# WHEN LEXICAL STATISTICS AND THE GRAMMAR CONFLICT: LEARNING AND REPAIRING WEIGHT EFFECTS ON STRESS*

Guilherme D. Garcia (Ball State University)
To appear in *Language*

ABSTRACT

In weight-sensitive languages, stress is influenced by syllable weight. As a result, heavy syllables should attract, not repel, stress. The Portuguese lexicon, however, presents a case where weight seems to negatively impact stress: antepenultimate stress is more frequent in light antepenultimate syllables than in heavy ones. This pattern is phonologically unexpected, and appears to contradict the typology of weight and stress: it is a case where lexical statistics and the grammar conflict. Portuguese also contains gradient, not categorical, weight effects, which weaken as we move away from the right edge of the word. In this paper, I examine how native speakers' grammars capture these subtle weight effects, and whether the negative antepenultimate weight effect is learned or repaired. I show that speakers learn the gradient weight effects in the language, but do not learn the unnatural negative effect. Instead, speakers repair this pattern, and generalize a positive weight effect to all syllables in the stress domain. This study thus provides empirical evidence that speakers may not only ignore unnatural patterns, but also learn the opposite pattern.

***Keywords***: stress, weight, lexical statistics, Bayes, probabilistic grammar, MaxEnt

## 1 Introduction

Phonological learning is a central topic in phonological theory. Given the lexicon of a particular language, we are interested in determining whether (and to what extent) speakers learn robust and subtle patterns in such a lexicon from the input to which they are exposed. In this context, the relationship between lexical (or frequency) statistics and speakers' grammars can help researchers understand how different patterns are learned, and, crucially, how particular phonological biases interact with linguistic information present in the input.

*For discussion and valuable comments, thanks to Heather Goad, Morgan Sonderegger, Natália Brambatti Guzzo, and Kie Zuraw. Thanks also to the audiences at the 47th Meeting of the North East Linguistic Society and the 41st Generative Linguistics in the Old World. Finally, I wish to thank the anonymous reviewers at *Language*, whose comments greatly improved this paper.

The role of lexical statistics is the focus of usage-based approaches to phonology (e.g., Hay et al. 2003, Bybee 2006), which assume that one's grammar is the result of the different patterns found in one's lexicon. In other words, "grammar is the cognitive organization of one's experience with language" (Bybee 2006, p. 711). At the other end of the spectrum are views which assume that frequency and usage are merely a product of performance; that is, lexical statistics do not affect the grammar (e.g., Chomsky and Halle 1968). Until recently, as pointed out by Coetzee (2008, p. 248), studies about the interaction between the lexicon and the grammar were largely lacking in the literature (though see, e.g., Frisch and Zawaydeh (2001) and Hayes and Londe (2006)). As a result, little was known about what happens when frequency statistics and the grammar conflict.

Since Coetzee (2008), however, the conflicting relationship between the lexicon and the grammar has been the object of investigation of different studies (e.g., Hayes et al. 2009, Carpenter 2010, Becker et al. 2011, Skoruppa and Peperkamp 2011, Becker et al. 2012, Moreton and Pater 2012, Hayes and White 2013, Glewwe 2017, Jarosz 2017, Beguš 2018). What some of these studies have shown is that certain unnatural lexical patterns are often underlearned by speakers. Overall, they suggest that learners are biased to favor prosodically natural generalizations. Thus, not all patterns in one's lexicon are necessarily learned, and the productivity of the patterns that are learned seems to rely on their phonological naturalness (defined in terms of analytic biases across languages).

One example of an analytic bias is explored by Carpenter (2010), who investigates how phonological learning is affected by the relationship between vowel sonority and stress. Low vowels are more sonorous, and tend to attract stress (Kenstowicz 1994), whereas high vowels are not targeted by stress rules cross-linguistically (de Lacy 2002). Based on these observations, Carpenter's study compares two possible stress rules in an artificial language setting: (i) stress the first *low* vowel, else the first vowel; and (ii) stress the first *high* vowel, else the first vowel. Carpenter shows that speakers of English and French learn rule (i) significantly better than rule (ii), thus confirming a bias for more natural prosodic patterns.

Another example comes from Becker et al. (2012), who examine the case of an unnatural lexical pattern in English involving laryngeal alternation (e.g., *leaf* → *leaves*). In the English lexicon, this type of alternation is more frequent in monosyllables than in polysyllables, thus violating

initial-syllable faithfulness, a cross-linguistically supported tendency to protect word-initial syllables (Steriade 1994, Beckman 1997). In a wug test, however, speakers treat both monosyllables and polysyllables equally, and therefore do not generalize the unnatural conditioning context present in the lexicon.

The present study provides new evidence that speakers not only underlearn unnatural patterns present in their lexicon, but also *repair* such patterns. The paper focuses on two central observations about Portuguese weight effects on stress (Garcia 2017; §2). First, that such effects are gradient in the lexicon.[1] In other words, stress is not categorically determined by weight, given that the effect size of weight-sensitivity on stress location gradually decreases as we move away from the right edge of the word. Second, that the weight effect on antepenultimate syllables is *negatively* correlated with antepenultimate stress. In other words, antepenultimate stress is more frequent in words with a light (L) antepenultimate syllable (e.g., *patético* 'pathetic') than in words with a heavy (H) antepenultimate syllable (*cántico* 'chant'), a relationship that can be represented as ĹLL ≻ ĤLL for short.[2] This lexical effect is unexpected insofar as heavy syllables should not repel stress in a weight-sensitive language where weight is a crucial predictor of stress (Garcia 2017).

Given the two observations above, the main objectives of this paper are (i) to examine whether native speakers generalize the gradient sensitivity to weight present in the lexicon, and (ii) to investigate whether speakers learn the negative weight effect in antepenultimate syllables. As we will see, two separate experiments show that the weight gradience observed in the lexicon is also observed in speakers' generalizations in nonce words, which attests to the role of lexical statistics. Crucially, however, speakers' grammars do not generalize the negative effect in question. Instead, a *positive* effect is learned: ĤLL ≻ ĹLL. In other words, this is a case where the grammar trumps lexical statistics when a conflict emerges. These findings are therefore consistent with the view that learners are biased to favor more natural generalizations (e.g., Hayes and White 2013). This, in turn, suggests that the interaction between lexical statistics and the grammar can be modulated by naturalness.

[1] Similar effects have been found for Spanish (Shelton 2007), English (Ryan 2014), as well as other stress systems of European languages.

[2] Throughout the paper, I use an acute accent to represent the location of primary stress.

The analysis provided in the present paper involves a probabilistic framework, where stress is not predicted to be categorical. I employ Bayesian hierarchical models to estimate the effect of a heavy syllable in different positions in the word. These models provide probability distributions of weight effects given the experimental data analyzed. In this probabilistic approach, stress patterns are assigned probabilities on the basis of the weight profile of a given word. As will be shown, this approach is underlyingly similar to a Maximum Entropy grammar (Goldwater and Johnson 2003, Wilson 2006, Hayes and Wilson 2008, Zuraw and Hayes 2017). The flexibility of such a framework allows for a more accurate and comprehensive characterization of the experimental data analyzed, which in turn helps us better understand the extent to which the empirical data reflect the lexical patterns in Portuguese.

This paper is organized as follows: in §2, I review the weight effects in the Portuguese lexicon, as well as previous (categorical) approaches to stress in this language. I also discuss the weight asymmetry found in the stress domain in Portuguese. In §3, I outline the methods used in the paper to establish a lexical baseline and the experimental design employed. I also present the fundamental concepts of Bayesian methods, which form the basis for the statistical analysis discussed in §4. Finally, in §5, I compare the probabilistic approach proposed in this paper with an optimality-theoretic framework which also allows for probabilistic outcomes. Even though a constraint-based formalization is underlyingly similar to the probabilistic analysis in §4, I discuss important differences between them.

## 2 Stress and weight in the Portuguese lexicon

In weight-sensitive languages such as Portuguese and English, heavy syllables are not expected to repel stress (i.e., to negatively affect the likelihood of stress). In English, for example, nouns tend to have stress on a heavy penultimate syllable (*agénda, Arizóna*). If the penultimate syllable is light, stress falls on the antepenultimate syllable (*Cánada, quálity*)—this is the same stress rule found in Latin, which is therefore also classified as a weight-sensitive language. In the vast majority of such languages, the weight distinction is reported to be binary, i.e., syllables are either heavy or light (Gordon 2006).

Even though not all studies agree that weight significantly affects stress in Portuguese (e.g., Lee 1994, Cantoni 2013), most previous approaches assume that the language, like English and Latin, is sensitive to weight (Wetzels 1992, 2007, Bisol 1992, 2013, Magalhães 2004, Araújo 2007, Garcia 2017): heavy syllables in word-final position in nouns and adjectives attract stress.[3] While weight-sensitivity in Portuguese is traditionally assumed to be constrained to the word-final syllable, the patterns found in the lexicon (Houaiss et al. 2001, Garcia 2014) contradict that assumption. As shown in Garcia (2017), once we investigate a sufficiently large word list, we find that weight effects are present in *all* positions in the trisyllabic stress domain in the language. More specifically, the effect of heavy syllables on stress depends on their position in said domain: it monotonically weakens as we move away from the right edge of the word.

A second—and more surprising—characteristic of stress in the Portuguese lexicon is that heavy antepenultimate syllables negatively affect antepenultimate stress (Garcia 2017). Unlike heavy final and penultimate syllables, which positively affect final and penultimate stress, respectively, the opposite is true of the leftmost position in the stress domain. In other words, in the Portuguese lexicon, ĹLL words (e.g., *prático* 'practical') are more common than ĤLL words (e.g., *plástico* 'plastic'). Even though this finding contradicts the very definition of weight-sensitivity, it could be a result of footing optimization, given that ĹLL words can be analyzed as having an extrametrical syllable preceded by a moraic trochee (see below in Table 1): (ĹL)⟨L⟩. In contrast, ĤLL words result in more marked metrical configurations, which contain either a medial unfooted syllable, i.e., (Ĥ)L⟨L⟩, or an uneven trochee, i.e., (ĤL)⟨L⟩.

Weight-based analyses of Portuguese stress have traditionally assumed that weight is a categorical phenomenon in the language (see Araújo (2007) for a comprehensive review of the literature). In other words, syllables are either heavy or light. Heavy syllables contain a coda consonant (nasal, liquid, or /s/), a diphthong, or a nasal vowel: *pomár* 'orchard', *chapéu* 'hat', *anã́* 'dwarf (fem)'. Previous analyses of stress in this language have also assumed that weight effects are positionally constrained to the word-final syllable. As a result, the diphthong in *gáita* 'harmonica' is not expected to increase the odds of penultimate stress relative to the light penultimate syllable in a word

[3]Stress in verbs, on the other hand, is determined mostly by morphological factors (Wetzels 2007).

such as *gáta* 'female cat'. Regular stress, then, is comprised of two groups of words: (A) words with a heavy final syllable and final stress, and (B) words with a light final syllable and penultimate stress—see (1). Exceptions are shown in (2), and consist of words which deviate from (A) and (B), namely, words with final stress and a light final syllable (e.g., *café* 'coffee'), and words with penultimate stress and a heavy final syllable (e.g., *jóvem* 'young'). The third group of words with an irregular stress pattern is comprised of all words with antepenultimate stress (e.g., *parabólica* 'parabolic').

Because previous weight-based analyses of stress in Portuguese typically assume that trochaic feet determine the location of stress, the positional constraint mentioned above implies one of two alternatives: (a) either Portuguese builds moraic trochees in all positions in the stress domain (Wetzels 1992), or (b) Portuguese builds moraic *and* syllabic trochees—for final and non-final stress, respectively (Bisol 1992). In both cases, extrametricality is assumed. Alternative (a) is exemplified in Table 1.

Table 1: Word stress derivation adapted from Wetzels (1992, p. 24)

| | *pomar* 'orchard' | *gata* 'female cat' | *jovem* 'young' | *parabolica* 'parabolic' |
|---|---|---|---|---|
| Morification | $\mu.\mu\mu$ | $\mu.\mu$ | $\mu.\mu\mu$ | $\mu.\mu.\mu$ |
| Extrametricality | - | - | Final mora | Final syllable |
| Footing | po(már) | (gáta) | (jóve)⟨m⟩ | para(bóli)⟨ca⟩ |

The widely held assumptions discussed above lie at the core of regular stress in Portuguese, given in (1)—"X" represents either "H" or "L". Assumption a (weight is binary) entails that the effect that two heavy syllables have on stress is identical. Assumption b (weight effects are restricted to the word-final syllable) entails that no weight effects should be found in penultimate or antepenultimate syllables (cf. Wetzels 2007). The weight profiles listed in (1) and (2) are accompanied by their lexical proportion, calculated over all non-monosyllabic non-verbs in Houaiss et al. (2001).

Words that do not follow the generalization in (1) are considered to be irregular—shown in (2). For example, antepenultimate stress is traditionally deemed to be unpredictable, regardless of the

weight of the antepenultimate syllable, given ASSUMPTION B in (1).

(1) **Regular stress in Portuguese**

ASSUMPTION A: weight is binary. Syllables are either heavy (H) or light (L)

ASSUMPTION B: weight effects are restricted to the word-final syllable

a. Final stress when the final syllable is heavy — XXH́ (15%)
*papél* 'paper', *motór* 'engine', *aventál* 'apron'

b. Penultimate stress when the final syllable is light — XX́L (58%)
*caválo* 'horse', *castélo* 'castle', *respéito* 'respect'

(2) **Exceptional stress in Portuguese**

a. Final stress when the final syllable is light — XXĹ (3%)
*café* 'coffee', *chulé* 'foot odor', *gurú* 'guru'

b. Penultimate stress when the final syllable is heavy — XX́H (11%)
*jóvem* 'young', *nível* 'level', *fácil* 'easy'

c. Antepenultimate stress — X́XX (13%)
*parabólica* 'parabolic', *patético* 'pathetic'

Even though (1) accounts for most words in the lexicon (≈ 73%, Houaiss et al. 2001), it does not capture important facts about the so-called exceptional cases in (2). For example, within the class of X́XX words (13%), X́LL words are much more common than X́HL, X́LH, and X́HH words combined (99.2% *vs.* 0.8%). Indeed, once we examine the entire lexicon (Garcia 2014), we find that weight effects are considerably more intricate than previously assumed.

As shown in Garcia (2017), once we take into account the entire lexicon, weight effects in Portuguese are neither categorical (*contra* ASSUMPTION A) nor restricted to the word-final syllable (*contra* ASSUMPTION B). Instead, weight-sensitivity is gradient between and within syllables, and can only be understood in relative terms. For example, weight effects are weaker in penultimate syllables relative to final syllables. In other words, heavy syllables affect stress differently depending on the position they occupy in the stress domain. As a result, in this paper I refer to heavy syllables

in isolation according to their position in the stress domain, defined as $[\sigma_3\sigma_2\sigma_1]$, where $\sigma_1$ demarcates the syllable at the right edge of the word. Therefore, final heavy syllables will be represented as $H_1$. Penultimate and antepenultimate heavy syllables will be represented as $H_2$ and $H_3$, respectively. If a heavy syllable in penultimate position has a stronger positive effect on penultimate stress (i.e., is more sensitive to weight) than a heavy syllable in antepenultimate position, we can represent that relation as $H_2 > H_3$. As we will see in the next section, however, some heavy syllables can have a *negative* impact on stress in the Portuguese lexicon. In other words, instead of attracting stress, some heavy syllables seem to repel stress in the lexicon.

## 2.1 Weight asymmetry: the case of antepenultimate stress

As mentioned earlier, in the vast majority of weight-sensitive languages, syllable weight is reportedly binary (see Gordon (2006) for a typological review of weight). As we saw in (1), this generalization also applies to Portuguese insofar as the language has traditionally been analyzed as having a two-way weight distinction. Furthermore, in any given weight-sensitive language, heavy syllables are by definition expected to attract stress. In the Portuguese lexicon, however, heavy antepenultimate syllables ($H_3$) have a *negative* effect, i.e., they significantly *lower* the odds of antepenultimate stress (Garcia 2017).

The weight effect observed for antepenultimate syllables in the Portuguese lexicon means that LLL words are more likely to bear antepenultimate stress than HLL words (i.e., ĹLL ≻ ĤLL). This can be observed in Table 2, where the percentage of antepenultimate stress in HLL words (21%) is lower than that in LLL words (24%)—an effect that is statistically credible (§3). In contrast, nearly 73% of LLH words have final stress—hence the generalization in (1). As we can see, the edges of the stress domain in the Portuguese lexicon present a remarkable asymmetry regarding weight effects. In (3), I summarize the three central observations regarding weight-sensitivity in the language.

Table 2: Stress patterns in the four most representative weight profiles (≥ 3 syllables) in the Portuguese lexicon ($N = 130,555$; Garcia (2014) based on Houaiss et al. (2001))

| **Weight profile** | **Stress** | $n$ | % |
|---|---|---|---|
| LLL | Antepenultimate | 16562 | 24 |
| | Penultimate | 49641 | 71 |
| | Final | 3374 | 5 |
| HLL | Antepenultimate | 3281 | 21 |
| | Penultimate | 11813 | 76 |
| | Final | 511 | 3 |
| LHL | Antepenultimate | 10 | 0 |
| | Penultimate | 26572 | 99 |
| | Final | 390 | 1 |
| LLH | Antepenultimate | 119 | 1 |
| | Penultimate | 4852 | 26 |
| | Final | 13430 | 73 |

(3) **Weight asymmetry in the Portuguese lexicon**

OBSERVATION 1: All three syllables in the stress domain contribute to weight

OBSERVATION 2: Weight effects weaken as we move away from the right edge of the word

OBSERVATION 3: $H_3$ has a negative effect on stress, i.e., ĹLL ≻ ĤLL

GRADIENT SENSITIVITY TO WEIGHT: $H_3 < H_2 < H_1$[4]

OBSERVATION 3 in (3) also impacts the different coda consonants in antepenultimate syllables: 32% of all words with /s/ in antepenultimate coda position have antepenultimate stress. Conversely, 8.5% of all words with antepenultimate stress have /s/ in the antepenultimate coda position (Table 3). In other words, it seems that the least sonorous antepenultimate coda consonant is the most stress-attracting in the lexicon. However, once we remove words containing the most common suffixes in Portuguese (see §3.3.3), the overall percentage of /s/ decreases considerably (4.3%) relative to that of /N/ (3.9%).[5]

[4]This can be seen in Table 2, where the percentage change of antepenultimate stress from LLL to HLL is -3%, whereas the change of penultimate stress from LLL to LHL is 28%, and the change of final stress from LLL to LLH is 68%.

[5]I thank an anonymous reviewer for suggesting that the relatively high percentage of /s/ could be an artifact of morphology.

Table 3: Antepenultimate stress by antepenultimate coda consonant

| **Coda** | ***n*** | **%** |
|---|---|---|
| /l/ | 238 | 1.19 |
| /r/ | 545 | 2.72 |
| /N/ | 691 | 3.45 |
| /s/ | 1707 | 8.53 |
| ∅ | 16830 | 84.10 |

The negative weight effect in the Portuguese lexicon could be compared to some extent to what is found in Fijian, Hawaiian and Tongan (Hayes 1995, 146–149, Zuraw 2018), where long vowels in some stressed syllables shorten in order for the word to achieve an optimal parsing into feet—a rule that Hayes refers to as *trochaic shortening*.[6] Unlike these languages, however, Portuguese has no long vowels, and the negative weight effect found in antepenultimate syllables stems from both vowels (diphthongs) and coda consonants (Garcia 2017). In addition, not only are stressed vowels in Portuguese phonetically longer (Major 1985), but coda consonants are not deleted due to stress. These observations suggest that trochaic shortening is not a likely explanation for the weight effects in question.[7] Finally, as we will see in §4.1, even if the negative weight effect in the Portuguese lexicon were due to parsing optimization, the fact is that speakers do not generalize such an effect.

In summary, the weight effects found in the Portuguese lexicon (Garcia 2017) contradict traditional assumptions insofar as they are (positionally) gradient, not categorical. Furthermore, the lexicon presents a typologically unexpected effect, namely, $H_3$, which negatively impacts antepenultimate stress. As previously mentioned, the objective of the present paper is to investigate whether the lexical patterns described above are reflected in speakers' grammars. The questions investigated in the present study are given in (4).

(4) **Questions**

a. To what extent do speakers learn the gradient weight-sensitivity in the lexicon?

b. How do speakers generalize the effect of $H_3$ (i.e., ĹLL ≻ ĤLL)?

[6] See also Prince 1990 on trisyllabic shortening in Middle English.

[7] To my knowledge, no one has ever suggested that Portuguese undergoes trochaic shortening.

# 3 Methods

To investigate the questions in (4), this paper first revisits the Portuguese lexicon to establish a realistic baseline to which experimental data can be compared. Secondly, I provide data from two forced-choice experiments. I will refer to these experiments as Version A and Version B—Version B is a replication of Version A. Below I explain in detail how a lexical baseline is defined, the experimental design, and the data analysis employed in the paper.

## 3.1 Statistical analysis

In this section, I describe the statistical methods employed in the paper. In §3.3, where a lexical baseline for $H_3$ effects will be provided, 10,000-word lexica are modeled using Frequentist logistic regressions. This results in a distribution of estimates ($\hat{\beta}$) for $H_3$, where each simulation ($n$ = 10,000) generates one value for $\hat{\beta}_{H_3}$. Section 3.3 will also provide Bayesian estimates of credible parameter values ($\hat{\beta}_{H_3}$) for the entire lexicon and for the word list containing only the most frequent words in the language (Tang 2012). As we will see below, all distributions of $\hat{\beta}_{H_3}$ are very similar, and confirm the negative weight effect discussed above.

The experimental data analyzed in the present study are modeled below using Bayesian logistic regressions. Unlike the lexical baseline mentioned above, however, these data are modeled with hierarchical regressions. More specifically, all models reported in §4.1 include by-speaker intercepts as well as random (weight) slopes, and by-item random intercepts. Below, I provide a brief overview of Bayesian data analysis, given that Bayesian methods are not widely used in linguistic research, and differ considerably from traditional statistical analysis.

### 3.1.1 Bayesian data analysis

Before providing the motivation for Bayesian analysis, it is useful to briefly discuss two central concepts in traditional (i.e., Frequentist) statistics, namely, $p$ values and confidence intervals. In Null Hypothesis Significance Testing (NHST), we are provided the probability ($p$ value) of observing data that are at least as extreme as the data we observe given a parameter value (assuming that the

null hypothesis is true). In other words, NHST provides the probability of the data given a specific statistic (e.g., a $z$ value) for a particular parameter value $\theta$. This is traditionally represented as $p(data|\theta)$. If $p(data|\theta)$ is above a certain threshold (e.g., $\alpha = 0.05$), we fail to reject the null hypothesis (e.g., that $\theta = 0$).

NHST also provides confidence intervals, which are based on hypothetical future sampling: if a given experiment were repeated on several samples, the confidence interval would encompass the true population parameter $x\%$ of the time (where $x$ is normally set to 90% or 95%). Confidence intervals are frequently misinterpreted as "the probability that the *true* parameter value lies within two values". Importantly, confidence intervals are not probability distributions (unlike Bayesian credible intervals, reviewed below), and would be different for every sample. Let us now turn to Bayesian data analysis.

Bayesian reasoning is the re-allocation of credibility across possibilities (Kruschke 2015). The possibilities in question are parameter values in a given model of data. Re-allocation of credibility implies a previous state of knowledge which is updated as new evidence is observed. This previous state is referred to as *prior*. The new evidence, i.e., the data, is modeled through a distribution, which is referred to as *likelihood*. Finally, the actual re-allocation of credibility is the *posterior*. This relationship is mathematically expressed in Bayes' rule (5), where the posterior is the product of the prior and the likelihood divided (normalized) by the evidence for the data observed.

The prior, represented in (5) as $p(\theta)$, is a crucial component in Bayesian data analysis.[8] The rationale is as follows: if previous research has consistently shown an effect of a particular condition, we can incorporate this body of knowledge into our model by using informative priors. When priors are strongly informative, more data are needed for the posterior to be affected, i.e., shifted from what is expected *a priori*. This is intuitive to the extent that if a single study wishes to challenge an entire body of consistent previous work, it will require an immense amount of data to do so. On the other hand, if little is known about a particular phenomenon, or if previous work shows inconsistent results, priors can be made non-informative, in which case their effects on the posterior are negligible.

[8] See Gelman (2008) for responses to common criticisms of the subjectivity of priors in Bayesian analysis.

(5) **Bayes' rule**

$$p(\theta|data) = \frac{p(data|\theta)p(\theta)}{p(data)}$$

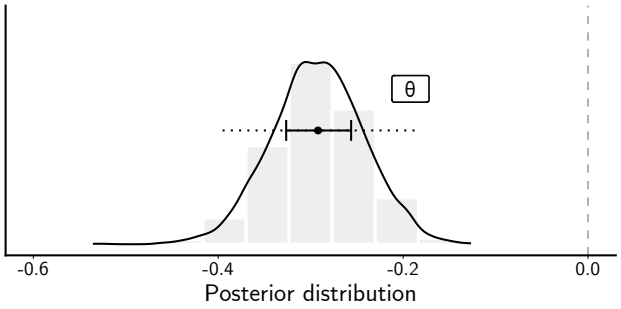
As we can see in (5), Bayesian data analysis provides the probability of a parameter value given the data, or $p(\theta|data)$, i.e., the posterior. Normally, this is in fact the question we are most interested in examining. In other words, assuming the data collected, what are the most credible parameter values? Instead of single estimates, a complete distribution of credible values is provided. The researcher can then specify a given credible interval, whose interpretation is straightforward: the values within that interval are more probable than the values outside of that interval. Crucially, credible intervals in Bayesian estimation are probability distributions, unlike confidence intervals, which means that parameter values that are located at the edges of the interval are less credible than parameter values in the center of the interval given the data (assuming that the posterior distribution is unimodal).

In realistic applications, where $n$ parameters need to be estimated, the posterior distribution cannot be analytically calculated, given that the parameter space is $n$-dimensional. For example, if we have eight parameters, each of which has 1,000 possible values, then the joint distribution has $1{,}000^8$ combinations of parameter values, which is too large a number to be computed. Instead, we approximate the distribution by randomly sampling several parameter values from it. To do that, we use sampling methods such as Markov Chain Monte Carlo (MCMC) or Hamiltonian Monte Carlo (HMC), which explore through different chains the possible values of a parameter (or combinations of parameters) in an $n$-dimensional space.[9] A typical distribution that results from a Monte Carlo simulation is shown in Fig. 1.

The mean of the distribution in Fig. 1 is $\theta = -0.29$. Because the distribution in question is practically normal, the mean is almost identical to the mode, and thus defines the most credible value in the distribution. Naturally, neighboring values such as $-0.285$ are also highly credible. For

[9]For further information on sampling methods and Bayesian data analysis more generally, see Gelman et al. (2014), Kruschke (2015) and McElreath (2016). For a general introduction to Bayesian data analysis as well as a comparison between Bayesian estimation and NHST, see Kruschke (2010, 2013).

Figure 1: Example of a posterior distribution of parameter $\theta$, and associated 50% (solid error bar) and 95% (dotted line) credible intervals




that reason, examining a distribution of credible parameter values is more informative (and realistic) than considering a single estimate: clearly −0.285 is practically just as credible as the mean in Fig. 1. In other words, examining a distribution provides a more comprehensive understanding of the credible parameter values—and it also reminds the researcher that a categorical answer tends to oversimplify the analysis.

The horizontal solid and dotted lines on top of the distribution in Fig. 1 represent the 50% and 95% credible intervals (CI) of the posterior distribution. By definition, parameter values within a given CI are more credible than parameter values outside of it—note that, because the distribution is approximately normal, we could remove the histogram and the density plot and focus solely on the CI lines and the mean of the distribution. As a decision tool, we can establish that values that are located outside of the CI are rejected (Kruschke et al. 2012).[10] In this particular case, all parameter values within the 95% CI exclude zero, i.e., we conclude that $\theta < 0$—indeed, all values in the entire distribution exclude zero. The posterior distribution in question can therefore be reported as $\theta = -0.29$, 95% CI = $[-0.40, -0.19]$.

Finally, the convergence of Bayesian models should be appropriately assessed. All Bayesian models analyzed below were diagnosed for chain convergence and Effective Sample Size (ESS; Kass et al. 1998). Furthermore, the Gelman-Rubin statistic (Brooks and Gelman 1998) was checked to ensure that between- and within-chain variance were the same. The Monte Carlo models were run using Stan (Carpenter et al. 2017) in R (R Core Team 2019).

[10] Note that any cut-off value used as the CI is *arbitrary*. In other words, there is no special reason to choose 95% over 87%, just like there is no special reason to choose $\alpha = 0.05$ over $\alpha = 0.06$ in Frequentist approaches. For that reason, categorical decisions should be interpreted with care.

In summary, Bayesian methods allow us to examine the credible parameter values given the data, which is a more meaningful and informative output than Frequentist estimates, $p$ values, and confidence intervals. The interpretation of posterior distributions is also more intuitive, in that the CIs provide the parameter values that are most consistent with the data modeled. In addition, CIs consist of actual probability distributions, unlike confidence intervals in NHST. As a result, Bayesian density intervals better estimate our uncertainty regarding parameter values given the data at hand: the wider the posterior distribution, the more uncertain we are about the parameter being modeled. Finally, a Bayesian framework is also more intuitively translated into a probabilistic grammar, where constraints weights are learned (or adjusted) given the input (e.g., Boersma 1998, Goldwater and Johnson 2003, Hayes and Wilson 2008; see also Griffiths and Tenenbaum (2006) and Lee and Wagenmakers (2014) for Bayesian applications in cognitive science). In such a grammar, lexical statistics are represented by the data, and biases to favor more natural generalizations could be represented by the prior.

### 3.2 Lexical baseline

Our starting point to define a lexical baseline is the Portuguese Stress Lexicon (Garcia 2014), which contains virtually all non-verbs in Portuguese ($n$ = 154,610). The first important question we need to ask is whether results based on such a comprehensive word list are a realistic reflection of speakers' lexica. Because speakers' lexica are, by definition, a subset of all the words in the language, we could hypothesize that a subset with a realistic number of words might not present the same weight effects found for the entire lexicon. For example, learned words and anachronisms may follow slightly different patterns, and may therefore not be present in the lexica of individual speakers. Even though the gradient weight-sensitivity patterns are overall robust (4a), the subtle effect of $H_3$ could be an artifact of an unrealistically large lexicon (4b).

One way to evaluate the weight effects in the lexicon modeled in Garcia (2017) is to simulate native speakers' lexica. For example, we can generate smaller (i.e., more realistic) lexica and model stress in each resulting subset. After $n$ simulations, we can then observe the distribution of $H_3$ effects across these subsets. If the vast majority of such smaller lexica still present weight effects

that are consistent with those observed in the comprehensive lexicon in Garcia (2017), then we have a more reliable lexical baseline that approximates a possible lexicon of an adult native speaker.

I will assume a realistic lexicon size of 10,000 words (non-verbs), which represents approximately 6% of the entire Houaiss dictionary (Houaiss et al. 2001). This is a considerably more conservative number compared to the estimate in Nagy and Anderson (1984) of 45,000 words for an average English-speaking high school graduate. The authors arrive at this number by sampling from 88,533 words that included both verbs and non-verbs. By assuming a considerably smaller lexicon, I intentionally lower the chances of finding the same weight effects that we observe for the entire lexicon.[11] At the same time, if even 10,000-word lexica show a negative effect of $H_3$, then we can be confident that speakers' lexica are highly likely to have such an effect as well.

Another simulation presented in §3.3 consists of a subset of frequent words in the Portuguese Stress Lexicon. This simulation helps us approximate the learners' input with regard to weight effects in Portuguese, which in turn allows us to determine how likely it is that learners are exposed to the negative effect of $H_3$ when building their own lexicon.

In the following section, I show the results for 10,000 simulated lexica, as well as the lexicon filtered by frequency alluded to above, both of which confirm the negative effect in question. Before we proceed, however, it is important to note that the effect of a given heavy syllable is relative not only to its position in the stress domain, but also to which stress patterns are being compared. For example, the effect of $H_1$ is stronger when final stress is compared to antepenultimate stress than when final stress is compared to penultimate stress. This is expected, given that ẊXH words are much less common than XẊH words—as shown in (2). Likewise, $H_2$ has a much stronger effect in antepenultimate *vs.* penultimate stress comparisons than in penultimate *vs.* final stress comparisons. These contrasts are crucial for interpreting estimates in statistical models, where a reference stress level is used.

[11] Note that, unlike Nagy and Anderson (1984), the simulated sublexica in question do not contain verbs.

### 3.3 Lexicon simulation

Two simulations will be examined in this section. First, to approximate adult speakers' lexica, stress will be modeled in 10,000 sublexica containing 10,000 words each. As discussed in §3.2 above, this is a conservative estimate of a speakers' lexicon size. These simulations will then be compared to the effects we find for the entire lexicon. Second, to approximate the input to which learners are exposed, stress will be modeled in frequent words.

As mentioned above (and as will be further discussed below), the statistical models employed in this paper are Bayesian logistic regressions. The 10,000 sublexica mentioned above, however, were modeled using Frequentist logistic regressions (see (6) and Fig. 2), due to the highly demanding computation involved in Bayesian models.

#### 3.3.1 Approximating speakers' lexica

Each simulated lexicon was modeled using a binomial logistic regression, where the response was either antepenultimate or penultimate stress—see examples in (6). In each model, the probability of antepenultimate stress was predicted in terms of the weight profile of each word. The effect of LHL (*vs.* LLL) is expected to disfavor antepenultimate stress (Garcia 2017). Indeed, LHL significantly disfavors antepenultimate stress in all simulated lexica (mean $\hat{\beta}_{H_2} = -12.36$).

(6) **Simple Logistic Regression ($\beta^0$ = intercept = LLL)**

$Pr(y_i = 1) = logit^{-1}(\beta^0 + HLL_i \cdot \beta_{H_3} + LHL_i \cdot \beta_{H_2})$

**Examples:**

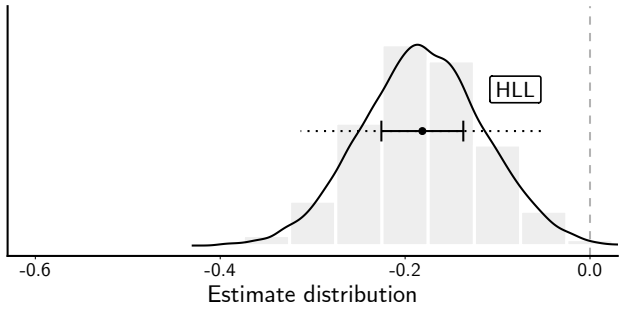
a. LLL *patético* 'pathetic': $Pr(\text{antepenultimate}) = logit^{-1}(\beta^0)$

b. HLL *cántico* 'chant': $Pr(\text{antepenultimate}) = logit^{-1}(\beta^0 + \beta_{H_3})$

c. LHL *gigánte* 'gigantic': $Pr(\text{antepenultimate}) = logit^{-1}(\beta^0 + \beta_{H_2})$

The crucial weight comparison in the simulated lexica is HLL *vs.* LLL. Recall that, in the Portuguese lexicon, HLL words are *less* likely to have antepenultimate stress relative to LLL. In other words, the estimate of HLL ($\hat{\beta}_{H_3}$) is negative, which is consistent with observation 3 in (3).

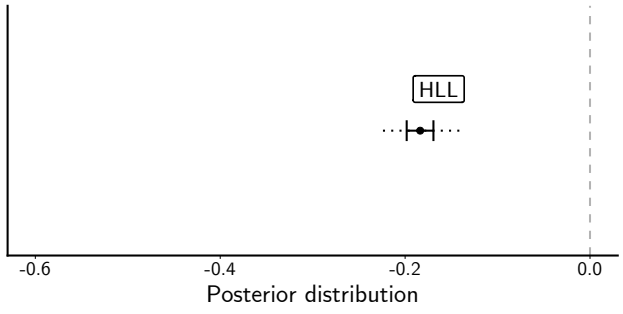
Figure 2: Simple logistic regression ($\hat{\beta}_{H_3}$) for 10,000 lexicon simulations (10,000 words each). Relative to LLL, HLL has a negative estimate in all simulations ($p < 0.0001$)




In Fig. 2, we can see a density plot of $\hat{\beta}_{H_3}$ effect sizes for all 10,000 simulated lexica. On the $x$-axis, we see a range of $\hat{\beta}$ values. Note that the estimate of HLL is negative for nearly all simulated lexica. Considering the mean of the distribution in Fig. 2 ($\hat{\beta}_{H_3} = -0.18$), $\hat{\beta}_{H_3}$ lowers the odds of antepenultimate stress by a factor of 1.2 ($e^{|-0.18|}$).

If we now compare the distribution in Fig. 2 to the posterior distribution of $\hat{\beta}_{H_3}$ effects for the entire lexicon in Fig. 3, what we see is practically the same pattern (Mean $\hat{\beta}_{H_3} = -0.18$), but a narrower distribution. The different widths in the distributions result mainly from the different samples of data used (several smaller lexica *vs.* a single large lexicon). As expected, when modeling the entire lexicon, we are more certain about the credible estimates of $\hat{\beta}_{H_3}$.[12]

Figure 3: Posterior distribution of $\hat{\beta}_{H_3}$ for the entire Portuguese lexicon and associated 50% and 95% credible intervals

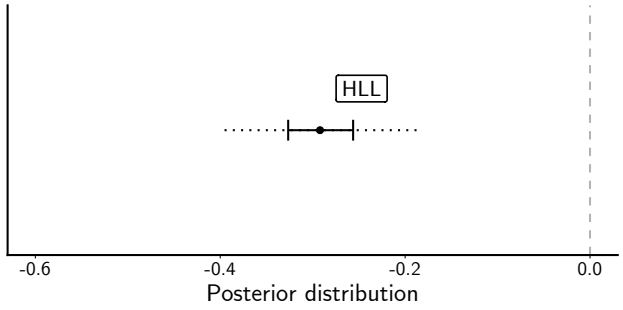



[12] A reviewer points out that vowel quality, in particular low-mid vowels (/ɛ,ɔ/), can be a confounding factor, given that such vowels must be stressed in standard Portuguese. That is correct, but even if we remove all words containing low-mid vowels, we still find the negative weight effect in question.

### 3.3.2 Approximating learners' input

We can also model only the most frequent words in the lexicon, in an attempt to approximate the input to which learners are exposed. To extract these words from the Portuguese lexicon, a frequency list was used (Tang 2012) as a filter, and the resulting frequency lexicon consisted of 22,634 words.[13] Like Fig. 3, Fig. 4[14] also plots the posterior distribution of $\hat{\beta}_{H_3}$. We can see that the negative effect of $\hat{\beta}_{H_3}$ is not only present, but is actually stronger (Mean $\hat{\beta}_{H_3} = -0.29$) relative to the effect found in the entire lexicon or in the simulated sublexica discussed above.

Figure 4: Posterior distribution of $\hat{\beta}_{H_3}$ for the most frequent non-verbs in the Portuguese lexicon (based on Tang 2012), and associated 50% and 95% credible intervals




In summary, whether we model (a) the entire lexicon or smaller and (more realistic) lexica to approximate speakers' lexica, or (b) the most frequent words to approximate the input to learners, we find the same negative effect of $\hat{\beta}_{H_3}$, i.e., ĹLL ≻ ĤLL. Thus, the negative weight effect in antepenultimate syllables is very likely to be reliably detectable in Portuguese (see discussion in §3.3.3 below). In the remainder of the paper, I examine how native speakers deal with this effect (question (4b)), which will help determine whether (and to what extent) speakers acquire the gradient sensitivity to weight referred to in §2, the topic of question (4a). Before moving on to the experimental design of the present study, I briefly discuss the possible role of morphology in the negative weight effect found in the Portuguese lexicon.

[13] These were the words in the Portuguese Stress Lexicon that were also present in the frequency list in question.
[14] The reader may remember this distribution as it was used as an example in Fig. 1.

### 3.3.3 The role of morphology

Because the lexicon of a language evolves through time and borrows words from different sources, it is expected that different lexical patterns and sub-patterns will emerge. A relevant question is what could be driving unnatural patterns such as the negative effect we observe in the Portuguese lexicon. One potential explanation could be morphology.

Even though morphology is typically associated with stress in verbs in Portuguese (§2), a number of non-verbal affixes in the language are pre-accenting, and will result in antepenultimate stress regardless of the phonotactic profile of the base to which they attach. One example is *-ico*: *báse* + *-ico* → *básico* (cf. **basíco*) 'base', 'basic'. Words ending with this suffix will have antepenultimate stress and, most of the time, a LLL weight profile—this is one of the confounding factors involving morphology, weight, and stress in the Portuguese lexicon (see below).[15] This fact about suffixes in Portuguese could be the reason why we observe more ĹLL words than ĤLL words—at least given the suffix in question.

In addition to pre-accenting suffixes, Portuguese also has a number of non-verbal suffixes which are stress-bearing, and will result in penultimate (or antepenultimate) stress—again regardless of the phonotactic profile of the base to which they attach. One example is *-óso*, which forms adjectives from nouns: *sabór* + *-óso* → *saboróso* (cf. **sabóroso*) 'taste' (*n*), 'tasty'.

Once we decide to investigate morphology in non-verbs, we inevitably run into additional confounding factors. First, highly frequent words often contain a common suffix: if we filter the 1,000 most frequent polysyllabic LLL and HLL words from the Portuguese lexicon using again the corpus in Tang (2012), nearly 62% of those words have a suffix. As a result, it seems that the more we remove morphology from a list of words, the less representative of speakers' lexica the list becomes. Second, monomorphemic words tend to be shorter, and we need at least three syllables to investigate antepenultimate weight effects. Longer monomorphemic words will be considerably less common, and thus less representative of speakers' input and lexica. Third, while some suffixes are transparent and common (e.g., *-ico*), some are definitely not (e.g., *-esía* in words like *maresía* 'sea breeze').

An important question is whether speakers can see past morphology and disentangle such con-

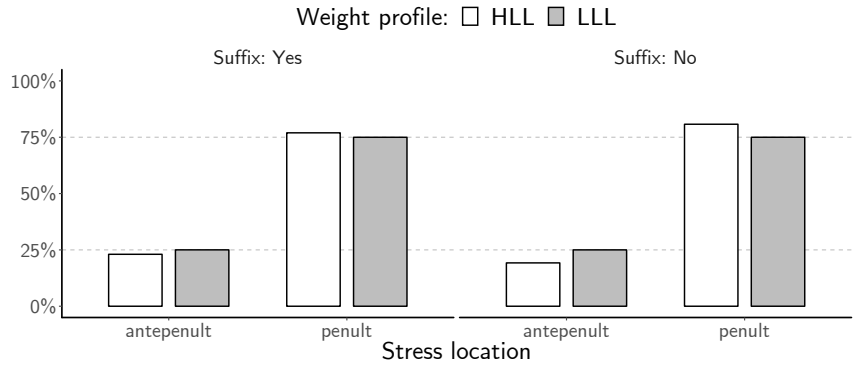
[15] HLL words ending in *-ico* include *báltico* and *asfáltico* 'Baltic', 'asphaltic'.

founding factors when they generalize weight patterns to monomorphemic nonce words. It is not clear that they can, given that it is not possible to completely isolate morphology in the lexicon without affecting crucial factors of interest at the same time—I return to this discussion in §6. With that in mind, let us now examine whether morphology is a likely cause for the negative weight effect in the Portuguese lexicon.

I will start this brief analysis by examining whether suffixes in general have a bias towards antepenultimate stress. In other words, we need to examine the representativeness of suffixes such as *-ico* above. Fig. 5 shows the distribution of stress patterns in Portuguese HLL and LLL words with and without a suffix—a comprehensive list of suffixes is provided in Appendix B. As we move from words without a suffix to words with a suffix, the percentage of antepenultimate stress goes up (by 1%), and the percentage of penultimate stress goes down (also by 1%). This shift seems to be consistent with an overall morphological bias towards antepenultimate stress in non-verbs.

The crucial question, however, is whether morphology can indirectly explain the negative weight effect observed in the Portuguese lexicon. The explanation would be indirect because suffixes will affect stress categorically, and are weight-blind. As a result, weight-sensitivity is not even computed when stress is assigned to a suffixed word. In Fig. 5, we can see that both HLL and LLL words behave similarly with regard to the presence or absence of a common suffix.

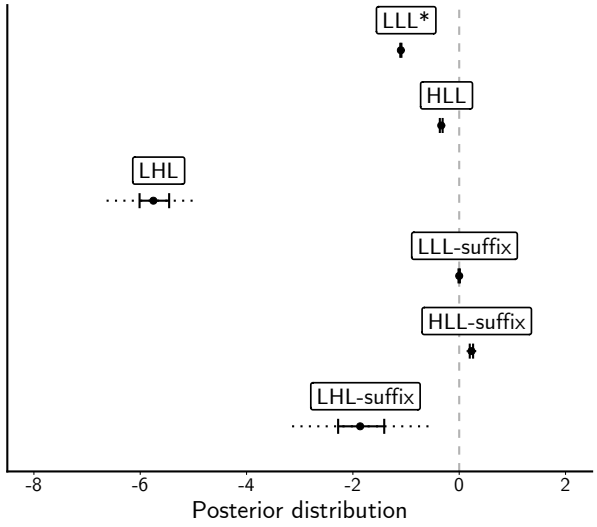
Figure 5: Stress patterns in HLL and LLL words with and without a common suffix ($n$ = 81,297)




An important pattern to note in Fig. 5 is the fact that ĹLL words are more common than ĤLL regardless of the presence of a suffix. More importantly, this trend seems to be even stronger

in words without a suffix. As a result, even if a morphological bias is statistically credible (which it is), it cannot be the main source of the negative weight effect in question. This is confirmed by the statistical model provided in Fig. 6, which includes weight, morphology, as well as their interaction as predictors of stress location.[16] In Fig. 6, the mean of the posterior distribution for LLL* (intercept; no suffix) is $\hat{\beta} = -1.10$, 95% CI = $[-1.13, -1.07]$. Relative to the intercept, HLL (no suffix) has a negative effect: mean $\hat{\beta} = -0.34$, 95% CI = $[-0.41, -0.26]$. LLL-suffix, in contrast, is centered around zero, which indicates that it is not credibly different from LLL*.

Figure 6: Posterior distributions with associated CIs (50% and 95%) for weight profile, morphology, and their interaction. Distributions must be interpreted relative to LLL* (intercept). Model specification: `stress ~ weight * suffix`




In summary, while the interaction between morphology and weight is a statistically credible predictor of stress location in the Portuguese lexicon, it does not completely account for the negative effect of a heavy antepenultimate syllable on antepenultimate stress. Therefore, it is unlikely that morphology alone would be the reason why negative weight effects exist in the Portuguese lexicon.

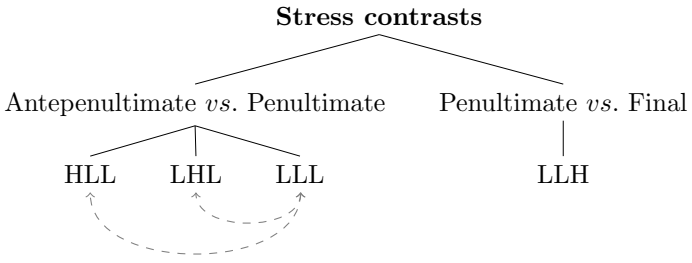
[16] Similar results were found in a separate model which was run using only the most frequent words in Tang (2012) with all common suffixes being removed from the dataset.

### 3.4 Experimental design

To examine speakers' behavior, an auditory forced-choice task was designed using Praat (Boersma and Weenink 2019) in which native Portuguese speakers ($n$ = 27 (Version A), $n$ = 32 (Version B)) were presented with pairs of trisyllabic nonce words that differed only in the position of stress. No orthographic forms were provided in order to keep participants from considering alternative pronunciations on the basis of vowel quality, which could bias their stress preference (see below). Target pairs contrasted antepenultimate and penultimate stress (e.g., *prísbade* vs. *prisbáde*). Final stress was also included (Penultimate *vs*. Final in Fig. 7) to verify whether speakers' judgments mirror the well-known robust effects of weight in word-final position (e.g., *bamésil* vs. *bamesíl*). Participants were asked which nonce word in each pair would sound more natural in Portuguese. They were also asked to judge how confident they were in their responses using a 6-point scale (1 = Not confident; 6 = Confident).

The weight profiles used in both Version A and Version B of the experiment are HLL, LHL and LLL (weight baseline) for antepenultimate *vs*. penultimate, and LLH for penultimate *vs*. final (see Fig. 7). The questions of interest are (i) whether penultimate stress is preferred in LHL words *relative to* LLL words, and (ii) whether antepenultimate stress is *dispreferred* in HLL relative to LLL words (dashed arrows in Fig. 7). By examining (i) and (ii), we address the question in (4a), namely, the extent to which speakers learn the gradient weight-sensitivity in the lexicon; by examining (ii), we address the question in (4b), namely, whether speakers' grammars generalize the negative weight effect of $H_3$. Finally, LLH words (penultimate *vs*. final) acted as controls, as they allow us to confirm the well-known word-final weight effects in Portuguese. In addition, we can also examine to what extent speakers' judgements for these words will be modulated by the fact that XX́H words are relatively common in the language (2), despite being traditionally classified as irregular.

Figure 7: Stress patterns and weight profiles of stimuli




### 3.4.1 Participants

All participants in the present study are native speakers of Brazilian Portuguese and were tested in person. Participants in Version A ($n$ = 27) were recruited in Montreal, Canada ($n$ = 14), and in southern Brazil ($n$ = 13). Those tested in Brazil had no or very little exposure to any foreign language. Those tested in Canada had relatively high levels of proficiency in English and/or French. This difference in linguistic background, however, had no effect on the results of Version A. Participants in this version of the experiment were selected on the basis of their performance on a short lexical decision task run before the actual experiment.[17] A threshold of 80% accuracy was used, which reduced the original sample size from 51 to 27. This selection criterion considerably increases the reliability of the data (e.g., as a proxy for participant attention during the experiment).

Participants in Version B (i.e., the replication of Version A) were all tested in southern Brazil ($n$ = 32). Like the participants in Version A who were tested in Brazil, none of these participants declared having fluency in any other language besides Portuguese at the time of the experiment. Importantly, Version B was not preceded by a lexical decision task: responses from all participants were analyzed, which may result in lower sensitivity relative to Version A. Thus, we can be certain that if Version A results are also replicated in Version B, the effects of interest are indeed reliable. Information on the profile of the participants is provided in Table 4.

[17] The lexical decision task was part of a different experiment, and contained trisyllabic nonce words with different stress patterns.

Table 4: Participants in Version A and Version B

| | Version A ($n$ = 27) | Version B ($n$ = 32) |
|---|---|---|
| Age | $\bar{x}$ = 30, $s$ = 8.3 | $\bar{x}$ = 26, $s$ = 5.8 |
| Gender | female = 15 | female = 18 |

### 3.4.2 Stimuli

All nonce words ($n$ = 240) contained at most one heavy syllable, and were generated by an R script (Garcia 2015). For each weight profile, approximately 200 nonce words were initially generated. These words were then ordered by their phonotactic naturalness on the basis of their bigram probabilities, which were calculated using the Portuguese Stress Lexicon (Garcia 2014). The words with the highest phonotactic probabilities in each weight profile group ($n$ = 60) were selected for the experiment (all the stimuli can be found in Appendix A). The syllabic shapes were constrained to C(C)V(C). Thus, all heavy syllables in the stimuli were either CVC or CCVC.

All the nonce words used in the experiment were preceded by a definite article: *o, a* 'the' (MASC, FEM). This ensured that the stimuli would be unambiguously interpreted as nouns. The quality of both vowels and onset/coda consonants was balanced across items, which allows us to examine whether sonority could affect stress preference in the language. All [article + nonce word] sequences were recorded in a carrier sentence by a female native speaker of Brazilian Portuguese with training in linguistics. This eliminates word-final lengthening and pitch falls, which could lead speakers to perceive light final syllables as heavy. The use of a carrier sentence also eliminates a list effect, which could result in similar problems. A template is provided in (7).

(7) **Stimulus template**

*O/A* [*palavra*] *também.* 'The [word] too.'

Brazilian Portuguese). This is important because low-mid vowels in standard Portuguese are only found in stressed syllables. For example, a nonce word such as *sostrole* was recorded as [ˈsos.tɾo.lɪ] and [sos.ˈtɾo.lɪ]. If /ɔ/ had been present in either version of the word in question, both stress *and* vowel quality would vary in these particular stimuli: [ˈsɔs.tɾo.lɪ] *vs.* [sos.ˈtɾɔ.lɪ]. By not having any low-mid vowels in the stimuli, stress was the only difference between the two versions of each nonce word in the experiment.

# 4 Data and analysis

The previous section briefly introduced Bayesian methods and their advantages over Frequentist statistics. Below, the experimental data are modeled with Bayesian logistic regressions. As will be shown, both Version A and Version B confirm that speakers generalize the gradient weight effects in Portuguese. Crucially, the results show that the negative effect of $H_3$ has not been learned. Rather, speakers' responses indicate a positive effect of $H_3$ in both experiments.

## 4.1 Experimental data

In this section, I explore and model the empirical results from Version A and Version B. As previously mentioned, these data are modeled using Bayesian hierarchical logistic regressions with by-speaker random slopes for weight effects, as well as random intercepts, and by-item random intercepts: `stress ~ weight + (1 + weight | speaker) + (1 | word)`. Segmental quality was inspected but no systematic trends were observed in the data. For example, both /u/ and /a/ are correlated with preference for antepenultimate stress in the data—but the correlations observed are not statistically credible. In addition, coda sonority does not seem to impact speakers' preferences for penultimate stress. In the case of antepenultimate stress, Version A and Version B show contradictory patterns: in the former, sonorant consonants are preferred in antepenultimate syllables when speakers choose antepenultimate stress; in the latter, it is sibilant codas which are preferred. In both cases, however, even though a trend is observed, the difference does not affect the overall effects of weight: both sonority groups favor antepenultimate stress relative to penultimate stress in Version A and Version B. Therefore, the weight effects discussed below are detected in the data

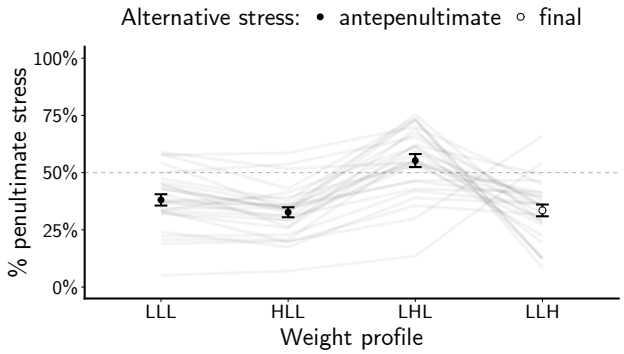
even when we take into account segmental quality and sonority in the stimuli.

#### 4.1.1 Version A

In Fig. 8, we can see the mean percentage of participants' preference for penultimate stress (and corresponding standard error bars) across the different weight profiles under consideration—gray lines represent the mean preference of each participant. As expected, speakers clearly favor final stress (over penultimate stress) in LLH words, confirming the well-known robust effect of $H_1$.

If we now turn to LHL *vs.* LLL, we can see that penultimate stress is favored by the presence of $H_2$. This can be contrasted with the preference for antepenultimate stress in LLL words—such a preference may be surprising given the literature on Portuguese stress, which often reports avoidance of antepenultimate stress (though see Araújo et al. (2011)).

Figure 8: Experimental results (Version A).
Mean response percentages by stress pattern and weight profile




It is possible that the preference for words with antepenultimate stress is associated with the more learned status of such words in the language. Given that these words are more commonly found in the speech of more educated speakers, we could ask ourselves whether participants' preferences were biased by extralinguistic factors.[18]

Even if novel words are associated with being more learned and, hence, with a preference for

[18] A reviewer suggests that the bias towards antepenultimate stress in LLL words could be, in reality, due to a possible lexical bias towards word-initial stress: if antepenultimate stress is usually word-initial in the lexicon, that could impact participants' responses, given that all stimuli were trisyllabic. However, antepenultimate stress is not initial in the vast majority of words in the lexicon (≈ 90%). The same trend is observed if we only examine LLL words (with or without common affixes from the lexicon).

antepenultimate stress, the crucial question is whether antepenultimate stress is more frequently favored in HLL relative to LLL words. In both cases antepenultimate stress is favored, but the data show a bias towards HLL words. Such a bias cannot be accounted for by extralinguistic factors: because the presence of $H_3$ is the only difference between HLL and LLL words, weight must be driving the stronger preference for antepenultimate stress in HLL words.

A question that arises is whether speakers' preference for ĤLL over ĹLL in Fig. 8 could be explained by a potential positive $H_3$ effect (ĤLL ≻ ĹLL) in the set of trisyllabic words in the lexicon. Given that all stimuli are trisyllabic, speakers could be generalizing a sub-pattern that contradicts the overall negative $H_3$ effect in the language. However, when we examine trisyllabic words in the lexicon, focusing on HLL and LLL words with antepenultimate or penultimate stress ($n$ = 17,480), we find that only 29% of words with antepenultimate stress are HLL, while 71% are LLL. As a result, if speakers' patterns mirrored the lexicon and were affected by weight effects in trisyllabic words, where antepenultimate stress is word-initial, they should prefer ĹLL to ĤLL.

To model the data in question, non-informative priors were used, as defined in (8). Because no previous experimental data exist regarding speakers' judgments of weight effects on stress in Portuguese, I assume that all regression coefficients are normally distributed around 0, with a standard deviation of $10^6$, and let the data obtained in the experiment determine what effects (i.e., parameter values) are more credible. The first model in (8) is an intercept-only model predicting final *vs.* penultimate stress in LLH words. In such a model, the intercept ($\beta^0$) indirectly reveals the effect of a final heavy syllable on stress location, and can therefore emulate the effects of $H_1$. The second model predicts antepenultimate *vs.* penultimate stress based on three different weight profiles, namely, LLL (intercept), LHL, and HLL.

(8) **Priors**

$$\textsc{Model}\text{: final } (\textit{vs.}\text{ penultimate}) = \beta^0 \sim \mathcal{N}(0, 10^6)$$

$$\textsc{Model}\text{: antepenultimate } (\textit{vs.}\text{ penultimate}) = \begin{cases} \beta^0 \sim \mathcal{N}(0, 10^6) \\ \beta_{H_3} \sim \mathcal{N}(0, 10^6) \\ \beta_{H_2} \sim \mathcal{N}(0, 10^6) \end{cases}$$

Fig. 9, shows the posterior distributions of effect sizes. Let us examine the model on the left, which predicts antepenultimate (*vs.* penultimate) stress based on the weight of antepenultimate and penultimate syllables ($H_3$, $H_2$). The baseline (intercept) is represented by LLL words, so the effects of HLL ($\hat{\beta}_{H_3}$) and LHL ($\hat{\beta}_{H_2}$) must be interpreted relative to LLL* ($\hat{\beta}^0$). The posterior distribution of LLL* is entirely positive, which means antepenultimate stress is favored relative to penultimate stress in LLL words. The posterior distribution of LHL, however, is entirely negative, which means antepenultimate stress is disfavored in LHL words (relative to LLL words). Finally, HLL has a positive posterior distribution, confirming that participants favor antepenultimate stress in HLL words (relative to LLL words). If HLL and LLL words with antepenultimate stress were seen as equally natural by participants, then the posterior distribution of HLL should be centered around zero (i.e., HLL − LLL* ≈ 0).

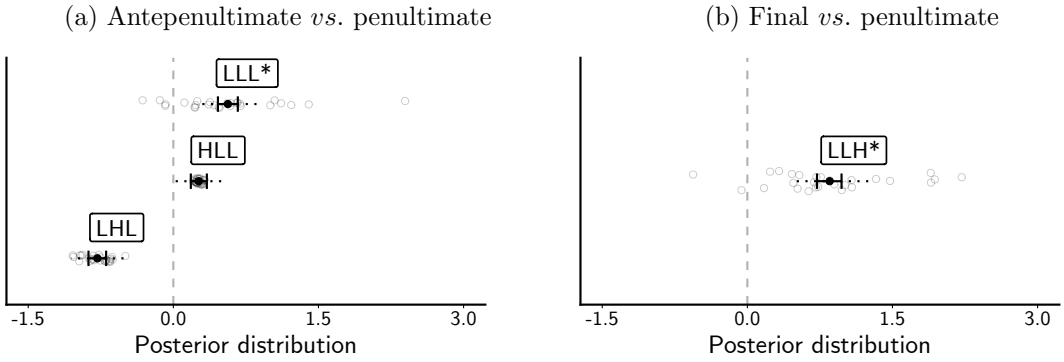
The intercept-only model on the right in Fig. 9 shows an entirely positive posterior distribution for LLH*, which indicates that, given only LLH words, participants favor final stress over penultimate stress. This is not surprising, and confirms the well-known effect of heavy final syllables in Portuguese.

The posterior distributions in Fig. 9 show that all weight effects are consistent with weight typology, and are statistically credible, i.e., all CIs exclude zero. The results show a clear gradient weight effect (i.e., $H_3 < H_2 < H_1$).[19] Note that the distribution of LLH* (Fig. 9b) is considerably wider when compared to LHL and HLL. This is exactly what we would predict if speakers' judgments mirrored the lexical statistics involving final and penultimate stress in (L)LH words, where

[19]The weight effect in LLH words is not directly comparable to words with other weight profiles, given the stress options available to participants (Fig. 7). However, I treat $H_1$ as the strongest effect given its status in the literature, which has been used to motivate the generalization in (1).

penultimate stress is relatively common in spite of its exceptional status (e.g., *jóvem* 'young', *nível* 'level', *fácil* 'easy')—as seen in (2). These results show that speakers capture the fact that the most robust weight effect in the domain ($H_1$) is also the most variable, as previously implied by (2), where XX́H words account for 11% of the lexicon.

Figure 9: Posterior distributions with associated CIs (50% and 95%) in Version A.
Distributions in (a) must be interpreted relative to LLL* (intercept).
Gray circles represent by-speaker random effects (mean estimates)



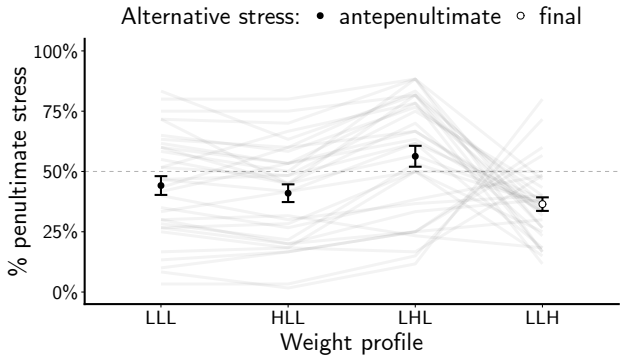

In summary, the results from Version A show that speakers' grammars generalize the gradient weight-sensitivity patterns in the language. In other words, the weight of a heavy syllable depends on its position in the stress domain (i.e., $H_3 < H_2 < H_1$). Importantly, $H_3$ has a *positive* effect on stress, which is consistent with the observation that Portuguese is sensitive to weight. Indeed, these results show that regularities can (and do) emerge from (arguably) exceptional patterns such as antepenultimate stress. I conclude that even though speakers may have a negative weight pattern in their lexica (approximated in §3.3 above), their grammars seem to override such a pattern in favor of a typologically consistent weight effect.

### 4.1.2 Version B

As mentioned in §3.1, the statistical methods employed in this paper provide a complete posterior distribution of credible parameter values given the data. Importantly, we obtain an intuitive interpretation regarding the level of certainty involved in the estimation of these parameter values (i.e., CIs). To test the reliability of the results discussed thus far, I now turn to a replication of

the experiment presented above. The replication (Version B) includes the same experimental design and statistical analysis as Version A, but consists of a new sample of native speakers of Portuguese ($n$ = 32), all of whom had had little or no exposure to foreign languages at the time of the experiment (cf. Version A). In addition, because exposure to a foreign language and level of instruction are typically correlated, this group of participants also had fewer years of formal education.

Figure 10: Experimental results (Version B).
Mean response percentages by stress pattern and weight profile

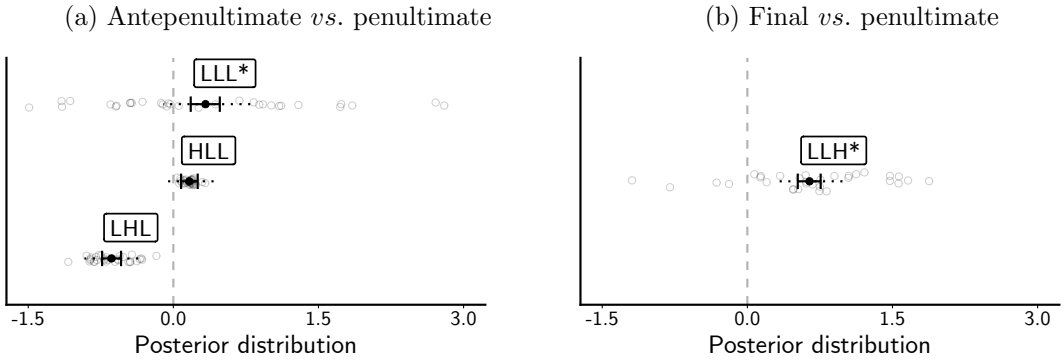



Version B results are shown in Fig. 10. We can see that speakers' responses are very similar to the responses we observed in Version A (Fig. 8). As expected from the literature, LLH words favor final stress. Crucially, as in Version A, HLL words seem to favor antepenultimate stress, and LHL words clearly favor penultimate stress. Note that the standard errors from the mean in Fig. 10 are higher relative to those in Version A, which is likely due to the fact that the group of speakers in question was not pre-tested and then filtered on the basis of their accuracy on a lexical decision task (as discussed in §3.4). To model the data in Version B, the same non-informative priors discussed above were used.

Fig. 11 provides the posterior distributions of all three weight estimates, namely, $\hat{\beta}_{H_{1-3}}$. As in Version A, the 50% CIs in all distributions in Version B exclude zero. The 95% CI for HLL is almost entirely positive, which is consistent with the results found in Version A—recall that these estimates take into account by-speaker[20] and by-item variation. In addition, the CI for LLH* is again wider

[20]Particularly important given the variation observed in Fig. 10.

Figure 11: Posterior distributions with associated CIs (50% and 95%) in Version B. Distributions in (a) must be interpreted relative to LLL* (intercept). Gray circles represent by-speaker random effects (mean estimates)




relative to LHL and HLL, which mirrors the fact that LĹH words are indeed not uncommon in the language. As we can see, Version B replicates the same statistically credible weight effects observed in Version A.

In summary, the empirical results from both Version A and Version B address the questions in (4), namely, (a) the extent to which speakers learn the gradient weight-sensitivity present in the lexicon, (b) whether speakers' grammars generalize the marked negative weight effect of $H_3$. First, speakers clearly generalize the weight-sensitivity gradience in the language to novel words. Second, the weight effect of $H_3$ shows a *positive* effect on antepenultimate stress, unlike what we see in both the entire Portuguese lexicon and in the simulated smaller lexica discussed in §3.2.

Let us assume for a moment that speakers in Version B showed a null effect of $H_3$, i.e., that ĹLL and ĤLL words were statistically equally likely (the 50% CI would unquestionably include zero in Fig. 11a). In that case, the results for Version B would mirror some of the results reported by Becker et al. (2012) for English laryngeal alternations, where polysyllables and monosyllables are treated equally by speakers in a wug test, even though alternations are more frequent among monosyllables in the English lexicon. We have seen, however, that speakers go beyond a null effect and learn the opposite pattern, ignoring the lexical statistics for $H_3$. Speakers' grammars show a predictable pattern of generalization, whereby stress is always positively and probabilistically affected by weight in the language.

## 5 Probabilistic weight effects and the grammar

In the previous sections, we saw that weight effects in Portuguese are positionally defined. This is the case if we examine the lexicon and, more importantly, it is also what we find once we investigate experimental data, where we clearly see probabilistic patterns. Crucially, we saw that speakers' grammars (i) mirror lexical patterns that are typologically consistent, and, at the same time, (ii) override a negative weight effect ($H_3$). Findings (i) and (ii) were the result of the statistical analysis in §4, where the probability of a given stress pattern depends on the weight profile under evaluation. This approach characterizes a probabilistic grammar, where different patterns are more or less likely.

One common framework employed to formalize probabilistic outcomes is a MaxEnt grammar (e.g., Goldwater and Johnson 2003, Wilson 2006, Hayes and Wilson 2008, Zuraw and Hayes 2017), which is underlyingly very similar to a logistic regression. As a result, we can easily map the analysis in §4 into a MaxEnt grammar, where we employ constraints instead of predictors. The remainder of this paper compares these two approaches and discusses important differences between them.

Unlike standard Optimality Theory (Prince and Smolensky 1993), where rankings involve strict constraint dominance, a MaxEnt grammar such as the one employed in Hayes and Wilson (2008) involves constraints with a non-negative weight $w$. The higher the weight associated to constraint $C$, the stronger its effect on lowering the probability of every candidate that violates $C$. In our case, $C$ will represent the effect of weight on stress. Our task, then, is to determine the importance (i.e., the weight) of $C$ in the grammar. Indeed, this is mathematically similar to estimating the effect size of weight (as a predictor) in a logistic regression (§4).

In order to calculate the probability of a given candidate in a Maximum Entropy Grammar, we first need to calculate the *score* of each possible output. The score of candidate $x$, denoted by $h(x)$, is the sum of all constraint violations incurred by $x$ multiplied by the weight of each violated constraint, as shown in (9).

Once each candidate has a score, a MaxEnt value ($P^*(x)$) is calculated by taking the exponential of the negated score $-h(x)$. As a result, candidates with more violations receive lower MaxEnt values. Finally, the actual probability of an output, denoted as $P(x)$, is calculated by dividing the

MaxEnt value of said output by the sum of all MaxEnt values in the candidate set $\Omega$. The definitions of a candidate's MaxEnt value and probability are provided in (10) and (11), respectively.

(9) **Candidate score in a MaxEnt grammar**

The **score** of candidate $x$, denoted by $h(x)$, is defined as $h(x) = \sum_{i=1}^{n} w_i C_i(x)$

where

$w_i$ is the weight of the $i$th constraint,

$C_i(x)$ is the number of times that $x$ violates the $i$th constraint, and

$\sum_{i=1}^{n}$ denotes the summation over all constraints $(C_1, C_2, ..., C_n)$

(10) **Candidate MaxEnt value in a MaxEnt grammar**

The **MaxEnt** value of candidate $x$, denoted as $P^*(x)$, is defined as $P^*(x) = e^{-h(x)}$

(11) **Candidate probability in a MaxEnt grammar**

The **probability** of candidate $x$, denoted as $P(x)$, is defined as $P(x) = \frac{P^*(x)}{Z}$

where $Z = \sum_{y \in \Omega} P^*(y)$

Following Ryan (2014), to represent weight effects in terms of weighted constraints, we can adapt a commonly used constraint based on the Weight-to-Stress Principle, WSP (Prince 1990), which states that heavy syllables are stressed. WSP captures weight-sensitivity by assigning violation marks to candidates that contain unstressed heavy syllables, and has been previously used in constraint-based analyses of Portuguese stress (Lee 2007).

I will, however, adapt the definition of WSP on the basis of its different *positional* effects (12). Let us assume that WSP is positionally defined in the stress domain: for example, $\text{WSP}_2$ requires that *penultimate* heavy syllables be stressed. The weight of each $\text{WSP}_n$ constraint is only meaningful in relation to the weight of other positions in the stress domain. Thus, the ranking $H_3 < H_2 < H_1$ is achieved by assigning weights to $\text{WSP}_n$ constraints such that $\text{WSP}_3 < \text{WSP}_2 < \text{WSP}_1$.

If the ranking $\text{WSP}_3 < \text{WSP}_2 < \text{WSP}_1$ is fixed across languages,[21] other constraints in the grammar can generate patterns which are not directly derivable from WSP alone—much like other predictors could be added to the statistical models in §4 to account for subtleties which are not the result of weight *per se*. For example, English stress in nouns and adjectives typically avoids word-final syllables (Hayes 1995). This can be achieved if a constraint such as NONFINALITY (*the final syllable is not stressed*) has a greater weight than $\text{WSP}_1$ in the grammar of English. Finally, note that the actual weight of a given constraint in a MaxEnt grammar only matters relative to the weight of other constraints under evaluation.

(12) **Incorporating gradient weight-sensitivity into the Weight-to-Stress Principle**

Let $n$ be a position in the stress domain.

$\text{WSP}_n$ A heavy syllable in position $n$ is stressed.

The definition in (12) requires that the stress domain be adequately defined, i.e., that only $\text{WSP}_{1-3}$ be active in the grammar of Portuguese. I will assume that this is achieved by other constraints in the grammar. The analysis that follows will therefore focus solely on the role of weight in defining stress probabilities. This will allow us to directly compare the statistical analysis in §3 with a MaxEnt implementation. The weights of $\text{WSP}_{1-3}$ used here were learned using the MaxEnt Tool (Hayes and Wilson 2008).

Let us see how stress in a LLH nonce word such as *dipramal* can be evaluated in a MaxEnt grammar. The input, in this case, contained all nonce words used in the experiment, as well as the associated frequencies of responses by stress pattern in Version A. Note, however, that such frequencies were collapsed across speakers, given that standard MaxEnt is not hierarchical (I return to this point below).

In *Tableau* 1, $\text{WSP}_1$ is violated by LĹH, where a heavy final syllable is not stressed. The more likely candidate given a LLH word is therefore LLĺ: $p(\text{LLĺ}|\text{LLH}) = 0.66$. For comparison, the mean effect size of $H_1$ (LLH*) in the posterior distribution shown in Fig. 9b is 0.85, which

[21] Where $\text{WSP}_1$ represents either the right or left edge of the word, depending on the location of the stress window in a given language (e.g., Kager 2012). Alternatively, one could employ cumulative constraints, e.g., $\{\text{WSP}_1, \text{WSP}_{1,2}, \text{WSP}_{1,2,3}\}$, in which case no fixed ranking would be required (cf. de Lacy 2004).

corresponds to $p(\text{LLH}|\text{LLH}) \approx 0.7$ ($\frac{e^{0.85}}{1+e^{0.85}}$). This leaves LĹH outputs with a probability of 0.34 ($p(\text{LĹH}|\text{LLH}) = 0.34$), which mirrors the observation that although penultimate stress in LLH words is dispreferred, it is judged as relatively natural by speakers in both Version A and Version B of the experiment discussed in the present paper.

Tableau 1: Stress assignment in a hypothetical LLH word.

| | $w = 0.68$ | | | |
|---|---|---|---|---|
| LLH | WSP$_1$ | $h(x)$ | $P^*(x)$ | $P(x)$ |
| LĹH | 1 | 0.68 | 0.51 | 0.34 |
| LLH́ | 0 | 0 | 1 | 0.66 |

Let us now turn to a LHL nonce word, *taclanda*. Recall that in these cases speakers were given two options, namely, *táclanda* and *taclánda*—for this reason, WSP$_1$ and WSP$_2$ cannot be directly compared, as noted in §4.1. The learned weight of WSP$_2$ is 0.20. In other words, this is the weight that maximizes the data observed. Given a LHL input, antepenultimate stress is predicted to win 45% of the time ($p(\text{ĹHL}|\text{LHL}) = 0.45$). This prediction does follow from the data: antepenultimate stress was chosen in 45% of the time in LHL words in Version A, and 44% of the time in Version B.

Tableau 2: Stress assignment in a hypothetical LHL word.

| | $w = 0.68$ | $w = 0.20$ | | | |
|---|---|---|---|---|---|
| LHL | WSP$_1$ | WSP$_2$ | $h(x)$ | $P^*(x)$ | $P(x)$ |
| ĹHL | 0 | 1 | 0.20 | 0.82 | 0.45 |
| LĹH | 0 | 0 | 0 | 1 | 0.55 |

One apparent difference between the statistical models in §4 and the MaxEnt implementation in this case is the absence of a baseline (though see below). In the statistical approach in §4, (non-final) stress preferences were modeled based on weight, a factor with three levels. Crucially, a reference level (LLL) was used as a baseline, which can be seen in Fig. 7. In other words, the difference between antepenultimate and penultimate stress in LHL was compared to the same difference in LLL words, where weight plays no role. This allowed us to subtract any positional bias that is

not due to weight before estimating the effects of weight *per se*. In the MaxEnt implementation discussed thus far, such a baseline level is not yet implemented, given that we have only assumed WSP: by definition, no constraint based on weight alone could be used to evaluate LLL outputs, since no candidate would violate such a constraint.

Note that the difference in probabilities between the candidate with antepenultimate stress and the candidate with penultimate stress in *Tableau* 2 is only 0.10. Taken in isolation, such a difference may not be considered substantial. However, consider that, given LLL words, antepenultimate stress is preferred 61% of the time overall in Version A. Thus, the existence of a heavy penultimate syllable decreases the probability of antepenultimate stress from 61% to 45%, and this is crucial information if we wish to accurately estimate the effects of a penultimate heavy syllable in this case. Naturally, *Tableau* 2 cannot provide such information.

The lack of a baseline becomes even more relevant when we evaluate output candidates for HLL inputs. Because LLL words show an overall preference for antepenultimate stress, removing such a baseline from the analysis results in the overestimation of $\text{WSP}_3$. In *Tableau* 3, we can see that the weight learned for $\text{WSP}_3$ is, in fact, greater than the weights learned for $\text{WSP}_{1,2}$, contradicting the statistical models in §4. For instance, given a HLL input such as *prisbade*, the probability of ĹLL (*prísbade*) is 0.67. In other words, antepenultimate stress is twice as likely to surface as penultimate stress. This is certainly a dramatic difference in probability if we do not take into account the fact that the probability of antepenultimate stress in LLL words is already high: $p(\text{ĹLL}|\text{LLL}) = 0.61$.

Tableau 3: Stress assignment in a hypothetical HLL word.

| | $w = 0.68$ | $w = 0.20$ | $w = 0.70$ | | | |
|---|---|---|---|---|---|---|
| HLL | $\text{WSP}_1$ | $\text{WSP}_2$ | $\text{WSP}_3$ | $h(x)$ | $P^*(x)$ | $P(x)$ |
| ĹLL | 0 | 0 | 0 | 0 | 1 | 0.67 |
| HĹL | 0 | 0 | 1 | 0.70 | 0.50 | 0.33 |

To fix the discrepancies observed above, other constraints in the grammar must be considered. For example, let us assume constraint $\mathcal{C}$ is introduced into our constraint set. Because $\mathcal{C}$ will simulate the positional bias in LLL words in the data, it will penalize outputs where stress does not fall on

the antepenultimate syllable.[22] Once $\mathcal{C}$ is included in our grammar ($w_{\mathcal{C}} = 0.50$), the weights of $\text{WSP}_{1-3}$ are adjusted accordingly: $w_{\text{WSP}_3} = 0.24$, $w_{\text{WSP}_2} = 0.70$, $w_{\text{WSP}_1} = 0.68$. Table 5 compares WSP weights before and after $\mathcal{C}$ with the estimates in the statistical models in §4—only $\text{WSP}_3$ and $\text{WSP}_2$ are shown as they are directly comparable (§3).

Table 5: Comparison between hierarchical regression estimates (Version A $\hat{\beta}$s in §4) and MaxEnt weights ($w$)

| *Statistical model* | | *MaxEnt Grammar* | | |
|---|---|---|---|---|
| **Predictor** | $\lvert\hat{\beta}\rvert$ | **Constraint** | $w$ **before** $\mathcal{C}$ | $w$ **after** $\mathcal{C}$ |
| HLL | 0.30 | $\text{WSP}_3$ | 0.70 | 0.24 |
| LHL | 0.79 | $\text{WSP}_2$ | 0.20 | 0.70 |

We can see that a MaxEnt grammar can formalize the effects observed in the data by employing positional WSP and additional constraints to emulate the reference level in the statistical models in §4. Naturally, the number of constraints in our MaxEnt grammar will likely be larger than the number of variables in the statistical models discussed in this paper, given that a reference level does not need to be linguistically motivated for a statistical model to be valid.

Another difference between a MaxEnt grammar and the statistical models in §4 is the hierarchical structure of the latter, which takes into account by-speaker and by-item variation—as shown in Figs. 8 and 10, by-speaker variation alone can be substantial. Given their hierarchical structure, the models in §4 also estimate weight effects for individual speakers, which are expected to exhibit some variation from the overall weight effects estimated. At the same time, speaker-level estimates are shrunk towards group-level estimates. This is desired insofar as native speakers are expected to share the grammar of Portuguese.

The hierarchical structure in the models discussed in §4 also increases our certainty regarding what is causing the stress preferences we observe in the data. By default, a MaxEnt grammar

[22] In this case, $\mathcal{C}$ could represent a group of constraints, such as FtBin (feet are binary), Align-Right(Ft, PWd) (the right edge of every foot coincides with the right edge of some prosodic word), and NonFinality. Whether the roles of these constraints are motivated in the grammar of Portuguese is beyond the scope of this paper.

does not consider such variation, even though possible implementations have been proposed in the recent literature which approximate MaxEnt and hierarchical models (e.g., Moore-Cantwell and Pater 2016, Shih and Inkelas 2016).

A third difference between the two methods compared above lies in the weight estimates themselves, which are a single point estimate in the case of standard MaxEnt, but an entire posterior distribution in the statistical methods in §4. Naturally, constraint weights can be estimated in a Bayesian fashion (e.g., Hayes et al. 2009), but such a comprehensive approach is not typically implemented in MaxEnt grammars. One advantage of a distribution of weights (rather than a single point estimate) is the possibility of inspecting the standard deviation of such a distribution, which in turn informs us about the robustness of the constraint weight being learned. This was particularly useful in interpreting the effects of $H_1$ in §4, where the wide posterior distribution reflected the fact that XX́H words are not uncommon in the Portuguese lexicon (§2).

An important fourth difference exists between a typical MaxEnt grammar and hierarchical logistic regressions: the former fails to learn statistical generalizations together with idiosyncrasies (Zymet 2018). This GRAMMAR-LEXICON BALANCING PROBLEM is the result of treating grammatical (generalizable) and lexical constraints as equally plausible (Zymet 2018, Ch. 6), which causes convergence problems in MaxEnt models as lexical constraints are quickly promoted in the grammar to explain idiosyncratic data points.

In spite of the differences highlighted above, both the statistical approach in §4 and the MaxEnt grammar discussed in this section assume that outputs are probabilistically selected given an input. Indeed, one can interpret the Bayesian models in §4 as a hierarchical version of standard MaxEnt. Both implementations capture the observation that weight effects are positionally defined based on a combination of lexical statistics and universal biases in the grammar. This combination was uncovered in the data examined above, where the grammar may or may not generalize lexical patterns in the language. On the one hand, lexical statistics play a major role in defining what weight effects are generalized: both $H_2$ and $H_1$ are reflected in speakers' grammars. On the other hand, when lexical statistics and the grammar conflict ($H_3$), we observe the prevalence of the latter.

# 6 Conclusion

The data analyzed in this paper show that both lexical statistics and the grammar play a role in phonological learning. We have seen that the gradient weight effects present in the Portuguese lexicon are acquired and generalized by native speakers. Such a generalization is, in and of itself, significant, given that morphology creates confounding factors which may affect whether weight effects are directly detectable in antepenultimate syllables. This confirms that the knowledge of stress patterns and weight-sensitivity in the language is at least in part due to lexical statistics. One example discussed above involves a sub-pattern which is traditionally classified as irregular in Portuguese, namely, (X)ẋH words such as *fácil* 'easy' and *jóvem* 'young'. These words are relatively common in the Portuguese lexicon, which is consistent with the wide posterior distribution of $H_1$: speakers generalize this patterned exception to novel words. Such effects were previously unknown, and contribute to a more accurate understanding of weight, its effects on stress in the language, as well as how sub-regular patterns are acquired and generalized.

While lexical statistics can explain some of the weight effects observed in Portuguese, it cannot explain why the negative weight effect in antepenultimate syllables ($H_3$) was not generalized to novel words in the data discussed above. What can explain such a repair is a bias that favors prosodic naturalness in the grammar: given that Portuguese stress is correlated with duration, and that heavy syllables tend to be longer, a bias towards heavy stressed syllables in *all* syllables in the stress domain is natural. These findings are therefore consistent with a view where learners are biased to favor natural generalizations (e.g., Hayes and White 2013): heavy syllables should not repel stress in a weight-sensitive language such as Portuguese. Crucially, unlike the weight effects in final and penultimate syllables, where heavy syllables attract stress, antepenultimate syllables offer a case where lexical statistics and the grammar are in conflict. As we saw above, naturalness seems to regulate the interaction between the two.

One question explored in §3.3.3 is whether the negative weight effect in the Portuguese lexicon is indeed the result of weight *per se* (i.e., an unnatural effect), or whether it is simply an artifact of morphology, given that suffixes have an effect on the location of stress in the language. Section

3.3.3 then examined the distribution of stress patterns, weight profiles, and presence or absence of a suffix. Such a distribution does not seem to support morphology as the main cause of the negative weight effect in question.

Section 3.3.3 also listed a number of confounding factors which may obfuscate the direct detection of weight effects in antepenultimate syllables. This, in turn, raises the question of whether speakers can perceive and learn any weight effects in that position. If, for example, morphology leads to indeterminacy of weight-sensitivity in antepenultimate position, then the results discussed in this paper show that speakers generalize weight effects beyond patterns present in the input (or lack thereof). Crucially, this generalization also shows a gradient sensitivity to weight—which is exactly what we find in the lexicon as a whole, despite the fact that the experimental data in the present study consisted only of monomorphemic words. In summary, speakers' grammars seem to resolve a lexical conflict (weight-sensitivity in antepenultimate position) by generalizing a pattern which is, at the same time, natural, and internally consistent with more robust patterns in the language, namely, weight-sensitivity in final and penultimate syllables.

This paper has also shown how the weight effects discussed above can be formalized in a MaxEnt grammar, where constraints are weighted and output candidates are probabilistic. We have seen that such an approach is underlyingly very similar to the statistical models in §4, which employed predictors rather than constraints. The analysis in §4 is equivalent to a hierarchical MaxEnt grammar, where by-speaker and by-item variation are taken into account to estimate the effect sizes of weight overall as well as for individual speakers. Importantly, both approaches assume a probabilistic grammar where stress patterns are more or less likely given a weight profile, which in turn provides a straightforward account of the weight-sensitivity gradience observed in the language.

One question for future research is whether negative weight effects such as those observed in the Portuguese lexicon are harder to learn in an artificial language experiment like the one in Carpenter (2010). The present paper predicts that a positive correlation between heavy syllables and stress should be easier to learn (given its naturalness) than a negative correlation. A second question is whether the findings of such an artificial language experiment would be modulated by the magnitude of the weight effects in one's native language.

## References


Araújo, G. A. (Ed.) (2007). *O acento em português: abordagens fonológicas*. São Paulo: Parábola.

Araújo, G. A., Z. O. Guimarães-Filho, L. Oliveira, and M. E. Viaro (2011). Algumas observações sobre as proparoxítonas e o sistema acentual do português. *Cadernos de Estudos Linguísticos 50*(1), 69–90.

Becker, M., N. Ketrez, and A. Nevins (2011). The surfeit of the stimulus: analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language 87*(1), 84–125.

Becker, M., A. Nevins, and J. Levine (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language 88*(2), 231–268.

Beckman, J. (1997). Positional faithfulness, positional neutralization, and Shona vowel harmony. *Phonology 14*(1), 1–46.

Beguš, G. (2018). *Unnatural phonology: A synchrony-diachrony interface approach*. Ph. D. thesis, Harvard University.

Bisol, L. (1992). O acento e o pé métrico binário. *Cadernos de Estudos Linguísticos 22*, 69–80.

Bisol, L. (2013). O acento: duas alternativas de análise. *Organon 28*(54), 281–321.

Boersma, P. (1998). *Functional Phonology: formalizing the interactions between articulatory and perceptual drives*. Ph. D. thesis, University of Amsterdam.

Boersma, P. and D. Weenink (2019). Praat: doing phonetics by computer [Computer program].

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language 82*(4), 711–733.

Cantoni, M. M. (2013). *O acento no português brasileiro: uma abordagem experimental*. Ph. D. thesis, Universidade Federal de Minas Gerais.

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology 27*(3), 345–392.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: a probabilistic programming language. *Journal of Statistical Software, Articles 76*(1), 1–32.

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

Coetzee, A. W. (2008). Grammaticality and ungrammaticality in phonology. *Language 84*(2), 218–257.

de Lacy, P. (2002). *The Formal Expression of Markedness*. Ph. D. thesis, University of Massachusetts, Amherst, MA. Distributed by GLSA.

de Lacy, P. (2004). Markedness conflation in Optimality Theory. *Phonology 21*(2), 145–199.

Frisch, S. A. and B. A. Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language* (77), 91–106.

Garcia, G. D. (2014). *Portuguese Stress Lexicon*. Comprehensive list of non-verbs in Portuguese. Available at http://guilhermegarcia.github.io/psl.html.

Garcia, G. D. (2015). *Word generator: an R script for generating nonce words*. Commented code available at http://guilhermegarcia.github.io/resources.html.

Garcia, G. D. (2017). Weight gradience and stress in Portuguese. *Phonology 34*(1), 41–79. Project materials available at http://guilhermegarcia.github.io/garcia2017.html.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis 3*(3), 445–449.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis* (3rd ed.), Volume 2. Boca Raton: Chapman & Hall/CRC.

Glewwe, E. (2017). Substantive bias in phonotactic learning: Positional extension of an obstruent voicing contrast. *Talk presented at the 53rd meeting of the Chicago Linguistic Society, Chicago, IL.*

Goldwater, S. and M. Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pp. 111–120.

Gordon, M. (2006). *Syllable weight: phonetics, phonology, typology.* New York: Routledge.

Griffiths, T. L. and J. B. Tenenbaum (2006). Optimal predictions in everyday cognition. *Psychological Science 17*(9), 767–773.

Hay, J., J. B. Pierrehumbert, and M. E. Beckman (2003). Speech perception, well-formedness, and the statistics of the lexicon. In J. Local, R. Ogden, and R. Temple (Eds.), *Papers in Laboratory Phonology*, Volume 6, pp. 58–74. Cambridge: Cambridge University Press.

Hayes, B. (1995). *Metrical Stress Theory: principles and case studies.* Chicago: University of Chicago Press.

Hayes, B. and Z. Londe (2006). Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology 23*(1), 59–104.

Hayes, B., P. Siptár, K. Zuraw, and Z. Londe (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language 85*(4), 822–863.

Hayes, B. and J. White (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry 44*(1), 45–75.

Hayes, B. and C. Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry 39*(3), 379–440.

Houaiss, A., M. Villar, and F. M. de Mello Franco (2001). *Dicionário eletrônico Houaiss da língua portuguesa.* Rio de Janeiro: Objetiva.

Jarosz, G. (2017). Defying the stimulus: acquisition of complex onsets in Polish. *Phonology 34*(2), 269–298.

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

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov Chain Monte Carlo in practice: a roundtable discussion. *The American Statistician 52*(2), 93–100.

Kenstowicz, M. (1994). *Phonology in generative grammar*. Oxford: Blackwell.

Kruschke, J. K. (2010). Bayesian Data Analysis. *Wiley Interdisciplinary Reviews: Cognitive Science 1*(5), 658–676.

Kruschke, J. K. (2013). Bayesian estimation supersedes the $t$ test. *Journal of Experimental Psychology: General 142*(2), 573–603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan* (3rd ed.). London: Academic Press.

Kruschke, J. K., H. Aguinis, and H. Joo (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods 15*(4), 722–752.

Lee, M. D. and E.-J. Wagenmakers (2014). *Bayesian cognitive modeling: a practical course*. Cambridge: Cambridge University Press.

Lee, S.-H. (1994). A regra de acento do português: outra alternativa. *Letras de Hoje 24*(4), 37–42.

Lee, S.-H. (2007). O acento primário no português: uma análise unificada na Teoria da Otimalidade. In G. A. Araújo (Ed.), *O acento em português: abordagens fonológicas*, pp. 120–143. São Paulo: Parábola.

Magalhães, J. S. d. (2004). *O plano multidimensional do acento na teoria da otimidade*. Ph. D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Major, R. C. (1985). Stress and rhythm in Brazilian Portuguese. *Language 61*(2), 259–282.

McElreath, R. (2016). *Statistical rethinking: a Bayesian course with examples in R and Stan*, Volume 122. Boca Raton: Chapman & Hall/CRC.

Moore-Cantwell, C. and J. Pater (2016). Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints. *Catalan Journal of Linguistics 15*, 53–66.

Moreton, E. and J. Pater (2012). Structure and substance in artificial-phonology learning, part I: Structure. *Language and Linguistics Compass 6*(11), 686–701.

Nagy, W. E. and R. C. Anderson (1984). How many words are there in printed school English? *Reading Research Quarterly 19*(3), 304–330.

Prince, A. (1990). Quantitative consequences of rhythmic organization. In M. Ziolkowski, M. Noske, and K. Deaton (Eds.), *Parasession on the Syllable in Phonetics and Phonology*, pp. 355–398. Chicago: Chicago Linguistic Society.

Prince, A. and P. Smolensky (1993). *Optimality Theory: constraint interaction in Generative Grammar*. Malden, MA & Oxford: Blackwell. A revision of the 1993 technical report was published in 2004 (Rutgers University Center for Cognitive Science). Available on Rutgers Optimality Archive, ROA-537.

R Core Team (2019). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ryan, K. M. (2014). Onsets contribute to syllable weight: statistical evidence from stress and meter. *Language 90*(2), 309–341.

Shelton, M. (2007). *An experimental approach to syllable weight and stress in Spanish*. Ph. D. thesis, Pennsylvania State University.

Shih, S. S. and S. Inkelas (2016). Morphologically-conditioned tonotactics in multilevel Maximum Entropy grammar. In *Proceedings of the Annual Meetings on Phonology*, Volume 3.

Skoruppa, K. and S. Peperkamp (2011). Adaptation to novel accents: feature-based learning of context-sensitive phonological regularities. *Cognitive Science 35*(2), 348–366.

Steriade, D. (1994). Positional neutralization. Unpublished manuscript. Los Angeles: University of California, Los Angeles.

Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics 24: 208–214*.

Wetzels, W. L. (1992). Mid vowel neutralization in Brazilian Portuguese. *Cadernos de Estudos Linguísticos 23*, 19–55.

Wetzels, W. L. (2007). Primary word stress in Brazilian Portuguese and the weight parameter. *Journal of Portuguese Linguistics 5*, 9–58.

Wilson, C. (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science 30*(5), 945–982.

Zuraw, K. (2018). Beyond trochaic shortening: a survey of Central Pacific languages. *Language 94*(1), e1–e42.

Zuraw, K. and B. Hayes (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Language 93*(3), 497–548.

Zymet, J. (2018). *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. Ph. D. thesis, UCLA.

# A Stimuli

Table 6: Stimuli used in Version A and Version B

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| babedral | balafo | bamezil | bamprere | baprabor | baronal | batasil | batrakoŋ |
| batratoŋ | beroko | bestedo | bibanko | bikrodar | binogal | blikanfo | bogrenda |
| bomblina | botablor | brabonko | bralino | branfora | brapesme | brasmare | brelero |
| brondale | brospeno | kabrervo | kabriza | kadrida | kaprifo | karipo | katrater |
| katrenir | semitur | siblanda | sikraba | sigroko | siminer | sinfrate | sirmika |
| sitredir | klasfika | klikarfe | klikumbo | kokuro | kogidar | kogotril | koltale |
| komadriŋ | kondito | konvade | koprobil | kortemo | kosavir | krafomo | krikombo |
| krikone | krizero | kritina | krobira | krolerpa | kuladil | kumbrosa | darnido |
| dekade | denoro | denzito | depebrir | detinsa | detubral | dikriba | dipigar |
| dipramal | dirana | ditradal | ditruspa | doroto | doteflar | dotorto | drangipa |
| drapeze | dundito | falegor | fankrela | fedado | figredo | fitilbo | fitrofa |
| fladeze | flandovo | flizonta | fontena | frakko | framera | frangile | freminso |
| frimpelo | frospato | fugrosto | fuliste | fuvosta | gadalo | ganomo | gapospe |
| ʒeninta | ʒerimul | ʒespрila | ʒestika | gimaro | glindebo | gofridor | grakolo |
| grozista | guplaro | ʒaklinko | ʒakomar | ʒapemba | ʒobarto | ʒoteror | makobaŋ |
| maglimbo | malgrodo | mampedo | marobrar | merotrer | meskriva | metanta | metribul |
| mikamiŋ | mipleska | momemir | monkriko | mopropir | mornola | mulopraŋ | muploste |
| pakrurol | padirto | padrenga | pafrizo | paleso | panvata | patrikoŋ | pekogo |
| pedenso | pempano | pengata | petritol | pibidral | pisipal | pifreno | pinalbo |
| piprogur | pizaprar | plaberme | plabunta | plarolo | plikurvo | plirame | podrido |
| popranva | porebros | potoplir | prabamo | prefanto | prempedo | prenkofa | primodo |
| prinisto | prinore | prisbade | prolurvo | putrenso | xadota | xavompa | xelaliŋ |
| xepreste | xidolo | xitelvo | xobitrir | xontruse | xumbraro | xuriko | sampodo |
| semalo | semprozo | senvide | sertrolo | setroko | sikresol | simbrime | siritral |
| sokrondo | sostrole | taklanda | tagrane | tarala | tarbita | taromil | tatremer |
| tergrame | tetrito | tetruko | tikrona | tifriro | tinalko | tinkrika | toblonso |
| tokronto | tompreda | toprizo | toputril | totrense | traduka | traduno | tredolto |
| tredonsa | trentode | trezime | tringabo | trizanga | tristuda | trolarto | trombafe |
| trostizo | truskome | tubradal | tunobral | vadronsa | vanispe | vasteko | velordo |
| venfrado | veralo | vesplako | vidatrir | volitriŋ | zagrente | zitrado | zorasto |

$N$ = 240

# B Suffixes

Table 7: List of Portuguese suffixes included in analysis

| | | | |
|---|---|---|---|
| *-aa* | *-aco* | *-aço* | *-ada* |
| *-ádego* | *-ádigo* | *-ado* | *-agem* |
| *-al* | *-algia* | *-alho* | *-alxia* |
| *-ame* | *-âmio* | *-ança* | *-ância* |
| *-ando* | *-andro* | *-ano* | *-ão* |
| *-ape* | *-ar* | *-aria* | *-ário* |
| *-aute* | *-beque* | *-bio* | *-bol* |
| *-bote* | *-caína* | *-ção* | *-cida* |
| *-cola* | *-cracia* | *-craft* | *-dade* |
| *-derma* | *-derme* | *-dermo* | *-diço* |
| *-dor* | *-ear* | *-eco* | *-ectomia* |
| *-edo* | *-eima* | *-eira* | *-eiro* |
| *-ela* | *-elo* | *-emia* | *-ença* |
| *-ência* | *-engo* | *-ense* | *-ento* |
| *-er* | *-es* | *-eta* | *-ete* |
| *-ez* | *-eza* | *-filia* | *-filo* |
| *-fobia* | *-fone* | *-fono* | *-forme* |
| *-geno* | *-grafia* | *-gráfico* | *-grafo* |
| *-ia* | *-iano* | *-ical* | *-ice* |
| *-ico* | *-iço* | *-idade* | *-ido* |
| *-ilho* | *-imento* | *-ina* | *-ingue* |
| *-inho* | *-ino* | *-isco* | *-ismo* |
| *-íssimo* | *-ista* | *-ístico* | *-ita* |
| *-ite* | *-ito* | *-ivo* | *-izar* |
| *-lândia* | *-latra* | *-latria* | *-látrico* |
| *-logia* | *-lógico* | *-logista* | *-logo* |
| *-mancia* | *-mania* | *-men* | *-mente* |
| *-mento* | *-mor* | *-móvel* | *-nomo* |
| *-nte* | *-oa* | *-oca* | *-oco* |
| *-ões* | *-oide* | *-oma* | *-onho* |
| *-orama* | *-ose* | *-oso* | *-pata* |
| *-patia* | *-pígio* | *-plastia* | *-poiese* |
| *-ptero* | *-reia* | *-rostro* | *-s* |
| *-sfera* | *-terapia* | *-tério* | *-tivo* |
| *-ucho* | *-uco* | *-uço* | *-udo* |
| *-ulho* | *-ume* | *-uo* | *-ura* |
| *-úria* | *-uxa* | *-uxo* | *-vel* |

$n$ = 148