Working Title:

Durational differences of word-final /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning

Dominic Schmitz[1], Ingo Plag[1], Dinah Baer-Henney[2], Simon David Stein[1]

[1] English Language and Linguistics, Heinrich Heine University, Düsseldorf, Germany

[2] Linguistics and Information Science, Heinrich Heine University, Düsseldorf, Germany

Abstract

Recent research has shown that seemingly identical suffixes such as word-final /s/ in English show systematic differences in their phonetic realizations. Most recently, durational differences between different types of /s/ have been found to also hold for pseudowords: the duration of /s/ is longest in non-morphemic contexts, shorter with suffixes, and shortest in clitics. At the theoretical level such systematic differences are unexpected and unaccounted for in current theories of speech production. Following this approach, we implemented a linear discriminative learning network trained on real word data in order to predict the duration of word-final non-morphemic and plural /s/ in pseudowords using production data by a previous production study. It is demonstrated that the duration of word-final /s/ in pseudowords can be predicted by LDL networks trained on real word data. That is, duration of word-final /s/ in pseudowords can be predicted based on their relations to the lexicon.

1. Introduction

Many studies on the acoustic properties of phonologically homophonous elements have shown unexpected effects of their morphological structure on their phonetic realization. Such effects were shown for seemingly homophonous lexemes (Drager, 2011; Gahl, 2008), for free and bound variants of stems (Kemps et al., 2005a; Kemps et al., 2005b), and for prefixes (Ben Hedia, 2019; Ben Hedia & Plag, 2017).

For the level of individual segments, a number of studies have shown that the acoustic realization of word-final /s/ and /z/ (henceforth S) in English depends on its morphological status and category. Corpus studies (Plag et al., 2017; Zimmermann, 2016) found that non-morphemic word-final S shows longest acoustic durations, followed by suffixes, which in turn are followed by clitics. Experimental studies (Hsieh et al., 1999; Plag et al., 2020; Seyfarth et al., 2017; Walsh & Parker, 1983) confirm durational differences between different types of S. However, their results are mostly not as clear as those by previous corpus studies. That is, only recently a study by Schmitz et al. (2020) on word-final S in pseudowords confirmed the pattern of durational differences found previously only in corpus studies.

Most importantly, none of the aforementioned studies on the matter of word-final S was able to explain found differences on a theoretical level. Traditional models of speech production come with the assumption of having no morphological information in phonetic processing (Levelt et al., 1999; Roelofs & Ferreira, 2019; Turk & Shattuck-Hufnagel, 2020), thus rendering an explanation on basis of differing morphological categories improbable. Other accounts, e.g. standard feed-forward theories of morphology-phonology interaction (e.g. Chomsky & Halle, 1968; Kiparsky, 1982) or prosodic phonology (e.g. Booij, 1983; Goad, 1998, 2002; Selkirk, 1996) do not offer a satisfying explanation for such durational differences, either.

Only recently, Tomaschek et al. (2019) analysed durational differences between types of S by means of an implementation of naïve discriminative learning (Baayen et al., 2011; Ramscar et al., 2010; Ramscar & Yarlett, 2007). Their results indicate that the duration of a word-final S in English can be sufficiently approximated by considering the support for its morphological function from the word's sublexical and collocational properties.

This paper continues this line of evidence by making use of the computational model of linear discriminative learning (Baayen et al., 2019b; Chuang et al., 2020), the more advanced successor of naïve discriminative learning. We analyse the durational differences between non-morphemic and plural word-final /s/ found not in real words, but

in pseudowords. By using nonce words, we want to rule out potentially confounding effects of the lexical and contextual properties of the individual utterances (e.g. Caselli et al., 2016). Making use of measures derived from this implementation of linear discriminative learning, the present study demonstrates that the effects found by Tomaschek et al. (2019) can be confirmed. Differences in phonetic duration emerge from differences in the strengths of associations between form and meaning.

We proceed as follows. The next section will give an overview on studies on the duration of word-final S, and possibilities and obstacles of theoretical accounts. Section 3 introduces linear discriminative learning on a theoretical level, while section 4 presents the implementation of linear discriminative learning used in the present study. The analysis and results of our study are given in section 5 and 6. A discussion of the obtained results and a conclusion follow in section 7.

2. Word-final /s/ and its duration

A number of morphological categories can take the phonological form of /s/ in English, i.e. plural, genitive, genitive plural, third person singular, and the clitics of *is*, *has*, and *us*. In itself, there is nothing in the phonological form of these morphological categories that indicates systematic differences in realization on the phonetic level between different S morphemes or a non-morphemic S. Yet, a number of studies report on durational differences between different types of S.

Corpus studies on word-final S in English find differences in duration between non-morphemic, suffix, and clitic variants. Zimmermann (2016) on New Zealand English, and Plag et al. (2017) and Tomaschek et al. (2019) on North American English find that non-morphemic S (as in *grace*, *cheese*, *bus*) shows longer durations than plural S and the clitic S of *has* and *is*, while plural S in turn shows longer durations than clitic S.

Turning to experimental studies, results are not as consistent. Walsh & Parker (1983) conducted a production experiment with three homophonous word pairs with all words ending in either a non-morphemic or morphemic word-final S. Tested in three different contexts, they find durational differences in two of them. They conclude that morphemic S in English is systematically lengthened by speakers (Walsh & Parker, 1983: 204). However, their conclusion relies on only a small number of 110 observations, a mixture of common and proper nouns as items, and lacks appropriate inferential statistical methods as well as an integration of covariates.

Hsieh et al. (1999) find that plural S is longer than third person singular S in child-directed speech. However, as their data was originally elicited for another study (Swanson & Leonard, 1994), half of all plural items occurred sentence-finally, while almost all third person singular items occurred sentence-medial. Thus, the durational differences found by Hsieh et al. (1999) may be attributed to effects of phrase-final lengthening (e.g. Klatt, 1976; Wightman et al., 1992) rather than to phonetic differences between different types of S.

In another study, Seyfarth et al. (2017) conducted a production experiment on word-final /s/ and /z/ in non-morphemic, plural, and third person singular contexts. Their results indicate that non-morphemic S is shorter than morphemic S. However, they do not find a difference between voiced and voiceless instances, even though previous studies confirm differences dependent on voicing (e.g. Plag et al., 2017). With only six items ending in /s/, but twenty items ending in /z/, it is questionable how meaningful their results on different types of S are.

Comparing affixes, Plag et al. (2020) find that plural and genitive plural S differ in duration. That is, in their study the genitive plural suffix shows a longer duration than the plural suffix.

Most recently, Schmitz et al. (2020) conducted a production experiment on pseudowords carrying either a non-morphemic, plural, or *is*- or *has*-clitic S. Their results are in line with those of aforementioned corpus studies. That is, non-morphemic S shows longest S durations, followed by plural S, which in turn is followed by clitic S, while there is no significant durational difference between the two clitics. An overview of the durational differences found in corpus and experimental studies is given in Table 1.

Table 1. Overview of durational differences of word-final /s/ found in previous studies.

Study	Findings
Zimmermann, 2016; Plag et al., 2017; Tomaschek et al., 2019; Schmitz et al., 2020	non-morphemic > plural > clitics
Walsh & Parker, 1983	plural > non-morphemic
Hsieh et al., 1999	plural > third person singular
Seyfarth et al., 2017	plural > non-morphemic
Plag et al., 2020	genitive plural > plural

There is a noteworthy discrepancy between experimental results and the results based on conversational speech data. One reason for the different findings could be that all experimental studies used homophones (such as *laps* and *lapse*). While homophones have the advantage of controlling for the phonetic environment in which S occurs, they may bring in particular problems in processing (e.g. competition at the form level, different part-of-speech) due to their being homophones. These problems may result in durational patterns different from non-homophonous word forms.

But even if the direction of durational differences between different types of S is not entirely clear yet, it appears that there are indeed durational differences of some sort. How is one to explain such differences? In standard feed-forward theories of morphology-phonology interaction (e.g. Chomsky & Halle, 1968; Kiparsky, 1982) all types of S, morphemic and non-morphemic, are treated in a similar way. For morphologically complex words, e.g. words ending in a morphological word-final S, a process named 'bracket erasure' is said to remove any morphological information. Thus, leaving speech production with no information on the morphology of a complex word (e.g. the plural form *cats*), rendering its morphological information equal to that of a morphologically simple word ending in a non-morphemic word-final S (e.g. the singular form *bus*). In such

a system, there is nothing that could account for realizational differences between phonologically identical forms of suffixes, clitics, and non-morphemic segments.

A similar distinction of lexical and post-lexical processing is also found in established theories of psycholinguistics. According to models of speech production (e.g. Levelt et al., 1999; Roelofs & Ferreira, 2019), morphemic types of word-final S do not differ in their realization from non-morphemic instances of word-final S. For a plural form, e.g. *cats*, the lemma of the lexical concept CAT and a plural specification are retrieved. Then, during morphological encoding, the plural specification is mapped onto the base lemma, i.e. *cat*, and the plural suffix, <-s>. During phonological encoding, phonemes are selected for the corresponding morphemes, i.e. /k/, /æ/, /t/, and /s/. Finally, the phonemes are syllabified, resulting in a phonological word representation. Such phonological forms are then forwarded and used in speech production. Thus, no information on the morphological origin of particular segments is contained in the phonetic realization, rendering an explanation on durational differences between types of S on morphological grounds improbable.

In prosodic phonology (e.g. Booij, 1983), differences in phonetic realization may arise from the position of sounds in different configurations of prosodic constituency. For instance, different types of word-final S can be analysed as being integrated at different levels of the hierarchical prosodic configuration. In the case of word-final S, different levels co-determine differing degrees of integration of an S to the word it belongs to. Non-morphemic S, uncontroversially, is an integral part of the prosodic word itself (Selkirk, 1996), see panel A of Figure 1. For plural S, Goad (1998) analyses it as an 'internal clitic', see panel B, while Goad (2002) analyses it as an 'affixal clitic', see panel C.

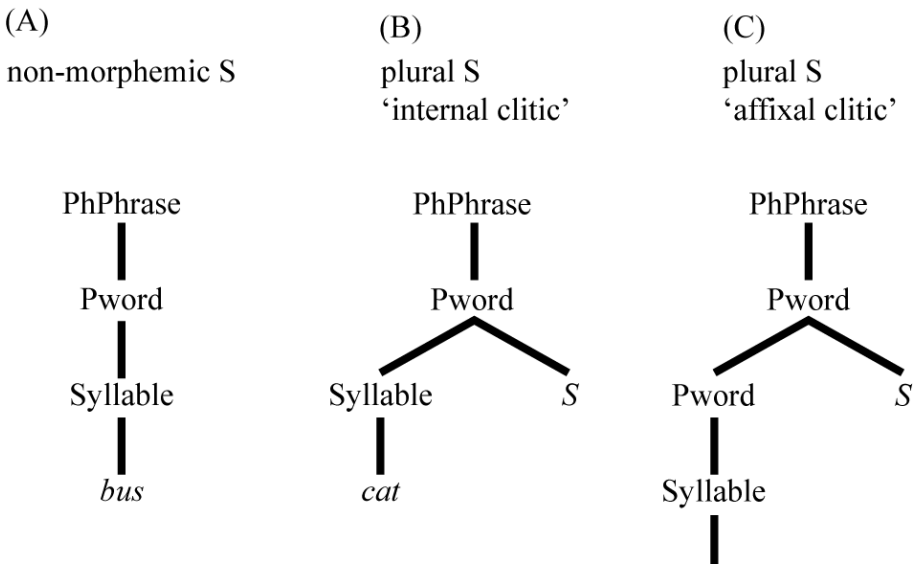


Figure 1. Prosodic configuration for non-morphemic and plural S.

Thus, the prosodic approach posits a structural prosodic difference between types of S. However, it is not so clear what particular phonetic effects these differences would predict. Most plausibly, a higher degree of integration would correlate with shorter durations, predicting shortest S durations in monomorphemic words. Yet, findings on S duration show the opposite (e.g. Plag et al., 2017; Schmitz et al., 2020; Tomaschek et al., 2019; Zimmermann, 2016), i.e. the duration of non-morphemic S is longest.

An alternative explanation for durational differences between different types of S can be found within the computational modelling framework of naïve discriminate learning (NDL; e.g. Baayen et al., 2011; Ramscar et al., 2010; Ramscar & Yarlett, 2007). NDL is based on simple but powerful principles of discriminative learning theory (Rescorla, 1988; Wagner & Rescorla, 1972), i.e. learning results from exposure to informative relations among events in the individual's environment. Such events are used to form associations between them, while the associations and their resulting representations are constantly updated on the basis of new experiences. Associations are built between features ('cues', e.g. biphones) and content lexemes or morphological functions ('outcomes', e.g. different types of S), which co-occur in events in which the individual is predicting outcomes from cues (Tomaschek et al., 2019: 11). Using the Rescorla-Wagner equations (Rescorla, 1988; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972), relations between cues and outcomes are modelled. That is, the weight of an association, i.e. its strength, increases every time a cue and an outcome co-occur, while it decreases if a cue occurs without the outcome. The result of this process is a continuous recalibration of association strengths, which is a crucial part of discriminative learning.

NDL has been used successfully to model various morphological phenomena, e.g. reaction times in studies on morphological processing (e.g. Baayen et al., 2011; Blevins et al., 2016; see Plag, 2018, chapter 7, for an introduction to NDL in morphological research). For word-final S, Tomaschek et al. (2019) reproduce the differences in duration found by Plag et al. (2017) by means of NDL measures. Their study shows that the duration of different types of S can be approximated by considering the support for these morphological functions from a word's sublexical and collocational properties. In the NDL network, all words and their diphones within a five word window centred on the target word that contained the S served as cues, and were associated with the morphological functions, which served as outcomes. Two main measurements from this network emerged as predictive for S duration. First, the so-called 'activation' as a measure of an outcome's baseline activation, i.e. of how well an outcome is entrenched in the

lexicon. Second, the so-called 'activation diversity' as a measure to quantify the extent to which the cues in a given context also support other targets. Taken together, the following pattern for S duration emerges: When the uncertainty about a targeted outcome increases, i.e. the level of 'activation' decreases and the level of 'activation diversity' increases, the duration of S decreases. In other words: The stronger the support for a morphological function is, both from long-term entrenchment and short-term from the context, the longer its duration.

While NDL implementations apparently offer some form of explanation for different durations of different types of S, they also come with shortcomings and limitations. In NDL, a word's meaning is defined in terms of the presence or absence of an outcome, i.e. NDL "adopted a stark form of naive realism" (Baayen et al., 2019b: 4) just for computational reasons. That is, NDL takes into account that words tend to have similar forms, but ignores that words are also similar in meaning. Thus, Baayen et al. (2019b) introduced semantic vectors of reals replacing the binarily coded row vectors of the semantic matrix (see 3.2), naming their new implementation *linear* discriminative learning (LDL) instead of *naïve* discriminative learning. Outcomes are no longer assumed to be independent, i.e. semantic similarities are now reflected, and networks are mathematically equivalent to linear mappings of matrices, i.e. vector spaces.

It is the implementation of such linear discriminative learning that the present paper makes use of for analysing the duration of word-final types of S. Our paper explores whether measures derived from an LDL implementation are predictive of different types of S and their durations. In order to better understand the relation between traditional psycholinguistic variables (such as lexical frequencies, neighbourhood densities, bigram probabilities, morphological category, etc.) and LDL measurements we also compare models that use measures derived from an LDL implementation with models that use traditional measures to predict S durations. Finally, we test whether measures derived from an LDL implementation render the specification of morphological structure proper (affix vs. no affix) as predictor variable for S duration unnecessary.

3. Introduction to LDL

3.1. Overview

Linear discriminative learning as a computational model implements a discriminative view of learning. In contrast to deep learning models that have multiple hidden layers based on non-linear functions, LDL networks are very simple two-layer networks and are linguistically transparent and interpretable. In LDL, the mental lexicon consists of five high-dimensional numeric matrices, each of which represents a different subsystem: the visual matrix, retina; the auditory matrix, cochlea; the semantic matrix; the speech matrix, speaking; and the spelling matrix, typing. For the current implementation, the semantic and the speech matrix are most important.

With regards to the mappings between vectors, linear mappings are implemented. These mappings are estimated using the linear algebra of multivariate regression. Thus, each mapping is defined by a matrix A that transforms the row vectors in a matrix X into the row vectors of a matrix Y, i.e. $Y = XA$. Then, $A = X'Y$, where X' is the generalized inverse of X. We will return to the mapping of matrices in section 3.4, and refer the interested reader to Baayen et al. (2019b) for an introduction to the mathematical details, as well as to Milin et al. (2017) for a detailed discussion on the restrictions and possibilities of linear mappings.

Another important feature of LDL is its notion of lexomes, i.e. basic semantic units corresponding to words or morphological functions. As outlined in Chuang et al. (2020), lexomes fall into two groups: content lexomes, and inflectional and derivational lexomes. Content lexomes can be morphologically simple or complex forms, i.e. *cat* and *cats*. Inflectional lexomes represent inflectional functions, e.g. number, tense, and aspect. Derivational lexomes represent derivational functions, e.g. morphological categories such as -NESS, -LESS, or UN-. Each lexome is paired with a vector of the aforementioned five subsystems. That is, for the semantic matrix, each lexome is paired with a semantic vector, making each lexome a pointer to a semantic vector on the one hand (Milin et al., 2017), and a location in a high-dimensional space on the other hand. For monomorphemic words, the semantic vector is identical to the semantic vector of the corresponding lexome. That is, the semantic vector of the word *cat*, $\overrightarrow{cat}$, is identical to the vector of the lexome CAT. For complex words, the semantic vector is the sum of its corresponding lexome vectors. That is, the semantic vector of the word *cats*, $\overrightarrow{cats}$, is the sum of the semantic vectors of the lexomes CAT and PLURAL, $\overrightarrow{cat} + \overrightarrow{plural}$. The implementation of LDL and the matrices necessary for the present paper are introduced in the subsequent sections. Please

refer to https://osf.io/zy7ar/?view_only=ef43a5caf6444270a56074027d7d6482 for a detailed implementation of LDL in R (R Core Team, 2020).

3.2. The S matrix: Semantic vectors

The semantic matrix S contains semantic vectors of word forms on basis of their corresponding lexomes. That is, the semantic vector $\vec{s}$ in S for a simplex word is identical to its corresponding lexome, while the semantic vector $\vec{s}$ in S for a complex word is the sum of its corresponding lexomes, e.g., $\overrightarrow{apple} + \overrightarrow{plural}$ for *apples* (Baayen et al. 2019b). Semantic vectors of lexomes can be derived in different ways (e.g. Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Shaoul & Westbury, 2010).

3.3. The C matrix: Form vectors

The present study use triphones to represent form, following Baayen et al. (2019b). Triphones are sequences of three phones within a word form. They overlap and can be understood as proxies for phonetic transitions. The cue matrix C encodes the forms of words in a binary fashion, giving information on which triphones are part of which word. In each word's individual form vector $\vec{c}$, the presence of a triphone is marked with 1, while the absence is marked with 0. The cue vectors of all words of a set of words constitute its C matrix. That is, each row in such a C matrix represents a word form, while the columns of the respective C matrix represent all triphones of its underlying word set.

3.4. Comprehension and Production

In LDL, comprehension refers to a model that has form vectors as input and semantic vectors as output. We illustrate the C matrix of a set of words with a toy lexicon containing the words *cat*, *bus*, and *eel*. Here, the DISC keyboard phonetic alphabet (the "Distinct Single Character" representation introduced by Burnage, 1988) is used for triphone representation. Word boundaries are marked by the # symbol.

$$C = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{c} \begin{array}{cccccccc} \#k\{ & k\{t & \{t\# & \#bV & bVs & Vs\# & \#il & il\# \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \end{array} \quad (1)$$

For the same toy lexicon, suppose that the semantic vectors for these three words are the row vectors of the following S matrix:

$$S = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{c} \begin{array}{ccc} cat & bus & eel \end{array} \\ \begin{pmatrix} 1 & 0.2 & 0.5 \\ 0.4 & 1 & 0.1 \\ 0.2 & 0.3 & 1 \end{pmatrix} \end{array}. \tag{2}$$

To map forms onto meanings we need transformation matrix F, such that

$$CF = S. \tag{3}$$

The transformation matrix F is straightforward to obtain. Let C' denote the Moore-Penrose generalized inverse[1] of C, available in R as the *ginv* function of the MASS package (Venables & Ripley, 2002). Then,

$$F = C'S. \tag{4}$$

For the toy lexicon example,

$$F = \begin{array}{c} \\ \#k\{ \\ k\{t \\ \{t\# \\ \#bV \\ bVs \\ Vs\# \\ \#il \\ il\# \end{array} \begin{array}{c} \begin{array}{ccc} cat & bus & eel \end{array} \\ \begin{pmatrix} 0.33 & 0.06 & 0.16 \\ 0.33 & 0.06 & 0.16 \\ 0.33 & 0.06 & 0.16 \\ 0.13 & 0.33 & 0.03 \\ 0.13 & 0.33 & 0.03 \\ 0.13 & 0.33 & 0.03 \\ 0.10 & 0.15 & 0.50 \\ 0.10 & 0.15 & 0.50 \end{pmatrix} \end{array}, \tag{5}$$

[1] The inverse of a matrix needs not exist, rendering such a matrix a singular one. Most matrices used in LDL implementations are singular matrices. Thus, an approximation of the inverse must be used instead of an inverse itself. One such approximation is the Moore-Penrose generalized inverse (Moore, 1920; Penrose, 1955).

with CF being exactly equal to S in this simple example. That is, taking form vectors as input for the prediction of semantic vectors as output, i.e. solving $\hat{S} = CF$, this toy example correctly predicts 100% of all (three) words' semantics, i.e. $\hat{s}_i = s_i$. In more complex cases, semantic vectors are only approximately identical, thus, for a word *i* and its predicted semantic vector $\hat{s}_i$, comprehension is successful if $\hat{s}_i$ shows the highest correlation with the targeted semantic vector s_i (Baayen et al., 2019b). Following this method, one can report the percentage of comprehension accuracy.

Production as modelled in LDL takes semantic vectors as input and delivers form vectors as output. Using the same toy lexicon as before, we adapt its C matrix, i.e. we borrow the notation by Baayen et al. (2019b) and henceforth call it T as is contains the Targeted triphones. For production, the transformation matrix G is of interest. Similar to F for comprehension, it is straightforward to obtain. Let S' denote the Moore-Penrose generalized inverse of S. Then,

$$G = S'T. \tag{6}$$

Given G, one can then predict the triphone matrix $\hat{T}$ from the semantic matrix S by solving

$$\hat{T} = SG. \tag{7}$$

For our toy lexicon example, the G transformation matrix is

$$G = \begin{array}{c} \\ cat \\ bus \\ eel \end{array} \begin{array}{c} \begin{array}{cccccccc} \#k\{ & k\{t & \{t\# & \#bV & bVs & Vs\# & \#il & il\# \end{array} \\ \begin{pmatrix} 1.14 & 1.14 & 1.14 & -0.06 & -0.06 & -0.06 & -0.56 & -0.56 \\ -0.44 & -0.44 & -0.44 & 1.05 & 1.05 & 1.05 & 0.12 & 0.12 \\ -0.09 & -0.09 & -0.09 & -0.30 & -0.30 & -0.30 & 1.08 & 1.08 \end{pmatrix}. \end{array} \tag{8}$$

As this is a toy example, SG is identical to T. For more complex cases, $\hat{T}$ will not be virtually identical to T "but will be an approximation of it that is optimal in the least squares sense" (Baayen et al., 2019b:21). Triphones with strongest support are expected to be the triphones making up a word's form. As triphones are not ordered, it is also checked whether the sequence of phones can be constructed correctly. Both, checking triphone support and sequence, are conveniently done by the functions of the

WpmWithLdl package (Baayen et al., 2019a). Following this method, one can report the percentage of production accuracy.

Figure 2 summarizes the mapping between form and meaning by the F and G transformation matrix for comprehension and production modelling.

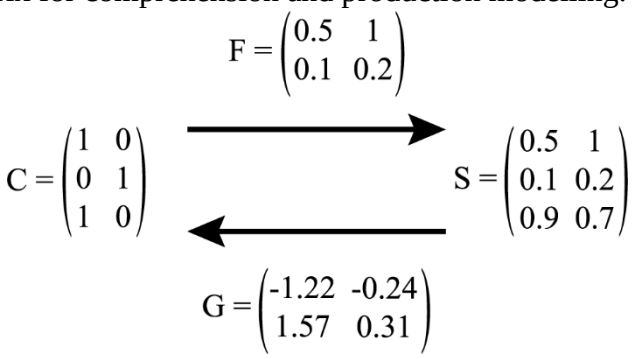


Figure 2. Illustration of mapping between C and S matrix via F (i.e. comprehension), and S and C matrix via G (i.e. production). Note: In production, C is referred to as T.

4. Combining real words and pseudowords in an LDL implementation

4.1. The semantics of pseudowords

The present paper follows the implementational basics outlined in section 3. However, as we are interested in /s/ durations in pseudowords (and not in real words), there are a number of complications. The most important complication arises from the widely shared belief that pseudowords do not have meaning. So how can we map form and meaning with forms that have no meaning? In a recent study (Chuang et al. 2020), it was shown that the assumption that pseudowords are bare of meaning is most probably wrong. Due to their formal similarity with existing words, pseudowords resonate with the lexicon. As a result, they may in fact carry meaning. The authors demonstrate that quantitative measures gauging the semantic neighbourhoods of pseudowords predict reaction times of lexical decisions and acoustic durations. The present study is inspired by these results and implements a similar architecture. To model resonance of pseudowords with the lexicon, both real words and pseudowords must be included in the networks. The following sections will detail the combined LDL implementation of real words and pseudowords.

4.2. Data sets: Real words and pseudowords

The pseudowords and their phonetic realizations that this paper is based on are taken from the study of word-final /s/ production by Schmitz et al. (2020). In total, 24 pairs of pseudowords were used in that study. Each pseudoword form can act as singular

or plural noun, e.g. *glaits* is either realized as singular, i.e. *a glaits*, or as plural, i.e. *two glaits*. Additionally, some pseudowords show a number of different realizations by the participants in the experiment, e.g. *prups* is sometimes produced as /pɹʌps/, and sometimes it is produced as /pɹups/. Thus, not 48 (i.e. 2 x 24) but 78 different phonological forms are included in the pseudoword set. Supplementary Table 1 gives an overview of all pseudowords and their phonological forms.

The second set of words contains real words and their phonetic realizations. Following Chuang et al. (2020) we extracted these words from the MALD corpus (Tucker et al., 2019). While the MALD corpus contains 26,793 real words, only a subset of 8,362 words is used for a number of reasons. First, some 7577 words in the corpus contain multiple affixes. As it was unclear how to handle such words, these were excluded. Second, only words for which we have semantic vectors could be used, leading to the exclusion of further 6751 words. Third, only words with transcriptions available in the CELEX corpus (Baayen et al., 1995) were retained, i.e. there was no transcription available for 818 words. Fourth, 3285 words showed ambiguities regarding their morphology, e.g. *walks* as a third person singular verb versus the plural of a noun. As huge numbers of words lead to extensive computation times, we decided to exclude such cases as well. The final set of real words contains 6242 simple and 2120 complex word forms.

4.3. Cue matrices

As introduced in section 3.3, cue matrices are coded in binary form, giving information on which triphones are part of which word. For the current implementation, two such cue matrices are created using the WpmWithLdl package's (Baayen et al., 2019a) *make_cue_matrix* function. First C_{rw}, the real word cue matrix, is created for the set of real words. Then, a second cue matrix, C_{pw}, is created for the set of pseudowords.

4.4. Semantic matrices

To introduce semantics, i.e. semantic vectors, for the present set of real words, a pre-built semantic matrix A from Baayen et al. (2019b) was used. These authors derived semantic vectors based on the TASA corpus (Ivens & Koslin, 1991). For this, words were parsed into their lexomes, i.e. inflected words were represented by their stem and sense-disambiguated labels for their respective inflectional functions. Ambiguous forms, e.g. *walks*, were disambiguated using part of speech tagging (Schmid, 1999). Derived words

were assigned a lexome for their stem and a lexome for derivational function. Then, following Baayen et al. (2016) and Milin et al. (2017), Naïve Discriminative Learning (Baayen et al., 2011; Sering et al., 2018) was used to build semantic vectors. The Rescorla-Wagner update rule (Rescorla, 1988; Rescorla & Wagner, 1972; Wagner & Rescorla, 1972) was applied incrementally to the sentences of the TASA corpus. That is, for each sentence the algorithm was given the task to predict the lexomes in that sentence from all lexomes of that sentence. This resulted in a 23562×23562 weight matrix A. This matrix lists all lexomes as rows and columns. Thus, for a given lexome at row *i*, the association strengths of this lexome with all other lexomes as given as columns is contained. In this state of the A matrix, lexomes predict themselves. Thus, the diagonal of the A matrix is set to zero (see Baayen et al., 2019b, for a discussion on this procedure). Lastly, columns which mostly contain zeros, i.e. no information, and show small variances ($\sigma < 3.4 * 10^{-8}$) are removed. The resulting A matrix is of dimension $23{,}562 \times 5030$. Following the method outlined in section 3.2, a semantic matrix for real words S_{rw} can be constructed based on A. That is, the semantic vector $\vec{s}$ in S_{rw} for a simplex word is identical to its corresponding lexome, while the semantic vector $\vec{s}$ in S_{rw} for a complex word is the sum of its corresponding lexomes. As a set of real words is used, S_{rw} contains only semantic vectors for this set of real words (instead of, e.g., all word forms of the TASA corpus). The final real word semantic matrix S_{rw} is of dimension 8362×5487.

While this procedure is rather straightforward, the creation of a pseudoword semantic matrix S_{pw} is not. Due to the nature of pseudowords, their lexomes are not contained within any corpus or our A matrix, for that matter. Instead, one can estimate a pseudoword's semantic content by utilizing the semantic and phonological information on real words, i.e. their C and S matrix (Chuang et al., 2020). That is, the same transformation matrix F that is used for mapping real word cues onto predicted real word meanings (see section 3.4) can be used to map pseudoword cues onto their estimated semantics. That is, one must first solve

$$F = C'_{rw} S_{rw} \tag{9}$$

to obtain F. Then, one can make use of the pseudoword cue matrix C_{pw}, and estimate pseudoword semantics, as

$$\hat{S}^{o}_{pw} = C_{pw}F, \quad (10)$$

with $\hat{S}^{o}_{pw}$ denoting the originally estimated semantic matrix for pseudowords. See the aforementioned complementary online material for a detailed implementation.

4.5. Comprehension and Production

Pseudoword comprehension and production are not computed and evaluated in isolation, but in combination with real words, simulating a real person's lexicon in a pseudoword comprehension and production situation, respectively. For this, we created a cue matrix C_{comb} based on a combined set of words, containing all aforementioned real words and pseudowords. In total, 8440 word forms are part of this set of words. A combined semantic matrix S_{comb} is created by attaching the estimated pseudoword semantic matrix to the real word semantic matrix, and reordering its rows to reflect the same order of words as found in C_{comb}.

Then, using the functions of the WpmWithLdl package (Baayen et al., 2019a), a comprehension model is trained and checked for accuracy. That is, taking form vectors as input for the prediction of semantic vectors as output, $\hat{S}_{comb} = C_{comb}F$ is solved. Comprehension is successfully modelled for a word *i* if its predicted semantic vector $\hat{s}_i$ is most highly correlated with its targeted semantic vector s_i. This is true for 74.41 % of cases (i.e. 6280 word forms) in our comprehension model. In total, 25.59 % of cases (i.e. 2160 word forms) are incorrectly predicted, with 1932 simple and 228 complex word forms. None of the incorrectly predicted word forms is a pseudoword.

Similarly, a production model is trained and checked for accuracy using functions of the aforementioned R package. Thus, semantic vectors are provided as input to predict form vectors as output, i.e. to solve $\hat{T}_{comb} = S_{comb}G$. Production is successfully modelled for a word *i* if its predicted triphones are those triphones present in its targeted cue vector in the correct sequence (possible sequences of triphones will be referred to below as 'paths') This is true for 97.3 % of cases (i.e. 8212 word forms) in our production model. In total, 2.7 % of cases (i.e. 228 word forms) are incorrectly predicted, with 100 simple and 128 complex word forms. None of the incorrectly predicted word forms is a pseudoword.

4.6. Measures

In order to explore the potential of different measures emerging from the network to predict phonetic duration, we extracted a whole range of measures, based on the measures introduced by the WpmWithLdl package (Baayen et al., 2019a) and by Chuang et al. (2020). As many of these measures correlate, we applied principle component analyses to get orthogonal measures that reflect interpretable dimensions of variation.

In the following, we first describe the semantic measures before we turn to the phonetic measures.

COR_MAX: To obtain this measure, correlations between a word's estimated semantic vector $\hat{s}$ and all other semantic vectors of S_{comb} are computed. Then, the highest of the correlation values is taken as COR_MAX. Higher values of this variable indicate the presence of a very close semantic neighbour, while lower values indicate a larger distance to a word's nearest semantic neighbour.

COR_TARGET: COR_TARGET describes the correlation between a word's predicted semantic vector $\hat{s}$ in $\hat{S}_{comb}$ and its targeted semantic vector s in the S matrix. Higher values indicate a good mapping from form to meaning, i.e. a higher semantic certainty.

L1NORM and L2NORM: The L1NORM is the sum of the absolute values of vector elements of a given word's predicted semantic vector $\hat{s}$, i.e. its city-block distance. The L2NORM is the square root of the sum of the squared values of a given word's predicted vector $\hat{s}$, i.e. its Euclidian distance. For both variables, higher values imply more strong links to many other lexomes. Thus, both measures may be interpreted as semantic activation diversity.

RANK: Given the correlation values of a word's predicted semantic vector $\hat{s}$ and its eight nearest neighbours' semantic vectors $s_{n1} \cdots s_{n8}$, ordinal values from 1 to 1+n are assigned, with a value of 1 given to the word with the highest correlation value. That is, if a word's predicted semantic vector $\hat{s}$ is most highly correlated with the targeted word's semantic vector s, its RANK will be 1. If a word's predicted semantic vector $\hat{s}$ is most highly correlated with any other but the targeted word's semantic vector, its RANK will be 1+n. Thus, a value of 1 indicates that a predicted semantic vector is closest to its targeted semantic vector, while values above 1 indicate a mismatch, i.e. a comprehension failure. This measure can be taken as an indicator of the goodness of comprehension.

DENSITY: Similar to RANK, the correlation values of a word's predicted semantic vector $\hat{s}$ and its eight nearest neighbours' semantic vectors $s_{n1} \cdots s_{n8}$ are taken into consideration. The mean of these eight correlation values describes DENSITY, with higher values indicating a denser semantic neighbourhood.

COR_AFFIX: This variable describes the correlation of a word's estimated semantic vector $\hat{s}$ with the vector of its affix s_{affix} in S. That is, only words containing affixes are ascribed a COR_AFFIX value. Higher values indicate a higher correlation of estimated word semantics and affix semantics, thus acting as a measure of semantic transparency.

CORRELATIONS: Given all candidate forms of a word, i.e. all word forms found as candidate production by the production model, and their estimated semantic vectors $\hat{s}_{candidate1} \cdots \hat{s}_{candidate1+n}$, correlation values between these estimated semantic vectors and a word's semantic vector s are computed. The highest of these correlation values is taken as value of CORRELATIONS. Similar to RANK for comprehension, this may be interpreted as a measure of goodness of production.

ALC: The Average Lexical Correlation, ALC, is the mean value of all correlation values of a pseudoword's estimated semantic vector with each of the real word semantic vectors. Higher ALC values indicate that a pseudoword's semantics are part of a denser semantic neighbourhood. Thus, ALC may be interpreted as a measure of semantic activation diversity for pseudowords.

EDNN: This variable describes the Euclidian Distance between a word's estimated semantic vector $\hat{s}$ and its Nearest semantic Neighbour. Thus, higher values indicate a larger distance to the nearest semantic neighbour. EDNN may be regarded as a measure of semantic neighbourhood density.

NNC: The Nearest Neighbour Correlation is computed by taking a pseudoword's estimated semantic vector and checking it for the highest correlation value against all real word semantic vectors. This highest correlation value is taken as NNC value. Thus, higher values indicate that a pseudoword is semantically close to a real word. Additionally, one can tell which real word a pseudoword's semantics are closest to. This measure may be interpreted as a measure of similarity between nonce and real words, indicating the co-activation of a real word when confronted with a pseudoword.

PATH_COUNTS: PATH_COUNTS describes the number of paths, i.e. possible sequences of triphones, detected for the production of a word by the production model. PATH_COUNTS may be interpreted as a measure of phonological activation diversity, as higher values indicate the existence of multiple candidates (and thus paths) in production.

PATH_SUM: PATH_SUM describes the summed support of paths for a predicted form. PATH_SUM may be interpreted as a measure of phonological certainty, with higher values indicating a higher certainty in the candidate form.

PATH_ENTROPIES: PATH_ENTROPIES contains the Shannon entropy values which are calculated over the path supports of the predicted form in $\hat{T}$. Thus, PATH_ENTROPIES

may be interpreted as a measure of phonological uncertainty, with higher values indicating a higher level of disorder, i.e. uncertainty.

LWLR: The length-weakest link ratio, LWLR, is computed by taking the number of path nodes divided by the value of the weakest link of that path. Higher values of LWLR indicate less support for a predicted form $\hat{t}$. Thus, LWLR may be interpreted as a measure of phonological uncertainty, with higher values indicating less certainty.

ALDC: The Average Levenshtein Distance of all Candidate productions, ALDC, is the mean of all Levenshtein distances of a word and its candidate forms. That is, for a word with only one candidate form, the Levenshtein distance between that word and its candidate form is its ALDC. For words with multiple candidates, the mean of the individual Levenshtein distances between candidates and targeted form constitutes the ALDC. Thus, higher values indicate that a word's candidate forms are very different from the intended pronunciation. ALDC may be interpreted as a measure of phonological neighbourhood density, i.e. large values indicate sparse neighbourhoods.

5. Analysis

The data set by Schmitz et al. (2020) contains non-morphemic, plural, or clitic word-final S as final segment of a pseudoword. As our LDL implementation does not include information on clitics, we only consider durational data on non-morphemic and plural S for the present study. A subset of 666 data points remains, with 303 observations with non-morphemic S and 363 observations with plural S. Due to some variable pronunciations requiring triphones not included in our LDL implementation, 13 data points had to be excluded, resulting in a final data set with non-morphemic and plural S durations of 653 data points, i.e. 300 entries on non-morphemic S and 353 entries on plural S. The data set and the following analysis can be found at https://osf.io/zy7ar/?view_only=ef43a5caf6444270a56074027d7d6482.

5.1. Covariates

Besides the aforementioned variables extracted and computed from the LDL implementation itself (see section 4.6), the following covariates, adopted from previous analyses of word-final S (e.g. Plag et al., 2017; Schmitz et al., 2020; Tomaschek et al., 2019), are included in the analysis. The main reason for this is to allow us to compare the performance of these predictors with the performance of LDL predictors. LDL measures often correlate with traditional measures (such as lexical frequencies, transitional probabilities, or neighborhood densities), but the traditional measures have no clear correlating mechanisms in learning or processing.

There are, however, also covariates that do not tap into lexical properties, but that control for other influences, such as speech rate, the speaker, gender, the order of stimuli in an experiment, etc. These will be referred to as 'non-lexical covariates' and they will also be included in our regression models.

AFFIX: This binary variable indicates whether a word contains an affix, i.e. whether the pertinent pseudoword is a singular or plural form. It takes the value `NM` for pseudowords without affix, and `PL` for pseudowords with affix.

SPEAKINGRATE: Analysing durational data, speech rate is a self-evident variable to consider. As speech rate is no inherent part of any LDL measure, we calculated speaking rate as the number of syllables in an utterance divided by the duration of the utterance (e.g. Schmitz et al., 2020; Tomaschek et al., 2019). This was done automatically using a script in Praat (Boersma & Weenink, 2019; de Jong & Wempe, 2008).

BASEDURLOG: Base duration was taken as a more local measure of speech rate (e.g. Plag et al., 2017, 2020b; Schmitz et al., 2020). Here, the term 'base' refers to the string of segments preceding the word-final S, for both non-morphemic and morphemic pseudowords. Base duration was then log-transformed to achieve a closer to normal distribution.

PAUSEBIN: To account for final-lengthening effects, stretches of silence between the offset of the word-final S and the onset of the following word were measured. Silence of 50 ms and above was considered as pause (Krivokapić, 2007; Lee & Oh, 1999). In order to make sure that closures of following plosives were not mistaken for pauses, their average closure duration (see Yao, 2007) was subtracted of the pertinent measured silence. Following the results by Schmitz et al. (2020), pause information was included as binary variable with the values PAUSE / NO PAUSE.

DISC: As some pseudowords were produced with multiple pronunciations, their transcription was incorporated as a categorical variable. This variable is called DISC after the DISC keyboard phonetic alphabet (Burnage, 1988).

BIPHONEPROBSUMBIN: The summed biphone probability for each pseudoword and its phonological variants is included as the binary variable BIPHONEPROBSUMBIN. It was calculated using the Phonotactic Probability Calculator (Vitevitch & Luce, 2004). The rationale for this variable is that more probable biphones should lead to shorter durations (e.g. Schmitz et al., 2020).

LIST & SLIDENUMBER: To account for priming effects, the list number (1-12) and the point of occurrence during the original experiment by Schmitz et al. (2020) are included.

PREC: To account for potential effects of the consonant preceding the word-final S (Umeda, 1977), it is included as PREC variable (similar to e.g. Tomaschek et al., 2019).

BIPHONEPROB: The probability of the final biphones /fs/, /ks/, /ps/ and /ts/ in monomorphemic words is included as covariate to account for potential effects of phonotactics (see Schmitz et al., 2020, for a detailed explanation).

FOLTYPE: As the segment following the word-final S is no part of the individual pseudoword, it is also not considered in LDL measures. Thus, the covariate FOLTYPE is introduced (similar to e.g. Tomaschek et al., 2019), coding the following segment by its segmental class (i.e. approximant APP for *listen*, fricative F for *find*, nasal N for *know*, plosive P for *cook*, and vowel V for *eat*), to account for potential effects of the following word (Klatt, 1976; Umeda, 1977).

SPEAKER, GENDER, AGE, LOCATION and MONOMULTILINGUAL: SPEAKER ID was included to account for general inter-speaker differences in production. GENDER, AGE, and LOCATION, i.e. the place in which the pertinent participant spent the bigger part of their life, were included as well. Additionally, participants who were early bilinguals were categorized as multilingual, while all other participants were categorized as monolingual in MONOMULTILINGUAL.

REAL: Some of the pseudowords in Schmitz et al.'s data set have an orthographically different, but phonologically identical real word counterpart. We introduced the variable REAL to control for this potential confound. This variable is TRUE for pseudowords with such a real word counterpart, and FALSE for those without. We considered the following real words as counterparts as given in Schmitz et al. (2020): *glits* corresponds to *glitz*, *glaiks* corresponds to *Gleicks*, *glifs* corresponds to *glyphs*, and *pleets* corresponds to *pleats*.

All of the following analyses make use of the following non-lexical covariates: BASEDURLOG, SPEAKINGRATE, SLIDENUMBER, and PAUSEBIN as variables concerning speech rate and continuity, PREC and FOLTYPE accounting for coarticulatory effects, LIST taking into consideration potential priming effects, MONOMULTILINGUAL, GENDER, LOCATION, AGE, and SPEAKER to account for speaker-individual differences, and REAL to include potential effects of real word counterparts.

5.2. Modelling strategy

We devised three kinds of model: First, a baseline model with the traditional predictor variables (plus the non-lexical covariates). Second, a model with LDL predictors that also includes AFFIX as a covariate (plus the non-lexical covariates). Third, a model that contains only the LDL predictors (plus the non-lexical covariates).

The three kinds of model will allow us to answer our research questions. Recall that our ultimate goal is to understand how systematic durational differences emerge between words of different, but homophonous morphological categories. Traditional lexical variables are predictive but cannot explain how morphology can make its way into durational differences. But these models can show that such differences exist by looking at the effect of the variable AFFIX. This is our baseline model. As an alternative we implement a model that uses LDL measures. If these measures are predictive, they offer an explanation of the morphologically-induced phonetic differences: they emerge as a by-

product of the association of form and meaning in the mental lexicon, and this association is the outcome of discriminative learning. By having a model that also includes AFFIX as an additional predictor, we can see whether the LDL measures completely capture the morphological effect, or whether there is a residue of morphological information that is predictive of duration but is still not captured by the LDL measures.

5.3. Model A: Traditional measures

This model is meant to resemble those in previous studies on word-final S duration (e.g. Plag et al., 2017; Schmitz et al., 2020). Thus, we make use of similar variables: AFFIX, BIPHONEPROBSUMBIN, and BIPHONEPROB, as well as those control variables included in all analyses of this paper. None of these covariates showed high correlation coefficients. Hence, no cautionary measures regarding collinearity were taken before an initial full model was constructed.

Models were fitted using linear mixed-effects regression in R (R Core Team, 2020) using RStudio (RStudio Team, 2020) and as implemented by lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), and LMERConvenienceFunctions (Tremblay & Ransijn, 2020) to analyse the data on non-morphemic and plural S duration. The dependent variable, duration of S, was log-transformed following standard procedures to reduce the potentially harmful effect of skewed distributions in linear regression models (e.g. Winter, 2019). The name of this variable is SDURLOG.

Following the standard backward step-wise selection process for model selection (e.g. Baayen, 2008), a first model containing the aforementioned variables as well the non-lexical covariates is created. GENDER, LOCATION, MONOMULTILINGUAL, SPEAKER, DISC, AGE, SLIDENUMBER, and LIST are included as random intercepts. Additionally, AFFIX is specified as random slope for SPEAKER. This full model was then continuously reduced through step-wise exclusion of non-significant variables. That is, a variable was considered as significant if it passed all of three tests. First, its F-value in the pertinent model had to yield a value below -2 or above 2. Second, the AIC value, i.e. the Akaike information criterion value, of the model including the variable had to be lower than the AIC value of a comparable model without the pertinent variable. Third, the results of log-likelihood tests comparing the model with to a model without the pertinent variable had to yield a p-value below the 0.05 threshold, thus indicating a significant improvement of model fit. This process was verified using the *step* function of R, which resulted in an identical model.

Then, variance inflation factors were checked. The covariates BIPHONEPROB and PREC showed high VIF values (i.e. 46.53 and 46.88, respectively), indicating potential overfitting of the model (e.g. Fox & Weisberg, 2019; Zuur et al,. 2010). Consequently, PREC was removed from the model as it showed the highest VIF value, following the procedure described by Zuur et al. (2010). Re-fitting the model without PREC and re-checking the new variance inflation factor values revealed only non-problematic values.

Finally, the resulting model's residuals were trimmed (e.g. Baayen & Milin, 2010). Data points with residuals larger than 2.5 standard deviations were removed, ensuring a satisfactory distribution of residuals. This procedure lead to a loss of 3 data points, i.e. 0.46 % of all data points. An overview of all variables used in the initial model is given in Supplementary Table 2.

5.4. Model B: LDL measures with affix specification

This model makes use of all LDL measures as well as of the AFFIX variable. Additionally, the non-lexical covariates are included. One issue to address when considering a model with such a multitude of variables is collinearity (e.g. Baayen, 2008; Tomaschek et al., 2018). To avoid collinearity related problems later on, all variables were tested for correlation using the languageR package (Baayen & Shafaei-Bajestan, 2019). This correlation check resulted in sixteen correlation coefficients indicating a high degree of correlation, for which we assume the threshold to be $rho \geq 0.5$. The pairs of correlated covariates as well as their correlation coefficients are given in Table 2.

Table 2. Correlated variables and their correlation coefficients.

variables		rho	variables		rho
COR_MAX	COR_TARGET	1.00	COR_TARGET	EDNN	-0.86
COR_MAX	CORRELATIONS	1.00	CORRELATIONS	EDNN	-0.86
COR_TARGET	CORRELATIONS	1.00	DENSITY	ALC	0.75
L1NORM	L2NORM	0.94	COR_AFFIX	ALC	0.72
DENSITY	COR_AFFIX	0.84	COR_AFFIX	NNC	0.84
PATH_COUNTS	PATH_ENTROPIES	0.95	PATH_COUNTS	ALDC	0.89
PATH_SUM	LWLR	-0.76	PATH_ENTROPIES	ALDC	0.90
COR_MAX	EDNN	-0.86	AFFIX	NNC	0.90

Due to the high number of correlated variables, we opted for a principal component analysis (PCA; e.g. Baayen, 2008; Tomaschek et al., 2018; Venables & Ripley,

2002) to address collinearity issues. In a PCA, the dimensionality of the data is reduced by transforming the included variables into principal components. These transformations result in linear combinations of the predictors that are orthogonal to each other. Thus, the resulting principal components are not correlated.

The PCA was carried out using the *PCAmix* function of the PCAmixdata package in R (Chavent et al., 2017), allowing the simultaneous integration of continuous and discrete variables. All variables (with the exception of COR_AFFIX as this variable only contains data on plural pseudowords) given in Table 2 were included in the computation of the principal component analysis, which yields thirteen principal components. The next step of the PCA is to determine how many of these principal components are meaningful and thus should be retained for further use. For this decision, we followed several rules of thumb (e.g. Baayen, 2008; O'Rourke et al., 2005). First, any component that displays an Eigenvalue greater than 1 accounts for a greater amount of variance than had been contributed by one variable. Such a component is therefore potentially meaningful. Second, one should retain enough components so that the cumulative percent of variance explained is equal to some minimal value. Following other implementations of principal component analyses, we aimed at a value of 80 % (e.g. O'Rourke et al., 2005). Third, only interpretable components are to be retained. That is, each component is made up out of loadings, i.e. parts of the variables included in the PCA's computation represented by correlation coefficient values. If none of these variables is strongly represented in a component, the interpretability of that component is extremely low, rendering the component of small interest for further analyses. Following these three criteria, we find that the first five of the principal components show an Eigenvalue of one or higher. Also, the first four components account for 77 % of variance, while including the fifth component raises the amount of variance explained to 88 %. However, considering the third criterion, the fifth component's loadings show only weak representations of variables (i.e. the highest value is $rho = 0.25$) which makes the interpretation of this component very difficult. We therefore retain components 1 to 4 for further analysis, all of which show an Eigenvalue greater than 1, account for almost eighty percent of variance, and contain strong representations of variables in their loadings.

But what do these principal components mean? The highest loadings of the principal components, i.e. the correlation of the original variables to the pertinent component, are given in Table 3.

Table 3. Loadings of original predictor variables in the four retained principal components of the first principal component analysis.

	Component1	Component2	Component3	Component4
cor_max	0.638			
cor_target	0.638			
correlations	0.638			
EDNN	0.450			
path_counts		0.448		
path_entropies		0.426		
l1norm			0.596	
l2norm			0.600	
ALC				0.523
Affix				0.516

Component1 is most strongly correlated with cor_max, cor_target, correlations, and EDNN. For cor_max, higher values indicate a close semantic neighbour, while high values for cor_target indicate a good mapping from form to meaning. Correlations can be interpreted as a measure of goodness of production by means of fitting semantics, with higher values representing a higher level of goodness. For EDNN, higher values indicate a larger distance to the nearest semantic real word neighbour. As Component1 is negatively correlated with EDNN (see Supplementary Figure 1), this effect goes together with the aforementioned effect of cor_max. Thus, Component1 can be seen as a measure of semantic certainty.

Component2 is most strongly correlated with path_counts and path_entropies. For path_counts, higher values indicate the existence of multiple candidates (and thus paths) in production, functioning as a measurement of phonological activation diversity. Values of path_entropies relate to the level of uncertainty concerning the path supports of the predicted candidate form, with higher values indicating a higher level of uncertainty. Component2 may thus be taken as a measure of phonological certainty.

Component3 is most strongly correlated with l1norm and l2norm. Both variables imply more strong links to many other lexomes with higher variables. Thus, Component3 may be interpreted as a measure of semantic activation diversity.

A closer inspection of the distribution of Component4 showed a near binary distribution. We therefore recoded it as a binary variable called Component4Bin, taking

as a threshold for the division of low/high the midpoint of the valley in-between the two peaks of the binary distribution at a value of COMPONENT4 $= 0.5$. As COMPONENT4 and COMPONENT4BIN show a high degree of correlation ($rho = -0.86$), simple models were fitted and compared to decide which variable to keep for further analyses. It turns out that COMPONENT4BIN has higher explanatory power ($p < 0.001$). We therefore discarded COMPONENT4 in favour of COMPONENT4BIN.

COMPONENT4BIN is special in that is not only strongly correlated with a continuous variable, i.e. ALC, but also with a categorical variable, AFFIX. For ALC, higher values indicate that a pseudoword's semantics are part of a denser real word semantic neighbourhood. This may be interpreted as a form of semantic activation diversity measurement. As for AFFIX, COMPONENT4BIN is positively correlated with the presence of plural S, while it is negatively correlated with the presence of non-morphemic S data points (see Supplementary Figure 2). We will come back to the interpretation of this correlation in section 6.2.

Following the modelling procedure introduced in section 5.3, a first model containing all remaining variables is created. That is, COMPONENT1, COMPONENT2, COMPONENT3, COMPONENT4BIN, BASEDURLOG, SPEAKINGRATE, PAUSEBIN, FOLTYPE, PREC, and REAL were included as fixed effects. Note that RANK, even though not part of the principal component analysis and thus a potential predictor variable, was not included as its value is the same for all pseudowords, rendering its explanatory power equal to zero. GENDER, LOCATION, MONOMULTILINGUAL, SPEAKER, DISC, AGE, SLIDENUMBER, and LIST are included as random intercepts

Then, variance inflation factors (VIFs) were computed for the model resulting from the step-wise exclusion of non-significant variables. Predictors showing variance inflation factor values equal or greater than 3 are to be excluded due to the high risk of introducing multicollinearity and thus overfitting of the model (e.g. Zuur et al., 2010). For the present model, all variance inflation factor values are below 3.

Finally, the resulting model needed trimming of its residuals (e.g. Baayen & Milin, 2010). Data points with residuals larger than 2.5 standard deviations were removed to ensure a more satisfactory residual distribution. This procedure resulted in a loss of three data points (0.46 %). An overview of all variables used in the initial model and their distribution is given in Supplementary Table 2.

5.5. Model C: LDL measures without affix specification

This model uses all LDL measures without the AFFIX covariate. As in the previous model, there was a high number of highly correlated variables (see Table 2 with the exception of the correlation of AFFIX and NNC, as AFFIX is not included in this analysis). We therefore again computed a principal component analysis, following the procedure outlined in section 5.4. Following the three criteria, we find that four principal components are to be retained: The first four components account for 82 % of variance, show an Eigenvalue of one or higher, and show interpretable loadings. We call these components COMPONENT.WOA.1 to COMPONENT.WOA.4, i.e. component WithOut Affix. The loadings of the four components are given in Table 4.

Table 4. Loadings of original predictor variables in the four retained principal components of the second principal component analysis.

	COMPONENT. WOA.1	COMPONENT. WOA.2	COMPONENT. WOA.3	COMPONENT. WOA.4
COR_MAX	0.433			
COR_TARGET	0.433			
CORRELATIONS	0.433			
EDNN	-0.373			
PATH_COUNTS		-0.373		
PATH_ENTROPIES		-0.362		
L1NORM			0.523	
L2NORM			0.523	
DENSITY				0.580
ALC				0.365
ALDC				-0.358

COMPONENT.WOA.1 is similar to COMPONENT1, COMPONENT.WOA.2 is similar to COMPONENT2, and COMPONENT.WOA.3 is similar TO COMPONENT.3 of the PCA introduced in section 5.4. That is, COMPONENT.WOA.1 can be seen as a measure of semantic certainty, while COMPONENT.WOA.2 can be taken as a measurement for phonological certainty. COMPONENT.WOA.3 may be interpreted as a measure of semantic activation diversity.

COMPONENT.WOA.4 is most strongly correlated with DENSITY, ALC, and ALDC. For DENSITY and ALC, higher values indicate a denser semantic neighbourhood. Phonological neighbourhood density is described by ALDC, for which higher values

indicate sparse neighbourhoods. Taking into account the sign of the three correlation coefficients, COMPONENT.WOA.4 can be seen as a measure of neighbourhood densities, with sparse semantic and phonological neighbourhoods at one end, and dense semantic and phonological neighbourhoods at the other end.

Linear mixed-effects regression models were fit according to the procedure given in section 5.3. That is, an initial full model was fit with the following variables: COMPONENT.WOA.1, COMPONENT.WOA.2, COMPONENT.WOA.3, COMPONENT.WOA.4, BASEDURLOG, SPEAKINGRATE, PAUSEBIN, FOLTYPE, PREC and REAL. The random effects specification is identical to the one given in section 5.4.

This full model was then continuously reduced through step-wise exclusion of non-significant variables, following the aforementioned criteria. Then, variance inflation factors were computed, resulting only in non-problematic values (e.g. Zuur et al., 2010). Finally, the resulting model needed trimming of its residuals (e.g. Baayen & Milin, 2010). That is, data points with residuals larger than 2.5 standard deviations were removed, ensuring a more satisfactory residual distribution. This procedure lead to a loss of 8 data points, i.e. 1.2 % of all data points. An overview of all variables used in the initial model and their distribution is given in Supplementary Table 2.

6. Results

6.1. Model A: Traditional measures

The final model of traditional measures includes main effects of the following variables: type of S (AFFIX), speaking rate (SPEAKINGRATE), log-transformed base duration (BASEDURLOG), pause (PAUSEBIN), the summed biphone probability (BIPHONEPROBSUMBIN), and following segmental type (FOLTYPE). As for random effects, random intercepts for SPEAKER and random slopes for AFFIX are included. The p-values of the analysis of variance of the final model are given in Table 5.

Table 5. p-values of fixed effects in the final 'traditional' model, fitted to the log-transformed durations of S.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr (>F)
AFFIX	0.781	0.781	1	37.48	15.092	0.000
SPEAKINGRATE	0.216	0.216	1	606.02	4.171	0.042
BASEDURLOG	6.205	6.205	1	577.86	119.852	0.000
PAUSEBIN	5.565	5.565	1	637.13	107.488	0.000
BIPHONEPROBSUMBIN	0.614	0.614	1	595.71	11.867	0.001
FOLTYPE	2.066	0.5165	4	606.20	9.975	0.000

The marginal R-squared value of the model is 0.43, i.e. fixed effects explain 43 % of variation in the data. Taking random effects into account as well, the conditional R-squared value is 0.62. That is, the model explains 62 % of data variation in total. (see Nakagawa et al., 2017, for details on marginal and conditional R-squared computation). Both R-squared values were computed using the MuMIn package (Barton, 2020). The R-squared values are similar to the values found by Schmitz et al. (2020) on their complete data set.

The coefficients of the final model and their p-values are given in Table 6. The reference levels for the categorical predictors are: for AFFIX it is `NM`, for PAUSEBIN it is `no-pause`, for BIPHONEPROBSUMBIN it is `high`, and for FOLTYPE it is `APP`.

Table 6. Fixed-effect coefficients and p-values as computed by the final 'traditional' model (mixed-effects model fitted to the log-transformed duration of S).

	Estimate	Std. Error	df	t-value	Pre (> |t|)
(Intercept)	-1.197	0.083	408.89	-14.388	0.000
AFFIXPL	-0.090	0.023	37.48	-3.885	0.000
SPEAKINGRATE	-0.026	0.013	606.02	-2.042	0.042
BASEDURLOG	0.631	0.058	577.86	10.948	0.000
PAUSEBINPAUSE	0.236	0.023	637.14	10.368	0.000
BIPHONEPROBSUMBINlow	-0.074	0.022	595.71	-3.445	0.001
FOLTYPEF	-0.004	0.073	611.44	-0.052	0.959

FOLTYPEN	-0.004	0.028	601.78	-0.119	0.906
FOLTYPEP	-0.027	0.025	600.47	-1.087	0.278
FOLTYPEV	-0.140	0.025	611.18	-5.647	0.000

The predictor strength of individual covariates was checked by taking the final model as template. For each predictor variable, a model was fitted lacking the particular variable. This resulted in seven models, each lacking a different predictor. Then, R-squared values were computed for these models and finally compared. The variable leading to the highest decrease in R-squared value as compared to the final model is thus the variable showing the highest predictor strength. The results of this comparison are reflected in the hierarchy given in (1). The decrease in R-squared is greatest when removing BASEDURLOG, followed by PAUSEBIN, and so forth. The resulting order is identical to the one found by Schmitz et al. (2020) for the complete data set.

(1) BASEDURLOG >> PAUSEBIN >> AFFIX >> FOLTYPE >> SPEAKINGRATE >> BIPHONEPROBSUMBIN

6.2. Model B: LDL measures with AFFIX specification

In the final model including LDL measures as well as the AFFIX covariate as parts of the individual components resulting from the principal component analysis, and fitted according to the procedure described in section 5.4, we find main effects of the third principal component (COMPONENT3), the fourth principal component in its binary version (COMPONENT4BIN), base duration (BASEDURLOG), speaking rate (SPEAKINGRATE), following pause (PAUSEBIN), following segmental type (FOLTYPE), and preceding consonant (PREC). Regarding random effects, only a SPEAKER-specific random intercept turns out to significantly improve model fit. The p-values of the analysis of variance of the final model are given in Table 7.

Table 7. p-values of fixed effects in the final 'LDL measures and Affix' model, fitted to the log-transformed durations of S.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr (>F)
COMPONENT3	0.291	0.291	1	613.29	5.519	0.019
COMPONENT4BIN	1.502	1.502	1	625.60	28.516	0.000
BASEDURLOG	5.481	5.481	1	633.62	104.053	0.000
SPEAKINGRATE	0.224	0.224	1	650.00	4.247	0.040
PAUSEBIN	5.419	5.419	1	637.40	102.875	0.000
FOLTYPE	2.053	0.513	4	614.43	9.745	0.000
PREC	0.608	0.203	3	612.76	3.847	0.010

The marginal R-squared value of the final model is 0.42, thus fixed effects explain 42 % of the variation in our data. The conditional R-squared value of the final model is 0.61, that is fixed and random effects taken together explain 61 % of variation.

The estimates of the final model and their p-values are given in Table 8. The reference levels for the categorical predictors are: for COMPONENT4BIN it is `high`, for PAUSEBIN it is `no-pause`, for FOLTYPE it is `APP`, and for PREC it is `f`.

Table 8. Fixed-effect coefficients and p-values as computed by the final 'LDL measures and Affix' model (mixed-effects model fitted to the log-transformed duration of S).

	Estimate	Std. Error	df	t-value	Pre (> |t|)
(Intercept)	-1.360	0.085	515.01	-15.933	0.000
COMPONENT3	-0.016	0.007	613.29	-2.349	0.019
COMPONENT4BINlow	0.105	0.020	625.60	5.340	0.000
BASEDURLOG	0.604	0.059	633.62	10.201	0.000
SPEAKINGRATE	-0.026	0.013	650.00	-2.061	0.040
PAUSEBINpause	0.232	0.023	637.40	10.416	0.000
FOLTYPEF	0.005	0.074	616.85	0.068	0.946
FOLTYPEN	0.002	0.028	611.28	0.082	0.934
FOLTYPEP	-0.016	0.025	612.45	-0.656	0.512
FOLTYPEV	-0.136	0.025	617.62	-5.404	0.000
PRECk	-0.027	0.027	612.31	-1.019	0.308
PRECp	-0.029	0.026	612.68	-1.093	0.275
PRECt	-0.088	0.027	612.50	-3.257	0.001

Similar to section 6.1., the predictor strength of individual covariates was checked by taking the final model as template. For each predictor variable, a model was fitted lacking the pertinent variable. This resulted in seven models, each missing a different covariate. Then, marginal R-squared values were computed and compared. The model showing the lowest of these values in turn missed the covariate with the highest predictor strength. The result of this procedure is reflected in the hierarchy in (2). The decrease in R-squared is greatest when removing BASEDURLOG, followed by PAUSEBIN, and so forth. In sum, variables containing measures obtained by our LDL analysis appear to be meaningful predictors of S duration.

(2) BASEDURLOG >> PAUSEBIN >> COMPONENT4BIN >> FOLTYPE >> SPEAKINGRATE >> COMPONENT3 >> PREC

Figure 3 shows the effect on S duration of the variables included in the model. The estimated values of the dependent variable SDURLOG, i.e. S duration, and BASEDURLOG, i.e. base duration, are back-transformed into seconds. Longer bases come with longer S

durations (panel A), and faster speaking rates come with shorter S durations (panel B). As for COMPONENT3, higher values lead to shorter S durations, while lower values lead to longer S durations (panel C).

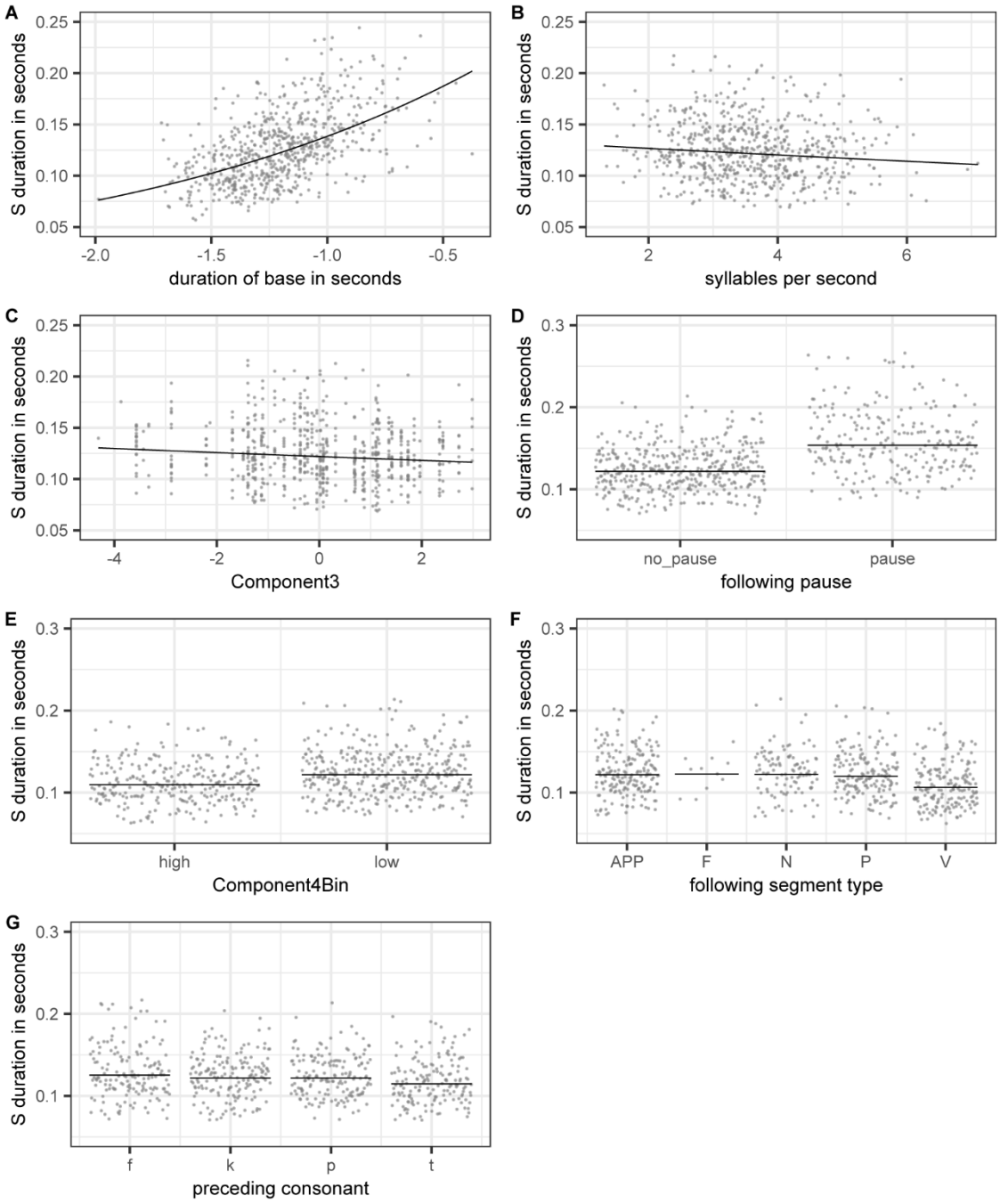


Figure 3. Partial effects of the numerical and categorical variables included in the final 'LDL measures and AFFIX' model, fitted to the log-transformed values of duration of S.

S duration is longer if the S is followed by a pause (panel D), which is most likely a case of phrase-final lengthening (e.g. Cooper & Danly, 1981). For COMPONENT4BIN,

higher values lead to longer S durations (panel E). There is also an effect of the following segment type, with S being shorter when followed by a vowel (panel F). This difference is significant for all consonant types being compared against vowels with the exception of fricatives. However, as there is only a small number of fricative cases in our data, this non-significant difference is potentially not meaningful. Lastly, there is an effect of preceding consonant on S duration (panel G). S duration is significantly longer if preceded by a voiceless labiodental fricative /f/ as compared to cases where S is preceded by a voiceless alveolar stop /t/. All other comparisons are non-significant.

Let us turn to the variables of interest, i.e. those containing LDL measures. COMPONENT3 acts as a general measure of semantic activation diversity. Higher diversity leads to shorter durations. Recall from section that COMPONENT4BIN relates to semantic activation diversity and to the presence or absence of the plural suffix. Higher values of COMPONENT4BIN indicate that a pseudoword's semantics are part of a denser real word semantic neighbourhood, which may be interpreted as a form of semantic activation diversity measurement. For the durations as shown in panel E of Figure 3 this means that high semantic activation diversity (equals high values of COMPONENT4BIN) goes together with shorter durations. High values of COMPONENT4BIN are positively correlated with the presence of plural S, while it is negatively correlated with the presence of non-morphemic S data points. The patterning of ALC and AFFIX within COMPONENT4BIN is internally consistent. The presence of plural makes words semantically more similar to each other as they share this meaning component. Hence it is to be expected that plural words live in a space of greater semantic activation diversity. COMPONENT4BIN is not only a measure of semantic activation diversity, but also indicates that plural pseudowords show a tendency of having a higher degree of semantic activation diversity as compared to monomorphemic pseudowords in general.

6.3. Model C: LDL measures without AFFIX specification

The final model of LDL measures without the AFFIX covariate is fitted with main effects of the following variables: the fourth principal component (COMPONENT.WOA.4), log-transformed base duration (BASEDURLOG), speaking rate (SPEAKINGRATE), pause (PAUSEBIN), following segmental type (FOLTYPE), and preceding consonant (PREC). The SPEAKER variable is included as random intercept. The p-values of the analysis of variance of the final model are given in Table 9.

Table 9. p-values of fixed effects in the final 'LDL measures' model, fitted to the log-transformed durations of S.

	Sum Sq	Mean Sq	NumDF	DenDF	F.value	Pr (>F)
COMPONENT.WOA.4	0.308	0.308	1	611.64	5.945	0.015
BASEDURLOG	5.351	5.351	1	634.09	103.259	0.000
SPEAKINGRATE	0.342	0.342	1	644.88	6.603	0.010
PAUSEBIN	5.106	5.106	1	631.54	98.534	0.000
FOLTYPE	2.446	0.612	4	608.68	11.801	0.000
PREC	0.543	0.181	3	607.08	3.495	0.015

With a marginal R-squared value of 0.40, the fixed effects of this model explain 40 % of variation within the data. The conditional R-squared value of the model is 0.61, that is the complete model accounts for 61 % of variation.

The coefficients of the final model and their p-values are given in Table 10. The reference levels for the categorical covariates are: for PAUSEBIN it is `no-pause`; for FOLTYPE it is `APP`, and for PREC it is `f`.

Table 10. Fixed-effect coefficients and p-values as computed by the final 'LDL measures' model (mixed-effects model fitted to the log-transformed duration of S).

	Estimate	Std. Error	df	t-value	Pre (> |t|)
(Intercept)	-1.301	0.084	497.52	-15.486	0.000
COMPONENT.WOA.4	-0.019	0.008	611.64	-2.436	0.015
baseDurLog	-0.592	0.058	634.09	10.162	0.000
SPEAKINGRATE	-0.032	0.013	644.88	-2.570	0.010
PAUSEBINPAUSE	0.225	0.023	631.54	9.926	0.000
FOLTYPEF	0.007	0.073	611.11	0.100	0.920
FOLTYPEN	0.008	0.028	605.56	0.291	0.771
FOLTYPEP	-0.013	0.025	606.93	-0.523	0.601
FOLTYPEV	-0.145	0.025	611.55	-5.840	0.000
PRECk	0.010	0.026	606.74	0.374	0.708
PRECp	-0.042	0.027	607.65	-1.529	0.127
PRECt	-0.064	0.026	607.51	-2.422	0.016

As for both other final models, the predictor strength of the individual predictors was checked. Models with one of the predictor variables were constructed based on the complete final model. Then, marginal R-squared values were computed for each of these six models. A comparison of R-squared values then revealed the hierarchy of predictor strength given in (3). That is, the decrease in R-squared is greatest when removing BASEDURLOG, followed by PAUSEBIN, and so forth.

(3) BASEDURLOG >> PAUSEBIN >> FOLTYPE >> SPEAKINGRATE >> COMPONENT.WOA.4 >> PREC

Base duration and speaking rate show identical effects as compared to the model fitted in section 6.2., i.e. longer base durations come with longer S durations, while higher speaking rates lead to shorter S durations. As for categorical variables, pauses again come with longer S durations, and S is shorter if followed by a vowel. There is also an effect of the preceding consonant, with S being shorter when followed by a voiceless alveolar stop /t/ as compared to a voiceless velar stop /k/, while all other comparisons do not yield significant differences. These results are generally in line with those by the analysis in the previous section.

Taking a closer look at the variable of interest, i.e. the remaining variable containing LDL measures, we find that for COMPONENT.WOA.4 higher values lead to shorter S durations. This effect is illustrated in Figure 4. COMPONENT.WOA.4 can be taken as a general measure of phonological and semantic neighbourhood density, with higher densities leading to shorter S durations.

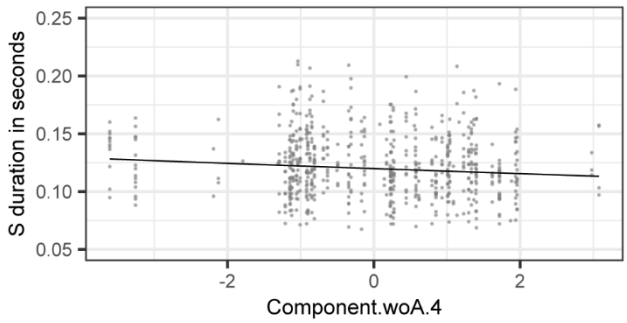


Figure 4. Partial effects of COMPONENT.WOA.4 in the final 'LDL measures without AFFIX' model, fitted to the log-transformed values of duration of S.

7. Discussion

7.1. The present results

Previous studies (Plag et al., 2020, 2017; Schmitz et al., 2020; Seyfarth et al., 2017; Tomaschek et al., 2019; Zimmermann, 2016) reported that there are significant differences in the acoustic duration between different types of word-final S in English. Such durational differences challenge established feed-forward theories of morphology-phonology interaction (e.g. Chomsky & Halle, 1968; Kiparsky, 1982) as well as theories of psycholinguistics (e.g. Levelt et al., 1999; Roelofs & Ferreira, 2019; Turk & Shattuck-

Hufnagel, 2020). The present study investigated whether measures derived on the basis of a discriminative learning theory are predictive of S durations in nonce words. In particular, we implemented LDL networks that model the production of a word based on its relation to the rest of the lexicon.

We explored the predictive possibilities of LDL measures by fitting three different models: a) a model based on the traditional predictors as used in previous studies (Plag et al., 2017; Schmitz et al., 2020; Tomaschek et al., 2019); b) a model with LDL measures and a variable AFFIX specifying the presence or absence of an affix; and c) a model with LDL measures but without AFFIX specification. Both models including LDL measures show that such measures are predictive of S durations. This result is the most important of our study. While traditional variables such as lexical frequencies, bigram frequencies, transitional probabilities or neighborhood densities measure important lexical properties, it is unclear why they would manifest themselves in a particular morphological effect in speech production. In LDL such effects can emerge through the mapping of form and meaning in a clearly defined process of discriminative learning.

All regression models showed a similar hierarchy of predictor strength for the variables included in the models. For the traditional model A, AFFIX is the third strongest predictor of S duration and for model B this spot is taken by COMPONENT4BIN, while there is no comparable variable included in model C. Comparing the variance explained by the fixed effects of the different models, we find that the traditional model accounts for most variation, i.e. 43 %, while the LDL model with AFFIX specification accounts for 42 %, and the LDL model without AFFIX specification accounts for 40 % of variation. Thus, in terms of marginal R-squared values, all three models are close to each other. To check whether these differences in marginal R-squared values are of significance, the three models were refitted to the untrimmed data set and then compared with an analysis of variance. The results suggest that there is no significant difference between the traditional model and the LDL model with AFFIX specification. However, the LDL model with AFFIX specification shows a significantly worse fit ($p<0.001$). This seems to indicate that the LDL measures do not capture the full amount of the variance that is captured by the variable AFFIX.

7.2. Comparison of results to other studies

The LDL measures included in our final models concern the semantic activation diversity and semantic neighbourhood density of the pertinent pseudoword (COMPONENT3 and COMPONENT4BIN of the LDL network with AFFIX specification

model), or present a combined measure of semantic and phonological neighbourhood density (COMPONENT.WOA.4 of the LDL without AFFIX specification model).

Higher degrees of semantic activation diversity come with shorter S durations. This effect is similar to the one which was reported by Tucker et al., (2019) in a study on stem vowels and Tomaschek et al. (2019) in their NDL study on S duration. A higher degree of activation diversity makes it "more difficult to discriminate the targeted outcome from its competitors" (Tomaschek et al., 2019:27), thus processing is slowed down (Arnold et al., 2017). As for production, a prolongation of the acoustic signal is dysfunctional if the prolongation maintains or increases the discrimination problem instead of contributing to resolving it (Tomaschek et al., 2019).

Neighbourhood density measures are also predictive for S durations. For semantic neighbourhood density, pseudowords which land in denser semantic neighbourhoods of real words show shorter S durations. This finding is in line with results on pseudoword durations modelled by LDL measures, in which pseudowords in more sparse semantic neighbourhoods show longer durations (Chuang et al., 2020). Similarly, phonological neighbourhood density is found to be predictive of S durations. Analogous to results by Chuang et al. (2020) for pseudoword word durations, pseudowords in denser phonological neighbourhood show shorter S durations. Such facilitatory effects of phonological neighbourhood density on articulation were found before (e.g. Gahl et al., 2012). In sum, a higher degree of semantic activation diversity, and denser semantic and phonological neighbourhoods lead to shorter S durations.

7.3. Directions for future research and conclusion

The results of the present study may bring up further questions. First, are the predictive measures found for word-final S duration in pseudowords also predictive for word-final S duration in real words? Tomaschek at al's (2017) NDL implementation suggests that it is, but LDL networks still need to be implemented. It would be especially interesting to model those data sets that have yielded seemingly contradictory effects. Second, taking into account that the specification of AFFIX in the modelling process leads to a significantly better model fit, one may ask what the underlying reasons for this significant effect are. This then automatically leads to another question: Is it possible to catch the effect of the AFFIX specification in terms of (new) LDL measures?

To summarize, this paper was the first to investigate durational differences between different types of word-final S (non-morphemic vs. plural S), in pseudowords by means of an LDL implementation, measures, and resulting statistical analyses. The

findings yielded important evidence on the question of how such durational difference come to be, i.e. they can be predicted based on their pseudoword's relations to the lexicon. We demonstrated that durational differences emerge from the pseudoword's resonance with the lexicon by way of differing degrees of semantic activation diversity and neighbourhood densities. These manifestations of the relations to other words in the lexicon in turn are the result of discriminative learning.

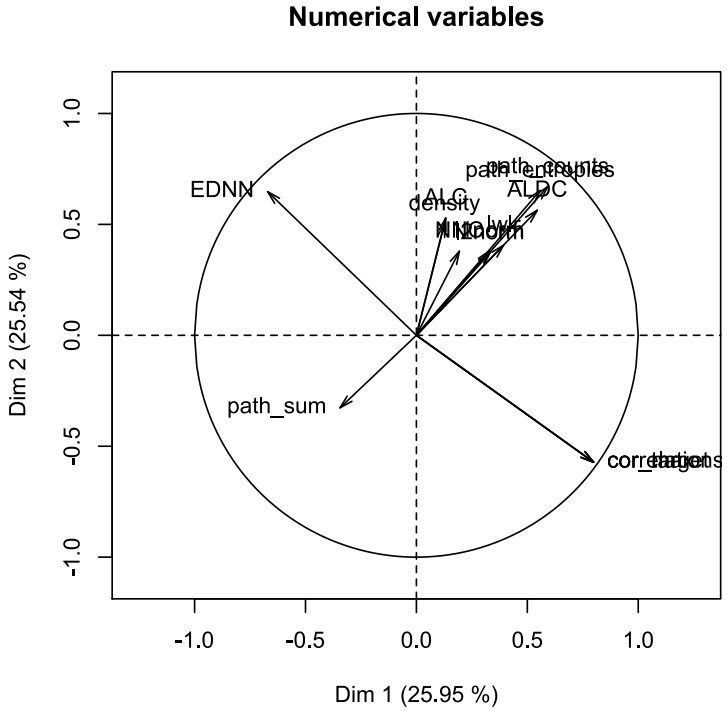
Appendix A

Overview of all pseudowords and their transcriptions used in the current LDL implementation.

Pseudoword		Transcription in DISC	Pseudoword		Transcription in DISC
blou-	fs	blufs	glai-	fs	gl1fs
	ks	bl{ks; bluks; blVks		ks	gl1ks; gl{ks
	ps	blups		ps	gl1ps; gl{ps
	ts	bl6ts; bluts		ts	gl1ts; gl{ts; gl2ts
cloo-fs; -ks; -ps; -ts		klufs; kluks; klups; kluts	plee-fs; -ks; -ps; -ts		plifs; pliks; plips; plits
gli-fs; -ks; -ps; -ts		glIfs; glIks; glIps; glIts glifs; gliks; glips; glits	pru-fs; -ks; -ps; -ts		prVfs; prVks; prVps; prVts; prufs; pruks; prups; pruts;

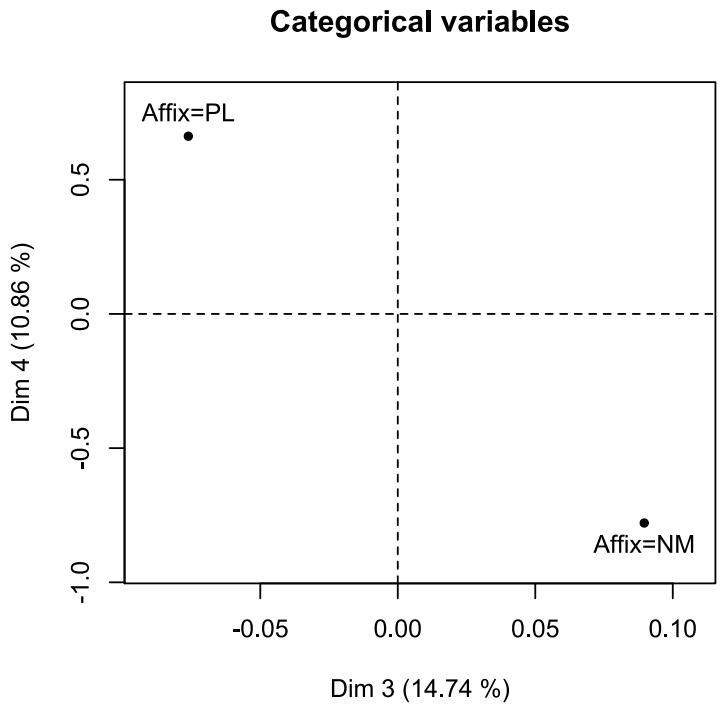
Appendix B

Loadings of the first and second principal component of the PCA described in section 5.4. The sign for EDNN in component 1 (‘Dim1’) is negative.



Appendix C

Categorical loading of the fourth principal component of the PCA described in section 5.4.



Appendix D

Summary of the dependent variable, numerical variables, and dependent variables used in the modelling processes.

Dependent variable	Mean	St. Dev.	Min	Max
sDurLog	-2.116	0.388	-3.361	-1.221
Numerical variables	**Mean**	**St. Dev.**	**Min**	**Max**
speakingRate	3,566	0,927	1,310	7,100
baseDurLog	-1,203	0,232	-1,987	-0,375
biphoneProb	0,001	0,002	0,000	0,004
age	28,470	9,323	19,000	58,000
Component1	0,000	1,975	-17,748	2,509
Component2	0.000	1,959	-2,832	11,989
Component3	0.000	1,488	-4,312	2,983
Component.woA.1	0,000	1,973	-18,860	2,178
Component.woA.2	0,000	1,957	-10,011	2,894
Component.woA.3	0,000	1,487	-4,175	2,928
Component.woA.4	0,000	1,269	-3,608	3,076
Categorical variables	**Levels**				
AFFIX	NM: 300	PL: 353			
PAUSEBIN	no: 412	yes: 241			
DISC	38				
BIPHONEPROBSUMBIN	high: 161	low: 492			
LIST	12				
SLIDENUMBER	48				
PREC	f: 156	k: 169	p: 164	t: 164	
FOLTYPE	APP: 190	F: 11	N: 106	P: 165	V: 181
SPEAKER	40				
GENDER	2				
LOCATION	London: 392		elsewhere: 261		
MONOMULTILINGUAL	monolingual: 532		bilingual: 121		
REAL	FALSE: 542	TRUE: 111			
COMPONENT4BIN	high:278	low: 375			

References

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. University of Philadelphia.

Baayen, R. H., & Shafaei-Bajestan, E. (2019). *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics* (1.5.0). https://cran.r-project.org/package=languageR

Baayen, R. H., Chuang, Y.-Y., & Heitmeier, M. (2019a). *WpmWithLdl: Implementation of Word and Paradigm Morphology with Linear Discriminative Learning* (1.3.17.1).

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & 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*, 1–39. https://doi.org/10.1155/2019/4895891

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience, 31*(1), 106–128. https://doi.org/10.1080/23273798.2015.1065336

Barton, K. (2020). *MuMIn: Multi-Model Inference* (1.43.17). https://cran.r-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Ben Hedia, S. (2019). *Gemination and degemination in English affixation: Investigating the interplay between morphology, phonology and phonetics*. Language Science Press. https://doi.org/10.5281/zenodo.3232849

Ben Hedia, S., & Plag, I. (2017). Gemination and degemination in English prefixation: Phonetic evidence for morphological organization. *Journal of Phonetics, 62*, 34–49. https://doi.org/10.1016/j.wocn.2017.02.002

Blevins, J. P., Ackerman, F., & Malouf, R. (2016). Morphology as an adaptive discriminative system. In D. Siddiqi & H. Harley (Eds.); *Morphological Metatheory* (pp. 271–302). John Benjamins. https://doi.org/10.1075/la.229

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer* (6.0.49, retrieved 03/02/2019). http://www.praat.org/

Booij, G. E. (1983). Principles and parameters in prosodic phonology. *Linguistics, 21*(1), 249–280. https://doi.org/10.1515/ling.1983.21.1.249

Burnage, G. (1988). *CELEX, A Guide for Users*. Centre for Lexical Information.

Chavent, M., Kuentz, V., Labenne, A., Liquet, B., & Saracco, J. (2017). *PCAmixdata: Multivariate Analysis of Mixed Data* (3.1). https://cran.r-project.org/package=PCAmixdata

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Harper and Row.

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & 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, 49*. https://doi.org/10.3758/s13428-020-01356-w

Cooper, W. E., & Danly, M. (1981). Segmental and Temporal Aspects of Utterance-Final Lengthening. *Phonetica, 38*, 106–115.

de Jong, N., & Wempe, T. (2008). *Praat Script Syllable Nuclei* [Praat script] (Retrieved November 2019). https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2

Drager, K. K. (2011). Sociophonetic variation and the lemma. *Journal of Phonetics, 39*(4), 694–707. https://doi.org/10.1016/j.wocn.2011.08.005

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage Publishing.

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. https://doi.org/10.1353/lan.0.0035

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language, 66*(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Goad, H. (1998). Plurals in SLI: Prosodic deficit or morphological deficit? *Language Acquisition, 7*, 247–284. https://doi.org/10.1207/s15327817la0702-4_6

Goad, H. (2002). Markedness in Right-edge Syllabification: Parallels across Populations. *Canadian Journal of Linguistics, 47*, 151–186.

Hsieh, L., Leonard, L. B., & Swanson, L. L. (1999). Some differences between English plural noun inflections and third singular verb inflections in the input: The contributions of frequency, sentence position, and duration. *Journal of Child Language, 26*(3), 531–543. https://doi.org/10.1017/S030500099900392X

Ivens, S. H., & Koslin, B. L. (1991). *Demands for Reading Literacy Require New Accountability Methods*. Touchstone Applied Science Associates.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kemps, R. J. J. K., Ernestus, M., Schreuder, R., & Harald Baayen, R. (2005a). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition*, *33*(3), 430–446. https://doi.org/10.3758/BF03193061

Kemps, R. J. J. K., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005b). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes*, *20*(1–2), 43–73. https://doi.org/10.1080/01690960444000223

Kiparsky, P. (1982). Lexical morphology and phonology. In I. Yang (Ed.), *Linguistics in the morning calm: Selected papers from SICOL1* (pp. 3–91). Hanshin.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208–1221. https://doi.org/10.1121/1.380986

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162–179. https://doi.org/10.1016/j.wocn.2006.04.001

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Landauer, T. K., & Dumais, S. T. (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. https://doi.org/10.1037/0033-295X.104.2.211

Lee, S., & Oh, Y. H. (1999). Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Communication*, *28*(4), 283–300. https://doi.org/10.1016/S0167-6393(99)00014-X

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1310.4546

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS ONE*, *12*(2), e0171935. https://doi.org/10.1371/journal.pone.0171935

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

O'Rourke, N, Hatcher, L., & Stepanski, E. J. (2005). *A Step-by-Step Approach to Using SAS for Univariate & Multivariate Statistics*. SAS Publishing.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, *51*(3), 406–413. https://doi.org/10.1017/S0305004100030401

Plag, I. (2018). *Word-Formation in English (Second Edition)*. Cambridge University Press. https://doi.org/10.1017/CBO9780511841323

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics, 53*(1), 181–216. https://doi.org/10.1017/S0022226715000183

Plag, I., Lohmann, A., Ben Hedia, S., & Zimmermann, J. (2020). An <s> is an <s'>, or is it? Plural and genitive-plural are not homophonous. In L. Körtvélyessy & P. Štekauer (Eds.), *Complex Words*. Cambridge University Press.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Ramscar, M., & 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. https://doi.org/10.1080/03640210701703576

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and Their Implications for Symbolic Learning. *Cognitive Science, 34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement BT - Clasical conditioning II: current research and theory. *Classical Conditioning II: Current Research and Theory*, 64–99.

Roelofs, A., & Ferreira, V. S. (2019). The architecture of speaking. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 35–50). MIT Press.

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Eds.), *Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology, 11*. Springer. https://doi.org/10.1007/978-94-017-2390-9_2

Schmitz, D., Baer-Henney, D., & Plag, I. (2020). *The duration of word-final /s/ differs across morphological categories in English: Evidence from pseudowords* [Manuscript submitted for publication]. English Language and Linguistics, Heinrich Heine University Düsseldorf, Germany.

Selkirk, E. (1996). The Prosodic Structure of Function Words. In K. Demuth & J. Morgan (Eds.), *Signal to Syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 187–213). Routledge.

Sering, T., Milin, P., & Baayen, R. H. (2018). Language comprehension as a multiple label classification problem. *Statistica Neerlandica*, 1-15.

Seyfarth, S., Garellek, M., Gillingham, G., Ackerman, F., & Malouf, R. (2017). Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience*, *33*(1), 32–49. https://doi.org/10.1080/23273798.2017.1359634

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413. https://doi.org/10.3758/BRM.42.2.393

Swanson, L. A., & Leonard, L. B. (1994). Duration of function-word vowels in mothers' speech to young children. *Journal of Speech and Hearing Research*, *37*(6), 1394–1405. https://doi.org/10.1044/jshr.3706.1394

Team, Rs. (2020). *R, RStudio: Integrated Development for R* (1.4.1103). http://www.rstudio.com/

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249–267. https://doi.org/10.1016/j.wocn.2018.09.004

Tomaschek, F., Plag, I., Ernestus, M., & 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, 2019*, 1–39. https://doi.org/10.1017/S0022226719000203

Tremblay, A., & Ransijn, J. (2020). *LMERConvenienceFunctions: Model Selection and Post-Hoc Analysis for (G)LMER Models* (3.0). https://cran.r-project.org/package=LMERConvenienceFunctions

Tucker, B., Sims, M., & Baayen, R. H. (2019). *Opposing forces on acoustic duration*. PsyArXiv, 1–38. https://doi.org/10.31234/osf.io/jc97w

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods, 51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Turk, A., & Shattuck-Hufnagel, S. (2020). *Speech Timing*. Oxford University Press. https://doi.org/10.1093/oso/9780198795421.001.0001

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America, 61*(3), 846–858. https://doi.org/10.1121/1.381374

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer New York. https://doi.org/10.1007/978-0-387-21706-2

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers, 36*(3), 481–487. https://doi.org/10.3758/BF03195594

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian Conditioning: Application of a Theory. *Inhibition and Learning*, 301–334.

Walsh, T., & Parker, F. (1983). The duration of morphemic and non-morphemic /s/ in English. *Journal of Phonetics, 11*(2), 201–206. https://doi.org/10.1016/s0095-4470(19)30816-2

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental Durations in The Vicinity Of Prosodic Phrase Boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707–1717. https://doi.org/10.1121/1.402450

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. Routledge. https://doi.org/10.4324/9781315165547

Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report, 8*, 183–225.

Zimmermann, J. (2016). Morphological status and acoustic realization. In C. Carignan & M. D. Tyler (Eds.), *Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology* (SST-2016) (pp. 201–204). ASSTA. https://assta.org/sst-2016-proceedings/

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x