Stratification without morphological strata, syllable counting without counts - modelling English stress assignment with Naive Discriminative Learning

February 15, 2021

Abstract Stress position in English words is well-known to correlate with both their morphological properties and their phonological organisation in terms of non-segmental, prosodic categories like syllable and foot structure. While two generalisations capturing this correlation, directionality and stratification, are well established, the exact nature of the interaction of phonological and morphological factors in English stress assignment is a much debated issue in the literature. The present study investigates if and how directionality and stratification effects in English can be learned by means of Naive Discriminative Learning, a computational model that is trained using error-driven learning and that does not make any a-priori assumptions about the higher-level phonological organisation and morphological structure of words. Based on a series of simulation studies we show that neither directionality nor stratification need to be stipulated as a-priori properties of words or constraints in the lexicon. Stress can be learned solely on the basis of very flat word representations. Morphological stratification emerges as an effect of the model learning that informativity with regard to stress position is unevenly distributed across all trigrams constituting a word. Morphological affix classes like stress-preserving and stress-shifting affixes are, hence, not predefined classes but sets of trigrams that have similar informativity values with regard to stress position. Directionality, by contrast, emerges as spurious in our simulations; no syllable counting or recourse to abstract prosodic representations seems to be necessary to learn stress position in English.

Keywords ndl · error-driven learning · modelling · stress assignment · morphological strata · directionality

1 Introduction

Stress position in English words is well-known to correlate with both their phonological and morphological properties. For example, stress is often penultimate in morphologically simplex nouns with a heavy penultimate syllable, as illustrated by the word ‘agenda’ *a.gén.da*. In derived words with a so-called stress-preserving suffix, stress is always on the same syllable as it is in the base word. For example, *éffort-less* is stressed on the same syllable as *éffort*, in spite of the fact that the word ‘effortless’ has a heavy penultimate syllable. By contrast, stress in derived words with so-called stress-shifting suffixes may be on a different syllable than it is in the base word (e.g *emplóy – employ-ée*).[1]

In the present paper, we will be concerned with two descriptive generalisations about English stress assignment that play a prominent role in virtually all formal accounts. The first is the principle of directionality (Hayes (1982); see (Pater, 2000) for an optimality-theoretic account for English; see Kager (2012) for a typological overview and a discussion of different modelling options within Optimality Theory; see Alber (2020) for an overview on Germanic languages). Phonological generalisations about stress position are usually thought to be directional in the sense that they count syllables from a word edge. This also means that they crucially rely on word representations that incorporate syllables as abstract units of prosodic organisation. The relevant word edge in English is usually assumed to be a three-syllable window at the right word edge.

The examples in Table 1 illustrate the principle. Main stress is indicated by an acute accent, syllable boundaries are marked by ’.’. A description of stress patterns that is in line with the principle of right directionality will refer to stress as being on the word-final syllable in (1a), on the penultimate syllable in (1b), and on the antepenultimate syllable in (1c). This generalisation captures the fact that, with the exception of compound words, main stress in English words always lands on one of the last three syllables of the word. However, the idea that main stress assignment is directional or even mono-directional in nature is not without problems.

Table 1: *Examples stress assignment in long English words. Main stress is indicated by an acute accent, syllable boundaries are marked by ’.’.*

a. Ka.la.ma.zóo
b. Mo.non.ga.hé.la
c. Ha.ma.me.li.dán.the.mum

For example, Hammond (e.g. 1999, 318ff) and McCully (2003) argue that both right alignment and left alignment play a role for main stress assignment in English. Furthermore, attempts to empirically verify edge alignment face the problem that the number of English morphologically simplex words that are longer than three syllables is rather low. Another, fundamental problem with the principle of directionality is that it is systematically constrained by morphological structure, in the sense that different affixes require different adjustments to the syllable counting generalisation about stress position. This phenomenon, among others, has led scholars to assume that the English lexicon is stratified, with the different strata representing different morphological categories.

Morphological stratification in this sense is the second generalisation that we will be concerned with in this article. Specifically, we will focus on suffixal strata, often referred to as ’stress-preserving’ and ’stress-shifting’ suffixes. Stress-shifting suffixes fall into two different subgroups: those which themselves attract stress (often called ’auto-stressed’) and those which do not; most of the latter suffixes are ’pre-stressing’, which means that main stress is on the syllable immediately preceding the suffix. Table 2 provides examples of all three classes. The suffixes *-ness* and *-ly* are examples of stress-preserving suffixes (2a), *-ity* and *-ical* are pre-stressing stress-shifting suffixes (2b), and *-ee* and *-ese* are auto-stressed stress-shifting suffixes.

The existence of stress-preserving and stress-shifting suffixes has prominently been used as evidence in favour of stratal approaches to the morphology-phonology interaction such as Lexical Phonology and Morphology (Kiparsky, 1982) and Stratal Phonology (Bermúdez-Otero and McMahon, 2006; Bermúdez-

[1] The distinction between ’stress preserving’ and ’stress shifting’ affixes (Fudge, 1984) is largely co-referent with other dichotomies, such as that of ’cohering’ - ’non-cohering’ affixes (Booij, 1983; Siegel, 1974; Booij and Rubach, 1987) and ’class I’ - ’class II’ affixes (SPE Siegel, 1974; Chomsky and Halle, 1968).

Table 2: *Examples of a) stress-preserving, b) stress-shifting and c) auto-stressed suffixation. The accent indicates the stressed syllable.*

	derived	base
a.	háppiness	háppy
	políteness	políte
	prodúctively	prodúctive
	extrémely	extréme
b.	productívity	prodúctive
	monstrósity	mónstrous
	metaphórical	métaphor
	symmétrical	sýmmetry
c.	employée	emplóy
	interviewée	ínterview
	Japanése	Japán
	Portuguése	Pórtugal

Otero, 2012, 2018). These approaches assume that English morphology is organised into two (or more) different strata, with interleaving phonological and morphological modules. The difference between stress preserving and stress shifting suffixes is then modelled in terms of the point in time when a suffix is attached to its base word or stem. So-called stress-shifting suffixes are attached before phonological stress rules have applied, stress-preserving suffixes are attached after stress rule application. Other approaches model the stress behaviour of different types of affixes in terms of affix-specific rule or constraint systems (esp. Co-phonology approaches, cf. e.g. Stanton and Steriade, 2014).

However, the exact nature of the interaction of phonological and morphological factors in a stratified lexicon is a much debated issue in the literature. Empirically, it is well-known that existing proposals (stratal or non-stratal) integrating phonological and morphological factors fall short of convincingly predicting stress position when tested on data sets of words, both actual and nonce words. Furthermore, attempts to quantify accuracy of predictions are very rare and often limited to subsets of the lexicon. One such attempt that focuses on derived words is Zamma (2012)'s study. The model developed in this study includes variable constraint rankings, and accuracy is measured in terms of the number of predicted rankings that conform to attested words (cf. Zamma, 2012, chpt. 6 for discussion). Domahs et al. (2014) provide a statistical analysis of the predictive power of syllable structural factors in morphologically simplex words, both nonce words and existing words. Simplex words are also studied by Moore-Cantwell (2016, chpt. 4); the study investigates the match between a constraint-based MaxEnt model (Goldwater and Johnson, 2003) that includes lexically specific constraints, and lexical distributions. Dabouis et al. (2017) investigate the predictive power of both phonological and morphological factors for stress in some 5,000 verbs extracted from Jones (2006)'s English Pronouncing Dictionary. All works cited show that the phonological and morphological factors they use in the analysis can explain a large portion of the data, but also admit to considerable leakage. In all pertinent accounts, it is thus assumed that stress assignment is subject to lexical idiosyncracy to some extent (cf. Alber (2020) for a recent overview of the literature on stress in Germanic languages, including English). It is also unclear, how these studies can be compared, since all of them use different kinds of baselines, constraints and evaluation metrics.

Another open question concerns how language users become aware of these principles. One potential answer is that learning takes place on the basis of abstract representations of the prosodic and morphological structure of words, and on the basis of constraints that operate on the basis of those abstract representations (cf. e.g. Moore-Cantwell (2016) for a recent OT model; cf. Pearl et al. (2016) for a comparison of the learnability of classic pertinent approaches). Abstract representations include syllables, morae, metrical feet, and the morphological stratum affiliation (e.g. level 1 or level 2) of affixes. Constraints include constraints on edge alignment, on the relation between syllable weight and stress, on extrametricality, as well as on the stressability of affixes. To what degree these representations and constraints are innate or learned is a matter of debate.

In the present paper, we will pursue an alternative answer, which is in line with usage-based theories of linguistic generalisation (Bybee, 2011, 2001, 2002). By 'usage-based' approaches we mean a group of theories that share the assumption that properties such as stress are associated with and may even emerge from the distributional characteristics of words and sub-word units in the Mental Lexicon. For stress assignment, this means that language users store words that they encounter with their stress pattern, and assign stress to words they have not encountered before on the basis of the distribution of stress patterns among stored words.

So far, only few attempts have been made to test this idea on stress assignment data with the help of computational implementations of usage-based models (see Daelemans et al., 1994, on Dutch for one of the few exceptions). One key challenge for a usage-based model of stress assignment is the definition and selection of input features provided to the computational model. Computational modelling approaches usually rely on flat and non-nested structures. This does not seem compatible with generalisations about stress assignment that, as we have seen above, rely on highly abstract and elaborate representations of phonological and morphological structure.

The present paper sets out to investigate if and how directionality and stratification effects in English can be learned by a computational model without any assumption about abstract phonological and morphological representations of words. The particular implementation that we will use is Naive Discriminative Learning ('NDL', Arppe et al., 2018). Based on a series of simulation studies we will show that neither directionality nor stratification need to be assumed to be a-priori properties of words or constraints in the lexicon. Stress can be learned solely on the basis of very flat word representations in terms of trigrams, by a system that is not given any explicit information about directionality or the morphological class affiliation of constituent affixes. Instead, morphological stratification emerges as an effect of the model learning that informativity with regard to stress position is unevenly distributed across all trigrams constituting a word. Morphological affix classes like stress-preserving and stress-shifting affixes are, hence, not predefined classes but sets of trigrams that have similar informativity values with regard to stress position. Directionality, by contrast, emerges as spurious in our simulations; no syllable counting or recourse to abstract prosodic representations seems to be necessary to learn stress position in English.

The paper is structured as follows. We will first introduce our computational framework in Section 2. Section 3 will then explain the methodology of our simulation experiments. The simulations will then be discussed in Section 4, in two steps. We will first be concerned with directionality (Section 4.1), and then with morphological strata (Section 4.2). In each section, we will present both general simulation outcomes and an in-depth analysis of our experiments, which shows why the algorithm makes the predictions it does. The paper ends with a summary and conclusion in Section 5, which will also discuss the implications for linguistic theory.

2 Discriminative learning and the error-driven learning rule

Different approaches to training a neural network are available. Due to hidden layers or complex learning algorithms, as is the case in deep neural networks and recurrent neural networks (Graves and Schmidhuber, 2005; Graves et al., 2013), the trained networks are usually hard to interpret from a cognitive perspective. We therefore used a two-layer neural network that is trained with a simple error-driven learning rule (Rescorla and Wagner, 1972; Rescorla, 1988; Ng and Jordan, 2002; Widrow and Hoff, 1960), implemented in Naive Discriminative Learning (the package 'NDL' as implemented in R, Arppe et al., 2018).

The error-driven learning rule mathematically formalizes general cognitive mechanisms assumed by the cognitive theory of *Discriminative Learning* (Ramscar and Yarlett, 2007; Ramscar et al., 2010; Ramscar, Dye, M. and Klein, J., 2013). According to the theory of Discriminative Learning, learners build cognitive representations of their environment by establishing associations between events in their environment on the basis prediction and prediction error. The algorithm formalizes this by establishing *association weights* between input features (henceforth *cues*) and classes or categories (henceforth *outcomes*) that co-occur in events. To name an example, in English the word final letter sequence '-ize' serves as a cue to the outcome 'verb', and the word final letter sequence '-ical' serves as a cue to the outcome 'adjective'.

According to Discriminative Learning, learning is shaped by prediction and prediction error. Error is positive and increases association weights between a cue and an outcome every time that the predicted outcome occurs (such as '-ize' in the word 'realize' indicating a verb). By contrast, error is negative and decreases association weights whenever the predicted outcome does not occur (such as 'ize' in the noun 'size'). As a result, weights and associations (and the resulting representations) are constantly updated on the basis of new experiences. The strength of the adjustment depends a) on the number of cues that are present in a learning event and b) on the size of the error between the prediction emerging from the cues and the actual outcome in the learning event. This gives rise to *cue competition*, during which cues compete for being informative about an outcome. As a result of learning through continuous prediction and error, cognitive representations emerge. An in-depth description of the theory can be found in (Ramscar, Dye, M. and Klein, J., 2013; Linke and Ramscar, 2020); a description of the NDL model can be found in Baayen et al. (2011); an overview how different cue-to-outcome structures affect learning can be found in Hoppe, Hendriks, Ramscar and Rij (2020).[2]

The error-driven learning rule has been shown to successfully model and predict a number of important effects observed in animal learning (Rescorla, 1988) and human learning. For example, Ramscar et al. (2010) demonstrated how the presentation order of cues and predicted events during learning affects the strength of learning. Learning is more effective when, for example in 'wug' experiments, the orthographic (or acoustic) word precedes the corresponding picture than when the picture precedes the word. This effect was also reported for phonetic learning (Nixon, 2020) and inflectional learning (Hoppe, van Rij, Hendriks and Ramscar, 2020). Nixon (2020) demonstrated that a new cue for an outcome is blocked from learning, once another cue has already been learned as informative about an outcome. This finding mirrors the 'blocking effect' in animal learning studies first demonstrated by Kamin (1968).

In addition, the error-driven learning rule successfully models aspects of child language acquisition (Ramscar et al., 2010, 2011; Ramscar, Dye and McCauley, 2013; Ramscar, Dye, M. and Klein, J., 2013), acquisition and usage of allomorphic suffixes (Divjak et al., 2020), reaction times in lexical decision tasks (Baayen et al., 2011; Milin, Feldman, Ramscar, Hendrix and Baayen, 2017), self-paced reading (Milin, Divjak and Baayen, 2017), acoustic duration of American English word final [s] depending on their morphological function (Tomaschek et al., 2019), auditory comprehension (Baayen et al., 2016; Arnold et al., 2017) and acoustic single-word recognition (Shafaei-Bajestan and Baayen, 2018).

To summarize, the association weight between a cue and an outcome is formed through the experience with other cues and outcomes that have been encountered during the learning history in both production and comprehension. The weight represents the support which a specific cue can provide for a specific outcome. Cognitive representations of grammatical structures emerge from the association weights between every encountered cue and every encountered outcome. In this model principles like the principle of directionality and stratification have no independent status as constraints on representations or grammatical outputs. The question then arises if and how the model can emulate and explain the empirical effects that have traditionally been ascribed to these mechanisms.

3 Methods

For our simulation experiments, we trained NDL to discriminate stress positions and then used the trained network to predict stress positions. The material for the simulations was obtained from the CELEX lexical database of English (N = 33,407 word forms, Baayen et al., 1993). This data set served as both the training set and the test set. We performed our analysis in two steps. In a first step we focused on directionality and investigated which cue structure best predicts the attested stress patterns. In a second step we focused on morphological stratification and studied if and how exactly morphological strata emerged in our model. In what follows we discuss the methodological details of our modelling approach.

The cues on which we trained the model were based on the orthographic transcriptions of all words in CELEX. We used orthographic transcriptions because English stress is strongly correlated with vowel quality. By presenting orthography to the model, we avoided the problem that in many English words, knowing the vowel quality is already predictive of stress. This is because only a very restricted set of vowels can occur in unstressed English syllables, a phenomenon that is usually accounted for in terms of 'vowel reduction'. The most common reduced vowel, schwa, is even restricted to exclusively occurring in unstressed syllables. For example, a common pronunciation of the word 'America' is [əmɛrəkə]. Given this sound structure, it is clear that the stress can only be on [ɛ], the only full vowel. Given the orthographic string <America>, however, all syllables are potentially stress-bearing. Providing the computational

[2] This formalization of learning differs from other theories of learning, such as Bayesian models (Kleinschmidt and Jaeger, 2015), or distributional learning models (Wanrooij et al., 2014, 2015; Werker et al., 2012; Terry et al., 2015).The latter class of models assumes that learners learn the frequency of occurrence of co-occurrences and the resulting distributions. For a review of the differences between distributional learning and error-driven learning in the context of language, see (Kapatsinski, 2018).

model with orthographic cues rather than with actual pronunciations, thus, serves to make its task more difficult.

One potential set of cues are letter monographs. However, letter monographs miss out on the informative properties of orthotactics, i.e. sequential information about adjacent letters in words. Accordingly, we decided to use higher-order n-grams, specifically bigrams and trigrams, as is common practice in linguistic studies using error-driven learning (Baayen et al., 2011, 2016; Milin, Feldman, Ramscar, Hendrix and Baayen, 2017; Tomaschek et al., 2019). The cue structure does not encode formally defined syllables or syllable positions. We tested which kind of cue structure best predicted stress position: letter bigrams (BG), or trigrams (TG), or both together (BGTG)[3].

Stress position was coded as outcomes in our simulations. We implemented three different types of outcome structures. The first is a representation of the traditional account that the stress position in a word is counted from the offset of the word (henceforth *stress from right* (e.g. Hayes, 1982; Pater, 2000; Alber, 2020, as discussed in Section 1 above)). In order to examine the validity of this claim, we also tested two other ways of representing stress as outcomes in our model. The first is to count the stress position from the onset of the word (henceforth *stress from left*). The second is to select the vowel letter present in the stressed syllable (henceforth *stress in the vowel*). The value of *stress from right* varied between one and seven. *Stress from left* contained six values, ranging between stress on syllable number one and stress on syllable number six. *Stress in the vowel* did not differentiate in which syllable the vowel was located and contained 59 different values in total.

Take the word ‘realize’ as an example. Its letter bigram cues are `#r, re, ea, al, li, iz, ze, e#`, its letter trigrams are `#re, rea, eal, ali, liz, ize, ze#` (where # represents the word boundary). Crucially, the model is unaware what phone sequence the letter n-grams represents. The bigram <ea> can represent the vowel [i] in ‘please’ but also the [iæ] hiatus in ‘reanalyse’. Similarly to a naive reader, the model has to learn to discriminate the outcomes on the basis of potentially ambiguous cues.

The outcomes of the models, – called ‘outputs’ in the terminology of neural networks – represent the position of the stress. For ‘realize’, this means that *Stress from left* is: 1; *stress from right* is: 3, and *stress in the vowel* is: ‘ea’.

We compared nine different networks in terms of how well they predict stress in our data set. Each network was trained on a different combination of cue and outcome structures (3×3, i.e. bigram cues, trigram cues, and a combination of bigram and trigram cues with the outcome *stress from left*, the outcome *stress from right*, and the outcome *stress in the vowel*). We use the Danks Equilibrium Equations (Danks, 2003) to train the model, as implemented in the NDL package.

After training, the network is evaluated in terms of whether it is able to discriminate among the outcomes on the basis of presented cues (typically from a word of interest). Thus, it is presented by a set of cues, e.g. `#re, rea, eal, ali, liz, ize, ze#`, and has to select which of the potential outcomes (e.g. for *stress from right* 1, 2, 3, 4, 5, 6, or 7) is best predicted by the cue set. This is achieved by means of an activation vector, summing up the association weights between the presented cues for each of the possible outcomes in the network. The outcome with the highest activation is the winner of the classification, thus the predicted stress position.

In formal approaches typically used to model stress assignment, such as Optimality Theory (e.g. Pater, 2000; Zamma, 2012; Moore-Cantwell, 2016), the selected outcome is one of the inputs (’candidates’ in OT) provided to the procedure. Note that this is not the case in neural networks. Instead, the test procedure decides among a set of possible outcomes provided to the model, not among inputs. This can be best exemplified by monosyllabic words. Naively, it should not be too hard to find the stressed position in monosyllabic words. Whereas this line of thought is of course plausible in the real world, it is not in our model. This is because this kind of reasoning follows the misconception that the model is aware of the number of syllables in the cue set that is presented during the classification procedure. This is not the case in the simulations presented in this paper. On the contrary, the model is absolutely unaware of how many syllables the word contains that the presented cue set is based on. The discrimination among the outcomes is based purely on the activation strength calculated on the basis of the presented cue set. It is therefore even possible that due to cue competition and due to the distribution of weights, the network predicts a stress position which is incompatible with the true number of syllables in the

[3] Theoretically, we could also use higher-order n-grams such as 4-grams. However, the longer the n-gram, the stronger the model is faced with a one-to-one mapping between cues and outcomes, which results in smaller cue-competition during training.

presented word. For example, it is possible that the network erroneously predicts stress on the penultima for a monosyllabic word.

4 Findings

4.1 Accuracy of prediction by cue-outcome structure

Each of our nine networks (cf. Section 3 above) was set the task of predicting stress position in all words from CELEX. As can be seen in Table 3, the prediction accuracy for all cue-and-outcome combinations ranges between 59.0% and 84.9%, i.e. highly above chance. As is clear from the table, the use of letter bigrams consistently yields a lower prediction accuracy than the use of letter trigrams. Also, a combination of bigrams and trigrams did not improve classification accuracy. This means that letter trigrams are sufficiently informative about stress positions[4].

Table 3: *Percentage of correctly categorized stress positions in whole data set.*

Cue structure	left	right	vowel
letter bigrams	71.4	59.0	72.1
letter trigrams	80.7	74.9	84.9
both together	80.6	74.9	84.8

Given that letter trigrams yield better prediction accuracy, we focus on this cue structure in what follows. We now inspect how it was used by the network to classify stress positions given different assumptions about directionality. Table 3 demonstrates that stress can be learned without syllable counting. The model trained to predict stress in terms of the orthographic vowel has the highest prediction accuracy, followed by the model that was trained to predict stress from the left word edge. The weakest model is the one that was trained to predict stress from the right word edge. All differences between model accuracies are significant (counting *stress from left* vs. counting *stress from right*: $\chi^2 = 246.4$, df = 1, p < 0.001; *stress in vowel* vs. counting *stress from left*: $\chi^2 = 961.5$, df = 1, p < 0.001). Using trigrams as cues, we tested the *stress from left* and *stress in vowel* models in twenty cross-validation runs. In each run, we trained the models on 70% of the data that were randomly selected and tested on the remaining unseen 30% of the data. The average prediction accuracy was 71.6% (sd = 0.005) in the *stress from left* model and 75.8% (sd = 0.003) in the *stress in vowel* model. Thus, even if the model has not encountered a word form, it was able to predict its stress position with a fairly high accuracy.

Since the model had no a-priori information about morphological structure, and since suffixing influences stress position in English, it is not surprising that the *stress from right* model showed only weak performance. This is because the descriptive generalisation that English stress always lands in a three-syllable window at the right edge is not true for complex words with so-called stress-preserving suffixes (cf. Section 1 above for discussion).[5] What is surprising, however, is that stress is best predicted by the *vowel* model, as none of the existing theories predicted this result.

Looking only at prediction accuracies, however, does not tell us much about why the models performed as well as they did. With regard to the vowel model, a very likely confounding factor is that orthographic

[4] The question arises why trigrams yield a higher accuracy than bigrams. Given that trigrams capture a larger portion of a word than bigrams, the uncertainty about the relation between cues and the stress position should be lower for trigrams than for bigrams. We assessed this uncertainty by calculating the entropy (Shannon, 1948) for each bigram and for each trigram in relation to the stress position. To obtain the entropy for each n-gram cue, we assessed how often each n-gram occurs with each stress position. To calculate entropy, we calculated the co-occurrence probability by dividing a cue's frequency by the summed frequencies of that cue and all stress positions. We found that the average entropy in relation to stress position is significantly lower for trigrams (H = 0.84) than for bigrams (H = 1.48, δH = 0.64, t = 24.54, df = 844.83, p-value < 0.001).

[5] Readers might wonder why the *stress from left* model performed so well, given that prefixes should also have an influence on stress assignment. This may be due to the fact that when a word is prefixed with one prefix it typically reoccurs with other, prosodically similar prefixes. For example, the adjective 'interpretable' occurs with 'un-', 're-', and 'mis-'. In all cases, the prefixed word has the third stress position, which provides the model with strong support for that position. A critical inspection of stress shifts due to prefixing is beyond the scope of this paper.

vowels may occur multiple times in words. It is thus unclear whether the high prediction accuracy of the *vowel* model results from the fact that vowel repetition increases the probability of finding the correct stress position. In the following section, we turn to a more detailed statistical analysis of our most successful model, the *vowel* model. We have two aims: The first is to learn more about the potential confounding factors mentioned. The second is to inspect how the emerging structure in the network affects the accuracy of stress assignment.

4.2 Prediction accuracy, model certainty, and the linguistic properties of words

We use linear logistic regression to study how the prediction accuracy (our dependent variable) is correlated with the word's linguistic properties, and with the network's certainty/uncertainty about the stress position. We first explain the linguistic predictors.

One interesting question that we will pursue here is how the vowel model predicts stress in words of different length. This is important because the vast majority of English words are short, with monosyllables having a particularly large share in the vocabulary. Model accuracy on short words will therefore also have a large share in the general accuracy score of the model. Recall from Section 3 that, in principle, NDL is ignorant of word length in our dataset and, hence, it is possible that a monosyllabic word, for example, is predicted to be stressed in other positions than the first. However, due to cue competition the association strength between cues and the stress position in a monosyllabic word is very likely to be stronger than in polysyllabic words. We thus expect that prediction accuracy will be very high for short words and will decrease in words with a greater number of syllables. This was tested with the predictor 'number of syllables'.

Since the outcomes in the vowel model did not differentiate in which syllable a vowel was located, the probability that the model correctly predicts stress is higher, when a word contains the identical vowel multiple times. Accordingly, we expect prediction accuracy to be higher if the word contains multiple instances of an identical vowel rather than different vowels. This was tested with the predictor 'double vowels' (with TRUE representing a word that contains the identical vowel multiple times).

In the upcoming analysis, we also wanted to gain an initial understanding of how morphological effects on stress assignment are represented in the model, and how they affect stress assignment. To do so, we used the information included in CELEX about whether words are derived, inflected, or simplex as a predictor variable in our regression model. Inflectional suffixes are generally stress-preserving in English. Thus, cues in the word stem should be good predictors for stress position. This is why we expect higher prediction accuracy for inflected words than for uninflected words. With respect to derivation, derivational processes can be stress-preserving or stress-shifting. This means that the variability of the stress position in derived words is very high, which should create more uncertainty about the stress position for the learning model. Accordingly, we expect accuracies for derived words to be lower than for underived words. These two hypotheses were tested with the predictors 'inflected word' and 'derived word' (with TRUE representing inflected or derived words). The distinction between stress-preserving and stress-shifting derivation is not part of following analysis. We will look at this issue in greater detail in Section 4.4.

In addition to linguistic properties of words, we analyze how prediction accuracy is affected by the network's certainty/uncertainty about a stress position. We do so with the help of two measures. The first is 'activation', i.e. the sum of the weights between a word's cues and the outcomes in the network (cf. section 2). Activation gauges the amount of support, or certainty, from a word's cues to its true stress position. Usually, activations are used as predictors in regression models. Higher activations have been shown to be correlated with faster response times and lower error rates (Baayen et al., 2011; Arnold et al., 2017; Milin, Feldman, Ramscar, Hendrix and Baayen, 2017). Accordingly, we predict that higher activations should be associated with better prediction accuracy.

The second measure we use to assess the network's certainty/uncertainty, is 'activation diversity'. This measure reflects the amount of competition among possible outcomes for a word's cue set. This competition is associated with the amount of uncertainty about an outcome. The stronger the activation of competing outcomes, the more uncertainty a cue set creates about the actual outcome. This is reflected by a higher 'activation diversity' (Arnold et al., 2017; Tucker et al., 2019; Tomaschek et al., 2019). Accordingly, we expect greater activation diversity to be associated with lower prediction accuracy.

4.3 Closeup on the ‘vowel model’

We tested these predictions with a linear logistic model. Activations and activation diversities were log transformed, centered and scaled to obtain a data set with a less skewed distribution. We excluded strong outliers in the NDL measures ($\sim$ 2.5 standard deviation away from the mean, loss of 2.65% of the data). We subtracted 1 from number of syllables to obtain an intercept located in the value space (which was back-transformed to original values in the plots).

In pilot analyses, we found that ‘number of syllables’ was collinear with ‘derived word’. This is because derived words have, on average, one syllable more than underived words (β = 1.1, sd = 0.011, t = 98.0, p < 0.0001). This significant correlation caused suppression in the regression model. i.e. a change in sign for one predictor, when the other, correlated, predictor was added (see Tomaschek et al., 2018, for inspection of collinearity in regression). Testing the effects of ‘derived word’ indicated that the stress position of derived words was less accurately predicted than that of underived words, when ‘derived word’ was used as a predictor on its own (β = -0.5, sd = 0.031, z = -15.89, p < 0.0001). However, due to the collinearity issue, ‘derived word’ was excluded from the following analysis.

We fitted prediction accuracy (logit) in the *stress in the vowel* model with an interaction between ‘number of syllables’ and ‘double vowels’, and main effect for ‘inflected word’, ‘activation’ and ‘activation diversity’. Table 4 presents the summary table. All predictors turned out to be significant. The intercept has a value of logit = 3.32 which corresponds to an accuracy of 96.4%.

Table 4: *Model summary of stress classification in the ‘stress in vowels’ model*

	Estimate	Std. Error	z value	p value
(Intercept)	3.32	0.06	58.09	< 0.001
DoubleVowels = TRUE	-0.64	0.08	-7.70	< 0.001
Number of Syllables	-0.56	0.03	-17.00	< 0.001
DoubleVowels = TRUE : Number of Syllables	0.28	0.04	7.10	< 0.001
Inflected = TRUE	0.76	0.13	5.86	< 0.001
Activation	1.26	0.02	56.52	< 0.001
Activation diversity	-0.91	0.02	-47.00	< 0.001

Figure 1 (a) illustrates the estimated interaction between ‘number of syllables’ and ‘double vowels’. The y-axis represents back-transformed accuracy. As expected, the number of syllables is negatively correlated with prediction accuracy. Note, however, that estimated accuracy of prediction drops below 80% only for words with more than five syllables. Words with more than five syllables are better predicted if they contain repetitions of the same orthographic vowel (cf. ’double vowels’ in Figure 1a) than if they do not (cf. ’different vowels’ in Figure 1a). In the former case, accuracy never drops substantially below 80% regardless of word length; in the latter case, accuracy drops to about 40% for words with eight syllables.

Figure 1 (b) shows that, as expected, prediction accuracy is higher for inflected words than for uninflected words. The difference is rather small (roughly 2.5%), which may be due to the fact that, with regard to their stress properties, uninflected words form a heterogenous group comprising derived words with different stress properties and simplex words. We will return to this problem in section 4.4 below.

Next we turn to the measures gauging the network’s certainty/uncertainty (Figure 1, c&d). Prediction accuracy is proportional to ‘activation’. It is very low for very low-activated stress positions, but reaches ceiling level very fast, i.e. an accuracy of almost 100%, when activation increases. This means that those cases in which prediction accuracy is (comparatively) low are characterised also by low activation of the stressed syllable, reflecting weaker support for its true stress position. Finally, prediction accuracy is inversely proportional to ‘activation diversity’, indicating that that when the uncertainty in the cues about the outcome increases, the model cannot make a well informed choice about stress. In conclusion, the higher the model’s certainty about a stress position, the better its prediction accuracy[6] .

[6] We inspected the performance of the *stress from left* model in the same way as we did for the performance of the *stress in vowel* model. As it turned out, effect sizes and directions are very similar between the two models.

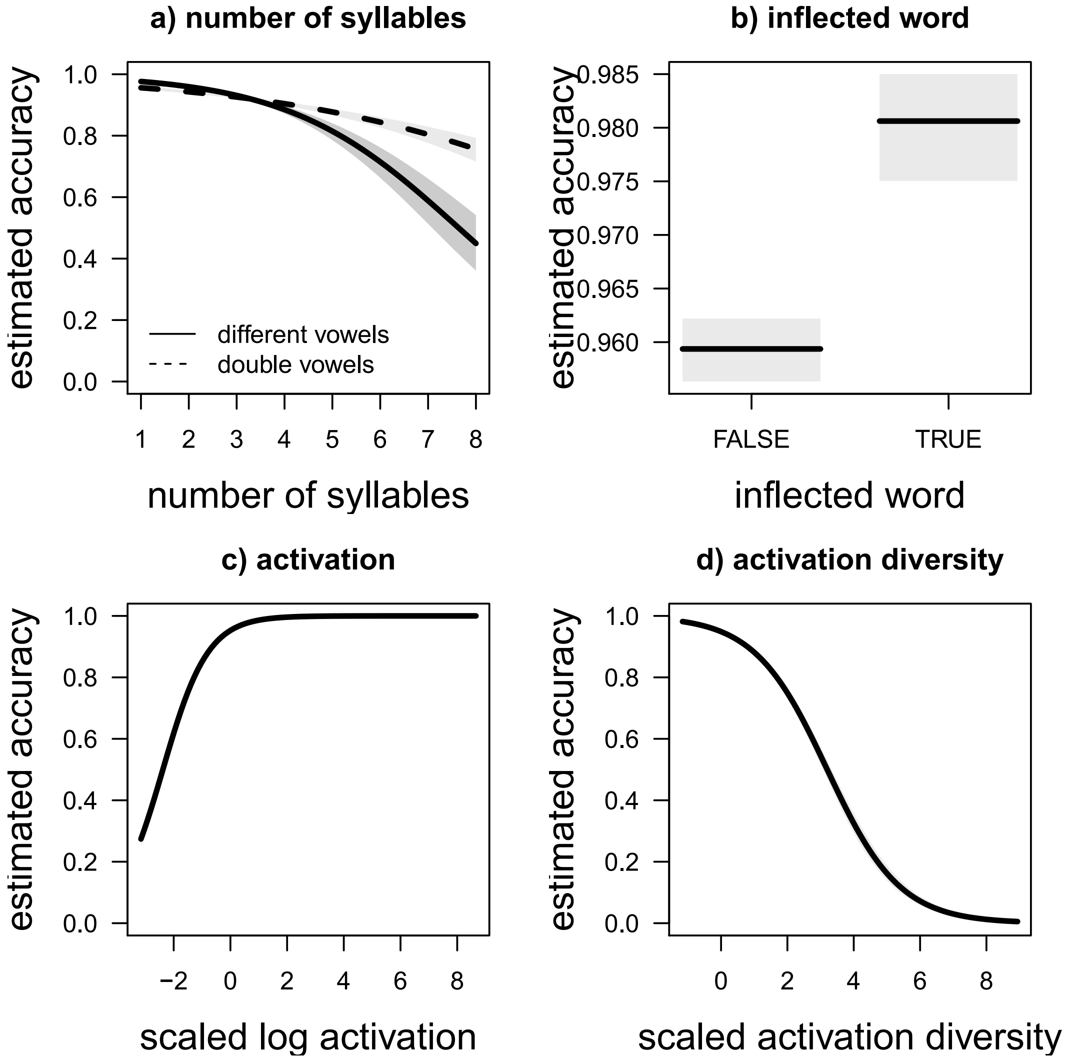

Figure 1: *Estimated prediction accuracy of the logistic regression for the 'stress in the vowel' NDL model. The y-axes represent the back-transformed prediction accuracy. The confidence intervals in c) and d) are so small that they are not visible in the plots. Note that scales may vary between plots.*

4.4 Derived words in the 'vowel model': a case study

Since we have excluded 'derived word' as a predictor from our models, the models described above ignore any complexities arising in stress assignment due to derivation. For example, suffixes such as '-ion', '-ity' or '-ical' (and their equivalent derivations) attract stress on the preceding syllable (*pre-stressing*). Suffixes such as '-ness' or '-less' preserve basal stress (*stress preserving*), whereas suffixes such as '-ese', '-teen' or '-ee' carry stress themselves (*auto-stressed*, cf. section 1 above for discussion).

We expect that prediction accuracy should be associated with the suffix type. In words with stress preserving suffixes, the stress position should be strongly supported by the cues in the base. By contrast, in words with pre-stressing and auto-stressed suffixes, the stress position is different in derivatives and corresponding bases, which should result in more uncertainty about the stress position. In words with stress-shifting suffixes, cues from the suffix support the stress position in the derived word. Thus, the cues in the base will have to support multiple stress positions (at least one for the derived word and one for the base word). This is why we expect higher predictive accuracy for stress preserving suffixes than for pre-stressing and auto-stressed suffixes. We do not make any predictions about the difference between auto-stressed and pre-stressing suffixes, as they both increase the uncertainty about the stress position.

We tested these hypotheses with the help of a case study of a subset of 4,626 words that contained only words with clearly *stress preserving*, *pre-stressing*, and *auto-stressed* suffixes.[7] The stress preserving suffixes that we considered were *-ness* (as in *happi-ness*), *-less* (as in *piti-less*), and *-ly* (as in *happi-ly*). The pre-stressing suffixes that we considered were *-ion* (as in *constrict-ion* or *informat-ion*), *-ity* (as in *divin-ity*), and *-ical* (as in *satir-ical*). The auto-stressed group comprised the largest number of different suffixes, as these suffixes occur in much fewer different words in English than the suffixes belonging to the other two groups. By including a larger number of different suffixes in this group we made sure that we would have a sufficient number of data for analysis. These suffixes are *-ese*, *-teen*, *-ee*, *-ana*, *-esque*, and *-ette* (as in e.g., *Japan-ese*, *seven-teen*, *employ-ee*, *Smithsoni-ana*, *Kafka-esque*).

Like in section 4.2, we ran a generalized linear regression model with ’prediction accuracy’ as a dependent variable. The model was based on the *stress in the vowel* model that used trigrams as cues. We used the same model structure as in the section above, but excluded the predictor ‘inflected words’ from the analysis and added ‘morphological class’, which represents the different stress conditions *stress preserving*, *pre-stressing* and *auto-stressed.* Table 5 provides a model summary. Results are illustrated in Figure 2.

We observe that words with stress preserving suffixes yielded the highest accuracy score (intercept, logit = 3.71, P = 0.98), followed by pre-stressing suffixes (logit = 3.4, P = 0.97). Auto-stressed suffixes caused indeed uncertainty during classification, reducing the accuracy slightly more (logit = 2.2, P = 0.90). The effect size and direction of the effects of the number of syllables, activation and activation diversity are very similar to the preceding model (Figure 2 b-d). This indicates that the effect of network measures is valid even for a smaller data set.

Table 5: *Model summary of stress classification accuracy on different morphological stress categories.*

	Estimate	Std. Error	z value	p value
(Intercept)	3.71	0.21	18.00	< 0.001
Stress Shift = auto	-1.45	0.26	-5.61	< 0.001
Stress Shift = prestressing	-0.28	0.10	-2.92	< 0.001
Number of Syllables	-0.95	0.12	-7.97	< 0.001
Double vowels = TRUE	-0.72	0.25	-2.93	< 0.001
Number of Syllables : Double vowels = TRUE	0.51	0.13	3.89	< 0.001
Activation	1.22	0.06	20.58	< 0.001
Activation Diversity	-0.83	0.05	-15.46	< 0.001

4.5 Learning morphological stratification

So far, we have shown that the NDL network is capable of learning stress position and that the network’s certainty/uncertainty of stress position is reflected in prediction accuracy. In the following, we turn our attention to the problem of how much morphological stratification has been learned by the network. We hypothesize that the network indeed learned stratification of suffixes. Specifically, we assume that stratification will be mirrored in differences in the activation of the stress position coming from the stem and coming from the suffix.

Suffixes that attract stress (*auto stressed* suffixes) and suffixes which attract stress to the preceding syllable (*prestressing* suffixes) systematically indicate the stress position, whereas the cues in the stem discriminate variable stress positions (i.e. those of the base word and those of its derivatives). Accordingly, suffix cues should be better cues for the stress position than the stem on its own. From this we predict that stress-shifting suffixes will have a relatively higher activation than their stems. By contrast, suffixes which preserve the stress position from the stem are worse cues for the stress position than the stem. Accordingly, these suffixes should yield a lower activation than the stem.

[7] Note that running the analysis on the whole set of derived words in our CELEX dataset rather than on a set of selected derivational categories was not an option. The reason is that assignment to stratal categories is not straightforward for all suffixes. Cf. e.g. Bauer et al. (2013, chpt. 9) for discussion.

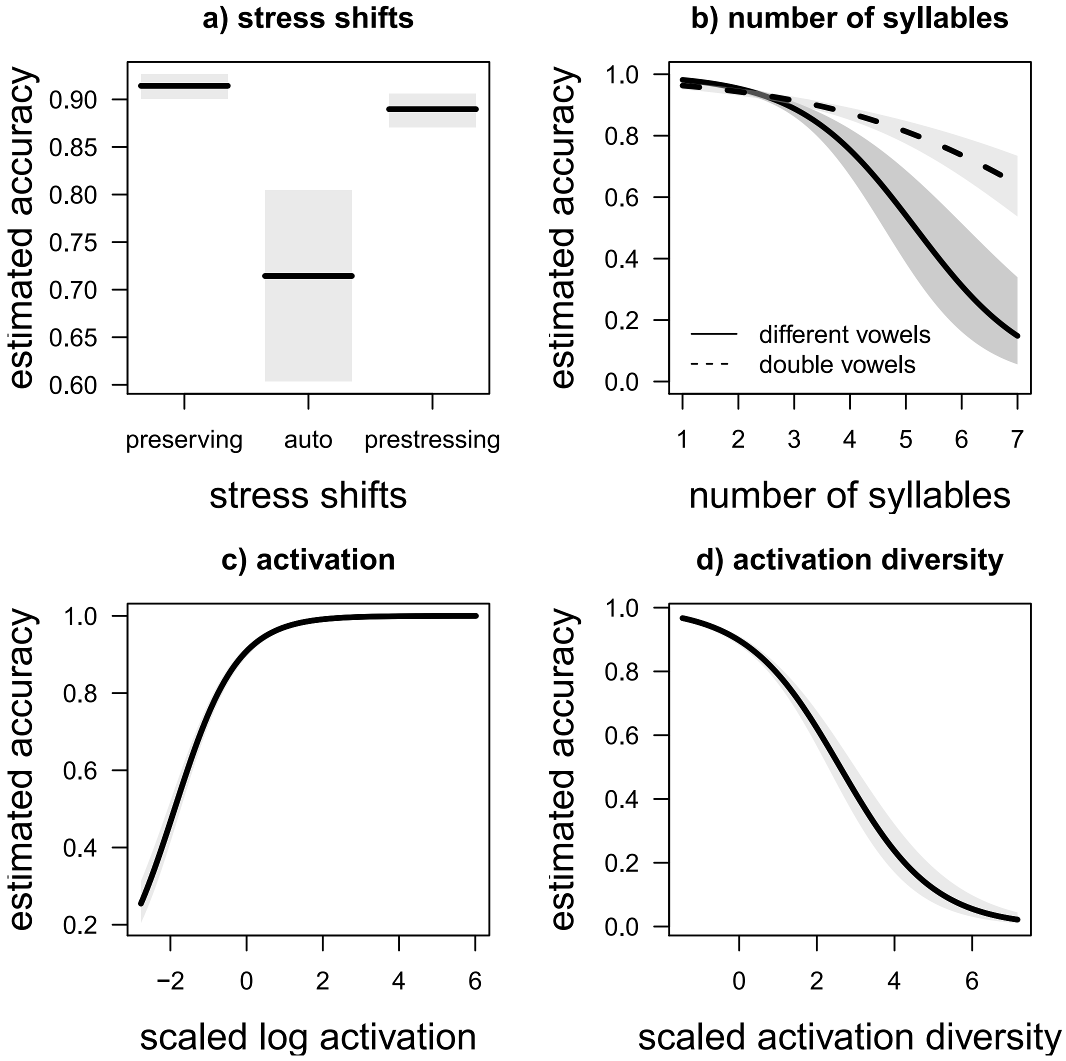

Figure 2: *Estimated prediction accuracy of the logistic regression for the 'stress in the vowel' NDL model, when the data is restricted only to derived, suffixed words. The y-axes represent the back-transformed prediction accuracy. Note that scales may vary between plots.*

We operationalized the relative support of stem and suffixes for the stress position by calculating the ratio between the activation of the stem and activation of the suffix for a word's stress position. Ratios larger than 1 indicate that the suffix has stronger activation than the stem. We based these calculations on 6,097 derived words, but excluding outliers with overly strong activation ratios (roughly 7.2%). The data were subjected to standard regression analysis, with activation ratio as the dependent variable, and stress position as a factorial predictor (with *preserving* as the reference level). Table 6 reports the model summary. Figure 3 visualizes the results.

The intercept of the model, i.e. the average activation ratio for *stress preserving* suffixes, is 0.06. We see that the levels *auto stressed* and *prestressing* yield significantly higher activation ratios than the level *stress preserving*. However, average activation ratios are always below 1, which means that the stem is more strongly activated for the word's stress position than the suffix, regardless of its stratal affiliation. A very likely explanation is that stems have on average more cues (μ = 7.2, sd = 2.5) than suffixes (μ = 2.9, sd = 1.0). As a consequence, they contribute more weights for summation than suffixes, yielding overall higher activation scores.

In spite of suffixes having smaller activations than stems, the direction of the observed effects supports our hypothesis. Stratification is indeed mirrored in the activation profiles of derived words. *Auto stressed*

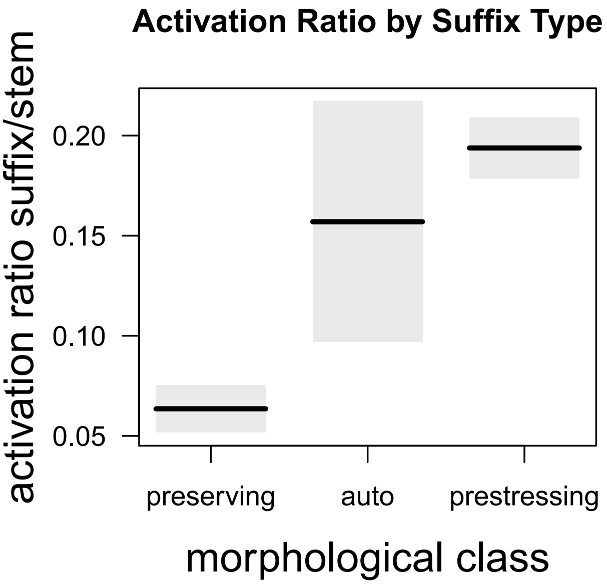

Figure 3: *Average activation ratio between suffixes and stems depending on stress shifts due to suffixation.*

and *prestressing* suffixes yield significantly higher activation ratios than *stress preserving* suffixes. In other words, stratification is reflected in the model in systematic differences in the activation of the stress positions.

Table 6: *Summary for model fitting the activation ratio (suffix/stem) as a function of stress position shifts depending on suffixation. Intercept represents 'preserving' stress.*

	Estimate	Std. Error	t value	p value
(Intercept)	0.06	0.01	10.52	< 0.001
morphological class = auto	0.09	0.03	2.99	< 0.001
morphological class = prestressing	0.13	0.01	13.23	< 0.001

5 Discussion

In the present study we set the Naive Discriminative Learner model (NDL, Arppe et al., 2018) to the task of classifying stress position in simplex and morphologically complex English words from the CELEX Lexical Database. The representation of words that the model was given as input comprised bigrams and trigrams, i.e. flat representations that encode sequences of sounds or letters and as such, intrinsically encode phonotactic information. The most important lesson to be learned from our modelling experiments is that stress position in English words can be learned extremely successfully without assuming an a-priori setting of a directionality parameter, and without an a-priori specification of morphological strata in the Mental Lexicon.

With regard to directionality, we saw that orthographic vowels provide better cues for stress position outcomes than a outcomes based on syllable count from either word edge. This finding provides a substantial challenge to existing formal accounts, which all assume that directionality is an indispensable parameter in stress assignment. The present findings also raise interesting questions about the role of orthography in stress assignment. In the present paper, orthographic representations were used as input simply because this offered a pragmatic solution to the problem that stress position and vowel quality are strongly correlated in English. Our simulations do, however, converge with previous work done in research on the acquisition of reading skills, which has provided support for the idea that orthography is indeed predictive of stress position (Arciuli et al., 2010; Abasq et al., 2019)

While this was not our aim, the present results suggest that it might be so on a larger scale than expected. English orthography has already been shown to provide informative cues about morphological structure in words (Berg, 2013). Our study indicates that its graphemic structure discriminates stress position. In order to explore this issue further, however, more research is needed to better understand how exactly trigrams encode information that is relevant for language processing. For example, a comparison of studies employing NDL to model language processing tasks seems to suggest that trigrams seem to be more successful cues than bigrams in some modelling tasks, but less successful in others (Baayen et al., 2011; Baayen and Smolka, 2020; Tucker et al., 2019). Why this is so, is not fully understood, and choice of input cues in pertinent studies (like the present one) is often opportunistic rather than motivated by considerations about theoretical plausibility. For example, to model auditory comprehension, Arnold et al. (2017) and Shafaei-Bajestan and Baayen (2018) used acoustic features, and Linke et al. (2017) used low-level visual features of letters to model visual word recognition by baboons.

With regard to morphological stratification, we saw that differences between morphological categories can be understood as differences in the activation profiles of pertinent words. Activation profiles refer to the way in which the distribution of stored weights are skewed within a word, as a result of linguistic experience when learning complex words with their stress patterns. According to this account, what speakers learn when they learn words with stress-preserving suffixes is that cues for stress are relatively stronger in the base than in the suffix. Conversely, learning stress shift in this account means learning that cues for stress position are relatively stronger in the suffix. The model therefore offers an articulate hypothesis about what underlies stratification effects. This hypothesis is testable. One prediction worth exploring is that, if stratum-specific stress behavior is emergent from activation profiles, the model should predict stress variation to occur exactly in cases in which both the suffix and its stem are strongly activated (cf. Bell (2015) for evidence that variation in English compound stress arises in similar situations). This prediction could be tested with the help of actual pronunciations of complex words, something that is clearly beyond the scope of this paper.

References

Abasq, V., Dabouis, Q., Fournier, J.-M. and Girard, I. (2019), ‘The Core of the English Lexicon: Stress and Graphophonology’, Anglophonia **27**.
URL: *https://journals.openedition.org/anglophonia/2317*

Alber, B. (2020), Word-stress in Germanic, in M. T. Putnam and B. M. Page, eds, ‘The Cambridge Handbook of Germanic Linguistics’, CUP, Cambridge.

Arciuli, J., Monaghan, P. and Seva, N. (2010), ‘Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations’, Journal of Memory and Language **63**, 180–196.

Arnold, D., Tomaschek, F., Sering, K., Lopez, F. and Baayen, R. H. (2017), ‘Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit’, PLOS ONE **12**(4), e0174623.
URL: *http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174623*

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T. and Shaoul, C. (2018), ‘ndl: Naive Discriminative Learning’.
URL: *https://CRAN.R-project.org/package=ndl*

Baayen, R. H., Milin, P., Durdevic, D. F., Hendrix, P. and Marelli, M. (2011), ‘An amorphous model for morphological processing in visual comprehension based on naive discriminative learning.’, Psychological review **118**(3), 438–481. Publisher: American Psychological Association.

Baayen, R. H., Piepenbrock, R. and van Rijn, H. (1993), The CELEX lexical database (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

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

Baayen, R. H. and Smolka, E. (2020), ‘Modeling morphological priming in German with naive discriminative learning’, Frontiers in Communication **5**, 17. Publisher: Frontiers.

Bauer, L., Lieber, R. and Plag, I. (2013), The Oxford Reference Guide to English Morphology, OUP, Oxford.

Bell, M. (2015), ‘Inter-speaker variation in compound prominence’, Lingue e Linguaggio **14**(1), 61–78.

Berg, K. (2013), ‘Graphemic alternations in English as a reflex of morphological structure’, Morphology **23**(4), 387–408. Publisher: Springer.

Bermúdez-Otero, R. (2012), The Architecture of Grammar and the Division of Labour in Exponence, in J. Trommer, ed., ‘The Phonology and Morphology of Exponence - the State of the Art’, OUP, Oxford, pp. 8–83.

Bermúdez-Otero, R. (2018), Stratal Phonology, in S. J. Hannahs and A. R. K. Bosch, eds, ‘The Routledge Handbook of Phonological Theory’, Routledge, Abingdon, OX, pp. 100–134.

Bermúdez-Otero, R. and McMahon, A. M. (2006), English phonology and morphology, in B. Aarts and A. M. McMahon, eds, ‘The Handbook of English Linguistics’, Blackwell, Oxford, pp. 382–410.

Booij, G. E. (1983), ‘Principles and parameters in prosodic phonology’, Linguistics **21**(1), 249–280.

Booij, G. and Rubach, J. (1987), ‘Postcyclic versus postlexical rules in lexical phonology’, Linguistic Inquiry **18**(1), 1–44.

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

Bybee, J. (2002), ‘Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change’, Language Variation and Change **14**(3), 261–290. Publisher: Cambridge University Press.

Bybee, J. (2011), Frequency of use and the organization of language, Oxford University Press, New York.

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

Dabouis, Q., Fournier, J.-M., Girard, I. and Lampitelli, N. (2017), Stress in English Long Verbs: Poster presented at the 25th Manchester Phonology Meeting, 25.-27.5.2017, Manchester.

Daelemans, W., Gillis, S. and Durieux, G. (1994), ‘The acquisition of stress: a data-oriented approach’, Computational Linguistics **20**(3), 421–451.

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

Divjak, D., Milin, P., Ez-zizi, A., Józefowski, J. and Adam, C. (2020), 'What is learned from exposure: an error-driven approach to productivity in language', Language, Cognition and Neuroscience **0**(0), 1–24. Publisher: Routledge _eprint: https://doi.org/10.1080/23273798.2020.1815813.
URL: *https://doi.org/10.1080/23273798.2020.1815813*

Domahs, U., Plag, I. and Carroll, R. (2014), 'Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited.', Journal of Comparative Germanic Linguistics **17**, 59–96.

Fudge, E. C. (1984), English Word-Stress, George Allen & Unwin, London.

Goldwater, S. and Johnson, M. (2003), 'Learning OT constraint rankings using a maximum entropy model'.

Graves, A., Mohamed, A.-r. and Hinton, G. (2013), 'Speech Recognition with Deep Recurrent Neural Networks', arXiv:1303.5778 [cs] . arXiv: 1303.5778.
URL: *http://arxiv.org/abs/1303.5778*

Graves, A. and Schmidhuber, J. (2005), 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures', Neural Networks **18**(5), 602–610.
URL: *http://www.sciencedirect.com/science/article/pii/S0893608005001206*

Hammond, M. (1999), The Phonology of English: A Prosodic Optimality-Theoretic Approach, Oxford University Press, Oxford, New York.

Hayes, B. (1982), 'Extrametricality and English stress', Linguistic Inquiry **13**(2), 227–276.

Hoppe, D. B., Hendriks, P., Ramscar, M. and Rij, J. v. (2020), An Exploration of Error-Driven Learning in Simple Two-Layer Networks From a Discriminative Learning Perspective, Technical report, PsyArXiv.
URL: *https://psyarxiv.com/py5kd/*

Hoppe, D. B., van Rij, J., Hendriks, P. and Ramscar, M. (2020), 'Order Matters! Influences of Linear Order on Linguistic Category Learning', Cognitive Science **44**(11).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7685149/*

Jones, D. (2006), Cambridge English Pronouncing Dictionary: Edited by Peter Roach, James Hartman & Jane Setter, CUP, Cambridge.

Kager, R. (2012), 'Stress in windows: Language typology and factorial typology', Lingua **122**(13), 1454–1493.

Kamin, L. J. (1968), Attention-like processes in classical conditioning, in M. R. Jones, ed., 'Miami symposium on the prediction of behavior', Miami University Press, Miami, pp. 9–31.

Kapatsinski, V. (2018), Changing minds changing tools: From learning theory to language acquisition to language change, MIT Press.

Kiparsky, P. (1982), 'Lexical phonology and morphology', Linguistics in the morning calm .

Kleinschmidt, D. F. and Jaeger, T. F. (2015), 'Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel.', Psychological review **122**(2), 148.

Linke, M., Bröker, F., Ramscar, M. and Baayen, R. H. (2017), 'Are baboons learning orthographic representations? Probably not', PLOS ONE **12**(8), 1–14. Publisher: Public Library of Science.

Linke, M. and Ramscar, M. (2020), 'How the Probabilistic Structure of Grammatical Context Shapes Speech', Entropy **22**(1), 90. Publisher: Multidisciplinary Digital Publishing Institute.

McCully, C. (2003), Left-hand word-stress in the history of English, in P. Fikkert and H. Jacobs, eds, 'Development in Prosodic Systems', de Gruyter, Berlin, Boston, pp. 349–393.

Milin, P., Divjak, D. and Baayen, R. H. (2017), 'A Learning Perspective on Individual Differences in Skilled Reading: Exploring and Exploiting Orthographic and Semantic Discrimination Cues', Journal of Experimental Psychology: Learning, Memory, and Cognition **43**(11), 1730–1751.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P. and Baayen, R. H. (2017), 'Discrimination in lexical decision', PLOS ONE .

Moore-Cantwell, C. (2016), The representation of probabilistic phonological patterns: Neurological, behavioral, and comp
Ph.D. dissertation, University of Massachusetts, Amherst.

Ng, A. Y. and Jordan, M. I. (2002), On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, in 'Advances in neural information processing systems', pp. 841–848.

Nixon, J. S. (2020), 'Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking', Cognition **197**, 104081.
URL: *http://www.sciencedirect.com/science/article/pii/S0010027719302549*

Pater, J. (2000), 'Non-uniformity in English secondary stress: the role of ranked and lexically specific constraints', Phonology **17**(237-274).

Pearl, L., Ho, T. and Detrano, Z. (2016), 'An argument from acquisition: Comparing English metrical stress representations by how learnable they are from child-directed speech', Language Acquisition **24**(4), 307–342.

Ramscar, M., Dye, M. and Klein, J. (2013), 'Children Value Informativity Over Logic in Word Learning -', Psychological Science **24**(6), 1017–1023.

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

Ramscar, M., Dye, M., Popick, H. M. and O'Donnell-McCarthy, F. (2011), 'The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better', PloS one **6**(7), e22501. Publisher: Public Library of Science.

Ramscar, M. and Yarlett, D. (2007), 'Linguistic Self-Correction in the Absence of Feedback: A New Approach to the Logical Problem of Language Acquisition', Cognitive Science **31**(6), 927–960. Publisher: Blackwell Publishing Ltd.

Ramscar, M., Yarlett, D., Dye, M., Denny, K. and Thorpe, K. (2010), 'The Effects of Feature-Label-Order and their implications for symbolic learning', Cognitive Science **34**(6), 909–957.

Rescorla, R. (1988), 'Pavlovian Conditioning - It's Not What You Think It Is', American Psychologist **43**(3), 151–160.

Rescorla, R. and Wagner, A. (1972), A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement, in A. H. Black and W. Prokasy, eds, 'Classical conditioning II: Current research and theory', Appleton Century Crofts, New York, pp. 64–69.

Shafaei-Bajestan, E. and Baayen, R. H. (2018), Wide Learning for Auditory Comprehension., in 'Interspeech', pp. 966–970.

Shannon, C. (1948), 'A mathematical theory of communication', The Bell System Technical Journal **27**, 379–423, 623–656.

Siegel, D. (1974), Topics in English Morphology, MIT, PhD. dissertation.

Stanton, J. and Steriade, D. (2014), 'Stress windows and Base Faithfulness in English suffixal derivatives'.

Terry, J., Ong, J. H. and Escudero, P. (2015), Passive distributional learning of non-native vowel contrasts does not work for all listeners., in 'ICPhS'.

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

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

Tucker, B. V., Sims, M. and Baayen, R. H. (2019), Opposing forces on acoustic duration, Technical report. Publisher: PsyArXiv.
URL: *psyarxiv.com/jc97w*

Wanrooij, K., Boersma, P. and van Zuijen, T. L. (2014), 'Distributional vowel training is less effective for adults than for infants. a study using the mismatch response', PloS one **9**(10), e109806.

Wanrooij, K. E. et al. (2015), Distributional learning of vowel categories in infants and adults.

Werker, J. F., Yeung, H. H. and Yoshida, K. A. (2012), 'How do infants become experts at native-speech perception?', Current Directions in Psychological Science **21**(4), 221–226.

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

Zamma, H. (2012), Patterns and Categories in English Suffixation and Stress Placement: A Theoretical and Quantitative
University of Tsukuba.