R^2 = 0.625$) outperforms this regression model by a wide margin, the regression model has the advantage of offering some insight in how predictors influence the response.

First consider the control variables in the upper part of subtable A. `Age` and `Sex` did not have main effects. Table 4 shows that, as expected, vowel durations were lengthened in phrase-final position. Tense vowels were produced with longer durations than were lax vowels. This difference was attenuated for male speakers. Vowel duration decreased as speech rate increased. This decrease was more prominent for tense vowels. Furthermore, for younger speakers, the decrease in vowel duration with speech rate was less marked than for the older speakers. As older speakers have sampled the language for a longer time than younger speakers (Ramscar et al., 2017), this effect of age possibly could be a frequency effect in disguise. Present tense words had stem vowels with longer durations than words in the past tense. Duration increased with the frequency of the next word (`Forward Frequency` but decreased with greater joint frequency of the verb and the next word (`Forward Bigram Frequency`), replicating the shortening effect of the conditional probability of the current word given the following word reported by Bell et al. (2003).

Next, consider the predictors of central interest. Frequency (in the Buckeye corpus) had a linear effect that interacted with regularity, such that irregulars had a positive slope and regulars a negative slope (further tests revealed that both slopes were significantly different from zero, $p < 0.0001$).

Neighborhood density also interacted with regularity. Its effect was nonlinear, with for the majority of data points a downward trend for irregulars and an upward trend for regulars (Figure 4). In other words, the effects of frequency and general trend of neighborhood density have opposite signs, which themselves change sign between irregulars and regulars. Why these two smooths have the observed undulating form is unclear to us at this time.

The gang size measure representing the number of verbs sharing the same vowel alternation was ranked by the random forest with the variables of medium importance. For regulars, the gangsize is many times larger than that for irregulars. Since it might be argued that the gangsize measure only comes into its own for irregulars, we defined a second gangsize measure that was set to zero for the regulars, and that was the scaled version of the gangsize measure for the irregulars. The partial effect of `Centered Gang Size` was U-shaped with a minimum near 0 and narrow confidence intervals for lower predictor values, as can be seen in the left panel of Figure 4. (The partial effects shown in Figure 4 and subsequent GAM plots present the effects of predictors centered around zero. Thus, these partial effects show how the response deviates as a function of the predictor from the group means as defined by the intercept and factorial predictors, with other covariates held constant at their median.) The Spearman correlation of `Centered Gang Size` and stem vowel duration was -0.03, from which we conclude that the strong downward trend in the smooth is not caused primarily by suppression or enhancement. This strong downward trend is exactly opposite to the prediction of the paradigmatic signal enhancement hypothesis.

The frequency and neighborhood density measures that we used as predictors were both calculated from the Buckeye corpus. For the neighborhood measure, this made it possible to base counts on what speakers actually said, rather than on the canonical forms as found in dictionaries. Furthermore, as the NDL learning measures are necessarily based on the Buckeye corpus, a fair comparison of these measures with frequency, neighborhood density, and gang size requires that the classic measures are also based on the Buckeye corpus.

Since frequency and density counts based on much larger corpora are available, we added subtitle frequency as available in the British Lexicon Project (Keuleers et al., 2012b) and neighborhood counts based on IPHOD (Vaden et al., 2009) as predictors, replacing the corresponding measures based on the Buckeye corpus. The interaction of frequency with regularity was again supported, with longer durations for irregulars and shorter durations for regulars for higher-frequency words. A qualitatively similar interaction of neighborhood density by regularity was observed as well, with the same undulating patterns as shown in Figure 4, albeit with some modulations of the relative magnitude of the undulations. To capture the main trends of these effects, we also fitted a model in which the effect of neighborhood density was restricted to a linear effect. The slopes for frequency and neighborhood density of this model are listed in the right column of Table 5. For both regulars and irregulars, a greater density predicts (under the restriction of linearity) a longer duration of the stem vowel.

These results are surprising in light of the findings of, e.g., Bell et al. (2003) and Gahl and Strand (2016) for word durations: They observed that word durations are shorter for more frequent words, and that likewise

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-2.7121	0.0336	-80.7306	< 0.0001
Speech Rate	-0.1014	0.0078	-12.9278	< 0.0001
Vowel Quality=Tense	0.3458	0.0123	28.0510	< 0.0001
Age=Young	-0.0451	0.0365	-1.2357	0.2166
Sex=Male	-0.0119	0.0379	-0.3146	0.7530
Tense=Present	0.0584	0.0096	6.1095	< 0.0001
Regularity=Regular	-0.0002	0.0130	-0.0174	0.9861
Position=Final	0.4913	0.0132	37.3118	< 0.0001
Forward Frequency	0.0718	0.0049	14.7268	< 0.0001
Forward Bigram Frequency	-0.0930	0.0057	-16.1887	< 0.0001
Speech Rate × V. Quality=Tense	-0.0440	0.0083	-5.3178	< 0.0001
Speech Rate × Age=Young	0.0389	0.0086	4.5310	< 0.0001
Vowel Quality=Tense × Sex=Male	-0.0511	0.0164	-3.1080	0.0019
Buckeye Frequency	0.0941	0.0090	10.4106	< 0.0001
Regularity=Regular × Buckeye Frequency	-0.1432	0.0108	-13.2636	< 0.0001
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(Centered Gang Size)	3.8311	3.9819	142.6382	< 0.0001
s(Buckeye N-count) × Reg=Irregular	3.9267	3.9953	113.4978	< 0.0001
s(Buckeye N-count) × Reg=Regular	3.9530	3.9982	129.8916	< 0.0001
re(Speaker)	34.9960	37.0000	605.0588	< 0.0001

Table 4: Coefficient table for a generalized additive mixed model fitted to the acoustic duration of the stem vowel, with the classic predictors, frequency neighborhood density, and (centered) gang size. Factors were dummy coded with treatment contrasts. `s()`: thin plate regression spline; `re()`: random effect. Adjusted R-squared: 0.390; -REML = 6113.3; AIC: 12012.33. The parameter estimates of the scaled t-distribution are $\hat{\sigma} = 17.691$ and $\hat{\nu} = 0.392$.

higher numbers of neighbors give rise to shorter word durations. However, for the duration of the stem vowel, it is only for word frequency and regular verbs that the expected negative correlation was present. To make sure that our results are not due to side-effects of collinearity, we calculated the Spearman correlations of the vowel duration and the frequency and neighborhood density measures. As shown in the third column of Table 5, the signs of these correlations (all $p < 0.0001$) mirrored those of the regression model. We can therefore rule out that suppression or enhancement (Friedman and Wall, 2005) are an issue. We also checked the correlations of frequency and neighborhood density with word duration. As expected given the literature, all correlations were negative, albeit not significant for the combination of irregular verbs and the neighborhood density.

Apparently, frequency and neighborhood density need not have the same general effect at the word level

		Spearman correlation		slope regression model
	regularity	word duration	vowel duration	vowel duration
subtitle frequency	regular	-0.375	-0.157	-0.045
subtitle frequency	irregular	-0.332	0.196	0.069
N-count UNS	regular	-0.097	0.100	0.090
N-count UNS	irregular	-0.004	0.205	0.072

Table 5: Spearman correlations and slopes in a regression model for subtitle frequency, as well as UNS neighborhood density, with stem vowel and word duration, subcategorized by regularity. The correlation of N-count UNS and word duration was not significant for the irregular verbs ($p = 0.72$); for all other correlations and slopes, $p < 0.0001$.

and the segment level, e.g., for the stem vowel of irregular verbs, larger frequency and neighborhood size give rise to longer durations, whereas for word durations, they have a shortening effect. It follows that for the segments of the word other than the stem vowel, the effects of these predictors must go in the opposite direction. Indeed, for the irregulars, the total duration of all segments except the stem vowel shows strong negative correlations with frequency ($r_s = -0.71, p < 0.0001$) and neighborhood size ($r_s = -0.24, p < 0.0001$). Note that both correlations are substantially larger than those listed for word duration in Table 5.

Frequency of occurrence and neighborhood density are lexical measures that gauge properties of the word as a whole. For a holistic measure, one would expect its effect to be uniform across the segments of the word. Since effects go in opposite directions for the irregulars, and in part for the regulars, it is unlikely that frequency of occurrence and neighborhood density are central driving forces shaping the acoustic duration of the stem vowel, nor those of other segments. What is more likely to be at issue is the functional load of these segments. Let us therefore consider whether measures based on discrimination learning provide enhanced insight into the forces shaping the duration of the stem vowel.

5.2.2. GAMM with NDL predictors

Table 6 presents the model summary of the GAMM fitted to the stem vowel duration using the NDL predictors introduced above (with non-predictive variables removed from the model specification). The adjusted R-squared of this model was 0.414, the AIC was 11553.55, and the (negative) REML score was 5878.5. The parameter estimates of the scaled t-distribution were $\hat{\sigma} = 14.131$ and $\hat{\nu} = 0.378$. (The model reported in Table 3 was identical to this model, except that estimation was based on ML instead of REML.) When `Vowel-Tense Activation` and `Activation Diversity` are calculated from an NDL weight matrix estimated from the "canonical" phonological transcriptions instead of from the transcription indicating what people actually said, the fit of the model becomes significantly worse (adjusted R-squared: 0.407, AIC: 11696.8, negative REML: 5949.4). In what follows, we therefore do not further consider the measures based on the 'canonical dictionary' forms.

`Semantic Activation Diversity` is the learning measure that is most strongly correlated with frequency of occurrence (Buckeye frequency: $r_s = 0.94$, subtitle frequency: $r_s = 0.91$). Unsurprisingly, given the

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-2.7073	0.0359	-75.3656	< 0.0001
Speech Rate	-0.1044	0.0076	-13.6678	< 0.0001
Vowel Quality=Tense	0.3260	0.0126	25.9177	< 0.0001
Age=Young	-0.0463	0.0366	-1.2645	0.2060
Sex=Male	-0.0246	0.0380	-0.6462	0.5181
Tense=Present	0.0271	0.0098	2.7645	0.0057
Regularity=Regular	-0.1208	0.0259	-4.6561	< 0.0001
Position=Final	0.4895	0.0129	38.0557	< 0.0001
Forward Frequency	0.0646	0.0047	13.6403	< 0.0001
Forward Bigram Frequency	-0.0870	0.0056	-15.6521	< 0.0001
Speech Rate × V. Quality=Tense	-0.0402	0.0081	-4.9885	< 0.0001
Speech Rate × Age=Young	0.0369	0.0084	4.4058	< 0.0001
Vowel Quality=Tense × Sex=Male	-0.0446	0.0161	-2.7784	0.0055
Semantic Typicality	-0.0435	0.0045	-9.5980	< 0.0001
Sem. Act. Diversity	0.0417	0.0073	5.6870	< 0.0001
Regularity=Regular × Sem. Act. Diversity	-0.1359	0.0152	-8.9311	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(Vowel-Tense Act) × Reg=Irregular	2.1914	2.6162	6.1131	0.1206
s(Vowel-Tense Act) × Reg=Regular	3.9148	3.9949	321.1747	< 0.0001
s(Activation Diversity)	3.5728	3.8971	389.4879	< 0.0001
re(Speaker)	35.1061	37.0000	636.9564	< 0.0001

Table 6: Coefficient table for a generalized additive mixed model fitted to the acoustic duration of the stem vowel, with NDL predictors. Factors were dummy coded with treatment contrasts. `s()`: thin plate regression spline; `re()`: random effect. Adjusted R-squared: 0.414, -REML = 5878.5, AIC = 11553.55. The parameter estimates of the scaled t-distribution are $\hat{\sigma} = 14.131$ and $\hat{\nu} = 0.378$.

interaction of frequency and regularity observed above, correlations of `Semantic Activation Diversity` with the duration of the vowel were likewise negative for regular verbs ($r_s = -0.22$) and positive for irregular verbs ($r_s = 0.12$). `Semantic Activation Diversity` also predicts the duration of the word itself. For both regulars ($r_s = -0.42$) and irregulars ($r_s = -0.35$), the correlation is negative, mirroring the effects of frequency of occurrence.

This pattern of results raises the question of why semantic activation diversity shows a negative correlation only for the stem vowel of regular verbs. To address this question, we begin with noting that semantic activation diversity is a measure of the extent to which a word perturbs the state of the lexical system. It is a measure of semantic uncertainty. For all but irregular stem vowels, we find that under conditions of increased semantic uncertainty, both word and stem vowel are articulated with shorter durations. From a communicative perspective, this makes perfect sense. If a word's meaning is highly confusable with the meanings of other words, investing in a long acoustic duration will not help resolve this confusion, to the contrary, it will prolong the confusion for the listener. Since the opposite sign of `Semantic Activation Diversity` for irregular verbs does not make sense, we must be missing out on a critical property of irregular verbs.

One missing factor turns out to be the semantic density of irregular verbs, which is greater than that of regular verbs (Baayen and Moscoso del Prado Martín, 2005). This study reports that irregular verbs have more synonym sets (synsets) than regular verbs, and that irregulars co-occur more often in these synsets. In other words, irregular words have richer semantics. Furthermore, the larger the gang size of an irregular verb is, the smaller the number of synsets it occurs in. Irregulars with larger gang sizes are somewhat less irregular, and approximate more the less rich semantics of regulars. In addition, it was observed that irregulars occur in more verb alternation classes in which there are more irregulars, and that irregulars cluster more in semantic space compared to regulars. Also, association norms indicate that irregulars are more strongly associated semantically than is the case for regulars. Across English, German and Dutch, irregulars are more often used to denote movement or position of the body (*sit, stand, climb, swim*, . . .).

That irregulars are semantically more similar to each other than regulars can also be established on the basis of the correlation matrix $\boldsymbol{C}$. The correlations c_{ij} between regular verbs i and j are smaller than the corresponding correlations between irregular verbs (mean regulars 0.003, mean irregulars 0.006, $t_{(1222.6)} = -3.99, p < 0.0001$). We therefore calculated, for each verb, the number of verbs with a correlation $c_{ij} \geq 0.005$, where 0.005 is a threshold value empirically found to be optimal. The mean number of such verbs was 124.1 for irregulars and 110.6 for regulars ($t_{(66.384)} = 2.74, p = 0.0079$).

When this count of semantic neighbors (after scaling) is added as a predictor to the model, it receives strong support with a negative slope $\hat{\beta} = -0.046, p < 0.0001$). Interestingly, the slopes of semantic activation diversity are reduced in this new model. For irregulars, the originally estimated slope (0.039) was more than halved (0.018), while for regulars, its absolute magnitude was also reduced, albeit less strongly (original slope: -0.141, new slope: -0.094). (The new model was estimated using maximum likelihood, and was based on a slightly smaller dataset due to the correlation matrix $\boldsymbol{C}$ providing coverage of a slightly smaller subset of verbs. Comparisons with the model without semantic neighbors are based on this smaller dataset.)

Since the count of semantic neighbors is strongly correlated with semantic activation diversity, with negative sign ($r_s = -0.66$, for the subset of irregulars, $r_s = -0.76$), and given that semantic activation diversity is much higher for irregulars (mean 2.03) than regulars (mean 0.76), it seems likely that the positive correlation of semantic activation diversity with stem vowel duration is, to a considerable extent, for the irregulars, an effect of semantic density. Below, we shall see that with the inclusion of two further predictors that are required on independent grounds, the effect of semantic activation diversity for irregulars is no longer significant.

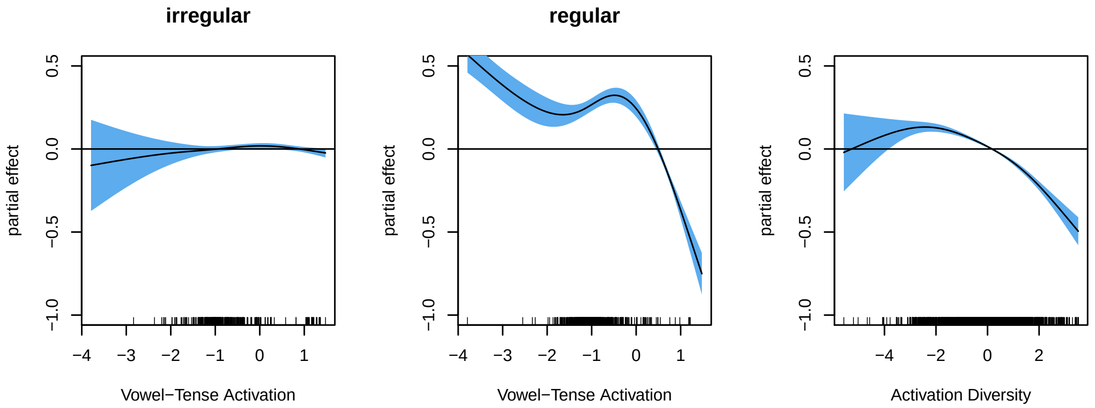

Figure 5: Partial effects of `Vowel-Tense Activation` for irregular (left panel) and regular (center panel) verbs and of `Activation Diversity` (right panel), according to the generalized additive mixed model fitted to the acoustic durations of stem vowels.

The effect of semantic typicality on the duration of the stem vowel was linear with negative slope. The more a verb is similar to other verbs, the more difficult it is to discriminate this verb from other verbs, and economy of discriminative effort again demands the acoustic duration of the stem vowel to be shorter.

The `Activation Diversity` measure, just as the semantic activation diversity measure, quantifies uncertainty. As expected for a measure of uncertainty, the duration of the stem vowel decreases for increasing uncertainty. The right panel of Figure 5 presents the partial effect of this predictor, the effect of which was attenuated for the smaller values of the predictor.

The `Vowel-Tense Activation` quantifies the support from the diphones into and out of the stem vowel for tense. There was no noticeable effect for the irregulars, whereas a complex wiggly pattern emerged for the regulars that is not straightforwardly interpretable (see Figure 5). Upon closer inspection, it turns out that the effect of `Vowel-Tense Activation` is confounded with an effect of regularity. Note that, first, knowledge of whether a word should be inflected regularly or irregularly is important for producing past-tense forms, given that the option exists for some verbs to use either a regular or irregular form (e.g., *dived, dove*). Second, the significant interactions with regularity as factorial predictor in our statistical models also support the relevance of regularity. We therefore added two further predictors, one for the activation of a ‘Regularity’ outcome by the vowel diphones (`Activation Supporting Regularity`), and one for the corresponding activation of the ‘Irregularity’ outcome (`Activation Supporting Irregularity`). We added the interaction of these two predictors to the NDL model using a tensor product smooth (constrained to 5 basis functions in each dimension), together with the semantic neighborhood measure introduced above. The resulting model is summarized in Table 7.

The smooths in this model are presented in Figure 6. The upper panels visualize the partial effects for `Vowel-Tense Activation` for irregulars (left) and regulars (center), as well as the partial effect of `Activation Diversity` (right). The latter effect is similar to the one in the preceding model (Figure 5), but the smooths

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-2.6615	0.0366	-72.6603	< 0.0001
Speech Rate	-0.1014	0.0075	-13.5465	< 0.0001
Vowel Quality=Tense	0.3164	0.0129	24.5643	< 0.0001
Age=Young	-0.0453	0.0348	-1.3008	0.1933
Sex=Male	-0.0212	0.0362	-0.5860	0.5579
Tense=Present	0.0240	0.0110	2.1732	0.0298
Regularity=Regular	-0.0072	0.0325	-0.2222	0.8242
Position=Final	0.4846	0.0126	38.3728	< 0.0001
Forward Frequency	0.0662	0.0046	14.2430	< 0.0001
Forward Bigram Frequency	-0.0927	0.0055	-16.9148	< 0.0001
Speech Rate × Vowel Quality=Tense	-0.0432	0.0079	-5.4598	< 0.0001
Speech Rate × Age=Young	0.0345	0.0082	4.2177	< 0.0001
Vowel Quality=Tense × Sex=Male	-0.0473	0.0157	-3.0119	0.0026
Semantic Activation Diversity	0.0131	0.0100	1.3042	0.1922
Semantic Activation Diversity × Regularity=Regular	-0.1406	0.0166	-8.4800	< 0.0001
Semantic Typicality	-0.0192	0.0053	-3.6018	0.0003
Semantic Neighborhood Density	-0.0419	0.0071	-5.8945	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(Act. Supporting Irregularity, Act. Supporting Regularity)	15.7191	17.8743	349.2792	< 0.0001
s(Vowel-Tense Activation) × Regularity=Irregular	3.1507	3.5455	42.9154	< 0.0001
s(Vowel-Tense Activation) × Regularity=Regular	3.5122	3.8523	50.3325	< 0.0001
s(Activation Diversity)	3.5111	3.8661	407.9239	< 0.0001
s(Speaker)	34.9937	37.0000	651.5788	< 0.0001

Table 7: Generalized additive model fitted to the acoustic durations of the stem vowel of regular and irregular verbs in the Buckeye corpus, using the final set of NDL predictors. Control variables are backgrounded in grey. Factors were dummy coded with treatment contrasts. `s()`: thin plate regression spline; `re()`: random effect. Adjusted R-squared: 0.432, -REML = 5624.2, AIC = 11146.16. The parameter estimates of the scaled t-distribution are $\hat{\sigma} = 11.49$ and $\hat{\nu} = 0.365$.

for `Vowel-Tense Activation` have changed considerably. For irregular verbs, the duration of the stem vowel first decreases (ignoring values below -2 where there is little or no data) and then increases for higher values of `Vowel-Tense Activation`. For regular verbs, the effect appears restricted to the tails, with an upward swing for the lowest values and a downward swing for the highest values, resulting in an overall pattern of decreasing duration for increasing `Vowel-Tense Activation`. For the majority of regular data points, in the middle range from -2 to 0, there is no clear effect.

The result for regulars fits well with the vowels of regular verbs being neutral with respect to tense. For words with roughly balanced frequency of occurrence for past and present tense, competition during learning will inhibit the development of strong weights for either tense. When frequencies are unbalanced, a higher log frequency (or a higher ratio of log present tense frequency and past tense frequency) predicts higher values of Vowel-Tense Activation ($r_s = 0.41$ and 0.33 respectively, both $p < 0.0001$). Therefore, this effect is the counterpart of the effect of Buckeye Frequency for regular verbs, which also had a negative slope (see Table 4).

For irregular verbs, frequency and `Vowel-Tense Activation` are also positively correlated ($r_s = 0.48$ for Buckeye frequency, and $r_s = 0.20$ for the frequency ratio of past and present tense forms, both $p < 0.0001$). One would therefore predict a downward trend for the irregulars, but such a downward trend is present only for values of Vowel-Tense Activation less than 0 (i.e., because `Vowel-Tense Activation` is centered, for values below the mean). For values above the mean, the downward trend reverses into an upward swing, which may perhaps explain why in the model with classic predictors, the coefficient for Buckeye Frequency was positive for irregular verbs. This upward swing fits with the paradigmatic signal enhancement hypothesis, in the sense that for irregulars with stronger than average support from the vowel diphones for the tense, longer durations are observed. Note that this U-shaped pattern was also observed for the centered gang size, albeit with wider confidence intervals for larger values. (A further complicating factor is that the support from the vowel diphones to tense varies systematically with the support from the vowel diphones to the verb's lexome. For irregulars, the two mainly increase in tandem, but for regulars, the main pattern is inverse U-shaped, with a flat peak for lexome support for values of tense support between -1 and 1. We have not included this variable in the analysis as it is too strongly related to `Vowel-Tense Activation`. We have also refrained from including interactions of `Vowel-Tense Activation` by `Regularity` by `Tense`, as data become too sparse and the risk of overfitting is too high.)

The lower panels of Figure 6 graph the partial effect of the interaction of Activation Supporting Irregularity by Activation Supporting Regularity. The left panel presents the regression surface with color coding, warmer colors denoting longer durations. The right panel presents the contour lines with 1 SE confidence regions, green dotted lines being at 1 SE distance up, and red dashed lines at 1 SE distance down. The horizontal and vertical lines in the left panel differentiate between positive and negative activation support. First note that most vowel diphones cluster close to the origin, where there is no clear support for either regularity or irregularity. There are two regions where durations are longer, namely, the rise for intermediate values of Activation Supporting Regularity and slightly negative values of Activation Supporting Irregularity, and a corresponding rise in the lower right of the plot. These are the points on the hills that peak around $(-0.05, 0.15)$ and (0.2, -0.05). These areas are perhaps best understood as areas of good congruent support, with intermediate values for Regularity (or Irregularity) being supplemented by small negative values for Irregularity (or Regularity). In other words, when support is solid but not extreme on both dimensions, durations are longer.

As can be seen in Table 7, the puzzling effect of Semantic Activation Diversity for irregular verbs is no

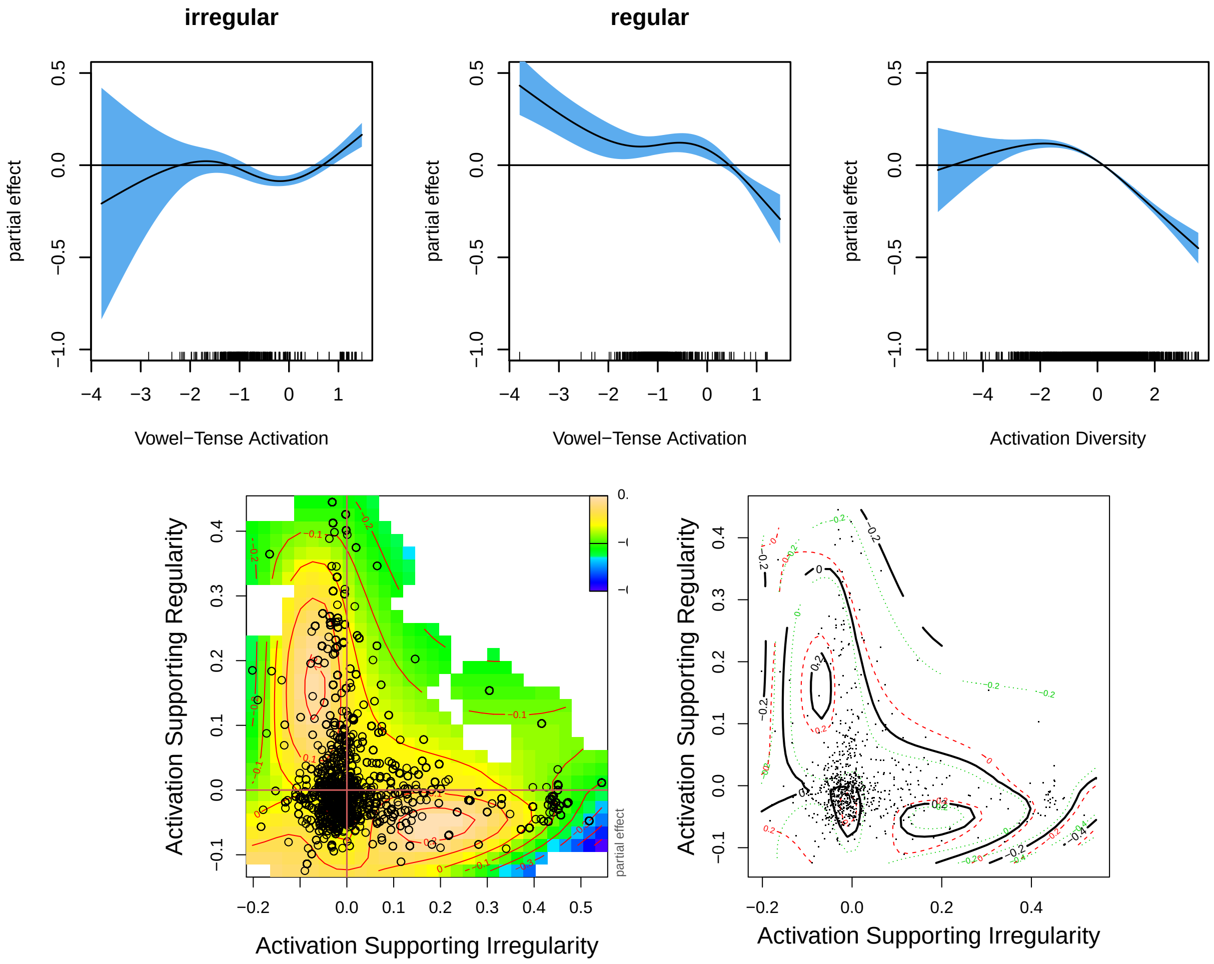

Figure 6: Partial effects for `Vowel-Tense Activation` and `Activation Diversity` (upper panels) and for the interaction of Activation Supporting Irregularity by Activation Supporting Regularity (lower panels) in the GAMM fitted to the acoustic duration of the stem vowel, with the full set of NDL predictors. The lower left panel shows the tensor product smooth with warmer colors denoting longer durations, the lower right panel present contour lines with 1SE confidence regions (dotted green lines 1 SE up, dashed red lines, 1 SE down).

longer supported once both semantic density and support for (ir)regularity are taken into account.

One variable that obtained good support in the random forest analysis, but that is not found in any of the above GAMMs is the `Prior Availability`. This measure is strongly correlated with `Semantic Activation Diversity` ($r = 0.84$), but it turns out that the latter measure provides superior predictivity, not only in the random forest analysis, but also when using the generalized additive model. We therefore did not further pursue `Prior Availability` as a predictor.

6. Discussion

The initial goal of the present study was to test the prediction of Kuperman et al. (2006) that the acoustic duration of the stem vowel of irregular verbs should increase with gang size. Statistical models using the classic predictors from word form competition (WFC) approaches (frequency, number of neighbors, gang size) did not provide an unequivocal answer, as the effect of gang size was U-shaped, partly supporting the paradigmatic signal enhancement hypothesis, and partly supporting the smooth signal redundancy hypothesis. The generalized additive mixed models based on WFC approaches yielded some puzzling results: frequency correlated positively with duration for irregulars, but negatively for regulars. To complicate matters further, neighborhood density correlated negatively with duration for irregulars, but positively for regulars. Why these word-form specific measures have such diverging and opposite effects for regulars and irregulars is unclear.

We also observed that word frequency and neighborhood density correlated negatively with the acoustic duration of the verb form for both regulars and irregulars. However, a negative correlation with stem vowel duration was present only for frequency, and only for the regular verbs. Interestingly, the effects of frequency and neighborhood density need not have uniform effects across all segments. Furthermore, since word frequency in WFC approaches is typically understood as reflecting words' resting activation levels or words' Bayesian priors, it remains unclear how, under these conceptualizations, the different effects of frequency at lexical and sublexical levels might be explained. However, more generally, the lengthening of the vowel and the shortening of consonants can perhaps be understood as instantiating a form of signal smoothing, with the role of the consonants being reduced more to the periphery of the syllable.

A second goal of this study was to examine the forces shaping the acoustic duration of a verbs' stem vowel from the perspective of discriminative learning. A preliminary issue that arises when considering the discriminatory force of diphone features is whether these diphones should be taken from the canonical dictionary forms, or from what speakers actually said. The answer here was unequivocal: Model fits substantially improve when the model weights are estimated from speakers' actual productions. This result is of theoretical importance for two reasons. On the one hand, it appears that neighborhood counts based on dictionary pronunciations may lack precision and not accuractely reflect the presumed competition. On the other hand, discrimination learning sidesteps the problem of whether words' pronunciation variants, including reduced forms, should or should not be counted as neighbors. The theoretical burden of explaining acoustic durations is shifted away from word forms and their neighbors to the discriminatory function of sublexical units, in this study, diphones (see Arnold et al. (2017) for features derived from the speech signal that can be explored using speech corpora that make the audio files available).

Another issue that arises when pitting discrimination learning against WFC approaches is whether discrimination learning actually provides more precise predictions. For the present data set, discrimation learning based predictors turned out to be more precise. Even though models with measures based on naive discriminative learning (NDL) required more parameters, significantly improved statistical fits were obtained, with increases in AIC ranging from around 40 to around 450. Formal model comparison supports NDL as providing enhanced precision. It is noteworthy that NDL's performance remained superior even when the WFC approach was granted access to frequency and neighborhood measures culled from much larger corpora and resources than just the Buckeye corpus — whereas NDL was always constrained to measures based on the Buckeye corpus.

Returning to the paradigmatic signal enhancement hypothesis, the measure closest to the gang size measure of WFC approaches is the `Vowel-Tense Activation` measure. For both `Gang Size` and `Vowel-Tense Activation`, we observed a U-shaped effect. Whereas the lengthening for higher values is in accordance with the prediction from the paradigmatic signal enhancement hypothesis, the shortening observed for smaller values is in contradiction with this hypothesis. Lengthening of acoustic duration was also observed when there was solid activation support for regularity or for irregularity. In other words, in those cases where acoustic duration increases with a predictor, this increase goes hand in hand with greater support for tense and regularity. Thus, NDL makes it possible to quantify to some extent the paradigmatic effects of tense and regularity. Although paradigms as such do not exist in NDL, when evidence gangs up to create pockets of certainty, strengthening of duration can occur.

For all other considered NDL measures, the vowel duration decreased with larger values. Larger values of these measures all express increased uncertainty. For activation diversity, greater values indicate greater uncertainty about the lexomes targeted by the speech signal (see also Arnold et al., 2017). For semantic neighborhood density, a greater density implies a greater number of verbs with similar semantic vectors, and hence an enhanced discrimination problem. Likewise, verbs with a high semantic typicality are verbs that are similar in meaning to many other words, and hence again difficult to tell apart. Finally, greater values of semantic activation diversity imply that a word has a rich collocational structure, and that a large number of words are all possible, which amounts to greater sentential uncertainty. Our hypothesis is that under increased uncertainty, less energy is invested in maintaining duration. Increasing duration would be disadvantageous for the speaker, as the speaker would have to maintain for a longer time a signal that is difficult to discriminate, thus increasing uncertainty in the production process. A longer duration would also be disadvantageous for the listener, as the listener would be confronted for a longer period of time with an ineffective signal that fails to properly reduce the listener's uncertainty about the message encoded in the speech signal. (We note here that in NDL, there is no winner-take-all mechanism as in TRACE and Shortlist-B (McClelland and Elman, 1986; Norris and McQueen, 2008). It is well-known for speech comprehension that words from spontaneous speech are often not recognized (Ernestus et al., 2002; Arnold et al., 2017). Thus, listeners can remain uncertain about what they have heard, and may fail to recognize the intended word when the signal is not sufficiently discriminative.)

The present study offers the following new insights. First, semantic measures co-determine acoustic duration: We observed effects of semantic activation diversity (regulars only), semantic typicality, and semantic neighborhood density.

Second, NDL makes it possible to examine the functional load of sublexical units in a much more fine-grained way than is possible on the basis of, for instance, phonological neighborhood density (Luce and Pisoni, 1998) or minimal pairs (Wedel et al., 2013).

Third, the present analysis indicates that it is not frequency of occurrence that is driving the acoustic duration of the stem vowel, but rather semantic activation diversity. Frequency, prior availability (the L1-norm of a lexome's column vector in $\boldsymbol{W}$), and semantic activation diversity are all strongly correlated, and we had originally expected prior availability (the measure most strongly correlated with frequency) to be the superior measure. Yet, even though it received good support also in the random forest analysis, it is outranked by semantic activation diversity (and diphone activation diversity as well). Interestingly, whereas the WFC approach produces an inexplicable positive correlation for irregular verbs of frequency and vowel duration, semantic activation diversity has no effect for irregulars.

Fourth, NDL makes it possible to approach the issue of lexical similarity in a novel, and we think more insightful way. Standard neighborhood counts, including recent extensions (Yarkoni et al., 2008), run into problems for words with variant forms, ranging from pronunciation variation (the possible realizations of /r/ in Dutch (Van de Velde and van Hout, 1999)) and reduced word forms (English *yesterday* realized as [jESe] (Tucker, 2007), or Dutch *natuurlijk* reducing to [tyk], [tək], [ntyk], [tylək], [tyrlək], among others (Ernestus et al., 2002)), to inflectional variants (*walks* as competitor of *walk*). The often arbitrary decisions that have to be made here to get neighborhood measures to work are not necessary in NDL. Putting this technical problem aside, it is worth noting that computationally, NDL activations and a computational measure for phonetic string similarity, a weighted edit distance, as developed by Wieling et al. (2012) are functionally equivalent (Wieling et al., 2014). Crucially, the weights used in this edit distance quantify the functional load of an edit across the vocabulary, which is the functional equivalent of what NDL networks accomplish using the Rescorla-Wagner learning rule. Of course, the next step forward is to move away from phonetic transcriptions, and to work from the speech signal itself. Possibly, the features developed in Arnold et al. (2017) will prove to be useful here.

Given the influence of semantic confusability on acoustic duration, we think the desideratum for speech is not so much a smooth signal with a constant information flow, but instead a signal that balances discrimination against articulatory effort (in line with Lindblom, 1990). We have shown that predictors grounded in error-driven learning can be used to investigate the forces that influence this balance, and to move beyond what can be accomplished with WFC approaches. An important challenge remains, however, namely to clarify algorithmically how the cognitive system achieves this balance. We have used GAMMs to chart, to some extent, the balance achieved by the opposing forces determining vowel duration, but as yet a computational theory that explains why the smooths and tensor products of the GAMMs take the form observed, is lacking.

The reader will have noted that the direction of learning, from form (diphones) to meaning (lexomes), is surprising given standard production models (Dell, 1986; Levelt et al., 1999), according to which the general flow of processing is from conceptualization to articulation. Nevertheless, the present set-up, with diphone cues and lexomic outcomes is well motivated, for several reasons. First, it is an empirical finding across several studies that durations are best predicted with measures based on networks predicting lexomes from form cues (Hendrix, 2015; Lensink et al., 2017). Second, this empirical result is unsurprising given that it is intrinsic to NDL approaches that learning is most effective when a large set of cues has to discriminate one particular outcome from other outcomes (Ramscar et al., 2010). For this reason, the model of reading aloud proposed in Hendrix (2015) trains from demi-syllables to lexomes. Third, the production system must have some form of feedback control, allowing it to evaluate the sensory consequences of speaking. Without such feedback, which comprises sensory feedback from the articulators as well as proprioceptive feedback from hearing one's own speech and bone conductance, learning cannot take place (see also Hickok, 2014, for detailed discussion). Importantly, for error-driven learning to be possible, distinct articulatory and acoustic targets must be set up before articulation, against which the feedback from the articulatory and auditory systems can be compared. In the present study, the set of diphones is a crude approximation of such acoustic and articulatory targets.

A note on the random forest analysis is also in order. The random forests not only provide independent support for the key predictors in our regression models, but they also outperform our best regression model ($R^2 = 0.63$ for the random forests, but only 0.43 for our most successful regression model). Conditional inference trees, and random forests of such trees, are known to be very good at picking up complex interactions that are difficult or impossible to capture with regression models. This raises the question of whether these high-order interactions are actually a true property of the actual processing system. It is possible that pervasive subtle interactions as detected by conditional inference trees and random forests are an artifact of the use of crude predictors. Under this view, the hope would be that as more refined predictors become available, simpler interactions will be required, and results obtained with machine learning and regression modeling will converge. It is equally likely that pervasive subtle interactions are a defining characteristic of the processing system, and that this is why random forests are able to outperform regression modeling. From this perspective, the GAMMs provide us with high-level summaries that, albeit more interpretable for the analyst, are a substantial step removed from the actual complexities of how lexical processing is shaped by the constraining factors represented by our predictors. Given the substantial margin by which the random forests outperform regression, the latter perspective remains one to be taken seriously.

This study has depended heavily on exploratory data analyses. Although we believe the statistical models that we have laid out are as solid as exploratory analyses can be, only replication studies can help clarify whether NDL truly improves on WFC approaches.

Acknowledgements

This work was funded by a grant to the first author from the Social Science and Humanities Research Council of Canada (#410-2011-0386), and by a grant from the Alexander von Humboldt foundation to the third author (#1141527). We are indebted to Susanne Gahl, Jessie Nixon and Yu-Ying Chuang for their comments on earlier versions of this manuscript.

Arnold, D., Tomaschek, F., Lopez, F., Sering, T., and Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, 12(4):e0174623.

Aylett, M. and Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, 47(1):31–56.

Aylett, M. and Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5):3048–3058.

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

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

Baayen, R. H. and Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, 81:666–698.

Baayen, R. H., Sering, T., Shaoul, C., and Milin, P. (2017a). Language comprehension as a multiple label classification problem. *Statistica Neerlandica*.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016b). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*, 31(1):106–128.

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. N. (2017b). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In Speelman, D., Heylen, K., and Geeraerts, D., editors, *Mixed Effects Regression Models in Linguistics*, page to appear. Springer, Berlin.

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

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

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

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B*, 26:211–252.

Bybee, J. L. and Moder, C. L. (1983). Morphological Classes as Natural Categories. *Language*, 59(2):251–270.

Bybee, J. L. and Slobin, D. I. (1982). Rules and Schemas in the Development and Use of the English past Tense. *Language*, 58(2):265–289.

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

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

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

Ellis, N. C. (2006). Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, 27(1):1–24.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. LOT, Utrecht.

Ernestus, M., Baayen, R. H., and Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81:162–173.

Flemming, E. (2010). Modeling listeners: Comments on Pluymaekers et al. and Scarborough. In Fougeron, C., Kuehnert, B., Imperio, M., and Vallee, N., editors, *Laboratory Phonology 10*, pages 587–606. De Gruyter, Berlin, Boston. DOI: 10.1515/9783110224917.5.587.

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

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

Gahl, S. (2008). *Time* and *Thyme* Are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. *Language*, 84(3):474–496.

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

Gahl, S. and Baayen, R. H. (2017). Twenty-eight years of vowels. *Manuscript submitted for publication.*

Gahl, S. and Strand, J. F. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language*, 89:162–178.

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

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

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. Data mining, inference, and prediction.* Springer, Berlin.

Hendrix, P. (2015). *Experimental explorations of a discrimination learning approach to language processing.* PhD thesis, University of Tübingen.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, 29(1):2–20.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.

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

Kemps, R., Wurm, L. H., Ernestus, M., Schreuder, R., and Baayen, R. H. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, 20:43–73.

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

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012b). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44:287–304.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208–1221.

Krott, A., Baayen, R. H., and Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics*, 39(1):51–93.

Krott, A., Schreuder, R., and Baayen, R. H. (2002). Linking elements in Dutch noun-noun compounds: constituent families as predictors for response latencies. *Brain and Language*, 81:708–722.

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

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

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

Lensink, S. E., Verhagen, A., Schiller, N., and Baayen, R. H. (2017). Keeping them apart: on using a discriminative approach to study the nature and processing of multi-word units. *manuscript, University of Leiden.*

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

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modeling*, pages 403–440. Kluwer, Dordrecht.

Luce, P. and Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, 19(1):1–36.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, and Computers*, 28(2):203–208.

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

McClelland, J. L. and Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18:1–86.

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

Milin, P., Divjak, D., and Baayen, R. H. (2017a). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

Mulder, K., Dijkstra, T., Schreuder, R., and Baayen, R. H. (2014). Effects of primary and secondary morphological family size in monolingual and bilingual word processing. *Journal of Memory and Language*, 72:59–84.

Munson, B. (2007). Lexical access, lexical representation, and vowel production. In Cole, J. and Hualde, J. I., editors, *Laboratory Phonology 9*, volume 9 of *Phonology and Phonetics*, pages 201–228. Mouton de Gruyter.

Munson, B. and Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *Journal of Speech, Language, and Hearing Research*, 47(5):1048–1058.

Norris, D. and McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395.

Nusbaum, H. C. (1985). A stochastic account of the relationship between lexical density and word frequency. Technical report, Indiana University. Research on Speech Perception, Progress Report #11.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., and Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release)[www. buckeyecorpus. osu. edu] Columbus, OH: Department of Psychology. *Ohio State University (Distributor).*

Plag, I., Homann, J., and Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, 53(1):181–216.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

Port, R. F. and Leary, A. P. (2005). Against formal phonology. *Language*, 81:927–964.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M. and Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53:58–74.

Ramscar, M., Sun, C. C., Hendrix, P., and Baayen, R. H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological Science*. https://doi.org/10.1177/0956797617706393.

Ramscar, M. and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6):927–960.

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

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

Rosenblatt, F. (1962). *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Spartan Books, Washington.

Scarborough, R. (2010). Lexical and contextual predictability: Confluent effects on the production of vowels. In Fougeron, C., Kuehnert, B., Imperio, M., and Nathalie, V., editors, *Laboratory Phonology 10*, pages 557–586. De Gruyter, Berlin, Boston. DOI: 10.1515/9783110224917.5.557.

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

Scarborough, R. A. (2004). Degree of Coarticulation and Lexical Confusability. In *Proceedings of the 29th Meeting of the Berkeley Linguistics Society, February 14-17.*

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, 90(13):413–422.

Strobl, C., Malley, J., and Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological methods*, 14(4):26.

Tagliamonte, S. and Baayen, R. H. (2012). Models, forests and trees of york english: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24:135–178.

Tomaschek, F., Tucker, B., and Baayen, R. H. (2017). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistic Vanguard*, page under revision.

Tomaschek, F., Tucker, B. V., Wieling, M., and Baayen, R. H. (2014). Vowel articulation affected by word frequency. In *Proceedings of 10th ISSP, Cologne*, pages 429–432.

Tomaschek, F., Wieling, M., Arnold, D., and Baayen, R. H. (2013). Frequency effects on the articulation of German i and u: evidence from articulography. In *Proceedings of Interspeech, Lyon*, pages 1302–1306.

Tremblay, A. and Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, 6(2):302–324.

Tucker, B. V. (2007). *Spoken word recognition of the reduced American English flap.* The University of Arizona.

Vaden, K. L., Halpin, H., and Hickok, G. (2009). Iphod: Irvine phonotactic online dictionary, version 1.4.[data file]. *URL¡http://www.iphod.com¿*.

Van de Velde, H. and van Hout, R. (1999). The pronunciation of (r) in standard dutch. *Linguistics in the Netherlands*, 16(1):177–188.

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2016). itsadug: Interpreting time series and autocorrelated data using GAMMs. R package version 2.2.

Vitevitch, M. S. (1997). The Neighborhood Characteristics of Malapropisms. *Language and Speech*, 40(3):211–228.

Vitevitch, M. S. (2002). The Influence of Phonological Similarity Neighborhoods on Speech Production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28(4):735–747.

Vitevitch, M. S. and Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4):491–504.

Wedel, A., Kaplan, A., and Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128(2):179–186.

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

Wieling, M., Margaretha, E., and Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., and Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PLOS-ONE*, 9(1):e75734.

Wightman, C. W., ShattuckHufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707–1717.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

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

Wright, R. (2004). Factors of lexical competition in vowel articulation. In Local, J., Ogden, R., and R, T., editors, *Papers in Laboratory Phonology 6*, pages 75–87. Cambridge University Press, Cambridge.

Yarkoni, T., Balota, D., and Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.

Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 15:1–95.