Morpho-phonetic effects in speech production: Modeling the acoustic duration of English derived words with linear discriminative learning

Simon David Stein[1*], Ingo Plag[1]

[1]English Language and Linguistics, Department of English and American Studies, Heinrich Heine University Düsseldorf, Germany

***Correspondence:**
Simon David Stein
simon.stein@uni-duesseldorf.de

Keywords: speech production, linear discriminative learning, acoustic duration, morphological theory, derivation, mental lexicon

Abstract

Recent evidence for the influence of morphological structure on the phonetic output goes unexplained by established models of speech production and by theories of the morphology-phonology interaction. Linear discriminative learning (LDL) is a recent computational approach in which such effects can be expected. We predict the acoustic duration of 4530 English derivative tokens with the morphological functions DIS, NESS, LESS, ATION and IZE in natural speech data by using predictors derived from a linear discriminative learning network. We find that the network is accurate in learning speech production and comprehension, and that the measures derived from it are successful in predicting duration. For example, words are lengthened when the semantic support of the word's predicted articulatory path is stronger. Importantly, differences between morphological categories emerge naturally from the network, even when no morphological information is provided. The results imply that morphological effects on duration can be explained without postulating theoretical units like the morpheme, and they provide further evidence that LDL is a promising alternative for modeling speech production.

Main text word count: 11,286/12,000 max. *Figures and tables: 8/15 max.*

1 Introduction

Recent findings in morpho-phonetic and psycholinguistic research have indicated that phonetic detail can vary by morphological structure. For example, the acoustic duration of English word-final [s] and [z] differs depending on morphological status and inflectional function (Plag et al., 2017; Seyfarth et al., 2017; Tomaschek et al., 2019; Plag et al., 2020). For derivation, too, studies have demonstrated effects of morphological structure on phonetic output. For example, morphological geminates in English differ in duration depending on morphological category and informativity (Ben Hedia and Plag, 2017; Ben Hedia, 2019), and phonetic reduction in various domains can depend on how easily speakers can decompose a complex word into its constituents (e.g. Hay, 2003, 2007; Plag and Ben Hedia, 2018; Zuraw et al., 2020).

These findings raise several problems at the theoretical level. The observation that phonetic detail varies systematically with morphological properties is unaccounted for by traditional and current models of the morphology-phonology interaction and of speech production (e.g., Chomsky and Halle, 1968; Kiparsky, 1982; Dell, 1986; Levelt et al., 1999; Roelofs and Ferreira, 2019; Turk and Shattuck-Hufnagel, 2020). This is because these models are either underspecified regarding the processing of complex words, or do not allow for post-lexical access of morphological information. For example, feed-forward models of the morphology-phonology interface (e.g., Kiparsky, 1982) assume that morphological brackets around constituents are "erased" in the process of passing on a word through morphological and phonological levels of processing. This means that no trace of morphological structure should be left at the level of phonetic realization. Similarly, established psycholinguistic models of speech production (e.g., Levelt et al., 1999) assume that morphological units select general phoneme templates which are then passed on to an articulator module to be realized phonetically. Again, no morphological information is encoded in these templates, meaning that no systematic differences between morphological properties are expected at the phonetic level.

Yet, morphological effects on the phonetic output have repeatedly been observed, which is incompatible with these assumptions. For example, the observation that complex words are more acoustically reduced when they are less decomposable into their constituents (Hay, 2003, 2007; Plag and Ben Hedia, 2018; Zuraw et al., 2020) seems to suggest that information about morphological boundaries must somehow still be present at phonetic level. From the perspective of the speech production models and theories of the morphology-phonology interaction outlined above, such effects are unexpected, and the mechanisms behind them are unclear. To better explain the morphology-phonetics interaction at the theoretical level and to understand the patterning of durations in complex words from a new perspective, we need alternative approaches.

One such approach is to model phonetic detail based on the principles of discriminative learning (see, e.g., Baayen et al., 2011; Ramscar and Yarlett, 2007; Ramscar et al., 2010). Such an approach sees form-meaning relations not as compositional, but as discriminatory instead. That is, form-meaning relations are created in a system of *difference*, which distinguishes between features based on their similarity and dissimilarity and connects them to each other in a learning process. In discriminative approaches, “signs” in the semiotic sense of relations of form and meaning (Saussure, 1916) are not fixed units. Discriminative models refrain from sub-lexical static representations such as morphemes or roots in the lexicon. Instead, speech comprehension and production are the result of a dynamic learning process where relations between form and meaning are constantly recalibrated based on the speaker’s experience. How strong associations between given forms and meanings are in the system depends on how often specific forms occur together with specific meanings, and on how often they fail to occur together with others. Each time a speaker makes a new experience, i.e., encounters a form together with a specific meaning, all associations of forms and meanings in the system are updated to reflect this new state of learning. An association strength increases when a ‘cue’ (such as a specific form) occurs together with an ‘outcome’ (such as a specific meaning), and an association strength decreases when a cue does not occur with the outcome.

Such an approach has clear advantages if we are to explain the evidence that morphology directly affects phonetic realization. A discriminative learning model lacks a feed-forward architecture which divides speech processing into separate levels. It is an end-to-end model that goes directly from form to meaning and from meaning to form. This means that the loss of morphological information between levels, e.g., through bracket erasure or phoneme template selection, is no longer an issue. Moreover, discriminative learning refrains from postulating morphemes or phonemes as psychologically relevant units in the first place. This opens the way for interpreting acoustic differences from a new perspective. In a discriminative approach, differences between morphological functions are expected to emerge naturally from sublexical and contextual cues. If we can model systematic acoustic variation between morphological functions with measures derived from a discriminative network, it is possible to explain potential effects by its theoretical principles of learning and experience.

While discriminative approaches have already been used to model other morphological correlates, such as reaction time (e.g., Baayen et al., 2011), the question arises whether a discriminative approach is able to successfully predict phonetic variation. Recently, Tomaschek et al. (2019) employed naïve discriminative learning (NDL) to model the duration of English word-final [s] and [z] of different morphological status. The measures derived from their network were predictive and indicated that a higher certainty in producing a morphological function leads to lengthening. While Tomaschek et al. (2019) focused on inflection, it is necessary to also test how well discriminative approaches can deal with derivational morphology. The present paper aims to account for this gap.

Our study investigates the durational properties of derived words in English. We modeled word durations for 4530 tokens with the derivational functions DIS, NESS, LESS, ATION and IZE from the AudioBNC (Coleman et al., 2012), using multiple linear regression models. The crucial predictors in our models are measures derived from the computational framework of linear discriminative learning (Baayen et al., 2019b).

Linear discriminative learning (LDL) is an improved version of naïve discriminative learning. Like NDL, it is *discriminative* because its system of form-meaning relations is generated by discriminating between different forms and meanings instead of building them from compositional units. Like NDL, LDL is a system of *learning* because the association strengths between forms and meanings are continuously recalibrated in a process of experience. This learning is simple and interpretable because, in contrast to deep learning, it features just two layers, an input layer and an output layer, both of which are linguistically transparent. Unlike NDL, however, LDL is *linear* and no longer "naïve". Its networks are linear mappings between form matrices and meaning matrices (which serve as either the input layer or the output layer, respectively). In this approach, forms are represented by vectors, and meanings are also represented by vectors, similarly to approaches in distributional semantics. The idea is that if we can express both forms and meanings numerically, we can mathematically connect form and meaning. In LDL, the network is no longer naïve because where NDL represents word meanings with binary vectors, LDL uses real-valued vectors, taking into account that words cannot only be similar in form, but also in meaning. How this is implemented is explained further below in Section 2.

Our aim in this study is, first, to investigate how well LDL can account for the durational variation in our data. Second, we investigate what the effects of the LDL-derived measures tell us about the mechanisms of speech production. How can we interpret potential effects conceptually? Third, as we are interested in exploring how these findings relate to morphological functions, we also investigate how the results differ depending on how much information the network has about these functions. For this purpose, we trained two different LDL networks: a first network with vectors that include semantic information about the derivative and the morphological category it belongs to (the M-Network), and a second network that does not include any information about morphological category and treats all derivatives as idiosyncratic (the I-Network).

We hypothesize that LDL-derived measures can successfully (i.e., significantly) predict derivative durations. If they do, the effects of LDL-derived measures should be interpretable with regards to speech production (for example, they should mirror the finding by Tomaschek et al. (2019) that higher certainty is associated with longer durations). Lastly, we explore whether there are differences between the network that contains information about the morphological category a derivative belongs to and the network that does not contain such information.

To preview our results, three key findings emerge from the analysis. First, both LDL networks achieve high learning accuracy and the proportion of variance in duration explained by the LDL-derived predictors is comparable to that explained by traditional predictors. Second, the effects of LDL measures highlight important patterns of speech production. For example, they suggest that words are lengthened in speech production when the semantic support of the word's predicted articulatory path is stronger (i.e., when certainty is higher), mirroring the finding by Tomaschek et al. (2019). Third, we find that, even though we did not provide the second network with any information about the morphological category a word belongs to, these categories still emerge from the network. For instance, the different morphological categories are reflected in the distributions of the correlation strength of a word's predicted semantics with the semantics of its neighbors. This corresponds to what we would traditionally describe as the differences in semantic transparency between affix categories.

The remainder of this paper is structured as follows. Section 2 describes our methodology, illustrating the procedure of collecting the speech data (Section 2.1), building the LDL networks (Section 2.2), the variables used (Section 2.3) and the modeling procedure (Section 2.4). Section 3 outlines our results, followed by a discussion and conclusion in Section 4.

2 Materials and Methods

Our methodology consists of three main steps: first, retrieving the speech data for the durational measurements for the response variable, second, building the LDL networks to retrieve LDL-derived predictors of interest, and third, devising regression models to predict derivative durations from various sets of predictors.

2.1 Speech data

The speech data was obtained from the AudioBNC (Coleman et al., 2012). This corpus consists of both monologues and dialogues from different speech genres of several British English varieties. It comes phonetically aligned by an automatic forced aligner. Containing about 7.5 million words, it is large enough to yield enough observations per derivational function. A corpus approach has the advantage that that we are not only able to analyze a lot of data, but also that the type of data is conversational speech. This enables us to investigate a more authentic process of language production than with carefully elicited speech. It has been argued (e.g. Tucker and Ernestus, 2016) that research on speech production in particular needs to shift its focus to spontaneous speech to be able to draw valid conclusions about language processing.

The morphological categories selected for investigation are DIS, NESS, LESS, ATION, and IZE. These were chosen, first, because they featured sufficient token counts in the AudioBNC and are attested in Baayen et al.'s (2019b) vector space (explained in Section 2.2.1). Second, they were chosen because they cover a wide spectrum of characteristics traditionally considered important for affix classification. For example, following Bauer et al. (2013) and Plag (2018), the affixes corresponding to those functions differ in their semantic transparency: *-ness*, *-less* and *dis-* produce mostly transparent derivatives, whereas *-ize* and *-ation* are overall a little less transparent in comparison. They vary in the range of their meanings, from relatively narrow and clearly definable semantics (e.g., the privative meaning of *-less* or the negative meaning of *dis-*) to more varied semantics (e.g., *-ness* denoting abstract states, traits, or properties) to highly multifaceted semantics (*-ize* can have locative, ornative, causative, resultative, inchoative, performative or similative meaning, *-ation* can denote events, states, locations, products or means). They also differ in their productivity, with *-ness* and *-less* being considered highly productive, and *-ize*, *-ation* and *dis-* being somewhat less productive. Lastly, they also differ phonologically. While *-ness*, *-less* and *dis-* are not (obligatorily) subject to phonological alternations and not involved in resyllabification processes, *-ize* and *-ation* can cause stress shifts and other phonological alternations within their bases, and resyllabification is commonplace.

We obtained speech data for these morphological categories by entering pertinent query strings into the web interface of the AudioBNC and extracting the resulting wordlist and associated recordings and textgrids. These query strings searched for all word tokens that begin or end in the orthographic and phonological representation of each of the investigated derivational function. We manually cleaned the datasets by excluding words which were monomorphemic (e.g., *bless*, *disk*, *station*), whose semantics or base were unclear (e.g., *harness*, *disrupt*, *dissertation*), or which were proper names or titles (e.g., *Guinness*, *Stenness*, *Stromness*).

Before starting the acoustic analysis, manual inspection of all items was necessary to exclude items that were not suitable for further analysis. This was done by visually and acoustically inspecting the items in the speech analysis software Praat (Boersma and Weenik, 2001). Items were excluded that fulfilled one or more of the following criteria: the textgrid was a duplicate or corrupted for technical reasons, the target word was not spoken or was inaudible due to background noise, the target word was interrupted by other acoustic material, laughing, or pauses, the target word was sung instead of spoken, the target word was not properly segmented or incorrectly aligned to the recording. In cases where the alignment did not seem satisfactory, we examined the word-initial boundary and the word-final boundary in order to decide whether to exclude the item. We considered an observation to be correctly aligned if none of these boundaries would have to be shifted to the left or right under application of the segmentation criteria in the pertinent phonetic literature (cf. Ladefoged and Johnson, 2011; Machač and Skarnitzl, 2009). Following Machač and Skarnitzl (2009), we considered the shape of the sound wave to be the most important cue, followed by the spectrogram, followed by listening.

In a final step, the dataset was reduced to only those words that were attested in the TASA corpus as well as in CELEX, and whose base was simplex (this step is explained in Section 2.2.1). The final dataset of derivatives that entered the models comprised 4530 tokens and 363 types. Table 1 gives an overview of the data in each morphological category.

2.2 Linear Discriminative Learning

Our aim is to predict the durational patterning in the 4530-token dataset described above with measures derived from an LDL network. These measures can be calculated on the basis of a transformation matrix that maps a cue matrix C for forms onto a semantic matrix S for meanings (for comprehension), and the semantic matrix S onto the cue matrix C (for production). The basic building blocks used to construct the meaning dimensions in matrix S are referred to as *lexomes*. Lexomes are atomic units of meaning in an LDL network. In comprehension, they are also the ‘outcomes’ in the S matrix, which are predicted from the ‘cues’ in the C matrix. Lexomes can for example correspond to words (content lexomes), but also to derivational or inflectional functions (function lexomes). How these lexomes and their vectors were obtained, how the matrices were constructed and how they were mapped onto each other is illustrated in the following sections.

2.2.1 Training data

To construct a linear discriminative learning network, it is necessary to obtain semantic vectors that represent the words' meanings (this will be explained in more detail in Section 2.2.2). For this, we made use of the vectors generated by Baayen et al. (2019b) from the TASA corpus. To make sure that we can use these semantic vectors for our derivatives, we first reduced our speech data set from the AudioBNC to those derivatives that are attested in TASA. In a second step, we used the CELEX lexical database (Baayen et al., 1995) to obtain phonological transcriptions for the words in our data set. These transcriptions are necessary for constructing the matrices. Since CELEX did not have transcriptions for all words, this step led to a slight reduction of our data set. In a final step, we excluded all derivatives whose bases were already complex, i.e., all derivatives that have more than one derivational function (e.g., *stabilization, specification, attractiveness, disclosure, disagreement*). One reason for excluding these derivatives is that it is currently not clear how to build their semantic vectors. Another reason is that multi-affixed words in corpora are comparatively infrequent. Too infrequent derivatives might require a corpus even bigger than TASA from which to construct reliable semantic vectors.

The resulting dataset contained 363 unique derivatives (i.e., types). One problem with this dataset is that it would be rather unrealistic as training data. A speaker encounters far more than just a few hundred words during their lifetime, and not all these encountered words contain one of the five investigated morphological categories DIS, NESS, LESS, ATION, and IZE. We therefore decided to merge this dataset with 4880 more words that are attested in TASA, which had already been coded in Baayen et al. (2019b) for derivational functions (function lexomes) and phonological transcriptions. This dataset contained 897 derivatives with the 25 derivational function lexomes AGAIN, AGENT, DIS, EE, ENCE, FUL, IC, INSTRUMENT, ATION, ISH, IST, IVE, IZE, LESS, LY, MENT, MIS, NESS, NOT, ORDINAL, OUS, OUT, SUB, UNDO, and Y, as well as 3983 monomorphemic words. Most of these words are not attested in our speech data and therefore not of interest for the durational modeling, but including them makes the training itself more realistic.

The resulting 5176 unique word forms were then used for the C matrix, and the 5201 unique lexomes (comprising the vectors for the 5176 content lexomes and the 25 derivational function lexomes) were used for the S matrix. The next section illustrates what these matrices are and how they are constructed.

2.2.2 Matrices for form and meaning

In an LDL network, features of a word are represented by a vector for this word in a multidimensional space. Each word has a vector that specifies its form features, and a vector that specifies its semantic features. We therefore need two matrices: a cue matrix C for the words' forms and a semantic matrix S for the words' meanings.

The cue matrix C contains in rows the words' phonological transcriptions, and in columns form indicators that are either present or absent in those words. As shown in Arnold et al. (2017) and Shafaei-Bajestan et al. (2020), it is possible to use real-valued features extracted directly from the speech signal instead of discrete features. In the present study, we use triphones as form indicators, following Baayen et al. (2019b). These triphones overlap and can be understood as proxies for transitions in the articulatory signal. Each cell in the matrix codes in a binary fashion (1 for present or 0 for absent) whether the respective triphone string (specified in the column) occurs in the phonological transcription of the word (specified in the row). An example of the layout of the C matrix is given in Table 2 on the left-hand side. For the C matrix in this study, we used the 5176 unique word forms mentioned in Section 2.2.1.

The semantic matrix S contains in its rows the words' phonological transcriptions, and in its columns the semantic dimensions, or lexomes, with which the words are associated. In the present study, these lexomes correspond to interpretable linguistic items, such as words and derivational functions. Each cell in the S matrix contains a real number, which represents the association strength of a word (specified in the row) to a lexome (specified in the column). As mentioned in Section 1, this is an important difference of LDL compared to NDL, where word meanings are initially coded as binary-valued vectors similar to the cue matrix. LDL, on the other hand, starts out with real-valued association weights. An example of the layout of the S matrix is given in Table 2 on the right-hand side. For the S matrix in this study, we used the 5201 unique lexomes mentioned in Section 2.2.1.

Where do these association weights come from? In the present study, we used association weights that were generated from word co-occurrence in real language data. For this, Baayen et al. (2019b) trained an NDL network on the TASA corpus (Ivens and Koslin, 1991). This NDL network operated on a simplified version of an established learning algorithm (Rescorla and Wagner, 1972; Widrow and Hoff, 1960) that incrementally learns association strengths between lexomes. In such an approach, words in a sentence are predicted from the words in that sentence. While the network goes through the sentences in the corpus, the associations strengths of the lexomes with each other are continuously adjusted over time. As language learning is about learning which connections are relevant, the association strength of lexomes that often occur together will be strengthened. As discriminative learning is also about *un*learning connections which are irrelevant, similarly, the association strength of lexomes will be weakened each time they do not occur together. For the implementational and mathematical details of this procedure, as well as for the validation of the resulting semantic vector space, the reader is referred to Baayen et al. (2019b). Importantly for the present study, Baayen and colleagues included lexomes not only for words, but also for derivational functions corresponding to suffixes and prefixes. This enables us to build an LDL network that takes into account morphological categories shared between derivatives (in addition to an LDL network that does not take these into account and treats all words as idiosyncratic).

The so-called *lexome-to-lexome matrix* resulting from this learning process is a vector space in which each lexome vector represents a certain association with the meanings of all other lexomes. According to the idea that “you shall know a word by the company it keeps” (Firth, 1957), each value in the vector of a lexome represents the association strength of this lexome to the meaning of another lexome in TASA. Following Baayen et al. (2019b), we used a version of their lexome-to-lexome matrix which was trimmed to about five thousand dimensions and whose main diagonal was set to zero. From this lexome-to-lexome matrix, we extracted the vectors for our 5201 unique lexomes (described in Section 2.2.1), which we then used for the S matrix.

For the present study, we built two different LDL networks: one in which the derivative vectors contain information about the morphological category the derivative belongs to, and one in which no such information is contained, but all derivatives are treated as idiosyncratic. For each of these networks we need a matrix S and a matrix C. We will refer to the matrices with information about the morphological category as matrix S_M and matrix C_M, and to the matrices with idiosyncratic derivatives as matrix S_I and matrix C_I. We will refer to the networks as a whole to the *M-Network* and the *I-Network*, respectively.

The M-Network with matrices S_M and C_M made use of the semantic vector of the content lexome of the derivative (e.g., the vector for HAPPINESS, which can be represented as $\overrightarrow{happiness}$) and the semantic vector of the corresponding derivational function lexome (e.g., the vector for NESS, which can be represented as $\overrightarrow{NESS}$). We took both these vectors from the lexome-to-lexome matrix, and the sum of these two vectors entered matrix S_M for each word. That is, the semantic vector associated with the word *happiness* was the sum of the vectors for HAPPINESS and NESS: $\overrightarrow{happiness} + \overrightarrow{NESS}$. This way, the resulting vector contains idiosyncratic information, but also information about the morphological category it shares with other derivatives.

The I-Network with matrices S_I and C_I considered only the semantic vector of the derivative lexome (e.g., only the vector for HAPPINESS, i.e., $\overrightarrow{happiness}$). This vector was taken as is from the lexome-to-lexome matrix and straightforwardly entered matrix S_I for each word. This way, the vector contains only idiosyncratic information, and no information about any shared morphological category.

We now have two matrices (for each morphological setup respectively) of the layout shown in Table 2. We have the C matrix, containing information about form, and the S matrix, containing information about meaning. These matrices can now be mapped onto each other.

2.2.3 Comprehension and production mapping

In speech comprehension, a speaker encounters a form and needs to arrive at the corresponding meaning. Therefore, for comprehension we calculate a transformation matrix F which maps the semantic matrix S onto the cue matrix C, so that

$$CF = S. \quad (1)$$

In speech production, on the other hand, a speaker starts out with a meaning and needs to find the right form to express this meaning. Therefore, for production we calculate a transformation matrix G which maps the cue matrix C onto the semantic matrix S, so that

$$SG = C. \quad (2)$$

Mathematically, the transformation matrices F and G can be calculated by multiplying the generalized inverse (Moore, 1920; Penrose, 1955) of C with S (for comprehension) and the generalized inverse of S with C (for production). The transformations are visually illustrated in Figure 1.

As soon as we have obtained the transformation matrices, we can use them to estimate what forms and meanings the network would predict. For this, we calculate the predicted matrices $\hat{S}$ and $\hat{C}$. For comprehension, we multiply the form matrix C with the transformation matrix F, i.e., we solve $\hat{S} = CF$. For production, we multiply the semantic matrix S with the transformation matrix G, i.e., we solve $\hat{C} = SG$. It is important to keep in mind that the mappings are simple linear transformations that are achieved by matrix multiplication (for an introduction in the context of LDL, see Baayen et al., 2019b). It is possible to think of the transformation matrices F and G like coefficients in linear regression, which try to approximate the target matrix but will not produce exactly the same values. This is true especially for large datasets like in the present study. The predicted matrices $\hat{S}$ and $\hat{C}$ are thus not exactly the same as the original matrices S and C.

We can also use the predicted matrices to evaluate model accuracy. To see how well the model predicts the semantics of an individual word in comprehension, we can multiply an observed form vector c from the cue matrix with the transformation matrix F to obtain a predicted semantic vector $\hat{s}$. We can then see how similar this predicted semantic vector $\hat{s}$ is to the target semantic vector s. For production, in turn, we can multiply an observed meaning vector s from the semantic matrix with the transformation matrix G to obtain the predicted form vector $\hat{c}$, which represents the estimated support for the triphones. We can then see how similar this predicted form vector $\hat{c}$ is to the target form vector c. If the correlation between the estimated vector and the targeted vector, i.e., between $\hat{s}$ and s or between $\hat{c}$ and c, respectively, is the highest among the correlations, a meaning or form is correctly recognized or produced. The overall percentage of correctly recognized meanings or forms is referred to as comprehension accuracy and production accuracy, respectively.

To obtain the mappings, we used the `learn_comprehension()` and `learn_production()` functions from the R package `WpmWithLDL` (Baayen et al., 2019a). Accuracy estimations were obtained with the functions `accuracy_comprehension()` and `accuracy_production()`. Finally, the measures of interest which we use to predict the durations were extracted from the networks with the help of the `comprehension_measures()` function and the `production_measures()` function. We will now describe these measures in more detail.

2.3 Variables

As described above, many potentially useful LDL measures can be extracted automatically from the matrices by the package `WpmWithLDL` (Baayen et al., 2019a). However, some of the variables provided by this package capture similar things and are strongly correlated with each other. Careful variable selection, and sometimes adaptation, was therefore necessary. Further below we illustrate our selection and explain the conceptual dimensions we aim to capture with each variable.

Conceptually, it is desirable to not have any traditional linguistic covariates in the models that are not derived from the network, such as lexical frequencies, neighborhood densities, or bigram frequencies. It is important to build models instead which contain LDL-derived variables only. This is because, first, we are interested in how well an LDL network fares on its own in predicting speech production. Second, many traditional covariates bring along implicit assumptions that LDL does not want to make, such as the existence of discrete phonemic and morphemic units. Third, the traditional measures have no clear correlating mechanisms in learning or processing, but at the same time they can be assumed to be reflected in a discriminative learning process. Hence, LDL measures often correlate with traditional measures.

There is, however, an important non-LDL variable that needs to be taken into account, SPEECH RATE. This is an influence that is beyond the control of the network.

2.3.1 Response variable

DURATION DIFFERENCE

One important problem in analyzing spontaneous speech is that which words are spoken is uncontrolled for phonological and segmental makeup. This problem is particularly pertinent for the present study, as our datasets feature different affixes whose derivatives vary in word length. To mitigate potential durational differences that arise simply because of the number and type of segments in each word, we refrained from using absolute observed duration as our response variable. Instead, we derived our duration measurement in the following way.

First, we measured the absolute acoustic duration of the word in milliseconds from the textgrid files with the help of scripts written in Python. Second, we calculated the mean duration of each segment in a large corpus (Walsh et al., 2013) and computed for each word the sum of the mean durations of its segments. This sum of the mean segment durations is also known as 'baseline duration,' a measure which has been successfully used as a covariate in other corpus-based studies (e.g. Gahl et al., 2012;

Caselli et al., 2016; Sóskuthy and Hay, 2017; Engemann and Plag, 2021). It would now be possible to subtract this baseline duration from the observed duration, giving us a new variable that represents only the difference in duration to what is expected based on segmental makeup. However, we found that this difference is not constant across longer and shorter words. Instead, the longer the word is on average, the smaller the difference between the baseline duration and the observed duration. In a third and final step, we therefore fitted a simple linear regression model predicting observed duration as a function of baseline duration. The residuals of this model represent our response variable. Using this method, we factor in the non-constant relationship between baseline duration and observed duration. We named this response variable DURATION DIFFERENCE, as it encodes the difference between the observed duration and a duration that is expected on the basis of the segmental makeup.

2.3.2 Predictor variables

MEAN WORD SUPPORT

MEAN WORD SUPPORT is a measure that we introduce to capture how well supported on average transitions from one triphone to the next are in the in production of a word. This variable is calculated based on the variable PATH SUM from the package `WpmWithLDL`. PATH SUM refers to the summed semantic support for the predicted articulatory path, i.e., the path from one triphone to the next in the predicted form of a word. Each node in the path, i.e., each triphone, has a certain probability of being selected against all the other possible triphones. The maximum value per transition is therefore 1, i.e., a hundred percent probability of being selected. However, with longer words, there are also more transitions. For example, if a word's form is perfectly predicted across all triphone transitions, but there are five such transitions, PATH SUM would take the value 5. Thus, the problem with PATH SUM is that it increases not only with higher support, but also with increasing segmental length of words. This would not be ideal as a measure of semantic support when modeling durations, since durations naturally increase with longer words. The interpretation of PATH SUM as a measure for mere semantic support would be difficult. Therefore, we decided to divide each value of PATH SUM, i.e., each summed support of a word's path, by the number of path nodes in a word. This new variable MEAN WORD SUPPORT controls for path length and only reflects the average transition support in each word. MEAN WORD SUPPORT can be read as a metaphor for certainty. The higher the average transition probabilities in a word, the more certain the speaker is in pronouncing this word based on its semantics.

PATH ENTROPIES

PATH ENTROPIES encode the Shannon entropy which is calculated over the path supports of the predicted semantic vector $\hat{s}$. Like MEAN WORD SUPPORT, this variable considers the transition probabilities between nodes in the path from one triphone to the next in the predicted form of a word. Higher entropy generally means more uniformity and disorder, in other words, less information. With higher entropy, the path supports vary less. Similarly to MEAN WORD SUPPORT, this measure is thus related to certainty, albeit in a conceptually different way. The higher the entropy, the less certain the speaker is in producing a word, because there is not much informational value in the path support differences. Higher PATH ENTROPIES thus indicate more uncertainty.

SEMANTIC VECTOR LENGTH

SEMANTIC VECTOR LENGTH refers to the L1 distance, also known as taxicab distance, Manhattan distance, or city-block distance, of $\hat{s}$. It thus measures the length of the predicted semantic vector by summing the vector's absolute values. We decided to use the L1 distance instead of the correlated L2 distance, as the former does not lose information by smoothing over the city-block distance. The longer the predicted semantic vector becomes, the stronger the links to other lexomes become. SEMANTIC VECTOR LENGTH can thus be understood as a measure of semantic activation diversity. It is the extent to which a given word predicts other words. As a result, it can also be understood as a measure of polysemy. The more semantic dimensions a speaker is active on for a word and the more other meanings the word can predict, the more collocational relations it has and the more varied and confusable the meanings of this word are (cf. Tucker et al., 2019).

SEMANTIC DENSITY

SEMANTIC DENSITY refers to the mean correlation of $\hat{s}$ with the semantic vectors of its top 8 neighbors' semantic vectors. A strong average correlation of the estimated semantic vector with the vectors of its neighbors means that the neighboring words are semantically very similar to the word in question. The higher the density, the more semantically similar these words are. SEMANTIC DENSITY applied to derived words is thus an important measure of semantic transparency: Words in a dissimilar neighborhood are idiosyncratic and their meaning is not predictable. Words in a semantically similar neighborhood are semantically transparent, i.e., mathematically shifted in the same direction.

TARGET CORRELATION

TARGET CORRELATION refers to the correlation between a word's predicted semantic vector $\hat{s}$ and the word's targeted semantic vector s. This is a measure for how accurate the network is in predicting meaning based on form. The closer the predicted meaning to the actual targeted meaning, the more successful the model is, and the better the listener is in making the correct connection between form and meaning.

SPEECH RATE

SPEECH RATE is the only covariate in our models, and the only predictor that is not derived from the LDL networks. The duration of a word is naturally influenced by how fast we speak. SPEECH RATE can be operationalized as the number of syllables a speaker produces in a given time interval (see, e.g., Pluymaekers et al., 2005; Plag et al., 2017). In the window containing the target word plus one second before and one second after it, we divided the number of syllables by the duration of this window. This is a good compromise between a maximally local speech rate which just includes the adjacent segments, but allows the target item to have much influence, and a maximally global speech rate, which includes larger stretches of speech but is vulnerable to changing speech rates during this larger window. The number of syllables in the window and the duration of this window were extracted from the textgrids with a Python script. A higher speech rate (i.e., more syllables being produced within the window) should lead to shortening.

2.4 Modeling word durations

We fitted multiple linear regression models to the data, using R (R Core Team, 2020). The use of mixed-effects regression with random intercepts for WORD or SPEAKER was precluded by the fact that many word types feature only one token and are produced by only one speaker in the corpus. The exclusion of these items would have resulted in a considerable loss of data.

In the course of fitting the regression models, we trimmed the dataset by removing observations from the models whose residuals were more than 2.5 standard deviations away from the mean, which led to a satisfactory distribution of the residuals (see, e.g., Baayen and Milin, 2010). This resulted in a data loss of 72 observations (1.6 % of the data) for the model based on the M-Network, and a loss of 80 observations (1.8 % of the data) for the model based on the I-Network.

From our experience, LDL-derived variables are often strongly correlated with each other. As explained in Section 2.3, we made sure to select variables that are not highly correlated and that had least conceptual overlap with each other, in terms of representing specific concepts such as certainty or semantic transparency. Still, we used variance inflation factors to test for possible multicollinearity of the remaining variables. All of the VIF values were smaller than 2, i.e., far below the critical value of 10 (Chatterjee and Hadi, 2006).

The initial models were fitted including all variables described in Section 2.3. The models were then simplified according to the standard procedure of removing non-significant terms in a stepwise fashion. An interaction term or a covariate was eligible for elimination when it was non-significant at the .05 alpha level. Non-significant terms with the highest p-value were eliminated first, followed by terms with the next-highest p-value. This was repeated until only variables remained in the models that reached significance at the .05 alpha level.

3 Results

Network accuracy was satisfactory, with comprehension accuracy at 82 % and production accuracy at 99 % for the M-Network, and a comprehension accuracy of 81 % and a production accuracy of 99 % for the I-Network.

Table 3 and Table 4 report the final models regressing duration difference against the LDL-derived variables and SPEECH RATE. The model in Table 3 includes the variables from the M-Network, while the model in Table 4 is based on the variables from the I-Network.

As we can see, of the LDL-derived variables, MEAN WORD SUPPORT, SEMANTIC DENSITY and PATH ENTROPIES significantly affect duration in both models. The variables SEMANTIC VECTOR LENGTH and TARGET CORRELATION, on the other hand, did not reach significance and were therefore excluded from these final models.

Before taking a look at the effects of individual variables, let us first examine how much variation is actually explained by the models. Table 3 and Table 4 show that for both models, the adjusted R^2 is about 0.37, i.e., about 37 % of the variance in duration is explained by the predictors. To put this number into perspective, we compared the explained variance of the two models to that of a model containing some predictor variables that are traditionally used in morpho-phonetic corpus studies of duration. We fitted a multiple linear regression model including the predictors RELATIVE FREQUENCY (a frequency-based measure for morphological decomposability, the frequency of the base word relative to its derivative from COCA; Davies, 2008), BIGRAM FREQUENCY (the frequency of the derivative occurring together with the following word, from COCA), MEAN BIPHONE PROBABILITY (the sum of all biphone probabilities in the derivative divided by the number of biphones, from the Phonotactic Probability Calculator; Vitevitch and Luce, 2004), AFFIX (which affix category the derivative belongs to) and SPEECH RATE (described in Section 2.3.2). These variables were fitted to the response variable DURATION DIFFERENCE. 76 observations, or 1.7 % of the data, were lost due to the same trimming procedure as explained in Section 2.4. For the sake of comparison of the explanatory power of individual predictors, we did not remove insignificant variables from the models. The model is reported in Table 5. We also report the ANOVA for this model in

Table 6 to summarize the effect of the AFFIX factor levels. RELATIVE FREQUENCY and BIGRAM FREQUENCY were not significant in the model, while MEAN BIPHONE PROBABILITY, AFFIX, and SPEECH RATE were. We can see that about the same proportion of the variance is explained by the traditional model (adjusted $R^2 = 0.37$).

Partitioning how much each of the predictors contributes to the proportion of explained variance, using the `lmg` metric (Lindeman et al., 1980) from the `relaimpo` package (Grömping, 2006), reveals that in both the traditional model and the LDL models, by far most of the variance is explained by SPEECH RATE (which alone explains about 35 % of the total variance in each model). The variables of interest MEAN WORD SUPPORT, SEMANTIC DENSITY, and PATH ENTROPIES are all comparable in their explanatory power to the categorical AFFIX variable and MEAN BIPHONE PROBABILITY (all between 0.2 % and 1.5 %), and considerably better than the two frequency measures RELATIVE FREQUENCY and BIGRAM FREQUENCY (<0.07 %). We can thus say with confidence that LDL-derived variables can compete against traditional variables from morpho-phonetic studies.

We can now take a closer look at the effects of each of the variables. Figure 2 plots the effects of all variables on duration (including the insignificant ones from the initial models) in the left panel, together with their distributions by derivational function in the left panel, for the M-Network and the I-Network, respectively.

Let us start with MEAN WORD SUPPORT. This variable has a significant effect on duration difference. We can see from the coefficients in Table 3 and Table 4 as well as from its positive slope in the top row of Figure 2 that higher MEAN WORD SUPPORT is significantly associated with longer durations. The higher the average semantic support of a word's predicted triphone path, the longer this word is pronounced. This means that the more certain the speaker is in producing the word, the more articulation is durationally enhanced. In other words, more certainty is associated with lengthening. Interestingly, if we look at the distribution of MEAN WORD SUPPORT in the top row of the second panel in Figure 2, we can see that mainly two derivational functions are responsible for this effect: Whereas the paths of IZE, DIS and ATION words are always very well supported, paths of NESS and LESS words often feature weaker transition probabilities between triphones. The distributional differences of each of these two categories compared to the others are significant (Mann-Whitney, $p<0.001$). This is true for both the M-Network and the I-Network. We will come back to these differences between morphological categories in the discussion.

If MEAN WORD SUPPORT indicates that with greater certainty, durations become longer, our next predictor PATH ENTROPIES should indicate that with greater uncertainty, durations become shorter. This is the case. Moving on to the second row in Figure 2, we can observe a negative slope for the effect of PATH ENTROPIES, which was significant in the models. The higher the Shannon entropy of the semantic support for the predicted articulatory paths becomes, i.e., the more disorder of support there is in the system, the shorter the durations are. More uncertainty is associated with reduction. In other words, a speaker's lower certainty in production means the articulatory signal is less strengthened or less enhanced. Again, there are differences between morphological categories both in the M-Network and the I-Network. For example, words with IZE are characterized by a more stable support, while the other categories often feature more varying supports across the paths, especially LESS and DIS. All differences in the distributions are significant at $p<0.001$ except for the non-significant difference between LESS and DIS.

The last significant LDL predictor of interest is SEMANTIC DENSITY. SEMANTIC DENSITY is significant in both models. However, its coefficients in Table 3 and Table 4 show that while it has a positive effect on duration when derived from the N-Network, it has a negative effect on duration when derived from the I-Network. This is illustrated in the third row of Figure 2. For the M-Network, the stronger an estimated semantic vector correlates with the semantic vectors of its neighbors, the longer the duration of a word becomes. For the I-Network, the stronger an estimated semantic vector correlates with its neighbors, the shorter the duration of a word becomes. High-density words are more semantically close to other surrounding words, i.e., they can be said to be less idiosyncratic and more semantically transparent. Higher transparency can thus lead to both lengthening and shortening, depending on how the network is constructed.

Moreover, SEMANTIC DENSITY does not only show differences between the networks, but also between derivational functions. Especially in the I-Network, this difference is very pronounced. This is again illustrated in Figure 2 (third row, last column). Words with LESS and IZE have particularly high densities, whereas densities are lower for DIS and NESS words, and lowest for ATION words. All of the distributions are significantly different from each other at $p < 0.001$. The fact that these morphological categories cluster so distinctly is particularly surprising, given that the I-Network was not provided with any information about these categories. We will return to the peculiar behavior of this variable in the discussion.

Let us now proceed with the remaining two variables. SEMANTIC VECTOR LENGTH and TARGET CORRELATION did not reach significance in the models, but it is still interesting to look at their distributions. For SEMANTIC VECTOR LENGTH (Figure 2, fourth row), we observe that the estimated semantic vectors are generally longer in the M-Network than in the I-Network. Not only are they longer on average, they also cluster more closely together in terms of their length: the L1 distance in the M-Network covers a range from about 2 to 3, while in the I-Network, it is spread out across a range from about 0 to 2.5. One reason for this may be purely mathematical: The vectors in the M-Network can often be longer because the vector for the derivational function lexome is added to the vector of the derived word's content lexome. However, the vectors are not just generally longer in the M-Network, but the spread of the datapoints is also narrower. This indicates that the words cluster more closely together. Since SEMANTIC VECTOR LENGTH can represent activation diversity, this is expected: If words share a morphological function with other words, they become more similar, hence are more likely to be semantically active when a member of their category is accessed. In the I-Network, words do not explicitly share a morphological category, hence members of this category are not as likely to be activated. Again, the distributions show that vector lengths cluster differently depending on derivational function, meaning that different morphological categories are characterized by different degrees of semantic activation diversity.

Finally, TARGET CORRELATION, while not significant either, tells us that the M-Network is on average more accurate than the I-Network. Looking at the distribution in the second-to-last row of Figure 2, we can observe that the correlation between the predicted semantic vector $\hat{s}$ and the target sematic vector s is slightly more condensed around the maximum value of 1 for the M-Network than for the I-Network, where distributions are gently left-skewed. This mirrors the slightly better comprehension accuracy for the M-Network of 82 % compared to an accuracy of 81 % for the I-Network. This means that the model is slightly better in predicting the correct meaning from the form when information about the morphological category is available.

The covariate SPEECH RATE in the bottom row of Figure 2 behaves as expected and requires no further investigation. We will now proceed to discuss the results in more theoretical detail.

4 Discussion and conclusion

This study set out to explore how morphological effects on the phonetic output, which were frequently observed in the literature, can be explained. From the perspective of current speech production models and theories of the morphology-phonology interaction, such effects are unexpected, and the mechanisms behind them are unclear. Our study investigated whether we can successfully model the durations of English derivatives with a new psycho-computational approach, linear discriminative learning. We hypothesized that measures derived from an LDL network are predictive of duration. We also explored what insight their effects can give us into the mechanisms of speech production, and whether the networks differ depending on the kind of information they have about morphological functions.

Our study demonstrated that LDL-derived variables can successfully predict derivative durations is supported. The mean semantic support of a word's articulatory path, the mean correlation of a word's predicted semantics with the semantics of its neighbors, and the entropy of a word's path supports all significantly affect duration. We have also shown that these measures explain a reasonable proportion of the durational variance, in the sense that their contribution to the explained variance is comparable to the contribution of traditional linguistic variables used in corpus studies of duration. The present study thus contributes to the growing literature that demonstrates LDL to be a promising alternative approach to speech production which can explain the variation in fine phonetic detail we find in different kinds of words, be they simplex, complex, or non-words (cf. Baayen et al., 2019b; Chuang et al., 2020).

Regarding the question what the effects of LDL-derived variables can tell us about speech production, we find that two important concepts relevant for production are certainty and semantic transparency. The positive effect of MEAN WORD SUPPORT and the negative effect of PATH ENTROPIES on duration both indicate that generally, higher certainty in the association of form and meaning is associated with longer durations. The better an articulatory path is on average semantically supported, and the less these supports vary over the path, the more strengthened the articulation becomes. It is important to note that the metaphor of "certainty" which is ascribed to these measures can generate two opposing expectations, both of which are intuitive in their own way. On the one hand, it could be assumed that the more certain a speaker is, the more strengthened the signal will be, leading to longer durations. This may be because a speaker invests more energy in maintaining duration when they are certain, and less energy when they are uncertain, in order to not prolong a state of uncertainty (Tucker et al., 2019). On the other hand, it could be assumed that the more certain a speaker is, the more efficient they can articulate, leading to shorter durations. This may be because more certainty could enable a speaker to select the correct path more quickly. The present results provide support for the first interpretation rather than the second one.

This is in line with the findings for other measures that have been interpreted with reference to the concept of certainty. Tomaschek et al. (2019), for instance, found that with higher functional certainty, gauged by the support for a word's inflectional lexome and the word's overall baseline support, segment durations of different types of English final S are lengthened. Kuperman et al. (2007) found that with higher certainty, gauged by the paradigmatic support (probability) of Dutch compound interfixes, these interfixes are realized longer. Cohen (2014) found that higher certainty, gauged by the paradigmatic probability of English suffixes, is associated with phonetic enhancement, i.e., again with longer durations. Cohen (2015) found that higher paradigmatic support can also enhance Russian vowels. Tucker et al. (2019) found that with higher support for tense and regularity (more certainty), acoustic duration of stem vowels increases, and with greater activation diversity (more uncertainty), acoustic duration decreases. In sum, regarding the question whether certainty has an effect of enhancement or reduction, recent evidence—including the present study—points towards enhancement.

The significant effects of SEMANTIC DENSITY indicate that a second relevant factor in the production of derivatives is the semantic relation of a word to other words. Depending on the architecture of the network, the average semantic similarity of a word's neighbors to this word can lead to both longer and shorter durations. If the network has information about the semantics of the morphological category of the derivative, higher densities are associated with longer durations. If the network has no such information and treats all words as idiosyncratic, higher densities are associated with shorter durations. In order to get a better understanding of this somewhat puzzling finding, three observations are helpful.

First, we can see in Figure 2 that SEMANTIC DENSITY is distributed very differently when derived from the M-Network than when derived from the I-Network (both the model results as well as the distributions are plotted on the same x-axis scale, respectively, for easier comparison). For the M-Network, the vast majority of data points show densities above 0.8, while for the I-Network, on the other hand, there are hardly any data points above 0.8 and the vast majority of data points have density values below 0.4. At the conceptual level this makes sense: We would expect words sharing the semantics of their morphological category to be closer to their neighboring words, i.e., to be more transparent and less idiosyncratic. This means that if the model has information about morphological categories, density should be generally higher. This is the case. In contrast, words in the I-Network are generally more dissimilar to each other because they do not share the semantic information that comes with belonging to a particular morphological category.

Returning to the relation between SEMANTIC DENSITY and duration, we can now see in Figure 2 that the two contradictory effects happen at different ends of the distribution. The negative effect found in the I-Network is carried by the low-density words, while the positive effect of semantic density on duration is carried by the high-density words. The positive effect of densities above 0.8 is even visible in the I-Network: the residuals in that range are clearly skewed towards higher durations. If we attempt an interpretation of the relation of SEMANTIC DENSITY and word duration across the two networks, we can say that the shortest durations are found in the middle of the semantic density range. Having many close semantic relatives speeds up articulation, and so does having very few relatives.

If our interpretation that SEMANTIC DENSITY captures semantic transparency is correct, we would expect higher densities to lead to longer durations. More transparent words should be more protected against phonetic reduction because they feature a stronger morphological "boundary", i.e., they are more decomposable. Such lengthening effects induced by supposed morphological boundaries have been observed in several studies (e.g., Hay, 2001, 2003, 2007; Plag and Ben Hedia, 2018). If we assume that the theoretical concept of a morphological boundary and the similarity of a word to its neighboring words capture the same underlying dimension of semantic transparency, we should still be able to replicate this effect. Since the M-Network knows that certain words share a morphological category, there are more words which are semantically very similar to each other than in the I-Network, hence many words that are semantically more transparent. However, it is not entirely clear why a higher degree of semantic transparency would lead to lengthening. Given that a higher semantic transparency means that more words will be more strongly activated, we would rather expect durations to shorten. This is because semantic activation diversity has been found to be associated with reduction (Tucker et al., 2019). It is thus unclear which of the effect directions of SEMANTIC DENSITY would be expected at the theoretical level. We leave this issue to be explored in future studies.

The discussion of SEMANTIC DENSITY leads us to another, more general issue, the nature of the network architecture that should be employed. So far, we have only discussed two kinds of network, one without any morphological information, the other with semantic information about the morphological categories involved, in addition to the information about the derived word as a whole. There is, however, a third possibility: a network that uses only the lexomes of the bases of derived words and the derivational function lexomes. In such a network, it is assumed (against our better knowledge) that the meaning of complex words is strictly compositional. This property makes this network unattractive and less suitable for predicting word durations, but it can be fruitfully used to gain further insights into the differences between architectures.

We therefore also trained this third network, which we call the *B-Network* (as it makes use of bases). Technically, instead of adding the derivational lexome vector to the lexome vector of the derivative as in the M-Network, in the B-Network we add the derivational lexome vector to the content lexome vector of the derivative's base. For instance, the semantic vector associated with the word *happiness* in matrix S_B is the sum of the vectors for HAPPY and NESS: $\overrightarrow{happy} + \overrightarrow{NESS}$. This way, the resulting vector contains information about the morphological category it shares with other derivatives, like in the M-Network. But unlike the M-Network, it contains no idiosyncratic information at all. Meaning in the B-Network is thus strictly compositional.

It is now possible to directly compare how different the three networks are with regards to their predicted semantic matrices $\hat{S}$. This can be done by calculating the correlation of each predicted semantic vector $\widehat{s_M}$ from the M-Network with its corresponding predicted semantic vector $\widehat{s_I}$ from the I-Network and $\widehat{s_B}$ from the B-Network, and then taking the mean of these correlations for all words. We find that the vectors in the M-Network and the I-Network are on average not very strongly correlated: the mean correlation between the vectors of the $\widehat{S_M}$ matrix and the $\widehat{S_I}$ matrix was r = 0.08. This means that the matrices are indeed rather different.

Interestingly, the mean correlation between the vectors of the $\widehat{S_B}$ matrix and the $\widehat{S_I}$ matrix is likewise weak (r = 0.1), but the mean correlation between the vectors of the $\widehat{S_B}$ matrix and the $\widehat{S_M}$ matrix is extremely high (r = 0.9). This indicates that it is indeed the information about derivational function that accounts for the difference between the networks. Morphological category matters.

Importantly, our results show that differences between morphological categories can emerge even from the network without any information about derivational functions. For example, semantic density is significantly higher for words with the derivational functions NESS, LESS and DIS than for words with ATION. This is in accordance with traditional descriptions of the semantic transparency of affixes, which posit *-ness*, *-less* and *dis-* as producing mostly transparent derivatives, while words with *-ation* are assumed to be less transparent (Bauer et al., 2013; Plag, 2018). Only IZE does not fit that pattern, as many IZE words are characterized by high densities but are considered about as transparent as *-ation* (however, *-ize* is considered to be more productive than *-ation*). Another interesting example of this is the distribution of SEMANTIC VECTOR LENGTH. The longer the vector of a word, the higher its semantic activation diversity becomes and the more collocational relations it has to other words, i.e., the more polysemous it is. The average vector length was highest for IZE and ATION words. This reflects traditional descriptions of *-ize* and *-ation* having highly multifaceted semantics (cf. the locative, ornative, causative, resultative, inchoative, performative or similative meaning of *-ize*, and the meanings of *-ation* denoting events, states, locations, products or means; Bauer et al., 2013; Plag, 2018). The affixes *-less*, *dis-*, and to a lesser extent *-ness*, on the other hand, have comparatively clearer and narrower semantics. In sum, these differences between morphological categories in the I-Network demonstrate that LDL can discriminate derivational functions from sublexical and contextual cues alone.

These results have implications for morphological theory and speech production models. First, the acoustic properties of morphologically complex words can be modeled successfully by implementing a discriminative learning approach. Traditional approaches were largely unable to accommodate effects of morphological structure on the phonetic output production (Chomsky and Halle, 1968; Kiparsky, 1982; Dell, 1986; Levelt et al., 1999; Roelofs and Ferreira, 2019; Turk and Shattuck-Hufnagel, 2020). Many theories of the morphology-phonology interaction assume that morphological boundaries are erased in the process of passing morphemic units on to phonological processing. And many models of speech production assume an articulator module that realizes phonemic representations with pre-programmed gesture templates independently of morphemic status. These approaches lack explanations for the fact that a word's morphological structure or semantics can cause differences in articulatory gestures, as they do not allow for a direct morphology-phonetics interaction. In LDL, however, such interaction is expected and can be explained by its underlying theoretical principles of learning and experience.

Second, our implementations show that morphological functions can emerge as a by-product of a morpheme-free learning process. Morphology is possible without morphemes. Given the many problems with the morpheme as a theoretical construct (see, e.g., Baayen et al., 2019b), this is a welcome finding. Finding morphological effects on phonetic realization need not lead to the conclusion that these effects must originate from morphemic structure. They can also come from elsewhere. As Divjak (2019) puts it, "it is not because a phenomenon can be described in a certain way that the description is psychologically realistic, let alone real" (p. 247). Of course, the success of LDL in this study and others does not allow us to infer that there is no cognitive plausibility to these structural units at all. If LDL is modeling rather how children learn languages, adult speakers may learn differently once they have explicit knowledge of morphemic structure. Such structure might also be acquired after-the-fact, when a speaker has seen enough words to start seeing analogies, or after learning about this structure explicitly. The morpheme might be epiphenomenal rather than superfluous. However, LDL does demonstrate that such fixed units of form and meaning are at the very least not obligatory. The connection between form and meaning can be dynamic and relational, allowing morphological theory to reframe its semiotic legacy. In fact, it has been argued that since its discriminative underpinnings emphasize that language is a system of *différence*, discriminative learning elegantly carries the discipline back to its Saussurean heritage (Blevins, 2016).

There are several potential future directions for discriminative learning studies on the phonetics of derived words. First, it would be interesting to model the durations of more derivational functions in a larger dataset. Investigating more than the five morphological categories of the present study might reveal further important differences between these categories. Second, one issue that we would like to resolve in future studies concerns the response variable. In a corpus study of duration with different word types, it is essential to control for length. This is why instead of duration, we decided to model duration difference, i.e., the residuals of a model regressing a word's absolute duration against the sum of its average segment durations. However, for an LDL implementation, this response variable is not optimal, since strictly speaking it still implicitly assuming segmental structure. It would be desirable to control for segmental makeup without actually having to refer to segments. Third, we think it could be fruitful to investigate how best to construct vectors for words with multiple derivational functions. This would enable us to gain more insight into the complex interplay of morphological categories.

To summarize, this study modeled the acoustic duration of 4530 English derivative tokens with the morphological functions DIS, NESS, LESS, ATION and IZE in natural speech data, using predictors derived from a linear discriminative learning network. We have demonstrated that these measures can successfully predict derivative durations. They reveal that more semantic certainty in pronunciation is associated with acoustic enhancement, i.e., longer durations, which is consistent with previous studies of paradigmatic probability and semantic support measures. We have also shown that differences between morphological categories emerge from the network, even without explicitly providing the network with such information. This further strengthens the position of LDL as a promising theoretical alternative for speech production, and provides further evidence that morphology is possible without morphemes.

Funding and acknowledgements

This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft), Research Unit FOR2373 'Spoken Morphology', project 'Morpho-Phonetic Variation in English', PL 151/7-1 and PL 151/8-1, which we gratefully acknowledge. We also thank Yu-Ying Chuang and Harald Baayen for their advice on the intricacies of linear discriminative learning.

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., 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, e0174623. doi: 10.1371/journal.pone.0174623.

Baayen, R. H., Chuang, Y.-Y., and Heitmeier, M. (2019a). *WpmWithLdl: Implementation of Word and Paradigm Morphology with Linear Discriminative Learning*.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019b). The discriminative lexicon. A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019, 1–39.

Baayen, R. H., and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research* 3, 12–28. doi: 10.21500/20112084.807.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118, 438–481. doi: 10.1037/a0023851.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database*. Philadelphia: Linguistic Data Consortium.

Bauer, L., Lieber, R., and Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford: Oxford University Press.

Ben Hedia, S. (2019). *Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics.* Berlin: Language Science Press.

Ben Hedia, S., and Plag, I. (2017). Gemination and degemination in English prefixation. Phonetic evidence for morphological organization. *Journal of Phonetics* 62, 34–49.

Blevins, J. P. (2016). "The minimal sign," in *The Cambridge handbook of morphology*, ed. A. Hippisley and G. Stump (Cambridge: Cambridge University Press), 50–69.

Boersma, P., and Weenik, D. J. M. (2001). *Praat: Doing phonetics by computer*.

Burnage, G. (1990). *CELEX: A guide for users*. Nijmegen: Centre for Lexical Information.

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

Chatterjee, S., and Hadi, A. S. (2006). *Regression analysis by example*. 4th ed. Hoboken: John Wiley & Sons.

Chomsky, N., and Halle, M. (1968). *The sound pattern of English*. New York, Evanston, London: Harper and Row.

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension. A computational modeling approach using linear discriminative learning. *Behavior Research Methods*. doi: 10.3758/s13428-020-01356-w.

Cohen, C. (2014). Probabilistic reduction and probabilistic enhancement. Contextual and paradigmatic effects on morpheme pronunciation. *Morphology* 24, 291–323.

Cohen, C. (2015). Context and paradigms. *The Mental Lexicon* 10, 313–338. doi: 10.1075/ml.10.3.01coh.

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). *Audio BNC: The audio edition of the Spoken British National Corpus*. University of Oxford.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990–present*.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93, 283–321. doi: 10.1037/0033-295X.93.3.283.

Divjak, D. (2019). *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press.

Engemann, U. M., and Plag, I. (2021). Phonetic reduction and paradigm uniformity effects in spontaneous speech. *The Mental Lexicon*.

Firth, J. R. (1957). "A synopsis of linguistic theory, 1930–1955," in *Studies in linguistic analysis (*Oxford: Blackwell), 1–31.

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, 789–806. doi: 10.1016/j.jml.2011.11.006.

Grömping, U. (2006). Relative Importance for Linear Regression in R. The Package relaimpo. *Journal of Statistical Software* 17. doi: 10.18637/jss.v017.i01.

Hay, J. (2001). Lexical frequency in morphology. Is everything relative? *Linguistics* 39, 1041–1070.

Hay, J. (2003). *Causes and consequences of word structure*. New York, London: Routledge.

Hay, J. (2007). "The phonetics of *un*," in *Lexical creativity, texts and contexts*, ed. J. Munat (Amsterdam, Philadelphia: John Benjamins), 39–57.

Ivens, S. H., and Koslin, B. L. (1991). *Demands for reading literacy require new accountability measures*. Brewster: Touchstone Applied Science Associates.

Kiparsky, P. (1982). "Lexical morphology and phonology," in *Linguistics in the morning calm: Selected papers from SICOL*, ed. I.-S. Yang (Seoul: Hanshin), 3–91.

Kuperman, V., Pluymaekers, M., Ernestus, M., and Baayen, R. H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America* 121, 2261–2271. doi: 10.1121/1.2537393.

Ladefoged, P., and Johnson, K. (2011). *A course in phonetics.* 6th ed. Boston: Wadsworth Cengage Learning.

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

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenview, London: Scott, Foresman and Company.

Machač, P., and Skarnitzl, R. (2009). *Principles of phonetic segmentation.* Prague: Epocha Publishing House.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society* 26, 394–395.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society* 51, 406–413. doi: 10.1017/S0305004100030401.

Plag, I. (2018). *Word-formation in English.* 2nd ed. Cambridge: Cambridge University Press.

Plag, I., and Ben Hedia, S. (2018). "The phonetics of newly derived words: Testing the effect of morphological segmentability on affix duration," in *Expanding the lexicon: Linguistic innovation, morphological productivity, and ludicity*, ed. S. Arndt-Lappe, A. Braun, C. Moulin, and E. Winter-Froemel (Berlin, New York: Mouton de Gruyter), 93–116.

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

Plag, I., Lohmann, A., Ben Hedia, S., and Zimmermann, J. (2020). "An <s> is an <s'>, or is it? Plural and genitive-plural are not homophonous," in *Complex words*, ed. L. Körtvélyessy and P. Stekauer (Cambridge: Cambridge University Press).

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, 2561–2569.

R Core Team (2020). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

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, 927–960. doi: 10.1080/03640210701703576.

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, 909–957. doi: 10.1111/j.1551-6709.2009.01092.x.

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

Roelofs, A., and Ferreira, V. S. (2019). "The architecture of speaking," in *Human language: From genes and brains to behavior*, ed. P. Hagoort (Cambridge, Massachusetts: MIT Press), 35–50.

Saussure, F. de (1916). *Cours de linguistique générale*. Paris: Payot.

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., and Malouf, R. (2017). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33, 32–49.

Shafaei-Bajestan, E., Moradipour-Tari, M., Uhrig, P., and Baayen, R. H. (2020). LDL-AURIS: Error-driven learning in modeling spoken word recognition. *PsyArXiv*, 1–36. doi: 10.31234/osf.io/v6cu4.

Sóskuthy, M., and Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand English. *Cognition* 166, 298–313. doi: 10.1016/j.cognition.2017.05.032.

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2019). Phonetic effects of morphology and context. Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, 1–39.

Tucker, B., Sims, M., and Baayen, R. H. (2019). Opposing forces on acoustic duration. Preprint submitted to Elsevier. *PsyArXiv*, 1–38. doi: 10.31234/osf.io/jc97w.

Tucker, B. V., and Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon* 11, 375–400. doi: 10.1075/ml.11.3.03tuc.

Turk, A., and Shattuck-Hufnagel, S. (2020). *Speech timing: Implications for theories of phonology, speech production, and speech motor control*. New York: Oxford University Press.

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

Walsh, L., Hay, J., Derek, B., Grant, L., King, J., Millar, P., Papp, V., and Watson, K. (2013). The UC QuakeBox Project. Creation of a community-focused research archive. *New Zealand English Journal* 27, 20–32.

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

Zuraw, K., Lin, I., Yang, M., and Peperkamp, S. (2020). Competition between whole-word and decomposed representations of English prefixed words. *Morphology*, 10.1007/s11525-020-09354-6. doi: 10.1007/s11525-020-09354-6.

Tables

Table 1: Overview of tokens and types per morphological category.

	DIS	NESS	LESS	ATION	IZE
tokens	233	344	145	3403	405
types	35	49	31	209	39

Table 2: Schematic examples of a cue matrix C (left) and a semantic matrix S (right) for the words *cat*, *happiness*, *walk*, and *lemon*. Note that for the triphones in the C matrix, word boundaries are also counted, represented by a hash (#). The DISC phonetic alphabet is used for computer-readable transcription (Burnage, 1990).

Schematic example of a C matrix

	#k{	k{t	{t#	#h{	h{p
k{t	1	1	1	0	0
h{pInIs	0	0	0	1	1
w$k	0	0	0	0	0
lEm@n	0	0	0	0	0

Schematic example of an S matrix

	CAT	HAPPINESS	NESS	WALK
k{t	0.000000	-6.24e-05	-0.0003179	4.71e-05
h{pInIs	-0.00056	0.0346008	0.032476	7.26e-05
w$k	0.000304	-0.0002335	-9.76e-06	0.00000
lEm@n	-7.28e-05	-2.41e-07	-0.0001247	-2.68e-05

Table 3: Final model reporting effects on duration difference with variables from the M-Network.

	Estimate	*Std. Err.*	*t-value*	*Pr(>|t|)*	
Intercept	0.090708	0.025887	3.504	0.000463	***
MEAN WORD SUPPORT	0.250262	0.020700	12.090	< 2e-16	***
SEMANTIC DENSITY	0.033868	0.012372	2.737	0.006217	**
PATH ENTROPIES	-0.008442	0.002309	-3.656	0.000259	***
SPEECH RATE	-0.058602	0.001159	-50.579	< 2e-16	***

R^2 multiple: 0.3748, adjusted: 0.3742

Table 4: Final model reporting effects on duration difference with variables from the I-Network.

	Estimate	*Std. Err.*	*t-value*	*Pr(>|t|)*	
Intercept	0.216901	0.026210	8.276	< 2e-16	***
MEAN WORD SUPPORT	0.170726	0.023507	7.263	4.45e-13	***
SEMANTIC DENSITY	-0.043545	0.008925	-4.879	1.10e-06	***
PATH ENTROPIES	-0.008688	0.002242	-3.875	0.000108	***
SPEECH RATE	-0.058757	0.001148	-51.186	< 2e-16	***

R^2 multiple: 0.3784, adjusted: 0.3778

Table 5: Model reporting effects on duration difference with traditional, non-LDL predictors.

	Estimate	*Std. Err.*	*t-value*	*Pr(>|t|)*	
Intercept	3.299e-01	1.086e-02	30.379	< 2e-16	***
RELATIVE FREQUENCY	-2.383e-05	4.167e-05	-0.572	0.567504	
BIGRAM FREQUENCY	-4.169e-07	6.135e-07	-0.680	0.496818	
MEAN BIPHONE PROBABILITY	-4.835e+00	8.661e-01	-5.583	2.51e-08	***
AFFIX `less`					
`ness`	2.921e-03	9.242e-03	0.316	0.751941	
`ation`	5.843e-02	8.201e-03	7.125	1.21e-12	***
`dis`	6.504e-02	1.016e-02	6.399	1.73e-10	***
`ize`	3.451e-02	9.222e-03	3.742	0.000185	***
SPEECH RATE	-5.885e-02	1.161e-03	-50.680	< 2e-16	***

R^2 multiple: 0.3736, adjusted: 0.3724

Table 6: Anova for model reporting effects on duration difference with traditional, non-LDL predictors.

	Df	*Sum Sq*	*Mean Sq*	*F-value*	*Pr(>F)*	
RELATIVE FREQUENCY	1	0.018	0.0182	2.1070	0.14669	
MEAN BIPHONE PROBABILITY	1	0.043	0.0433	5.0118	0.02522	*
AFFIX	4	0.581	0.1452	16.8251	1.069e-13	***
SPEECH RATE	1	22.223	22.2229	2574.5115	< 2.2e-16	***
BIGRAM FREQUENCY	1	0.004	0.0040	0.4618	0.49682	

Figures

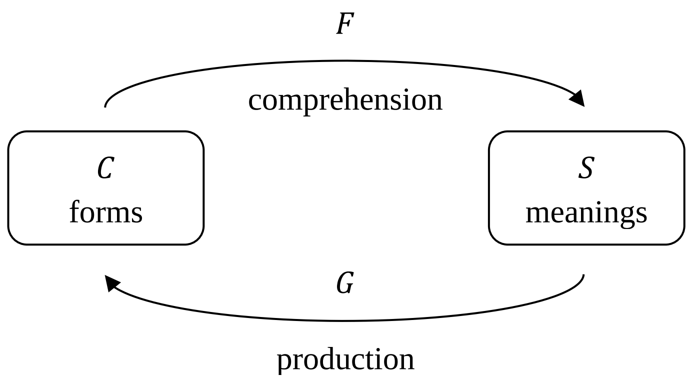

Figure 1: Comprehension and production mapping, adapted from Baayen et al. (2019b). For comprehension, transformation matrix F transforms the cue matrix C into the semantic matrix S. For production, transformation matrix G transforms the semantic matrix S into the cue matrix C.

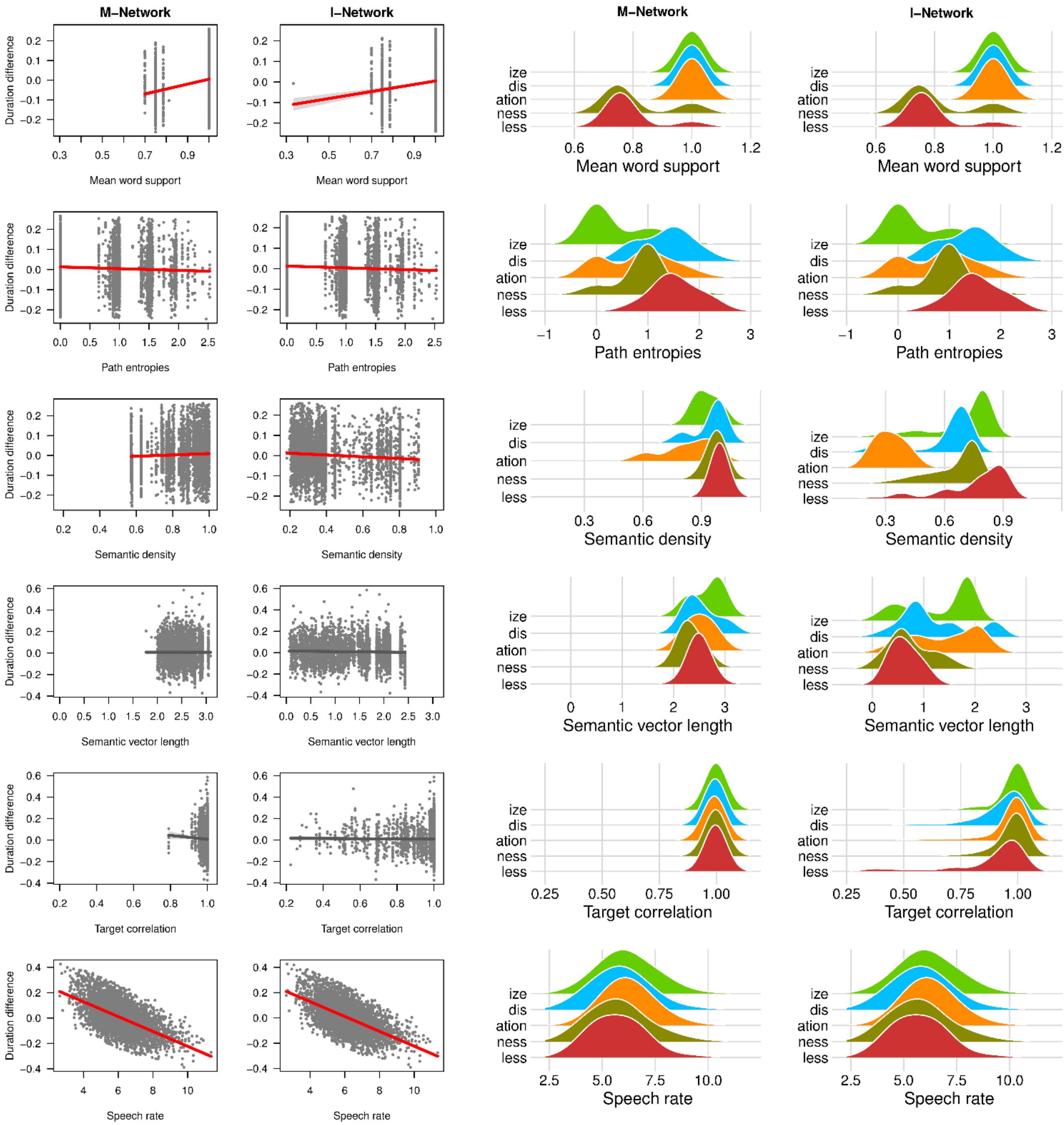

Figure 2: *Left panel*: Effects on duration difference for the M-Network variables (left column) and the I-Network variables (right column). Red regression lines indicate significant effects from the final models, grey regression lines indicate non-significant effects from the initial models before the non-significant predictors were excluded. *Right panel*: Density distributions of variables by derivational function in the M-Network models (left column) and in the I-Network models (right column).