Weight Gradience and Stress in Portuguese*

Guilherme D. Garcia
McGill University

Abstract

This paper examines the role of weight in stress assignment in the Portuguese lexicon, and proposes a probabilistic approach to stress. I show that weight effects are gradient and monotonically weaken as we move away from the right edge of the word. Such effects depend on the position of a syllable in the word as well as the number of segments the syllable contains. The probabilistic model proposed in this paper is based on a single predictor, namely, weight, and yields more accurate results than a categorical analysis, where weight is treated as binary. Finally, I discuss implications for the grammar of Portuguese.

Keywords: stress, weight, onsets, probabilistic grammar, Portuguese

1 Introduction

This paper examines Brazilian Portuguese (BP) primary stress in non-verbs,[1] and proposes a probabilistic analysis based on weight gradience in the language. Portuguese stress is constrained to the final three syllables of the word ('trisyllabic window'), although only final and penultimate stress are typically analysed as regular and productive (Hermans and Wetzels 2012). Previous research has proposed that weight-sensitivity in the language is constrained to the word-final syllable, i.e., that stress is influenced by the weight of the final syllable, but not the weight of syllables located earlier in the word (Bisol 1992, 1994). Additionally, weight-sensitivity is seen to be categorical and binary (a syllable is either heavy or light according to the shape of its rhyme).

Primary stress placement in Portuguese non-verbs can be largely explained by weight, in terms of the following generalisations (Bisol 1994): stress is final (U) if the word-final syllable is heavy—where *heavy* is defined as containing a diphthong, a nasal vowel or a coda consonant (1a)—Portuguese has no long vowels. If the word-final syllable is light, stress falls on the penult (PU) syllable (1b). Taken together, these are

[*] Thanks to Heather Goad, Morgan Sonderegger, Kie Zuraw, Michael Wagner, Natália Brambatti Guzzo, and the anonymous *Phonology* reviewers for valuable suggestions. Thanks also to the audiences at the 33rd West Coast Conference on Formal Linguistics, and the 2nd Workshop on Stress and Accent (Universiteit Leiden).

[1] BP and European Portuguese (EP) are nearly identical vis-à-vis primary stress; thus most of what follows can in principle be applied to both varieties. The main differences between the two lie in phonetics (see Frota and Vigário (2001) for a comprehensive comparison). Phonologically, both BP and EP have an almost identical phonemic inventory (see Mateus and d'Andrade (2000)), even though they respect different syllabification constraints. All transcriptions are in BP, but I use 'BP' and 'Portuguese' interchangeably in this paper, as the lexicon examined here is not limited to Brazilian Portuguese.

the regular stress patterns in the language, which are found in 72% of the lexicon (Houaiss et al. 2001). Phonetically, stress in Portuguese is highly correlated with duration (Major 1985).

(1) **Regular stress in Portuguese non-verbs**

a. *cacau* [kaˈkaw] 'cocoa' *anã* [aˈnã] 'dwarf' (f) *pomar* [poˈmaɾ] 'orchard'

b. *boca* [ˈbokɐ] 'mouth' *tonto* [ˈtõntʊ] 'dizzy' *pátio* [ˈpatʃjʊ] 'patio'

There are, however, three types of irregular stress patterns in the language: final stress when the word-final syllable is light (2a); penult stress when the word-final syllable is heavy (2b); and antepenult (APU) stress (2c).

(2) **Irregular stress in Portuguese non-verbs**

a. *café* [kaˈfɛ] 'coffee' *sofá* [soˈfa] 'sofa'

b. *nível* [ˈnivew] 'level' *míssil* [ˈmisiw] 'missile'

c. *fósforo* [ˈfɔsfoɾʊ] 'match' *n* *pérola* [ˈpɛɾolɐ] 'pearl'

Researchers have employed different mechanisms in order to accommodate the cases in (2) (Bisol 1992, Bisol 1994, Lee 2007). For example, cases (2b) and (2c) have been accounted for by segmental and syllabic extrametricality, respectively (discussed in §2). The pattern in (2a) has been explained via consonantal catalexis: *café* [ka$_\mu$ˈfɛ$_\mu$C$_\mu$]. Even though the catalectic consonant is only phonetically realised in derived forms (e.g., *cafe**t**-eira* [kafeˈtejɾa] 'coffee maker'), it bears its own mora (Hyman 1985), and stressed light word-final syllables are thus underlyingly heavy according to such analyses.

Cases such as (2a) have motivated some researchers to propose that morphological factors govern the location of stress—as an alternative to catalexis. In particular, the presence or absence of theme vowels has been argued to play an important role in determining where stress falls: most non-verbs in Portuguese are composed of a stem and a theme vowel (TV) (3b), but words such as (3a) are exceptions to that pattern, in that no theme vowel is present. By positing that regular stress in Portuguese falls on the stem-final vowel, such forms are no longer irregular.

(3) a. *jacaré* [ʒakaˈɾɛ]$_{stem}$ 'alligator'

b. *boca* [ˈbok]$_{stem}$[-ɐ]$_{\text{TV}}$ 'mouth'

Thus, existing accounts explain the location of stress in most of the lexicon (regular stress) largely by a single phonological factor, namely, syllable weight, with exceptions generally accounted for by mechanisms not directly involving weight. When we examine the lexicon of the language more closely, however, the relationship between weight and stress becomes less clear than what is traditionally assumed. As will be shown in §3, weight seems to affect stress in all syllables in the stress domain, including the irregular cases in (2), though to different degrees. For instance, antepenult stress is almost always found in words that contain light penult and light final syllables. If penult syllables are not sensitive to weight, this is an unexpected correlation. Furthermore, onsets seem to affect stress location in the lexicon (§3), which indicates that weight computation in Portuguese may not be restricted to the rhyme.

A more accurate measure of how weight is computed in Portuguese is naturally important if one wishes to have a more comprehensive understanding of how stress and weight interact in said language. In this paper, I present a probabilistic analysis that accounts for the vast majority of cases that fall into the patterns in (1) and (2). I propose that weight in Portuguese has a gradient effect on stress, which is positionally[2] and quantitatively determined. In other words, the weight effects of a given syllable depend on the position of said syllable within the word as well as the number of segments present in the syllable. As we will see, weight effects in Portuguese go beyond 'heavy' and 'light' syllables.

The analysis in this paper is developed by addressing three questions, provided in (4). Question (4a) examines whether weight in fact only plays a role word-finally in Portuguese. In the lexicon investigated here, weight seems to have some influence on all three syllables in the stress domain. Statistical models (§4) indicate that weight-sensitivity gradiently weakens as we move away from the right edge of the word. The observation that final, penultimate and antepenultimate stress are sensitive to weight (4a) shows that antepenultimate stress is not as idiosyncratic as one might think, *contra* standard views on Portuguese.

(4) a. Is weight-sensitivity only found word-finally in Portuguese?
 b. Is weight-sensitivity categorical or gradient?
 c. Do onsets contribute to weight, affecting stress likelihood in Portuguese?

Question (4b) refers to whether weight is categorical, as assumed in standard views. I show that weight is in fact *gradient*: how much each syllable is affected varies considerably, but the effects are statistically significant. Weight-sensitivity depends on the position of a given syllable within the word, and, crucially, its effect on stress monotonically weakens as we move away from the right edge of the word.

Previous research in BP is based on the assumption that onsets do not influence stress—following the

[2] For positional weight, see Gordon (2004) and Ryan (2014).

traditional view that weight is a property of the rhyme (Chomsky and Halle 1968, Liberman and Prince 1977, Halle and Vergnaud 1987, Halle and Kenstowicz 1991, Hayes 1995, among many others). Question (4c) investigates whether this assumption is appropriate for Portuguese, and, if not, how onsets might affect stress in the lexicon. Onsets do show statistically significant effects in Portuguese (§4), but not in the way we would expect from more recent studies, which have shown that onsets have a positive effect on stress in other languages (Gordon 2005, Topintzi 2010, Ryan 2014).

This paper is organised as follows: in section 2, I discuss Portuguese stress in detail and revisit analyses proposed to account for both the regular and irregular patterns found in the language. In section 3, I analyse weight and stress in a comprehensive corpus in order to answer the questions in (4). In section 4, I model the patterns in the lexicon using Binomial Logistic Regressions. Crucially, given their probabilistic nature, the predictions of the models presented here are more consistent with the actual lexical patterns than are previous analyses, which assumed categoricity. Finally, section 6 summarises the findings of the paper, and discusses directions for future work.

2 Stress in Portuguese

In this section, I discuss stress in Portuguese non-verbs, and examine both morphological (§2.1) and phonological approaches (§2.2) previously proposed to account for irregular patterns in the language. I argue that there is no compelling argument for morphological influence on non-verb stress, and therefore the analysis presented in this paper is solely based on phonological factors.

Stress in many Indo-European languages is constrained to the final three syllables of the word.[3] This is the case in Romance languages such as Italian, Portuguese, Catalan and Spanish—a trait inherited from Latin. Unlike Latin, however, stressed word-final syllables are relatively common in modern Romance languages, including Portuguese (Roca 1999). Stress in German, English and Dutch monomorphemic words also falls within a trisyllabic window (Domahs et al. 2014).

Several studies on stress in Portuguese (Câmara 1970, Major 1985, Bisol 1994, Lee 1994, Collischonn 1994, Araújo 2007, Wetzels 2007, among others) agree that primary stress in the language is relatively predictable in non-verbs with final or penult stress. On the other hand, antepenultimate stress is regarded as idiosyncratic (i.e., unpredictable), and represents less than 15% of all non-verbs in the Houaiss Dictionary corpus (Houaiss

[3] In this paper, 'word' is to be equated with Prosodic Word (PWd), defined as 'a single root plus any additional morphemes within the 'grammatical word' such that the resulting constituent exhibits the properties determined to be the crucial PWd domain properties for the language in question [...]' (Vogel 2008, p. 212). Theme vowels, for example, fall within the PWd.

et al. 2001), the most comprehensive dictionary of the Portuguese language. Words with antepenult stress have always existed in Portuguese, and although their stress profile is not regular in the language, there is no evidence suggesting that such forms are completely avoided (Araújo et al. 2007, p. 58)—though in some northeastern varieties 'this pattern has completely vanished in non-verbs' (Wetzels 2007, p. 29). Finally, antepenult stress is sometimes 'repaired' via syncope and resyllabification—as long as the resulting form obeys the phonotactic patterns in the language (see Amaral 1999): *fósforo* ⇒ [ˈfɔsfrʊ] 'match' *n.* This type of repair is found in most dialects of Brazilian Portuguese.

Antepenult stress is therefore phonologically more peripheral in the language when compared to final and penult stress, which are more common and much more productive (≈18% and ≈68% in the Houaiss corpus, respectively). As a result, it is normally assumed that a new word in the language is not likely to have antepenultimate stress (Hermans and Wetzels 2012). Rather, new words tend to have either final or penultimate stress, aside from some borrowings. The words *penalty* [ˈpenaltʃi] and *performance* [peɾˈfɔɾmãnsi], for example, are present in Portuguese dictionaries with the original stressed syllable, even though this results in stress on the antepenult syllable in both cases (once the final cluster in 'performance' is repaired). This preservation of the source language's stressed syllable is respected in the spoken language as well, despite following a disfavoured pattern in Portuguese.

Across the entire Portuguese lexicon (Houaiss et al. 2001), primary stress has both morphological and phonological components: whereas stress in verbs is lexically defined by mood, tense, person and number morphemes (see Wetzels (2007) for a review), stress in non-verbs is heavily influenced by weight (cf. Mateus and d'Andrade (2000)). The morphological aspect of stress in verbs is undisputed, but some researchers have suggested that morphological factors also play a role in stress in non-verbs (Pereira (2007) and Lee (2007), among others). These researchers assume stress in non-verbs is sensitive to both morphological and phonological factors.

Table 1 summarises the stress patterns in non-verbs. As mentioned earlier, heavy syllables ('H') may have a nasal vowel, a coda consonant, and/or a complex nucleus: *pagã* [paˈgã] 'pagan'; *valor* [vaˈloɾ] 'value'; *funil* [fuˈniw] 'funnel'. Light syllables ('L') are open and contain only one segment in the nucleus: *abacaxi* [abakaˈʃi] 'pineapple' ('X' stands for either 'H' or 'L').

Note that very few words have antepenult stress and a heavy penult or final syllable (also noted in Wetzels (2007) for a subset of cases, discussed in §2.2)—this situation is similar to what we find in Dutch (Oostendorp 2012). Almost all these cases consist of borrowings, such as *performance* [peɾ.ˈfɔɾ.mãn.si] and *propolis* [ˈpɾɔ.pʊ.lis] 'propolis'. Some of these words undergo syncope in spoken BP: *óculos* [ˈɔ.kʊ.lʊs] ⇒

[ˈɔ.klʊs] ‘glasses’.

Table 1: Portuguese stress patterns (> 1σ non-verbs) in the Houaiss lexicon (N= 163,625)

Stress pattern	**Regular**	n	%	**Irregular**	n	%
Final (U)	...X$\acute{\text{H}}$]$_{PWd}$	24,060	14.7%	...X$\acute{\text{L}}$]$_{PWd}$	5,662	3.46%
Penult (PU)	...$\acute{\text{X}}$L]$_{PWd}$	93,715	57.27%	...$\acute{\text{X}}$H]$_{PWd}$	18,546	11.33%
Antepenult (APU)				...$\acute{\text{X}}$LL]$_{PWd}$	21,367	13.05%
				...$\acute{\text{X}}$LH]$_{PWd}$	233	0.14%
				...$\acute{\text{X}}$HL]$_{PWd}$	35	0.02%
				...$\acute{\text{X}}$HH]$_{PWd}$	7	0.004%
		117,775	≈ 72%		45,850	≈ 28%

2.1 Morphological approaches to Portuguese stress in non-verbs

In this section, I review the arguments for morphological influence in non-verb stress, and argue that there is no unambiguous evidence for such an influence. Previous research has proposed that morphology plays an important role in Portuguese stress, in that theme vowels are never stressed. I show that, whether or not theme vowels have an active role in the synchronic grammar of Portuguese non-verbs, there is no convincing evidence suggesting that such vowels actually influence stress: effects often attributed to theme vowels can be accounted for by phonological factors alone.

Morphological influence on Portuguese non-verb stress has been proposed by Mateus (1983), Lee (1995, 2007) and Pereira (2007). These analyses assume that the stress domain in Portuguese non-verbs is the stem—that is, number, gender and theme vowels are not visible to stress, and therefore these morphemes are never stressed in Portuguese.

(5) a. jacaré -s [ʒakaˈrɛs] (singular: *jacaré*)
STEM PL
alligators

b. boc -a -s [ˈbokas] (singular: *boca*)
STEM FEM.TV PL
mouths

As a result, irregular final stress in Table 1 is accounted for in the following way: in a word like *jacaré*[4] (5a), for example, stress falls on the stem-final vowel (/ɛ/)—this approach entails that all words with irregular final stress are monomorphemic. A word like *boca* (5b), on the other hand, has a theme vowel (/a/), and therefore stress falls on /o/, the only vowel in the stem.

The main argument for this proposal lies in derived forms. If we add a suffix to both words above, the theme vowel is typically deleted, whereas the stem-final vowel cannot be. In (6), the diminutive suffix *-inho* [-iɲʊ] is attached to *pato* and *sofá*. In (6a), the theme vowel is deleted, yielding *patinho*; in (6b), since the word-final vowel is part of the stem, an epenthetic consonant (/z/) is inserted to avoid hiatus (Bachrach and Wagner 2007).

(6) a. pat -o -inh -o *patinho* (cf. **patoinho*) [paˈtʃiɲʊ]
STEM MASC.TV DIM MASC
'Small duck'

b. sofá -inh -o *sofazinho* (cf. **sofinho*) [ˌsofaˈziɲʊ]
STEM DIM MASC
'Small sofa'

However, example (7) shows that the situation is not as straight-forward as implied by (6). Whereas /livr-o/ should pattern exactly like /pat-o/, two forms are instead accepted, indicating the optionality of TV deletion. Such cases are less common but not rare. In addition, they seem to be more acceptable with certain lexical items than others (de Freitas and Barbosa 2013).

(7) a. livr -o -inh -o *livrinho* or *livrozinho* [liˈvɾiɲʊ] ~ [livɾʊˈziɲʊ]
STEM MASC.TV DIM MASC
'Small book'

A stem-based analysis of stress seems to be more comprehensive than a purely phonological analysis, in that it accounts for more patterns: $\ldots \mathrm{X}\text{ˈ}\mathrm{L}]_{PWd}$ words are no longer irregular, as they are in phonological approaches—rather, they simply lack a theme vowel. However, the assumptions of such an analysis are problematic. The argument in question is circular: a given vowel is stressed because it is not thematic, and

[4] The use of a diacritic (´ or ˆ) in BP orthography denotes stress irregularity—hence all three irregular patterns in Table 1 are accented, except for words with final stress ending in /u/ or /i/, as these vowels cannot be thematic.

it is not thematic because it is stressed. Note that there is nothing in the pair presented in example (6) that motivates the presence/absence of TV in present-day Portuguese—except for the location of stress. In addition, the three nominal TVs in Portuguese {a, e, o} also appear stem-finally in words like *sofá*, *dendê* and *metrô*, which have word-final stress ('sofa', 'palm oil', 'metro'). Thus, stress placement is the only way to determine whether a given vowel is (or is not) thematic.

A purely phonological alternative to theme vowels follows from the observation that, cross-linguistically, prominent segments are more likely to be preserved (Harris 2011). In Portuguese, stressed vowels are never deleted in monomorphemic or derived forms. Consequently, a word like 'sofá' could not possibly lose its stressed vowel in any derived form (see (6)). Theme vowels, on the other hand, are not stressed, which explains why they may be deleted.

There are other phonological processes in BP often said to be associated with theme vowels, such as vowel raising and external sandhi.[5] Theme vowels may raise in the language, whereas stem-final vowels cannot: *mergulh-o* [meɾˈguʎo] ⇒ [meɾˈguʎʊ] 'dive' (n), but *robô* [xoˈbo] ⇏ *[xoˈbu] 'robot'. Likewise, external sandhi is only allowed in words with a theme vowel: *camisa usada* [kaˈmizɐ uˈzada] ⇒ [kaˈmizuˈzadɐ] 'used shirt', but *jacaré amarelo* [ʒakaˈɾɛ amaˈɾɛlʊ] ⇏ *[ʒakaˈɾamaˈɾɛlʊ] 'yellow alligator'. Like vowel deletion in derived forms ((6a) and (7)), both vowel raising and sandhi can be accounted for without additional mechanisms: stressed vowels are protected, and therefore they cannot raise, be deleted in derivations, nor undergo external sandhi.

The question, thus, is whether stressed vowels are maintained because they are more prominent or because they are part of the stem. Given the facts, it is not possible to tell these two alternatives apart. The same question can be posed for other Romance languages, where the problem also arises. In fact, Roca (1999, p. 673) proposes an extrametricality rule for all Romance languages to capture the observation that theme vowels are 'invisible' to stress.

(8) **Romance Extrametricality Rule**:

Assign extrametricality to the (metrical projection of the) desinence

Roca prefaces the rule as follows: 'In the absence of evidence to the contrary, however, it is reasonable to assume that final stressless vowels are desinential'. What motivates the rule in (8) is exactly the fact that theme vowels seem to be frequently deleted in Romance languages (unlike stressed stem-final vowels).

Whether or not theme vowels exist in present-day Portuguese is beyond the scope of this paper, but their

[5] Vowel deletion across word boundaries.

alleged relevance to stress clearly bears on the questions examined here. In this section, however, I argued that there is no solid evidence that such vowels have a role in Portuguese stress. Therefore, this paper is based only on phonological factors, discussed in the next section.

2.2 Phonological approaches to Portuguese stress in non-verbs

Even if we assumed that morphological factors did impact stress in Portuguese, we would still need to consider phonological factors, which heavily influence stress in the language. In this section, I examine such factors in more detail, focusing on weight and how it affects the stress patterns found in the language. I briefly review previous analyses of stress in Portuguese, which employ different mechanisms to account for stress irregularities. Finally, I provide independent evidence for weight effects in Portuguese.

Previous analyses of Portuguese stress all make reference to syllabic constituency. In view of this, I first describe syllable shape in the language. In Brazilian Portuguese, only two segments can occupy the onset position (see Fig. 1). Onset clusters consist of stop+liquid or labial fricative+liquid sequences. A word such as *macabro* ‘macabre’, for example, can only be syllabified as [ma.ˈka.bɾo] (cf. *[ma.ˈkab.ɾo]). In other words, stop+liquid clusters in Portuguese are not ambiguous vis-à-vis their syllabification (Cristófaro-Silva 2005). Finally, rhymes in Portuguese normally contain up to four segments.

Very few words violate these syllabic restrictions (borrowings, proper names etc.), some of which are listed in the Houaiss Dictionary (Houaiss et al. 2001). These cases, however, are phonotactically adapted in spoken BP, mostly via epenthesis (e.g., the borrowing *skate* is produced as [isˈkejt͡ʃi]). Recent borrowings are not the only words that are repaired: well-established words also undergo epenthesis and resyllabification if they violate the syllabic template in Portuguese: *advogado* ‘lawyer’ and *obstetra* ‘obstetrician’, for example, are normally produced as [ad͡ʒi.vo.ˈga.dʊ]/[ade.vo.ˈga.dʊ] and [o.bis.ˈtɛ.tɾɐ] in BP, respectively.

The syllabification algorithm in Portuguese is straightforward and unambiguous, given the restricted number (and quality) of segments in complex onsets and codas (see Thomas (1974) for a comprehensive description and Neto et al. (2015) for a computational implementation). A nonce word such as *pantridocra*, for example, is unambiguously syllabified as /pan.tɾi.do.kɾa/. How such a word is actually produced will vary considerably between BP and EP (Mateus and d’Andrade 1998). Take the word *devedor* ‘debtor’, which is syllabified as /de.ve.ˈdoɾ/. In colloquial EP, where vowels are frequently deleted, such a word is often produced as [dvdor]. This type of reduction never happens in BP (as we have seen, certain coda-onset sequences often undergo epenthesis).

In Fig. 1, on-glides and off-glides are treated identically: both contain two × slots. Under a categorical view, this entails that both rising and falling diphthongs are heavy. This contrasts with standard approaches to stress in Portuguese, which treat rising diphthongs as light (Bisol 2013). The present paper, however, does not distinguish rising and falling diphthongs. One reason is empirical: As Harris (1983, p. 11) shows for Spanish, rising diphthongs also seem to contribute to weight in some positions in the word. Antepenult stress in Spanish (e.g., *teléfono* 'telephone') is blocked when another syllable in the stress domain is heavy. As a result, a word such as **teléfosno* or **teléfoino* is not found in Spanish. Interestingly, a word such as **teléfiono* ([te.ˈle.fjo.no]) is not found in the language either. All these generalisations, and critically the last one, also hold for Portuguese. This indicates that, like coda consonants, both types of diphthongs affect weight to some degree, a fact which is consistent with the representation in Fig. 1. Even though the overall findings of the present paper do not hinge on the particular choice of on-glide treatment, I return to this discussion in §4.1.3, where I provide another reason for not differentiating on-glides and off-glides in the probabilistic approach proposed in this paper.

Figure 1: Syllabic structure in Portuguese

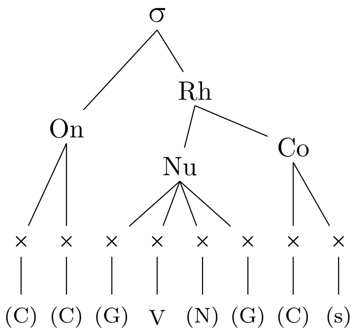


Fig. 1 implies that at most four segments can occupy the nucleus of the syllable in Portuguese: *gato*, *gaita*, *guaipeca*, *bastião* ('cat', 'harmonica', 'mongrel', 'bastion'). This, however, depends on how one treats nasality (*fã* 'fan'), in particular nasal diphthongs (*patrão* 'boss') and triphthongs (*bastião*). This paper assumes the standard approach to nasality in Portuguese, according to which a word such as *fã* is underlyingly bisegmental: /faN/ (see Battisti (1997) for a review). Since nasality is realised on the vowel, if the same assumption is extended to diphthongs and triphthongs, then certain syllable nuclei in Portuguese may contain up to four segments (e.g., *bastião*: /bas.tja~w/). Assuming this standard representation, minimal pairs such as *mão* ('hand') and *mau* ('bad') can be quantitatively differentiated: the former contains three nuclear segments, while the latter contains two nuclear segments.[6]

[6] The analysis proposed in §4 takes this quantitative difference into account. However, this is not a crucial aspect of the

Traditionally, the concept of weight has been seen as relevant only for the presence of rhyme segments—thus excluding onsets from the computation of weight (Halle and Vergnaud 1980, Hyman 1985, Hayes 1989, among others). Portuguese is an example of a language that is analysed as such: as mentioned earlier, a heavy syllable contains a diphthong, a nasal vowel or a coda consonant; onset structure is seen to be irrelevant. However, some studies show that onsets also have an impact on stress in several languages, suggesting at least some contribution to the calculation of weight (Davis 1988, Gordon 2005, Topintzi 2010, Ryan 2011, 2014).

To my knowledge, thus far no researcher has proposed a role for onsets in Portuguese stress. However, in southeastern varieties of BP, onset clusters are often simplified in unstressed syllables (Harris 1997): *prato* [ˈpɾatʊ] *-inho* [iɲʊ] ⇒ [paˈt͡ʃiɲʊ] ‘plate’, ‘small plate’. In other words, complex onsets are preferred in more prominent positions. This simplification is relatively common in some spoken BP varieties: words such as *próprio* [ˈpɾɔpɾjʊ] are sometimes produced as [ˈpɾɔpjʊ] ‘proper’. In addition, onset metathesis is observed in words such as *obstetra* [ob(i)sˈtɛtɾɐ] ⇒ [ob(i)sˈtɾɛtɐ] ‘obstetrician’. Despite the apparent correlation between onset clusters and stressed syllables in such processes, Cristófaro-Silva (2005) argues that cluster reduction is not in fact phonologically conditioned. She shows that cluster simplification may occur in both stressed and unstressed syllables, which suggests that word-level prominence is not the underlying cause for the process in question.

The proposal that onset cluster simplification is not related to stress does not necessarily mean that onsets do not affect stress. Since no study has directly examined the impact of onsets on Portuguese stress, all weight-based analyses thus far only focus on rhymes (Bisol 1994, Lee 1994, Wetzels 2007, Bisol 2013, among others), given the assumptions of standard Moraic Theory (Hyman 1985, Hayes 1989). Under such a view, a CV.CV.CV word (*macaco* ‘monkey’) and a CV.CCV.CV word (*catraca* ‘turnstile’) are both predicted to bear penult stress, as they have exactly the same moraic representation: $\sigma_\mu.\sigma_\mu]_{PWd}$. As onsets are outside the rhyme, these constituents are not moraic, and therefore are not predicted to affect stress likelihood. However, in §3 I show that the onset patterns found in the Portuguese lexicon deviate from these predictions.

As mentioned in §2, previous studies of Portuguese stress only consider weight in word-final rhymes (Bisol 1994, Collischonn 1996, Araújo 2007 and others). The standard claim that only word-final syllables are weight-sensitive is mostly based on the observation that antepenult, penult and final syllables behave very differently from one another regarding syllable shape (open *vs.* closed) and stress, as can be seen in Table 2.

probabilistic approach presented in this paper for two reasons. First, very few words contain nasal triphthongs (n=15). Second, nasal triphthongs are only found in word-final syllables, where weight effects are already known to be robust.

If Portuguese is in fact only weight-sensitive word-finally, its weight profile could be classified as *combined*. Combined systems have distinct weight computations for different positions or circumstances. There are 42 languages (out of 500) in the WALS database with a combined weight system (Goedemans and van der Hulst 2013). Among these languages, we find Spanish and Romansch, both closely related to Portuguese.

Wetzels (2007), however, argues that weight may also play a role word-internally, given the behaviour of palatal consonants in Brazilian Portuguese.[7] Although consonantal quality does not have an evident effect on stress in the language, {[ɲ], [ʎ]} are an exception. These consonants are never found in final onsets in words with antepenult stress (≈ 3.8% of the corpus contain such onsets). Wetzels (2007, p. 25) analyses such consonants as geminated, which therefore occupy both onset and (preceding) coda slots: *baralho* ⇒ [ba.ˈɾaʎ.ʎo] (*[ˈba.ɾaʎ.ʎo]) 'deck of cards' (see Fig. 1). This analysis is consistent with the fact that very few words with antepenult stress have a heavy penult syllable: in both cases, weight in the penult syllable would block antepenult stress.

Standard views on stress in Portuguese non-verbs tend to rely on more frequent/robust patterns in the lexicon, such as the distribution of open *vs.* closed syllables across stress locations. Table 2, for instance, provides a clear positive correlation between final closed syllables and final stress: 80.98% of all words with final stress have a closed word-final syllable. On the other hand, antepenult closed syllables and antepenult stress show a negative correlation, as only 20.33% of words in that category have a closed antepenult syllable. A similar pattern is found for penult stress, given that only 35.4% of stressed penult syllables are heavy. These facts have been the motivation for most phonological analyses of Portuguese stress. Such analyses typically conclude that weight-sensitivity is only present word-finally.

Table 2: Stressed syllable profiles by stress pattern in the Houaiss corpus (n=164,291)

	Open σ		**Closed σ**	
Pattern	n	%	n	%
Final stress	5780	19.02%	24608	80.98%
Penult stress	72531	64.60%	39730	35.40%
Antepenult stress	17242	79.67%	4400	20.33%

What is missing from Table 2 is whether or not the unstressed syllables in a given word are closed or

[7] See Wetzels (1997) for a comprehensive discussion on the distribution of final and penult syllabic shapes. A similar discussion for Spanish is found in Harris (1983).

open. In other words, what do the penult syllables look like in words with final stress? This is an important gap in traditional analyses of weight in BP. If penult syllables are not weight-sensitive, then having heavy or light syllables in that position should not alter the probability of antepenult stress for a given word. However, we have just seen that the weight profile of penult syllables does affect how likely antepenult stress is. §3 explores this and other patterns in detail.

Thus far, we have seen that weight clearly has an impact on the distribution of stress patterns in Portuguese. However, stress is not the only context where weight plays a role in the language: weight also influences mid vowel contrasts when stress is held constant on the penultimate syllable. This is known as spondaic lowering (SL), and was first formalised by Wetzels (1992). SL neutralises the mid vowel contrast in the stressed syllables of non-verbs with penult stress. Crucially, SL is conditioned by weight—more specifically, by the weight of the word-final syllable (see Table 3), a fact which is consistent with the claim that weight effects in the language are restricted to this position. Spondaic lowering can be formalised as follows: /ɛ, e, ɔ, o/ ⇒ [ɛ, ɔ] / ˈ__ H]$_{PWd}$, where the stressed syllable may be either open or closed. Therefore, the relevance of weight to Portuguese goes beyond stress.

Table 3: Spondaic lowering (Wetzels 1992)

$\ldots\acute{\mathbf{V}}\mathbf{L}]_{PWd}$	Gloss	$\ldots\acute{\mathbf{V}}\mathbf{H}]_{PWd}$	Gloss
[ˈɛli] *vs.* [ˈeli]	‘letter L’, ‘he’	[ˈfɛzis] *vs.* ∅	‘feces’
[ˈsɛd͡ʒi] *vs.* [ˈsed͡ʒi]	‘head office’, ‘thirst’	[eˈlɛtɾoŋ] *vs.* ∅	‘electron’
[ˈbɔxa] *vs.* [ˈboxa]	‘bird species’, ‘sediment’	[ˈdɔɾis] *vs.* ∅	‘Doris’
[ˈmɔʎʊ] *vs.* [ˈmoʎʊ]	‘bundle’, ‘sauce’	[ˈmɔvew] *vs.* ∅	‘furniture’

Given the positional bias of weight effects discussed above, Bisol (1992) proposes that BP builds moraic and syllabic trochees (the former applying only word-finally). Thus, *papel* [paˈpɛw] ‘paper’ is parsed as [pa(ˈpɛ$_{\mu}$w$_{\mu}$)] and *sapato* [saˈpatʊ] ‘shoe’ is parsed as [sa(ˈpa$_{\sigma}$tʊ$_{\sigma}$)]. Let us now briefly look into how the moraic approach[8] deals with irregularities in stress, and what issues arise from such an approach.

Previous approaches to stress in Portuguese are categorical; that is, a set of rules or constraints generates predictable patterns only. As a result, ‘exceptions’ are explained with mechanisms such as extrametricality and catalexis: Bisol (1992), d’Andrade (1994) and Massini-Cagliari (1999) employ exceptional syllable extrametricality to account for antepenult stress, in which case final syllables are skipped and a syllabic trochee is

[8] See Lee (2007) and Hermans and Wetzels (2012) for recent moraic approaches to stress in Portuguese.

built from the right edge of the word: (ˈσ σ) ⟨σ⟩. Likewise, words with penult stress and a heavy final syllable ($\ldots$ˈXH$]_{PWd}$) are explained with segment extrametricality, which makes the (heavy) final syllable light: ˈCV.CV⟨C⟩. For $\ldots$XˈL$]_{PWd}$ words, Bisol (1992) proposes a catalectic consonant, which is only phonetically realised in derivations: *café* [kaˈfɛ] ‘coffee’ would then be represented as [kafɛC]. The catalectic consonant C makes the final syllable heavy, and the moraic pattern is maintained: ka(ˈf$\text{ɛ}_\mu$$\text{C}_\mu$). We can see such a consonant in derived forms: *cafe<u>t</u>eira* ‘coffee pot’ *vs. cafe<u>z</u>al* ‘coffee plantation’—note that the quality of the catalectic consonant can vary in derivations of the same stem.

In sum, we have seen that stress as well as spondaic lowering provide strong evidence for the role of weight in Portuguese. To investigate in detail *how* weight affects stress in the language, I now turn to §3, which explores the patterns found in the Portuguese lexicon. We will see that the weight effects on stress are considerably more intricate than previously thought.

3 Data

This section probes the Portuguese lexicon in an attempt to answer the three questions posed in §1, repeated in (9) for convenience.

(9) a. Is weight-sensitivity only found word-finally in Portuguese?
 b. Is weight-sensitivity *categorical* or *gradient*?
 c. Do onsets contribute to weight, affecting stress likelihood in Portuguese?

The questions in (9) are clearly connected, since (9a) examines where weight-sensitivity is found and (9b) examines how it affects stress. Likewise, question (9c) also affects the answer to question (9b).

The data examined in this paper is based on the most comprehensive corpus available in the Portuguese language: the Houaiss Dictionary (Houaiss et al. 2001). The Houaiss corpus contains 442,000 entries/lemmas, of which 164,291 are non-verbs, including monosyllables. Even though such a corpus contains nearly all words in the language, it lacks the necessary components needed for a thorough phonological analysis, e.g., syllabification, stress location, segmental information etc. As a result, the list of words present in Houaiss et al. (2001) was used as a starting point for the elaboration of a stress corpus in Portuguese (see below). The final corpus (*Portuguese Stress Corpus*) is freely available (Garcia 2014), and contains over fifty analysable variables, which range from syllabification, stress location, weight and segmental profiles to neighbourhood density and bigram probabilities.

Given its large size, the Houaiss corpus also includes many words that are rarely used in spoken Portuguese. Some words are borrowings whose phonotactic patterns do not match those found in the language—e.g., German words such as *schnitzel* and *Bretschneidera* (the sequences [ʃn] and [tʃn] are not allowed in Portuguese, and undergo [i]-epenthesis). Words with more than two onset segments or two coda segments, as well as words that violate the phonotactic patterns in the language were excluded from the Portuguese Stress Corpus (≈ 5.6%), as were monosyllables (≈ 0.4%).

No constraints were imposed on word length (aside from a lower bound of two syllables). The median number of syllables in the whole corpus is four, but spoken Portuguese contains very few words with more than five syllables. If we examine the FrePOP database of spontaneous speech (Frota et al. 2010), for example, more than 90% of the words listed ($n = 188,269$) contain fewer than four syllables. Thus, a separate analysis was implemented where only words with fewer than six syllables were considered. The results of this separate analysis did not differ significantly from the results presented in this paper. Therefore, the more comprehensive analysis was preferred, where no length constraints were imposed.

One further adaptation was necessary: approximately 0.12% of the words in the corpus have antepenult stress *and* word-final hiatus, which is always resolved through diphthongisation in Portuguese:. . . ˈCV.CV.V ⇒ . . . ˈCV.CGV. For example, *terráqueo* /te.ˈxa.ke.o/ is realised as [te.ˈxa.kjʊ] 'earthling'. Diphthongisation is not categorical when the second V in a VV sequence is stressed: *piada* [pi.ˈa.da] ~ [ˈpja.da] 'joke'. This directly affects stress, since the diphthongisation yields penult stress in a word such as *terráqueo*. As a result, these data could potentially bias the analysis.[9] Thus, words such as *terráqueo* were removed from the data. Finally, words with more than one coda segment in any syllable in the stress domain ($n = 1216$) were removed, given that the vast majority of such words are either borrowings or contain a prefix such as *trans-*. The final version of the Portuguese Stress Corpus (Garcia 2014) contains 154,610 entries/lemmas (Table 4).

Grapheme-phoneme conversion was done by different scripts and regular expression substitutions. Some cases, however, are ambiguous. For example, the grapheme x can be realised as [s], [z], [k.s] and[ʃ]: *máximo* 'maximum', *exato* 'exact', *oxigênio* 'oxygen', *coxa* 'thigh', respectively—note that in all four examples x is in intervocalic position. Besides a qualitative difference, this grapheme is particularly important because one of its phonemic realisations involves a different syllabic configuration ([k.s]), i.e., a quantitative difference. All words containing this type of mismatch (n=2399), as well as other grapheme-phoneme idiosyncrasies, were manually checked and corrected.

[9] In fact, statistical models were run with and without such words, and the predicted negative correlation was confirmed between antepenult stress and word-final hiatus. No other effects were influenced by these forms.

Among the rare words in the corpus, many are technical terms, which often have antepenult stress. This could mean that the corpus used here is not representative of everyday Portuguese vis-à-vis stress patterns. Although the analysis in this paper is concerned with the lexicon *per se*, it would be ideal if the distribution of stress patterns in the lexicon did not deviate much from what speakers would normally experience in their language use. To verify this, two frequency corpora were examined, both of which contain only the most frequent words in the language: the Invoke Limited (IL) corpus (Dave 2012) and the LaPS corpus (Klautau 2013), from the Federal University of Pará, in Brazil—unlike the Houaiss Dictionary, the IL and LaPS corpora are based solely on Brazilian Portuguese. In all three corpora, the proportions of each pattern are relatively similar across all non-verbs considered. More importantly, the order *penult* > *final* > *antepenult* is observed in all three cases.

Table 4: Portuguese/BP* corpora

Stress pattern	**Houaiss**	**IL***	**LaPS***
Final	18%	21%	27%
Penult	69%	71%	62%
Antepenult	13%	8%	11%
	n=154,610	n=39,705	n=8,468

3.1 Weight-sensitivity: the Portuguese lexicon

In this subsection, I examine how weight-sensitivity affects stress placement in the Portuguese Stress Corpus (Garcia 2014). Firstly, I show that segmental quality does not have a clear correlation with stress in Portuguese. Secondly, I explore how the size of each syllabic constituent (§3.1.1) may affect stress: both subtle and robust effects are found in all three syllabic positions, namely, onset, nucleus and coda. In section 4, I present statistical models that capture such trends in the lexicon.

3.1.1 Segmental quality and stress

The corpus described above was analysed in terms of stress patterns based on number of segments as well as consonantal quality for all three possible positions, namely, final, penult and antepenult syllables. Consonantal quality in codas or onsets does not seem to affect stress likelihood in a consistent way. Even though correlations do exist, their effects are not systematic. For example, [ɲ], which is only possible in onset position, is significantly correlated with penult stress when in penult position ($p < 0.0001$), but negatively correlated with final stress when in final position ($p < 0.0001$). On the other hand, (onset) [k] is negatively correlated with final stress in final position ($p < 0.0001$), and also negatively correlated with penult stress in penult position ($p < 0.0001$). Different trends are found for other consonants, and no systematic pattern is observed—the same can be said for vowel quality.

When we observe the distribution of the most frequent consonants in onset and coda position, we also see no consistent pattern (Table 5). In fact, the distribution of such consonants is as unsystematic as their correlation with stress mentioned above. For example, it could be the case that more sonorous onset segments appear more frequently in stressed syllables (shaded cells in Table 5). In other words, stressed positions could be more frequently occupied by more sonorous segments. That is simply not the case when we look at consonantal distributions (in Table 5) or consonantal correlations with stress.

Table 5: Most frequent onset and coda segments by stress pattern

	Final σ		**Penult σ**		**Antepenult σ**	
Stress pattern	Onset	Coda	Onset	Coda	Onset	Coda
Final	/d,s,ɾ/	/ɾ,l,s/	/k,t,ɾ/	/n,ɾ,m/	/k,t,s/	/n,ɾ,s/
Penult	/t,d,s/	/l,m,s/	/t,d,n/	/n,s,ɾ/	/l,k,m/	/n,ɾ,s/
Antepenult	/k,l,ɾ/	/s,n,ɾ/	/t,f,n/	/n,ɾ,l/	/t,l,n/	/s,n,ɾ/

3.1.2 Onset size effects

Let us now explore the data by examining the impact of onset size on stress. Onsets may be absent (0), as in *árvore* 'tree', singleton (1), as in *cólica* 'spasm', or complex (2), as in *prático* 'practical'—all three words have antepenult stress in this particular case, and are therefore represented by the darker bars in Fig. 2 (Antepenult σ). The primary focus of the exploratory data analysis that follows is to visualise how properties of a given syllable affect stress on that syllable, as opposed to stress on the other two syllables in the stress domain. The plots in Fig. 2 show the percentage of words with a given stress pattern according to the onset size in each syllable. All three stress patterns are shown in the top legend. For convenience, in each figure, the darker bars represent the stress pattern directly affected by the position of the onset being analysed (Antepenult σ, Penult σand Final σ, respectively).

Figure 2: Onset size effects by syllable and stress pattern

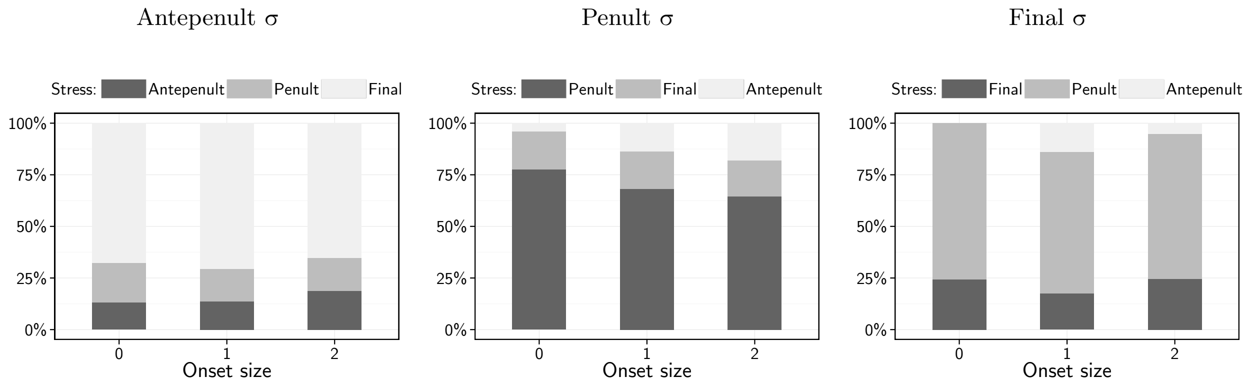


Fig. 2 suggests that onsets are positively correlated with stress in the antepenult and final syllables. The number of words with antepenult stress does not seem to be affected in different ways when the antepenult onset size is either 0 or 1. Rather, the difference in the Antepenult σ plot lies between {0,1} and 2 segments. For both antepenult and final syllables, onset effects on stress are not clear from the figures.

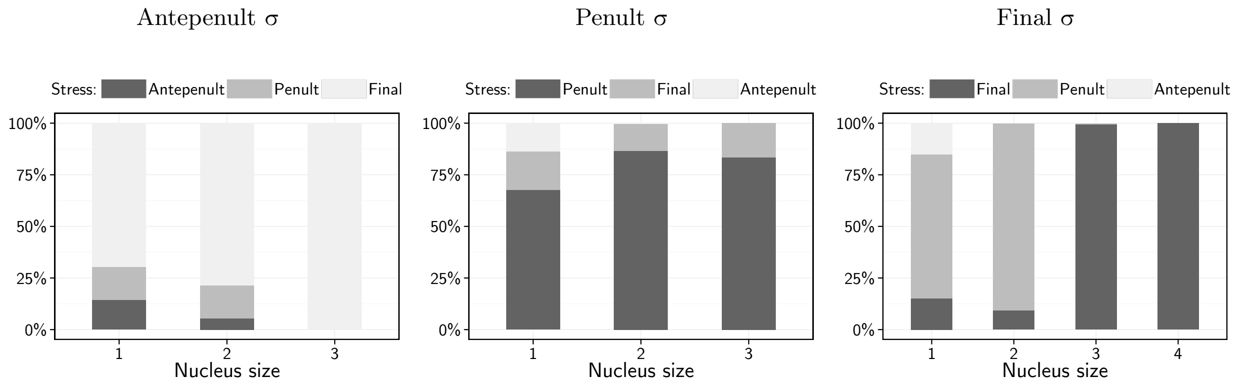
We can see in Fig. 2 that onset size is negatively correlated with stress in penult syllables. In other words, as we increase the number of onset segments in the penult syllable, we observe a decrease in the number of words with penult stress. Interestingly, it is the number of words with *antepenult* stress that increases as a function of penult onset size. As we will see below, these effects become clearer once we control for coda size. The importance of these effects will be examined in §4.

3.1.3 Nucleus size effects

Nuclei and codas are expected to have stronger effects on stress than onsets. In Fig. 3, we can see that words with penult and final stress seem to be affected by penult and final nucleus size, respectively. Longer nuclei seem to have a strong effect on stress, which is consistent with typological weight distinctions, where complex nuclei are heavier than V nuclei. For example, words such as *bastião* [basˈtjãw̃] 'bastion' always bear final stress. In these cases, the final nucleus is coded as [ja~w], and contains four segments, which means nasality is counted as a segment in and of itself (as mentioned in §2.2).

Note that the effect of nuclei on stress is visible not only word-finally, but also for the penult syllable, contrary to what we would expect if weight-sensitivity were constrained to the right edge of the word in Portuguese (according to the traditional view discussed in §2). Interestingly, antepenult nuclei seem to have a *negative* effect on antepenult stress, which is clearly unexpected.

Figure 3: Nucleus size effects by syllable and stress pattern


Antepenult σ

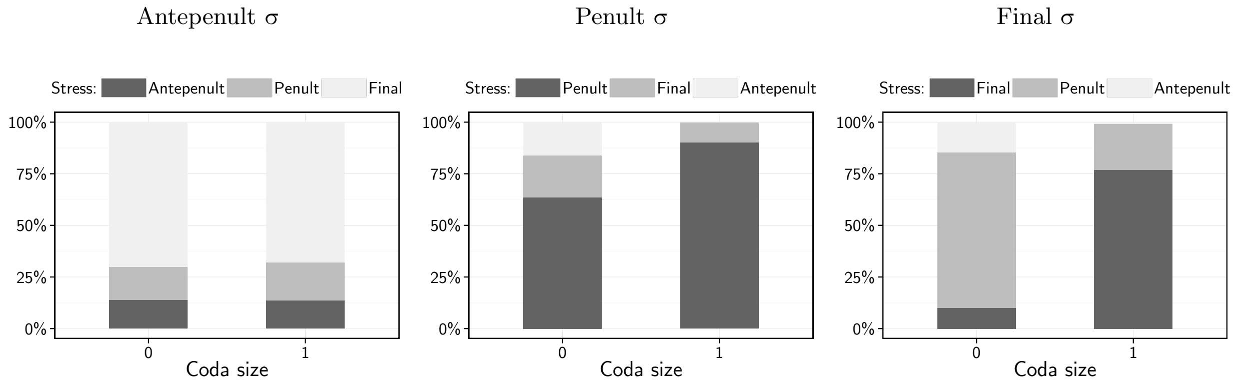
Penult σ

Final σ

3.1.4 Coda size effects

Let us now examine the effect of coda size on stress placement. Fig. 4 shows a very strong effect of the presence of a final coda on stress placement, consistent with the standard approaches to stress in Portuguese discussed in §2: final stress is far more frequent when the final syllable has a coda. On the other hand, the presence of a coda in the antepenult or penult syllables does not seem to strongly affect stress placement. Penult codas still suggest a positive effect on penult stress, at least if we compare *no* coda and *some* coda segments (the same trend is observed in final syllables). Antepenult syllables, however, suggest a null effect, given that the presence of a coda segment does not seem to affect antepenult stress. Recall, however, that in almost all words with antepenult stress, only the antepenult syllable can be heavy (see Table 1). In other words, though the antepenult rhyme may not affect the likelihood of antepenult stress, the presence of penult and final codas is expected to have a very strong (negative) effect on antepenult stress.

Figure 4: Coda size effects by syllable and stress pattern


Antepenult σ

Penult σ

Final σ

The trends observed above suggest that the effect of syllable weight is gradient, not categorical: coda effects are overall stronger than nucleus effects, which is unexpected, but both seem to have a substantial impact on stress. One of the possible reasons for the weaker effect of nuclei may be the fact that rising diphthongs are traditionally considered to be light in Portuguese, but such cases count as complex nuclei in Fig. 3 (see §2.2; I return to this discussion in §4.1.3). How much weight influences stress also depends on which syllable one examines: final stress is more strongly affected by nuclei and codas than penult stress. In other words, weight effects seem to vary considerably across (and within) syllables, and are not only found word-finally. Onsets also show some effect on stress, though the trends observed here indicate that these segments may be *negatively* correlated with stress in a given syllable. These trends are statistically analysed in §4 below.

Given the trisyllabic window in which stress falls in Portuguese, we can verify the onset-stress relation in the two final syllables. Considering the coda effects in Fig. 4, $\ldots\text{LL}]_{PWd}$ words will most likely have pre-final stress regardless of onset size, as the absence of a final coda will definitely impact stress on that syllable. Still, how much stress is affected could vary as the number of onset segments increases. Thus, let us examine whether final onset size affects penult/final stress.

Figure 5: Stress patterns by final onset size in ...LLL words

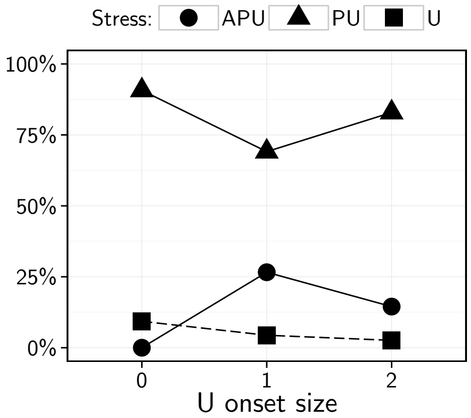


Fig. 5 suggests that larger final onset sizes (specifically from 1 to 2) are more highly correlated with penult stress than final stress. It should be noted that singleton onsets are much more frequent in the Portuguese Stress Corpus than complex onsets: 84.3% *vs.* 4.7% in words with final stress, 89.1% *vs.* 2.1% in words with penult stress, and 81% *vs.* 10.2% in words with antepenult stress. These data refer to stressed syllables in each pattern, but unstressed syllables also have more singleton onsets than complex onsets—Portuguese, like other Romance languages, has a relatively low frequency of onset clusters.

The trend in Fig. 5 is problematic, given that onsets of a given syllable are not expected to negatively impact stress on that syllable (see, for example, Ryan (2014)). If this particular trend is statistically credible, however, the data would be consistent with a different theory of weight computation, namely, Interval Theory (Steriade 2012).[10] Unlike syllables, intervals are rhythmic units that span from a given vowel up to (but not including) the following vowel (i.e., a V-to-(V) interval). Since all intervals begin with a vowel, it follows that the string CCVCCVC is parsed into intervals as ⟨CC⟩VCC.VC (word-initial consonants are treated as extrametrical in this theory). The longer an interval, the heavier it is—and, as a result, the more likely it is to attract stress.

The crucial parsing difference between syllables and intervals lies with onset segments: onsets of syllable i are parsed into interval $i - 1$. A coda and its preceding nucleus, which in Syllable Theory belong to the same syllable, also belong to the same interval in Interval Theory. It follows that, if we transition from syllables to intervals, more onset segments in the final syllable result in a longer (and therefore heavier) *penult* interval—all else being equal. Consequently, the negative onset effects observed in Fig. 5 would be predicted by Interval Theory,[11] even though the negative effect of antepenult nucleus size in Fig. 3 would still be unexpected.

Penult and antepenult syllables are locations where coda effects are less apparent (standard analyses assume there is no such effect in these positions, as discussed in §2). Fig. 6 presents the proportion of such words for different onset sizes in the penultimate syllable. Under Syllable Theory, increases in onset size in the penult syllable should increase the amount of material in that constituent, positively impacting its duration, which should in turn affect the likelihood of penult stress (assuming onsets play a role in stress assignment). Interval Theory, on the other hand, predicts an increase in the likelihood of antepenult stress.

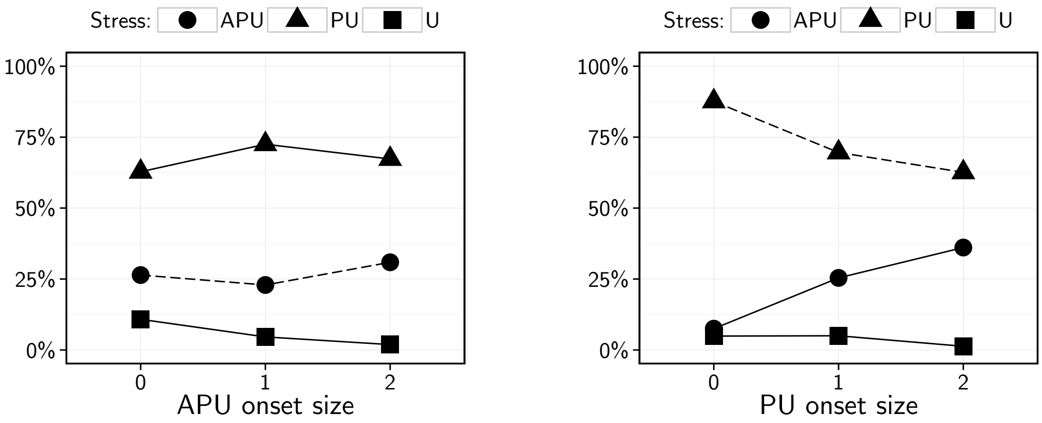
We see in Fig. 6b that the likelihood of antepenult stress increases when the penult syllable contains onset segments. Figs. 5 and 6 show a clear pattern, which is consistent with intervals. Antepenult onset size (Fig. 6a), on the other hand, presents a less clear pattern (recall that antepenult onsets are assumed to be extrametrical, thus no particular pattern is expected): the presence of onset clusters in this syllable seems to favour antepenult stress when compared to singleton onsets, but not when compared to onsetless syllables. The high degree of unpredictability of antepenult syllables (relative to penult and final syllables) might be one of the reasons behind the unexpected patterns that we find. The onset effects observed in Fig. 6 are

[10] A statistical comparison between syllables and intervals for Portuguese can be found in Garcia (2016).

[11] This assumes that complex onsets are indeed longer than singleton onsets, given that the relevant dimension of interval weight is duration (Steriade 2012).

Figure 6: Stress patterns by antepenult and penult onset size in ...LLL words

(a) Antepenult onset (b) Penult onset



also unexpected given more recent work by Kelly (2004) and Ryan (2011), for example, who show a positive effect of onset size on stress—note, however, that these studies focus on word-initial onsets.[12] As we will see in the next section, antepenult syllables show a pattern that is not accounted for under syllables nor under intervals.

4 Statistical analysis

In the previous section, we observed that the patterns in the Portuguese Stress Corpus show gradient weight effects concerning stress location in the language. In this section, I test whether the correlations in the data are supported (i.e., are significant) using statistical models that predict the location of stress based on the different syllabic constituents in the stress domain. In §4.1, I describe each statistical model proposed, analyse the results, and examine how they relate to the main questions in this paper, stated in (9). In §5, these models are compared to previous approaches, which serve as the baseline for the present analysis. Even though some of the patterns observed support intervals, the statistical models proposed in this paper are based on syllables, as the representational assumptions encoded in the predictors of syllable-based models are, by definition, a superset of those encoded by interval-based predictors, given the structural differences that hold between the two theories. In other words, one can evaluate intervals by considering syllable-based results, but not vice-versa (the reader can easily make direct comparisons between the two theories by examining the effect sizes of onsets in the models provided).

[12] A recent study by Olejarczuk and Kapatsinski (2014) shows that stress preference in English is also affected by the phonotactic profile of word-medial clusters.

The factors examined in §3 are listed in Table 6. The number of predictors in the statistical models proposed in this paper is proportional to the size of the stress domain (3 syllables × 3 constituents per syllable = 9 predictors). Antepenult constituents are coded as `NA` in disyllabic words.

Table 6: Predictors and response

Syllables	`onset.fin`	Number of onset segments in the **final** σ (`0-2`)
	`nucleus.fin`	Number of segments in the nucleus of the **final** σ (`1-4`)
	`coda.fin`	Number of coda segments in the **final** σ (`0,1`)
	`onset.pen`	Number of onset segments in the **penult** σ (`0-2`)
	`nucleus.pen`	Number of segments in the nucleus of the **penult** σ (`1-3`)
	`coda.pen`	Number of coda segments in the **penult** σ (`0,1`)
	`onset.ant`	Number of onset segments in the **antepenult** σ (`0-2`)
	`nucleus.ant`	Number of segments in the nucleus of the **antepenult** σ (`1-3`)
	`coda.ant`	Number of coda segments in the **antepenult** σ (`0-1`)
	Response	`antepenult`, `penult`, `final`

The analysis presented in this section employs two Binomial Logistic Regressions to model the Portuguese lexicon. Given that the stress patterns found in the language involve more than two categorical responses, a Multinomial Logistic Regression could be employed. However, goodness of fit and diagnostics become more intricate in such a model; i.e., it is less straight-forward to assess the model's accuracy and interpret the meaning of coefficients, for instance, since outcomes are interpreted in relation to a reference response. Furthermore, the literature on multinomial models applied to linguistic data is scarce when compared to binomial models.

A more parsimonious alternative would be to model the data using *Ordinal Regression* (see Agresti 2010), also known as *Cumulative Link Model.* In this case, the stress domain in the data would need to be treated as a three-point scale, where final (1) and antepenult (3) positions demarcate the end-points of the domain: $3 > 2 > 1]_{PWd}$. This scale mirrors the stress domain, in terms of ordering as well as end-points (i.e., stress cannot be later than final nor earlier than antepenult). A single Ordinal Regression for the stress domain in Portuguese can be understood as equivalent to two (Binomial) Logistic Regressions. Another advantage of ordinal regressions is that predictors in such models tend to have lower standard errors when compared to equivalent binomial regressions (Christensen 2013, p. 6).

Despite the advantages of Ordinal Regressions, their interpretation is also less trivial (much like Multinomial Regressions). Because a single coefficient is generated, its interpretation depends on multiple thresholds, which act as intercepts along the scale assumed. More importantly, it is not clear that the stress domain should be treated as a scale. In other words, it is not intuitive why penult stress should be a higher (or lower) point in the scale when compared to final stress.

A third option is to analyse the data using Logistic Regressions (`glm()` in R (R Core Team 2016)). As mentioned above, this is the option employed in this paper. Because standard logistic models involve binary responses (i.e., binomial), two such models are necessary to accommodate the stress domain in Portuguese. As a result, interpreting the effect of individual predictors is more straight-forward, and no scale needs to be assumed (cf. Ordinal Regressions). In fact, all three options just described were compared, and the results did not differ substantially with regard to the central focus of the present study, i.e., weight gradience and its effect on stress.

In the analysis proposed in this paper, one model (`antPenFin`) will predict `antepenult` *vs.* `penult/final` stress, and another model (`penFin`) will predict `penult` *vs.* `final` stress ('Response' in Table 6). This division is aligned with traditional analyses, which classify antepenult stress as irregular, and penult/final stress as (mostly) regular (§2).

Logistic Regressions predict the log-odds of $y = 1/0$ based on a set of predictors. In this case, $y = antepenult$ *vs.* $\{penult, final\}$ in one model and $y = penult$ *vs.* $final$ in another model. The fitted model is given in (10), where $Pr(y_i = 1)$ denotes the probability that response $y = 1$; β^0 represents the intercept, which can only be interpreted when all other variables are set to zero (this is not meaningful for the purposes of the present analysis); $(\beta^{1\cdots n})$ represents the regression coefficients for each predictor; and X_i stands for the values of the i^{th} data point (i.e., the segmental count at each syllabic constituent). For example, assume we have a CVCCVCV word such as *martelo* 'hammer', which is syllabified as CVC.CV.CV (mar.te.lo). In the `antPenFin` model, we would predict the probability of antepenult stress (*vs.* penult/final stress) based on the segmental count in each syllable in the stress domain: $Pr(y_i = APU \text{ } vs. \text{ } \{PU, U\}) = logit^{-1}(\beta^0 + [1_m \cdot \beta^1 + 1_a \cdot \beta^2 + 1_r \cdot \beta^3]_\sigma + [1_t \cdot \beta^4 + 1_e \cdot \beta^5 + 0_\varnothing \cdot \beta^6]_\sigma + [1_l \cdot \beta^7 + 1_o \cdot \beta^8 + 0_\varnothing \cdot \beta^9]_\sigma)$. In this case, we are interested in how much each predictor in the set $\{\beta^{1\cdots 9}\}$ affects such a probability.

(10) **Logistic Regression**

$$Pr(y_i = 1) = logit^{-1}(\beta^0 + X_i^1 \cdot \beta^1 + X_i^2 \cdot \beta^2 + ... + X_i^n \cdot \beta^n)$$

As we will see, both models (`antPenFin` and `penFin`) capture the weight gradience in Portuguese. Besides, given the probabilistic nature of the approach proposed in this paper, the models are more accurate than previous categorical analyses in predicting the weight-stress patterns present in the lexicon. In the subsection that follows, I present the models and examine their results and predictions. In §5, I contrast these predictions with the actual patterns in the lexicon, and discuss how the present analysis differs from previous approaches.

4.1 Models of stress

4.1.1 Model A: `antPenFin`

In this model, stress (antepenult or penult/final) is predicted based on syllabic constituents in all positions in the stress domain. The `antPenFin` model is presented in Table 7, where we can see that all nine predictors have a highly significant effect on stress ($p < 0.00001$), which confirms that weight effects are not limited to word-final syllables. In addition, we can see that effect sizes weaken as we move away from the right edge of the word. All coefficient values in Table 7 have been centred and scaled, and are therefore directly comparable to one another (each $\hat{\beta}$ unit corresponds to one standard deviation of a given predictor).

Table 7: Scaled (and unscaled) coefficient values for `antPenFin` model ($\hat{\beta} > 0 \Rightarrow$ higher likelihood of antepenult stress), with associated odds ratio ($\mathbf{OR} = e^{|\hat{\beta}|}$), standard errors, Wald z values and significances

Predictor	`scale(`$\hat{\beta}$`)`	$\hat{\beta}$	`scale(`**OR**`)`	**OR**	**SE**	***z* value**	***p* value**
`onset.ant`	0.109	0.27	1.115	1.31	0.009	12.540	< 0.00001
`nucleus.ant`	-0.220	-1.22	1.246	3.38	0.012	-18.380	< 0.00001
`coda.ant`	-0.051	-0.14	1.052	1.15	0.008	-5.974	< 0.00001
`onset.pen`	0.334	0.89	1.396	2.43	0.009	36.755	< 0.00001
`nucleus.pen`	-1.107	-4.71	3.025	111.05	0.047	-23.460	< 0.00001
`coda.pen`	-2.724	-6.89	15.241	982.40	0.119	-22.840	< 0.00001
`onset.fin`	0.624	2.31	1.87	10.07	0.014	45.799	< 0.00001
`nucleus.fin`	-2.773	-5.82	16.007	336.97	0.132	-20.968	< 0.00001
`coda.fin`	-1.169	-3.68	3.219	39.65	0.026	-44.450	< 0.00001
							κ = 28.19

The results in Table 7 indicate key trends. First, we find divergent weight effects between rhymes and onsets across all three predictor positions in question. For example, whereas both the nucleus size and the coda size in the antepenult syllable negatively affect the likelihood of antepenult stress, the size of antepenult onsets *positively* affects antepenult stress.

The onset effects we observe in Table 7 are consistent with the data trends discussed in §3.1.1, i.e., increasing the penult onset size increases the likelihood of antepenult stress. It is also possible to see that onset effects (in absolute terms) weaken as we move away from the right edge of the word—the same is true for nucleus effects. If we combine the penult onset effect with the antepenult rhyme effect discussed above, we can conclude that a word such as CV.CCV.CV could likely be the optimal candidate for antepenult stress (multiple onset clusters in the same word are uncommon in Portuguese).

Unsurprisingly, both penult and final rhymes negatively affect antepenult stress. In other words, LLL is the ideal weight profile for this particular stress pattern. Interestingly, the effect size of nuclei and codas is different when penult and final syllables are compared: in final position, nuclei have a stronger effect than codas, while in penult position codas have a stronger effect. In fact, the presence of a word-final coda reduces the odds of antepenult stress by a factor of nearly 40, whereas the presence of a penult coda reduces the odds of antepenult stress by a factor of 982.40 (see §4.1.2 for a discussion of nucleus-coda effects). These observed differences in effect size also capture a particular lexical pattern in the language, namely, that ĹHL is less common than ĹLH (Table 1).

One important characteristic of an optimal data set is that the predictors involved are orthogonal, i.e., uncorrelated—although this is rare in practice, predictors should ideally be as uncorrelated as possible. The more non-orthogonal predictors are, the more difficult it becomes to explain exactly which predictors are responsible for a given effect—this is a phenomenon known as *collinearity*[13] (Belsley et al. 1980). The predictors included in the model in Table 7 have medium-high collinearity ($\kappa = 28.19$).

The syllabic shapes found in Portuguese explain why collinearity is not low between onsets, nuclei and codas: although both heavy nuclei and codas are allowed, GVC/VGC syllables are rare in the language—i.e., syllabic predictors are not completely orthogonal. Furthermore, words with coda segments in multiple syllables are uncommon in the Portuguese lexicon. A Spearman ρ^2 test reveals that the most collinear pair of predictors included in the `antPenFin` model is `onset.pen` and `coda.ant` ($\rho^2 = 0.15, p < 0.00001$). Higher collinearity does not affect the model's coefficients; rather, it raises standard errors, which in turn lower the significance of a given effect (Baayen 2008). However, all the effects in question are highly significant ($p < 0.00001$), and therefore even relatively high collinearity should not pose problems for the analysis.

4.1.2 Model B: `penFin`

The `penFin` model in Table 8 shows that only penult onsets have no significant effect on penult (*vs.* final) stress—all other predictors are highly significant ($p < 0.00001$). Let us begin by examining the three predictors in the final syllable (positive $\hat{\beta}$ values indicate a higher likelihood of penult stress). First and foremost, we

[13] Represented here by κ. A model with $\kappa \leq 6$ has no collinearity; $\kappa \approx 15$ indicates moderate collinearity; and $\kappa \geq 30$ points to high collinearity (Baayen 2008, p. 182).

can see that most of the trends discussed in §3.1.1 are also confirmed in this model. For example, final onsets do have a positive effect on *penult* stress. In fact, adding an onset segment to the final syllable increases the odds of penult stress by a factor of 1.57 (note that this effect is inconsistent with the typical representational assumptions of Syllable Theory, as mentioned in §3.1). We also see that both `nucleus.fin` ($\hat{\beta} = -1.103, p < 0.00001$) and `coda.fin` ($\hat{\beta} = -1.49, p < 0.00001$) have negative effects on penult stress, which is expected, given that this is known to be a very robust aspect of Portuguese stress (§2).

Surprisingly, `nucleus.fin` has a weaker effect than `coda.fin`—a pattern also found for penult syllables in the `antPenFin` model discussed above. This contradicts a strong typological tendency, whereby VV is heavier than VC (Gordon 2011). Recall that Portuguese has no long vowels, and, importantly, not all complex nuclei in the language are assumed to affect stress, as rising diphthongs are traditionally treated as light. The model presented in Table 8 makes no distinction between rising and falling diphthongs (as discussed in §2.2), since `nucleus.fin` and `nucleus.pen` simply count the number of segments (× slots in Fig. 1) in the domain (this is further motivated in §4.1.3). This could explain why the effect of final nuclei is smaller than that of final codas in this model. To check whether this was the case, alternative models (`*`) were run where only falling diphthongs were considered to be heavy. In the `penFin*` model, `nucleus.fin` ($\hat{\beta} = 1.00$) still has a smaller effect size than coda.fin ($\hat{\beta} = 1.39$), and `nucleus.pen` ($\hat{\beta} = 0.08$) still has a smaller effect size than `coda.pen` ($\hat{\beta} = 0.17$). The same pattern is found in the `antPenFin*` model.

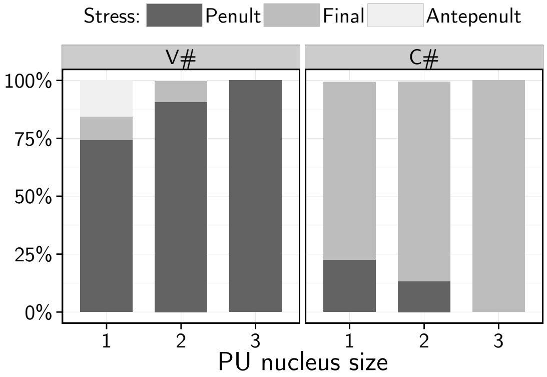
Table 8: Scaled (and unscaled) coefficient values for `penFin` model ($\hat{\beta} > 0 \Rightarrow$ higher likelihood of penult stress), with associated odds ratio ($\mathbf{OR} = e^{|\hat{\beta}|}$), standard errors, Wald z values and significances

σ **predictor**	`scale(`$\hat{\beta}$`)`	$\hat{\beta}$	`scale(`**OR**`)`	**OR**	se($\hat{\beta}$)	z **value**	p **value**
`onset.pen`	0.010	0.03	1.01	1.03	0.01	1.09	0.259
`nucleus.pen`	−0.084	-0.33	1.09	1.39	0.01	−8.04	< 0.00001
`coda.pen`	0.141	0.33	1.15	1.39	0.01	12.28	< 0.00001
`onset.fin`	0.134	0.45	1.14	1.57	0.01	14.61	< 0.00001
`nucleus.fin`	−1.103	-2.14	3.01	8.50	0.01	−135.75	< 0.00001
`coda.fin`	−1.490	-4.29	4.43	72.97	0.01	−181.14	< 0.00001
							$\kappa = 18.23$

Let us now examine the results of `nucleus.pen` and `coda.pen`. First, `nucleus.pen` shows a negative effect on penult stress, which is unexpected. This, again, could be connected to the distinction between rising and falling diphthongs discussed above: if most diphthongs in `nucleus.pen` happen to be *rising* diphthongs, this pattern could be explained. However, that is not the case. In fact, if we only examine words with a complex penult nucleus, 52% of such words contain the falling diphthong [ej], almost all of which have penult stress.

One potential reason behind the negative effect of `nucleus.pen` is another variable in the model: `coda.fin`. These two variables are negatively correlated, and removing `coda.fin` makes the effect of `nucleus.pen` turn positive—which is what we would expect given the trends in Fig. 3. The interaction between these two variables, however, is not captured in Fig. 3, since nuclei are plotted independently. Once we visually inspect these two variables (Fig. 7), we can clearly see that penult diphthongs have different effects depending on whether the word-final syllable contains a coda consonant (C#) or not (V#). Particularly, once the word-final syllable contains a coda consonant, the more segments a word has in its penult nucleus, the less likely penult stress becomes (dark bars in Fig. 7). For example, words such as *fácil* 'easy' are more frequent in the Portuguese lexicon than words such as *lêucon* [ˈlew.koŋ] 'leucon'[14] (23.2% *vs.* 12.4%). In other words, if we only look at disyllables that contain no penult coda but which do contain a word-final coda (n=1,871), those with a monophthong in penult position are two times more likely to bear penult stress when compared to those with a diphthong in penult position.

Figure 7: Penult nucleus size by word-final profile (V# *vs.* C#)



Because this paper assumes that theoretical premises should guide the statistical analysis, the model presented in Table 8 does not include the interaction in question. A syllabic representation does not predict that nuclei and codas in different syllables should interact. In other words, there is no principled reason to believe these two variables should affect each other (see §4.1.3 for a discussion)—in fact, other interactions could also exist in the language. The objective of the present analysis is not to build the best statistical model, which could include a number of unprincipled interactions. Rather, the objective is to build a theoretically principled model that is able to best capture weight gradience in Portuguese.

[14] Interestingly, almost all CV́G.CVC words are borrowings, and are rarely used in spoken Portuguese.

Let us now turn to coda.pen, which had a significant effect in the penFin model. The positive coefficient value of this predictor ($\hat{\beta} = 0.141$) indicates that adding a coda segment to the penult syllable increases the odds of penult stress by a factor of 1.39. This is naturally a much smaller effect than, for example, coda.fin, but it is highly significant. The effect sizes listed in Table 8 clearly show a gradient effect, whereby predictors in the final syllable have a greater absolute effect than predictors in the penult syllable.

In Table 8, onset.pen had no significant effect on stress. A relevant question is whether this null effect is also found once we model only disyllabic words. Indeed, if we restrict the penFin model to disyllables only (n=11,475), we do find that onset.pen has a positive effect on penult stress ($\hat{\beta} = 0.16, p < 0.00001$).

4.1.3 Model assessment

The models above have both expected and unexpected results. In the penFin model, for example, the effects of nucleus.pen and onset.fin go against what a syllabic representation would predict. On the other hand, the expected strong effect of final nuclei and codas possibly explains why previous analyses of Portuguese stress have constrained weight effects to the right edge of the word: such analyses have concentrated on word-final syllables only most likely because of the considerably different coefficient values between final and penult syllables ($\frac{\hat{\beta}_{\texttt{coda.fin}}}{\hat{\beta}_{\texttt{coda.pen}}} \approx 10$ in the penFin model). Therefore, though the structure of earlier syllables does affect stress placement, these effects are small compared to the structure of the final syllable, and may not be noticed unless a large enough subset of the Portuguese lexicon is examined.

In §2.2, we saw that, contrary to most analyses of Portuguese stress, rising diphthongs may not always pattern as light (following Harris (1983)). In particular, given that both CV́.CVG.CV and CV́.CGV.CV words are unattested in the language, it is not clear that a categorical weight difference can be determined for on-glides *vs.* off-glides. This is one of the reasons why the models above do not differentiate rising and falling diphthongs. A second reason, discussed below, is conceptual.

Should a model that only includes quantitative predictors be sensitive to the difference between rising and falling diphthongs? Should this distinction be 'visible' to the model? How rich a model is has to do with the types of theoretical and representational assumptions said model should encode. We are interested in a model that predicts stress based on quantitative information. One of the main objectives of the model is to determine how weight affects stress. Such a model should be as unbiased as possible. By differentiating rising and falling diphthongs, we would be adding information to the model that goes beyond a neutral segmental count—in fact, this would inform the model of a specific weight effect in the language (an effect which should be unknown *a priori*). In other words, we would be telling the model that a specific sequence of segments is light, even though the purpose of the model is to inform us about weight effects.

The two models presented and discussed above show that the weight patterns in the Portuguese lexicon are much more intricate than one would expect—and far from categorical. Firstly, such effects go in two directions. Whereas in the `penFin` model penult stress becomes less likely when final syllables are heavy, in the `antPenFin` model antepenult stress becomes less likely when *antepenult* syllables are heavy. In fact, we also see positive and negative weight effects in the penult rhyme (`penFin` model), where `nucleus.pen` and `coda.pen` have opposite effects on penult stress.

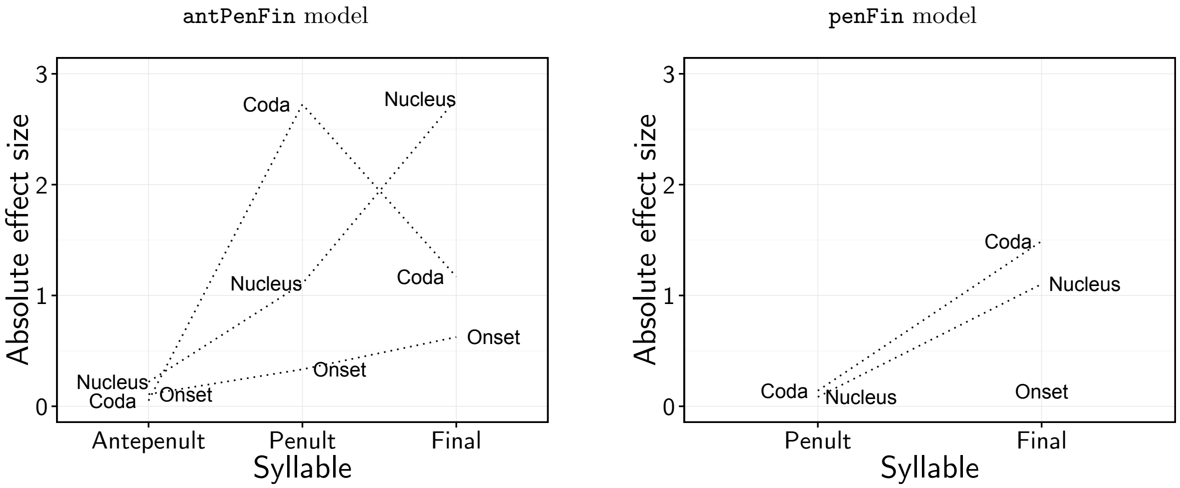
One could argue that some of these facts may be related to the footing patterns in Portuguese. The language is traditionally classified as trochaic (see Bisol (2000) for a review), and therefore (ĹL) and (H́) feet should be preferred (Hayes 1995). In addition, recall that previous analyses have argued that the final syllable is extrametrical in words with antepenult stress (Bisol 1994 and many others). If we now combine these two facts, we can partially explain why both `nucleus.ant` and `coda.ant` are negatively correlated with antepenult stress: given that ĹL trochees are preferred to H́L trochees, (ĹL)⟨X⟩ is better than (H́L)⟨X⟩, and therefore the former should be more likely than the latter.[15] A third footing option, namely, (H́)L⟨X⟩, is preferred to (H́L)⟨X⟩. However, this leaves a syllable unparsed in the middle of the stress domain, which not only contradicts traditional foot-based analyses of Portuguese, but is also highly marked. This approach thus assumes that light antepenult syllables are more stress-attracting for footing reasons, and not weight *per se*.

Not all facts are accounted for by extrametricality and footing patterns, however. For example, whereas the negative effect of `nucleus.pen` would be explained, the positive effect of `coda.pen` would not. Furthermore, the onset effects found in both models would require an additional explanation, as one would not expect such effects in a standard foot-based analysis. Indeed, there does not seem to be a theoretically unified way of accounting for all the effects found in the syllable models under discussion.

Let us now turn to the main focus of the present analysis, namely, weight gradience. The absolute coefficient values in the `antPenFin` and `penFin` models argue for a clear *gradient* notion of weight-sensitivity in Portuguese. Contrary to what previous analyses assume, the models discussed above show that weight is not a categorical phenomenon in the language. In Fig. 8, the absolute effect size of each predictor (i.e., syllable constituent) is plotted for each of the two models (`antPenFin` and `penFin`). These figures provide a more evident gradient trend (dotted lines): predictors in the final syllable have a stronger effect on stress when compared to predictors in the penult syllable (`penFin` model), which in turn have stronger effects on stress than predictors in the antepenult syllable (`antPenFin` model).

[15] Thanks to an anonymous reviewer for pointing out the possible connection between footing and stressed light antepenult syllables.

Figure 8: Absolute effect sizes in the syllable models


antPenFin model

penFin model

As can be seen in Fig. 8, the effects of predictors in the penult syllable are relative to the statistical model. In other words, the absolute difference between penult and final predictors is smaller than that of penult and antepenult predictors. This trend indicates that the antepenult syllable is the least weight-sensitive position in the stress domain in Portuguese. In addition to the weight gradience across syllables, we also observe gradual effects within syllables: Coda > Nucleus > Onset for final syllables in the `penFin` model and penult syllables in the `antPenFin` model, but Nucleus > Coda > Onset for final syllables in the `antPenFin` model. For antepenult syllables, the absolute effect sizes indicate a different trend, namely, Nucleus > Onset > Coda. Although this trend is highly significant, it is surprising and difficult to interpret; that is, it is not clear how such a pattern could be accommodated by any representational assumptions regarding rhythmic units.

5 Discussion

In this section, I summarise and discuss the main results presented in this paper. Section 5.1 evaluates the accuracy of the probabilistic analysis I propose, and section 5.2 briefly explores the implications of the approach adopted here for the grammar of Portuguese.

The models discussed in §4 clearly answer the questions in (9). First, weight-sensitivity is found in all positions in the stress domain, not only word-finally. Second, weight effects are gradient, not categorical. These two facts are evident in both statistical models discussed above. Third, onsets do contribute to weight in the Portuguese lexicon. However, the latter effect manifests itself in an unexpected way, given that in penult and final syllables onset size is *negatively* correlated with stress.

Both models examined in this paper clearly show that the relationship between stress and weight in Portuguese is far more intricate than previously assumed. Inconsistencies and surprising effects are not only limited to onsets: (i) penult codas have a stronger effect than penult nuclei; (ii) final codas have a stronger effect than final nuclei in predicting penult stress (`penFin` model); (iii) heavy antepenult rhymes disfavour antepenult stress.

The most important characteristic of the present approach is its probabilistic nature. A categorical approach cannot predict that a certain irregular pattern exists (e.g., ĹLL), given that it deviates from traditional generalisations about the language (XXH́ else XX́L). The present proposal, however, predicts that all licit stress patterns are possible (including so-called irregular cases), that some are more likely than others. Crucially, and perhaps most importantly, it is no longer the case that all irregular forms are *equally* unlikely, an implication of standard analyses. As we will see below, the probabilistic nature of the present approach results in a more accurate characterisation of stress in the Portuguese lexicon.

5.1 Accuracy

In this section, I briefly compare the predictions of the present approach with those of traditional categorical analyses. First, let us examine the predictions of the `antPenFin` model in Fig. 9, which plots the proportion (or probability) of words with antepenult stress (*vs.* penult or final stress) across sets of words that mirror the different weight profiles (i.e., sequences of light (L) and heavy (H) syllables) in the language.[16] The dotted line represents the predicted probability of antepenult stress based on traditional (categorical) approaches (i.e., 0%, since antepenult stress is considered to be irregular). Actual lexical proportions are represented by ● (where the size of the circle corresponds to lexical representativeness). As we can see in Fig. 9, in some cases (e.g., HHH, HHL, LHL), categorical predictions accurately match the actual lexical proportions. However, a clear mismatch is observed for HLL and LLL words. Finally, σ represents the mean predicted probability of antepenult stress based on the present approach.

Overall predicted proportions (σ) in Fig. 9 approximate the actual lexical values.[17] In other words, given a new LLL word, the present analysis predicts that there is a ≈25% probability that such a word will be assigned antepenult stress, and a 75% probability that stress will be either penult or final. Traditional approaches, on the other hand, would not predict antepenult stress in this (or any other) case.

Assuming that a word is not assigned antepenult stress, we now need to consider penult *vs.* final stress, which account for the vast majority of words in the lexicon (Table 1). Fig. 10 plots the percentage (or proba-

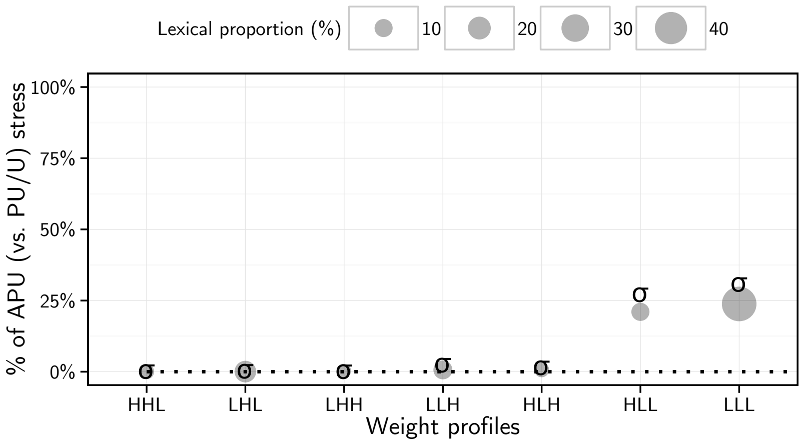
[16] Predicted probabilities are averaged across all words with a given weight profile.

[17] Lexical proportions are based on the set of words being modelled in each model: `antPenFin`: n=143,136; `penFin`: n=134,600.

bility) of words with penult stress (*vs.* final stress) across the different weight profiles in the language. Recall that traditional approaches predict final stress for all words with a heavy final syllable (XXH́) and penult stress elsewhere (XX́L). Clearly, these predictions deviate considerably from the actual lexical proportions of penult stress (●).

Like Fig. 9, Fig. 10 shows that probabilistic predictions are substantially more accurate than a categorical approach. Even though we observe a clear distinction between XXH and XXL words, a gradient effect *within* each group is also visible. For example, H́L words are more frequent than ĹL words—and this difference is mirrored in the models' mean predicted probabilities.

Figure 9: `antPenFin` model's accuracy: Mean predicted probabilities (σ) of antepenult (*vs.* penult/final) stress by weight profile as well as actual lexical frequencies (●) are plotted. Dotted lines indicate predicted stress based on a standard categorical analysis



Whereas Figs. 9 and 10 both provide a means to visually compare the present proposal to traditional analyses of Portuguese stress, Table 9 presents a numerical comparison, namely, the weighted mean deviation of predicted probabilities from actual lexical percentages. The mean deviation takes into account the representativeness of each weight profile in the lexicon (● in the plots). Not only is the weighted mean deviation lower in the probabilistic approach presented here, the weighted standard deviations are also lower when compared to a categorical approach, a fact which mirrors the trends in Figs. 9 and 10.

Figure 10: `penFin` model's accuracy: Mean predicted probabilities (σ) of penult (*vs.* final) stress by weight profile as well as actual lexical frequencies (●) are plotted. Dotted lines indicate predicted stress based on a categorical analysis

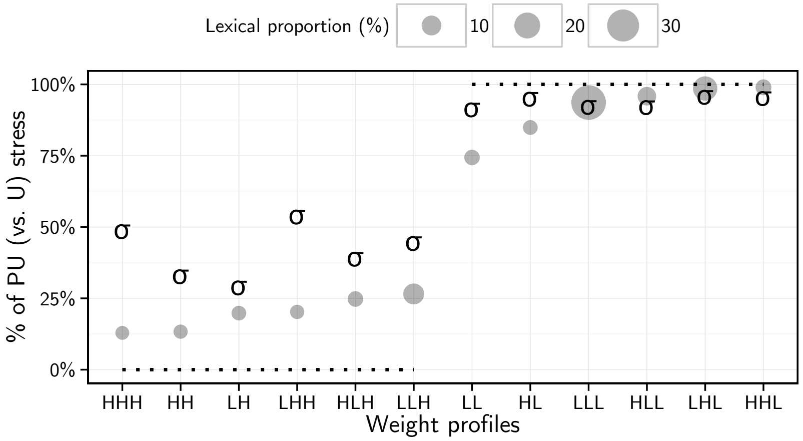


Table 9: Weighted mean deviation of mean predicted probabilities from actual lexical proportions: probabilistic *vs.* categorical approaches

	Probabilistic approach		**Categorical approach**	
	Mean	SD	Mean	SD
`antPenFin`	**4.2%**	3.6%	**14%**	13.6%
`penFin`	**6.4%**	7.9%	**9.7%**	10.4%

5.2 A probabilistic grammar

Thus far, we have investigated the stress patterns in the Portuguese lexicon by employing different statistical models. Little has been said, however, about what these patterns mean for the *grammar* of Portuguese speakers. If the lexical patterns explored in this paper are psychologically real, an important question is (i) how such patterns could be implemented in a phonological grammar and (ii) how the lexicon and grammar interact. We will not construct such a grammar here, given that at present we do not know how closely speakers' grammars mirror the lexical patterns modelled in this paper, but will sketch what such a grammar (henceforth $\mathcal{G}$) could look like.

A first step towards modelling $\mathcal{G}$ would be to determine whether the patterns presented in this paper reflect what is in the minds of speakers. If that is the case, $\mathcal{G}$ could be modelled within probabilistic versions of Optimality Theory (Prince and Smolensky 1993) where constraints are weighted (Pater 2009), such as MaxEnt Grammar (Hayes and Wilson 2008) or Noisy Harmonic Grammar (Boersma and Pater 2008). A

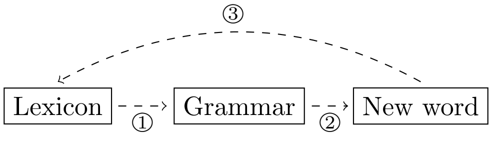
MaxEnt Grammar would make particular sense, given that constraints correspond to different predictors (Goldwater and Johnson 2003).

To map the present analysis into a MaxEnt Grammar, the predictors discussed thus far would be equivalent to Markedness constraints that enforce weight-stress mappings based on the lexical patterns observed in the language. For example, the positional constraint $\textsc{wsp}_n$ (Weight-to-Stress Principle, Prince (1990)) would penalise an unstressed syllable in position n according to the number of segments present in σ_n—where n represents the possible positions in the stress domain.

In $\mathcal{G}$ (Fig. 11), the lexical distributions of stress determine how $\mathcal{G}$ will assign stress probabilistically to a novel word. Once an output is selected (probabilistically), stress will remain lexically marked on the word, which correctly ensures that stress in existing words does not vary.[18] Finally, this novel word will now be part of the lexicon (③ in Fig. 11).[19] Therefore, only words without stress information (i.e., novel words) will be assigned stress probabilistically. This type of distinction between existing and novel words draws on Zuraw (2000, p. 48), who employs 'listedness' as a means to differentiate the two types of words vis-à-vis the application of nasal substitution in Tagalog.

As a result of the probabilistic approach presented here, patterns are no longer treated as regular or irregular, but rather as *more* or *less* likely. For example, in a new word such as *setamira*, penult stress is most likely, but final (and antepenult) stress is also possible for a word of this shape. If the (less likely) candidate with final stress is chosen by the grammar, it will enter the lexicon as *setamirá*. Other constraints in $\mathcal{G}$ will ensure that (i) illicit stress patterns are not generated, e.g., pre-antepenult stress, and that (ii) stress does not shift once it has been assigned (i.e., stress is required to be faithfully realised in the output).

Figure 11: Relationship between lexicon and grammar ($\mathcal{G}$) assumed in the present analysis. Lexical patterns generate constraint weights ①. Stress in new words is assigned based on probabilities ②. Once stressed, a new word enters the lexicon ③.



Because stress is lexically marked in the present approach, speakers need to learn a word with its particular stress position. Under categorical analyses, only irregular cases were lexically marked, since regular cases

[18] An exception is derivationally related words, where stress shifts to obey the trisyllabic window. Although an examination in stress location in such cases is beyond the scope of this paper, it appears to not be probabilistically assigned.

[19] For an alternative which assumes lexically specific constraints, see Moore-Cantwell and Pater, to appear.

were derived based on the generalisations already discussed. The latter approach entails that speakers would memorise only the additional mechanisms responsible for irregular stress (e.g., extrametricality). Crucially, the present approach provides an explanation as to how lexical stress is assigned to *all* words (probabilistically, based on the stress patterns already present in the lexicon). In other words, particular groups of words (e.g., words with antepenult stress) do not require a different explanation.

The probabilistic approach presented here is solely based on the quantitative aspect of weight, which means segmental quality was not part of the model employed in the analysis. Likewise, metrical representation is not included in the model, as the objective was to evaluate how accurate a model solely based on weight could be. Naturally, the absence of such a representation does not imply that a metrical component plays no role in the grammar of Portuguese. For example, even if feet do not play a direct role in primary stress assignment *per se*, they could still play a role in restricting the stress domain to the final three syllables in a word and in assigning secondary stress.

In addition to WSP discussed above, other constraints in $\mathcal{G}$ will play an important role if speakers' grammars in fact encode all the lexical patterns found in this paper. For example, suppose that the specific effect where antepenult rhymes negatively impact antepenult stress is indeed generalised to novel words by speakers. This effect would not be captured by a constraint such as WSP, given that constraints in a MaxEnt framework cannot have negative weights. Instead, positionally defined constraints against marked structures (e.g., *Complex) could indirectly capture the observation that antepenult syllables are more likely to bear stress if their rhymes are minimally complex.

In sum, at present we do not know how speakers' grammars and the lexical patterns modelled in this paper compare. Previous work has shown that statistically significant trends in the lexicon are not necessarily generalised by speakers (Albright and Hayes 2006, Hayes et al. 2009, Becker et al. 2011). For example, the negative effects observed in antepenult syllables mentioned above may not be reflected in the minds of speakers and therefore encoded in the grammar. Likewise, the negative onset effects found for penult and final syllables may be restricted to the lexicon, and may not be generalised to novel forms by native speakers. Crucially, the model presented here formalises the lexicon as a hypothetical baseline, which is a necessary step if one wishes to examine whether the lexicon mirrors speakers' grammars. Future work is needed to investigate how the grammar and lexicon compare vis-à-vis the probabilistic assumptions made in this paper.

6 Conclusion

This paper examined the role of weight in stress assignment in Portuguese. I proposed a probabilistic model that focuses on weight as the only predictor of stress location. The objective of such a model was to show that weight effects are gradient, not categorical as assumed in previous literature (Bisol 1994, Lee 1994, 2007, Wetzels 2007, Mateus and d'Andrade 2000). Likewise, these effects are shown to be more intricate than what traditional approaches presume, given that some effects are negatively correlated with stress (e.g., antepenult nuclei; penult and final onsets).

Assuming that the lexicon does indeed reflect the grammar, the probabilistic grammar implied in this paper considers that stress is assigned based on a probability distribution derived from the patterns present in the lexicon. Stress remains lexically marked once assigned. This approach is substantially different from traditional analyses. First, a formal distinction between regular and irregular patterns no longer exists. Rather, a given stress location is more or less likely. Likewise, weight is not categorically defined (e.g., heavy or light). Instead, a weight continuum is assumed, whereby the notion of weight-sensitivity is understood as inherently gradient (cf. Albright and Hayes (2006), Ryan (2011)).

Unlike previous approaches, the probabilistic analysis proposed in this paper predicts that speakers could in principle assign antepenult stress to a new LLL word, for instance. In contrast, categorical studies predict that so-called irregular cases are not generalisable. Future research is needed to test which of these predictions is confirmed, and whether the weight effects in the Portuguese lexicon are reflected in speakers' grammars. This will also provide a means to compare to what extent the subtleties found in the Portuguese lexicon are in fact captured (and generalised) by speakers. As the relationship between the Portuguese lexicon and speakers' grammars becomes clearer, the probabilistic approach to weight assumed here can be further developed, and its impact on other aspects of the grammar can be evaluated.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. New Jersey: John Wiley & Sons.

Albright, A. and Hayes, B. (2006). Modeling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations. In Fanselow, G., Féry, C., Vogel, R., and Schlesewsky, M., editors, *Gradience in grammar: Generative perspectives*, chapter 10, pages 185–204. Oxford: Oxford University Press.

Amaral, M. P. d. (1999). *As proparoxítonas: teoria e variação*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Araújo, G. A., editor (2007). *O Acento em Português: abordagens fonológicas*. São Paulo: Parábola.

Araújo, G. A., Zwinglio, O. G.-F., Oliveira, L., and Viaro, M. (2007). As proparoxítonas e o sistema acentual do português. In Araújo, G. A., editor, *O Acento em Português: abordagens fonológicas*, pages 37–60. São Paulo: Parábola.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. New York: Cambridge University Press.

Bachrach, A. and Wagner, M. (2007). Syntactically driven cyclicity vs. output-output correspondence: the case of adjunction in diminutive morphology. *U. Penn Working Papers in Linguistics*, 10(1).

Battisti, E. (1997). *A nasalização no português brasileiro e a redução dos ditongos nasais átonos: uma abordagem baseada em restrições*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

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

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Bisol, L. (1992). O Acento: Duas Alternativas de Análise. Unpublished manuscript.

Bisol, L. (1994). The stress in Portuguese. *Actas do Workshop sobre Fonologia*. Universidade de Lisboa.

Bisol, L. (2000). O troqueu silábico no sistema fonológico (um adendo ao artigo de Plínio Barbosa). *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, 16(2):403–413.

Bisol, L. (2013). O Acento: Duas Alternativas de Análise (Stress: two alternative analyses). *Organon*, 28(54):281–321.

Boersma, P. and Pater, J. (2008). Convergence properties of a gradual learning algorithm for harmonic grammar. *Rutgers Optimality Archive*, 970.

Câmara, J. M. (1970). *Estrutura da língua portuguesa*. Petrópolis: Editora Vozes.

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

Christensen, R. H. B. (2013). Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal.

Collischonn, G. (1994). Acento secundário em português. *Letras de Hoje–Estudos e debates de assuntos de linguística, literatura e língua portuguesa*, 29(4):43–55.

Collischonn, G. (1996). Acento em português. In Bisol, L., editor, *Introdução a estudos de fonologia do português brasileiro*, pages 132–165. Porto Alegre: EDIPUCRS, 5th edition.

Cristófaro-Silva, T. (2005). Fonologia probabilística: estudos de caso do português brasileiro. *Lingua(gem)*, 2(2):223–248.

d'Andrade, E. (1994). *Temas de fonologia*, volume 4. Lisboa: Edições Colibri.

Dave, H. (2012). *Frequency word lists: Brazilian Portuguese*. Frequency corpus available at `https://invokeit.wordpress.com/frequency-word-lists/`.

Davis, S. (1988). Syllable onsets as a factor in stress rules. *Phonology*, 5(01):1–19.

de Freitas, M. A. and Barbosa, M. F. M. (2013). A alternância do diminutivo-inho/-zinho no português brasileiro: um enfoque variacionista. *ALFA: Revista de Linguística*, 57(2).

Domahs, U., Plag, I., and Carroll, R. (2014). Word stress assignment in German, English and Dutch: quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, pages 1–38.

Frota, S. and Vigário, M. (2001). On the correlates of rhythmic distinctions: The European/Brazilian Portuguese case. *Probus*, 13(2):247–275.

Frota, S., Vigário, M., Martins, F., and Cruz, M. (2010). Frepop–frequency of phonological objects in portuguese (version 1.0). *Laboratório de Fonética da Faculdade de Letras de Lisboa.*

Garcia, G. D. (2014). *Portuguese Stress Corpus.* Available at `http://www.guilherme.ca/psc`.

Garcia, G. D. (2016). The computation of weight in Portuguese: Syllables and Intervals. In Kim, K.-m., Umbal, P., Block, T., Chan, Q., Cheng, T., Finney, K., Katz, M., Nickel-Thompson, S., and Shorten, L., editors, *Proceedings of the 33rd West Coast Conference on Formal Linguistics*, pages 137–145. Cascadilla Proceedings Project.

Goedemans, R. and van der Hulst, H. (2013). *Weight Factors in Weight-Sensitive Stress Systems.* Leipzig. Available at `http://wals.info/chapter/16`.

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

Gordon, M. (2004). Positional weight constraints in OT. *Linguistic Inquiry*, 35(4):692–703.

Gordon, M. (2005). A perceptually-driven account of onset-sensitive stress. *Natural Language & Linguistic Theory*, 23(3):595–653.

Gordon, M. (2011). Stress systems. In Goldsmith, J. A., Riggle, J., and Alan, C. L., editors, *The handbook of phonological theory*, volume 75. Hoboken: John Wiley & Sons.

Halle, M. and Kenstowicz, M. (1991). The free element condition and cyclic versus noncyclic stress. *Linguistic Inquiry*, 22(3):457–501.

Halle, M. and Vergnaud, J.-R. (1980). Three dimensional phonology. *Journal of linguistic research*, 1(1):83–105.

Halle, M. and Vergnaud, J.-R. (1987). *An essay on stress.* MIT Press, Cambridge, MA.

Harris, J. (1997). Licensing inheritance: an integrated theory of neutralisation. *Phonology*, 14:315–370.

Harris, J. (2011). Deletion. In Oostendorp, M. v., Ewen, C., Hume, E., and Rice, K., editors, *The Blackwell Companion to Phonology.* Oxford: Wiley-Blackwell.

Harris, J. W. (1983). Syllable structure and stress in Spanish: a non-linear analysis. *Linguistic Inquiry Monographs Cambridge, Mass.*, (8):1–158.

Hayes, B. (1989). Compensatory lengthening in moraic phonology. *Linguistic Inquiry*, 20(2):253–306.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: University Of Chicago Press.

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

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

Hermans, B. and Wetzels, L. (2012). Productive and unproductive stress patterns in Brazilian Portuguese. *Revista Letras*, 28:77–114.

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

Hyman, L. M. (1985). *A theory of phonological weight*, volume 19. Dordrecht: Foris Publications.

Kelly, M. H. (2004). Word onset patterns and lexical stress in English. *Journal of Memory and Language*, 50(3):231–244.

Klautau, A. (2013). *UFPADic 3.0*. Retrieved from `http://www.laps.ufpa.br/falabrasil` on 14 Sep, 2013.

Lee, S.-H. (1994). A regra de acento do português: outra alternativa. *Letras de Hoje*, 98:37–42.

Lee, S.-H. (1995). *Morfologia e fonologia lexical do português do Brasil*. PhD thesis, Unicamp.

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

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2):249–336.

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

Massini-Cagliari, G. (1999). *Do poético ao linguístico no ritmo dos trovadores: três momentos da história do acento*. FCL, Laboratório Editorial, UNESP.

Mateus, M. H. and d'Andrade, E. (1998). The syllable structure in European Portuguese. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 14(1):13–32.

Mateus, M. H. and d'Andrade, E. (2000). *The phonology of Portuguese*. Oxford: Oxford University Press.

Mateus, M. H. M. (1983). O acento da palavra em português: uma nova proposta. *Boletim de Filologia*, 28:211–229.

Moore-Cantwell, C. and Pater, J. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. To appear in *Catalan Journal of Linguistics*.

Neto, N., Rocha, W., and Sousa, G. (2015). An open-source rule-based syllabification tool for Brazilian Portuguese. *Journal of the Brazilian Computer Society*, 21(1):1–10.

Olejarczuk, P. and Kapatsinski, V. (2014). The syllabification of medial clusters: evidence from stress assignment. In *Poster from LSA Annual Meeting, Minneapolis*.

Oostendorp, M. v. (2012). Quantity and the three-syllable window in Dutch word stress. *Language and Linguistics Compass*, 6(6):343–358.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, 33(6):999–1035.

Pereira, M. I. (2007). Acento latino e acento em português: que parentesco? In Araújo, G. A., editor, *O acento em português: abordagens fonológicas*, pages 61–83. São Paulo: Parábola.

Prince, A. (1990). Quantitative consequences of rhythmic organization. *Cls*, 26(2):355–398.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roca, I. M. (1999). Stress in the Romance languages. In Hulst, H. v. d., editor, *Word Prosodic Systems in the Languages of Europe*, pages 672–811. Berlin: Mouton de Gruyter.

Ryan, K. M. (2011). Gradient syllable weight and weight universals in quantitative metrics. *Phonology*, 28(03):413–454.

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

Steriade, D. (2012). Intervals vs. syllables as units of linguistic rhythm. Handouts, EALING, Paris.

Thomas, E. W. (1974). *A grammar of spoken Brazilian Portuguese*. Vanderbilt University Press.

Topintzi, N. (2010). *Onsets: suprasegmental and prosodic behaviour.* New York: Cambridge University Press.

Vogel, I. (2008). The morphology-phonology interface: Isolating to polysynthetic languages. *Acta Linguistica Hungarica*, 55(1):205–226.

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

Wetzels, W. L. (1997). The lexical representation of nasality in Brazilian Portuguese. *Probus*, 9(2):203–232.

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

Zuraw, K. (2000). *Patterned exceptions in phonology.* PhD thesis, University of California at Los Angeles.