# Weight Gradience and Stress in Portuguese*

Guilherme D. Garcia
*McGill University*

**Abstract**

This paper investigates how weight influences the location of stress in Portuguese. In this language, weight effects are traditionally seen as categorical, and only word-final syllables are considered to be weight-sensitive. I examine stress placement across the most comprehensive lexicon of Portuguese, and show that weight effects are also present word-internally, and gradually weaken as we move away from the right edge of the word. Additionally, I compare two theories of weight computation, the syllable and the interval—defined as a rhythmic unit that spans from one vowel up to (but not including) the next vowel. Under both theories, weight effects on stress are gradient, and not limited to the right edge of the word. Finally, the probabilistic analysis proposed is shown to be more accurate and parsimonious than traditional approaches, as it accounts for the vast majority of words. Unlike previous analyses, irregularities are captured without the use of any extra mechanisms.

***Keywords***: stress, weight, interval theory, 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 regular and productive (Hermans & 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, unlike ternary systems).

Primary stress placement in Portuguese non-verbs is highly correlated with duration (Major 1985), and can be largely explained by weight, in terms of the following generalizations: stress is final (U) if the word-final

*Thanks to Heather Goad, Morgan Sonderegger, Kie Zuraw, Michael Wagner, 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 & Vigário (2001) for a comprehensive comparison). Phonologically, both BP and EP have an almost identical phonemic inventory (see Mateus & 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.

syllable is heavy—where *heavy* is defined as containing a falling diphthong, a nasal vowel or a coda consonant (1a). Rising diphthongs are traditionally considered to be light, and Portuguese has no long vowels. If the word-final syllable is light, stress falls on the penult (PU) syllable (1b). Taken together, these are the regular stress patterns in the language, which are found in 72% of the lexicon (Houaiss et al. 2001).

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

| | | | |
|---|---|---|---|
| a. | *cacau* [kaˈkaw] 'cocoa' | *anã* [aˈnã] 'dwarf' (f) | *pomar* [poˈmaɾ] 'orchard' |
| b. | *boca* [ˈboka] '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', *jacaré* [ʒakaˈɾɛ] 'alligator'

b. *nível* [ˈnivew] 'level', *míssel* [ˈmisew] 'missile'

c. *fósforo* [ˈfɔsfoɾʊ] 'match' *n*, *pérola* [ˈpɛɾola] '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 realized in derived forms, it bears its own mora, 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 should fall: 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]$_{stem}$[-a]$_{\text{TV}}$ '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 §4, 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, 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) without the use of catalexis, extrametricality or morphological factors. Instead, I propose that weight in Portuguese has a gradient effect on stress, which is positionally[2] and quantitatively determined; i.e., weight effects are gradient across and within each position in the stress domain.

The analysis in this paper is developed in this paper 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. 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.

(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?

Statistical models (§5) 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.

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

Previous research in BP is based on the assumption that onsets do not influence stress—following the traditional view that weight is a property of the rhyme (Chomsky & Halle 1968, Liberman & Prince 1977, Halle & Vergnaud 1987, Halle & 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 (§5), a result that is in line with more recent studies, which have shown that onsets also contribute to weight in other languages (Gordon 2005, Topintzi 2010, Ryan 2014).

In answering question (4c), one would anticipate that onsets have either a positive or a null effect on stress. The lexicon examined here, however, shows that neither seems to be the case: onsets in Portuguese are negatively correlated with stress. Such effects are small (in size) when compared to coda effects (4b), but are highly significant, and cannot be accounted for under a theory of weight computation that assumes a syllabic representation.

Onset effects are not the only inconsistencies found in the Portuguese lexicon (§5.1), but are the main motivation for considering an alternative theory of weight domain, namely, *interval theory* (Steriade 2012). In this theory, the domain of weight is the interval, defined as a rhythmic unit that spans from one vowel up to (but not including) the next vowel. Onset segments in a given syllable are, under interval theory, grouped with the previous interval, in contrast with their organization in syllable theory ($\mathrm{VC}_{\sigma}\mathrm{CCVC}_{\sigma}$ *vs.* $\mathrm{VCCC}_{\iota}\mathrm{VC}_{\iota}$).

Both syllables and intervals account for the weight gradience in the language (§5). This is confirmed by different statistical models,[3] which are presented in §5. Interval-based models are overall more economical and more internally consistent, but not always more accurate than syllable-based models.

This paper is organized 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 briefly review different approaches to weight computation and the role of onsets. In section 4, I analyse the Portuguese lexicon (Houaiss et al. 2001) vis-à-vis weight and stress in order to answer the questions in (4). In section 5, I model the patterns in the lexicon based on syllable theory and interval theory, contrasting the assumptions, results and implications of each theory in different statistical models. The models based on intervals and on syllables show effects that are consistent with a gradient notion of weight-sensitivity in Portuguese. Crucially, given their probabilistic nature, the predictions of both interval and syllable models

[3]Throughout this paper, 'model' is to be equated with 'statistical model'.

are more consistent with the actual lexical patterns than are previous analyses, which assumed categoricity. Finally, section 7 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.[4] 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 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). Some of these forms, however, are repaired via syncope and resyllabification, as long as the resulting form obeys the phonotactic patterns in the language (see Amaral 1999): *fósforo* ⇒ [ˈfɔsfru] ‘match’ *n.* This is the case for most dialects of Brazilian Portuguese, though in some northeastern varieties ‘this pattern has completely vanished in non-verbs’ (Wetzels 2007, p. 29).

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). It is normally assumed that a new word in the language is not likely to have antepenultimate stress (Hermans & Wetzels 2012). Rather, new words tend to have either final or penultimate stress, aside from some borrowings. The words ‘penalty’ [ˈpenaltʃi] and ‘performance’ [pɐˈfɔɾmãnsi], for example, are

[4] 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.

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 (but see Mateus & d'Andrade (2000), who argue that Portuguese stress is in fact not sensitive to weight). 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.

Whether or not morphology influences stress in non-verbs, the systematic patterns in BP stress make it very difficult to assume that Portuguese has unpredictable stress. Table 1 summarizes the stress patterns in non-verbs—'H' stands for a heavy syllable. As mentioned earlier, heavy syllables 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'). The table does not differentiate rising from falling diphthongs, since it only takes into account quantitative information in each syllable. Thus, both rising and falling diphthongs are treated as heavy.

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

## 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, 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

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

| **Stress pattern** | **Regular** | $n$ | **%** | **Irregular** | $n$ | **%** |
|---|---|---|---|---|---|---|
| Final (U) | ...XH́]$_{PWd}$ | 24,060 | 14.7% | ...XĹ]$_{PWd}$ | 5,662 | 3.46% |
| Penult (PU) | ...X́L]$_{PWd}$ | 93,715 | 57.27% | ...X́H]$_{PWd}$ | 18,546 | 11.33% |
| Antepenult (APU) | | | | ...X́LL]$_{PWd}$ | 21,367 | 13.05% |
| | | | | ...X́LH]$_{PWd}$ | 233 | 0.14% |
| | | | | ...X́HL]$_{PWd}$ | 35 | 0.02% |
| | | | | ...X́HH]$_{PWd}$ | 7 | 0.004% |
| | | 117,775 | ≈ 72% | | 45,850 | ≈ 28% |

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 [ʒakaˈɾɛs] (singular: *jacaré*)
STEM PL
*alligators*

b. boc -a -s ['bokas] (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*[5] (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,

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

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 & Wagner 2007).

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

b. sofá -inh -o *sofazinho* (cf. **sofinho*) [ˌsofaˈziɲu]
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 & Barbosa 2013).

(7) a. livr -o -inh -o *livrinho* or *livrozinho* [liˈvɾiɲu] ~ [ˌlivɾʊˈziɲu]
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: $\ldots \mathrm{X}\text{'}\mathrm{L}]_{PWd}$ 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 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, more 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)). On the other hand, theme vowels may be deleted. Since TVs are semantically vacuous, the optionality in (7) is not at all surprising.

There are other phonological processes in BP often said to be associated with theme vowels, such as vowel raising and external sandhi.[6] 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* [kaˌmizɐ uˈzada] ⇒ [kamizuˈzadɐ] 'used shirt', but *jacaré amarelo* [ʒakaˈɾɛ amaˈɾɛlʊ] ⇏ *[ʒakaɾamaˈɾɛlʊ] '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 actually tell these two alternatives apart. The same question can be posed for other Romance languages, where the same problem 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).

A final argument for a morphological effect on stress in non-verbs could be that derivational suffixes in Portuguese affect stress—this includes the diminutive suffix in (6). This suffix, however, is traditionally treated as a PWd on its own—see Vigário (2003). The main motivation for such an analysis is related to vowel raising: in standard Portuguese (Brazilian and European), low-mid vowels are only contrastive in stressed position. Thus, we have [paˈpɛw] *-ada* ⇒ [ˌpapeˈladɐ] 'paper(work)', but [paˈpɛw] *(z)inho* ⇒ [ˌpapɛwˈziɲu] (cf. *[ˌpapewˈziɲu]). This is because most such suffixes do cause stress shifting. However, as pointed out by Garcia (2012), these stress shifts can be phonologically motivated, and, thus, can be accounted for without invoking morphological information: in (9a), stress shifts to respect the trisyllabic window constraint. In this

[6] Vowel deletion across word boundaries.

case, stress is shifted to avoid ungrammaticality. (9b), on the other hand, simply follows the stress patterns listed in (1).

(9) a. [ˈatomo -iko] ⇒ [aˈtomiko] ($\sigma$ [$\sigma$ $\sigma$ $\sigma$)
atom -ic/-ical
'Atomic'

b. [kaˈfɛ -(z)al] ⇒ [ˌkafeˈzaw] (LLH́)
coffee 'place'
'Coffee plantation'

However we approach these suffixes, the fact remains that almost all derivational suffixes in the language follow phonological patterns that one would expect (e.g., (9b)). In other words, we see the exact same stress patterns as displayed by monomorphemic forms.

The only compelling reason to propose a morphological component for non-verbs was to accommodate irregular cases where words with a light final syllable had final stress. Though a valid attempt, there is no independent evidence for a morphological effect here. …XĹ]$_{PWd}$ forms clearly deviate from regular cases in Portuguese. Recall that only 3.46% of words in the lexicon (5.7% of all …XL words) fall into that category (Table 1).

Whether or not theme vowels exist in present-day Portuguese is beyond the scope of this paper, but their 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. Because such analyses make reference to syllabic constituency, I first describe syllable shape in Portuguese.

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.ˈca.bro]. In other words, stop+liquid clusters in Portuguese are not ambiguous vis-à-vis their syllabification (Cristófaro-Silva 2005).

In the rhyme, up to four segments (including the nucleus) may be present.[7] Rising diphthongs are traditionally treated as light, and usually arise through hiatus resolution: *piada* 'joke' [pi.ˈa.da] ⇒ [ˈpja.da]. In the very few cases where two coda consonants are present, the second element is an /s/, and is almost always found in prefixes (*trans-*). Therefore, only one coda consonant is commonly found in Portuguese (a nasal, a liquid, or an /s/).

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 *crisp* is produced as [ˈkrispi]). Recent borrowings are not the only words that are repaired: common words also undergo epenthesis and resyllabification if they violate the syllabic template in Portuguese: *advogado* 'lawyer' and *obstetra* 'obstetrician', for example, are produced as [ad{i,e}.vo.ˈga.do] and [o.bis.ˈtɛ.tɾa] in BP, respectively.

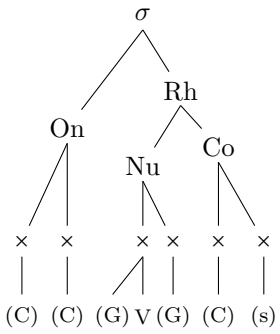
The syllabification algorithm in Portuguese is fairly straightforward and unambiguous, given the restricted number (and quality) of segments in complex onsets and codas (see Thomas (1974) for a comprehensive description and Neto et al. (2015) for a computational implementation). A nonce word such as *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 & d'Andrade 1998). Take the word *devedor* 'debtor', which is syllabified as /de.ve.ˈdoɾ/. In colloquial EP, however, vowels are frequently deleted, which results in more complex onsets: [dvdor]. This type of reduction never happens in BP (as we have seen, certain coda-onset sequences often undergo epenthesis).

Traditionally, the concept of weight has been tied to the presence of rhyme segments only—thus excluding onsets from the domain of weight (Halle & 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 falling 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*

[7] Rhymes with five segments (or more) are unattested.

Figure 1: Syllabic structure in Portuguese




[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 indicates 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 traditional view mentioned above.

As seen in §2, Portuguese stress in non-verbs is weight-sensitive (Wetzels 2007). However, previous studies only consider weight in rhymes, and most assume that weight only affects stress word-finally (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. Wetzels (2007), however, argues that weight may also play a role word-internally, given the behaviour of some palatal sonorants in Brazilian Portuguese.[8] Although consonantal quality does not have an evident effect on stress in the language, the one clear exception is the palatal consonants [ɲ, ʎ], which are never found in final onsets

[8]In the stress domain. 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).

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.ˈraʎ.ʎo] (*[ˈba.ɾaʎ.ʎo]) 'deck of cards' (see Fig. 1). This would be 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 often 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 $\sigma$** | | **Closed $\sigma$** | |
|---|---|---|---|---|
| **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, however, is whether or not the unstressed syllables in a given word are closed or 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 predicted stress pattern for a given word. §4 will examine whether this is the case.

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 & van der Hulst 2013). Among these languages, we find Spanish and Romansch, both closely related to Portuguese.

Stress is not the only context where weight effects are found in Portuguese: 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 applies to non-verbs only, and neutralizes the mid vowel contrast in the stressed syllables of words with antepenult stress. SL is conditioned by weight—more specifically, by the weight of the word-final syllable (see Table 3). This pattern also suggests that weight effects in the language are restricted to the word-final syllable, as SL is not observed in words with a heavy penult syllable. 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)

| **. . . V́L]**$_{PWd}$ | Gloss | **. . . V́H]**$_{PWd}$ | 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’ |

Syllable weight is traditionally formalized using Moraic Theory (Hyman 1985). Some previous analyses of stress in Portuguese indeed assume (or imply) such a framework to account for weight effects in the language (see Lee (2007) and Hermans & Wetzels (2012) for recent approaches). Given the positional bias 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* [saˈpatu] ‘shoe’ is parsed as [sa(ˈpa$_\sigma$to$_\sigma$)]. Let us now briefly look into how the moraic approach 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 trochee is built from the right edge of the word: (ˈ$\sigma$ $\sigma$) ⟨$\sigma$⟩. Likewise, words with penult stress and a heavy final syllable (. . . ˈXH]$_{PWd}$) are explained with segment extrametricality, which makes the (heavy) final syllable light: ˈCV.CV⟨C⟩. For . . . XˈL]$_{PWd}$ words, Bisol (1992) proposes a catalectic consonant, which is only phonetically

realized in derivations: *café* [kaˈfɛ] 'coffee' would then be represented as [ka.fɛC]. The catalectic consonant C makes the final syllable heavy, and the moraic pattern is maintained. We can see such a consonant in derived forms: *cafe̲teira* 'coffee pot' *vs. cafe̲zal* 'coffee plantation'—note that the quality of the catalectic consonant can vary in derivations of the same stem.

With respect to onsets, under Moraic Theory, a . . . CV.ˈCV word (*jacaré*) and a . . . CV.ˈCCV word (*colibri* 'hummingbird') 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 §4 I show that the Houaiss corpus deviates from these predictions. In view of this, let us now examine how weight can be affected by onsets, in an attempt to understand why the predictions of Moraic Theory just discussed do not hold once we examine the lexicon.

## 3 Onsets and weight: two alternative views

In the previous section, I reviewed both morphological and phonological factors proposed in traditional analyses to account for the stress patterns found in Portuguese. Most researchers agree on the major role weight plays in determining stress in the language. Unlike the morphological factors examined in §2.1, weight has a clear effect on stress likelihood and spondaic lowering. Thus, the analysis proposed in this paper focuses on phonological factors only, in particular weight.

Examining the effects of onsets on weight is relatively recent in the literature (though see Davis (1988)). The classic view that onsets are outside the domain of weight becomes problematic as more cross-linguistic evidence is brought to light: some Australian languages, such as Agwamin and Aranda, seem to differentiate between V and CV, favouring the latter in stressed syllables (see Topintzi (2010)). Likewise, geminate onsets (C:V) attract stress more than CV syllables in Bellonese and Trukese, among other languages (Topintzi 2010, Ryan 2014). Onset effects have also been found in several well-studied languages in the last decade, including English and Russian, where syllables with larger onsets are more likely to attract stress (Gordon 2005, Topintzi 2010, Ryan 2011).

In light of the cross-linguistic evidence, the question arises as to whether onset effects are also observed in Portuguese. We could consider an alternative scenario to the classic view that onsets play no role in Portuguese, namely, that the number of onset segments positively correlates with stress. The question, thus, is *do onsets affect stress in the Portuguese lexicon and, if so, how?*

### 3.1 P-center theory

An alternative view for the computation of weight that can include a role for onsets is provided by Ryan (2014), who argues that the domain of weight does not start at the left edge of the rhyme. Rather, it begins with the p-center (perceptual center; Morton et al. 1976), which is influenced by the onset segments in a given syllable. The p-center corresponds to the moment when speakers perceive that a rhythmic unit begins. This perception can be empirically seen when beats and speech are aligned: beats normally align with the left edge of the rhyme, not the left edge of the syllable. However, as more material is inserted between the left of the syllable and the left edge of the rhyme (i.e., in onset position), the p-center is perturbed, and is perceived to be earlier. For example, if we compare *spa* and *ba*, the former has an earlier p-center than the second: if we clap to the beat of both words, the clap on *spa* will be slightly earlier in the word than the clap on *ba*. This indicates that the [s] has some influence on where the perceptual centre will be (Ryan 2014, p. 328).

According to the p-center model, thus, the more onset material a syllable has, the earlier the p-center in that syllable will be perceived. Ryan (2014, p. 329) finds that in English 'onset consonants affect weight by roughly 35-47% as much as coda consonants do'. Onsets have a weaker effect than codas, given that the alignment between rhythmic unit and perception favours rhymes over onsets—i.e., the latter only perturbs the p-center, whereas the former is included in it *a priori*. It is important to note that this theory is not necessarily tied to a traditional view of syllabic representation—onsets and codas could simply be referred to as pre- and post-nuclear segments (where 'nuclear' simply stands for the most prominent position in a given grouping of segments). As Ryan (2014, p. 21) puts it, '...the p-center is not a syllabically defined event, but a perceptual function whose exact characterization remains unclear'.

The bigger question, however, is whether the Portuguese lexicon actually contains the patterns predicted by p-center theory, i.e., if the patterns of primary stress placement in the language are compatible with such a theory. If the answer is *yes*, then we should find that complex onsets have a stronger effect on weight than singleton onsets. What we actually find in Portuguese, however, is a more intricate pattern, where onsets have a negative effect on stress likelihood. This is unexpected under syllable theory, where we would expect a positive effect. In view of this, in the following section, I consider an alternative domain for weight computation, namely, intervals.

## 3.2 Interval theory

Given how syllabic constituency is understood, it would be phonologically surprising if onset size *negatively* correlated with stress in a given language (when rhymes are controlled for), i.e., if increasing onset size in syllable $j$ increased stress likelihood on syllable $j-1$, but *decreased* stress likelihood on syllable $j$. Yet, the Portuguese lexicon seems to present such a pattern (§5). Onset effects in the language thus contradict the representational assumptions of syllable theory. Because of this, in this section I briefly review an alternative weight domain, namely, the interval.

Steriade (2012) examines whether the domain of weight computation, referred to as '$\pi$',[9] is the syllable ($\sigma$) or the interval ($\iota$) between two vowels. Simply put, an interval is a rhythmic unit that spans from a given vowel up to (but not including) the following vowel (i.e., a V-to-(V) interval). Thus, segments preceding the leftmost vowel in a word are not included in any interval (i.e., they are extrametrical). All intervals begin with a vowel, i.e., the sonority peak of a rhythmic unit, which means that onset segments in $\sigma_{j-1}$ in syllable theory are part of the interval $\iota_{j-2}$ (see Fig. 2).[10] Unlike standard syllable theory, intervals have no well-defined internal constituency. Thus, onsets and codas are no longer formally distinguished.

Steriade (2012) suggests that intervals capture cross-linguistic weight hierarchies more accurately than syllables. In her motivation for intervals, Steriade illustrates different weight hierarchies present in several languages, such as Finnish, Norwegian, Greek, Latin, Bhojpuri and Estonian. In Bhojpuri (Shukla 1981), for example, (C)VCC and (C)VVC syllables are allowed in all positions (including word-finally). In certain words, VV syllables lose primary stress to VC syllables ([pán.càː.ì.tî][11] 'assembly'); in some other words, VC syllables lose stress to VV syllables ([màh.táː.rì] 'mother')—the difference being the presence or absence of an onset following the syllable with the long vowel. This paradox (i.e., VV > < VC) is resolved if we assume a weight hierarchy computed from intervals, where [páncàːìtî] and [màhtáːrì] are parsed as [⟨p⟩ánc$_\iota$àː$_\iota$ìt$_\iota$î] and [⟨m⟩àht$_\iota$áːr$_\iota$ì], respectively. Because intervals include segments that would otherwise be part of the onset of the following syllable, the distinction in question is captured.

The weight hierarchy arising from interval theory is thus able to capture the patterns in the data examined by Steriade: VVC > VCC > VV > VC. This hierarchy would not be possible in a theory based on syllables, where the weight of a given syllable is not affected by the onset of the following syllable. As a result, the VV syllables in VV.CV and VV.V have equivalent weight, and subtle weight distinctions such as those discussed above are not predicted. Crucially, the different structural assumptions of syllables versus intervals result in

[9] $\pi$ is intended to be theory-neutral notation (see Steriade (2012)).
[10] It follows that, in a rising diphthong, the glide is parsed into the previous interval: VCG$_\iota$VC$_\iota$.
[11] Where '^' stands for *no stress* (Steriade 2012, p. 8).

different alignments between the rhythmic unit and segmental material. This, in turn, yields differences in the computation of weight. Because VCC is heavier than VV but lighter than VVC, both 'páncà꞉ìtî' and 'màhtá꞉rì' are accounted for. Simply put, intervals can be thought of as a particular algorithm for parsing segments.

Fig. 2 exemplifies how $\pi$ differs in both theories (syllables and intervals) in a CVCCCVVC word. The longer the interval, the heavier it is: VCCC > VCC > VC > V. Therefore, syllables and intervals in Fig. 2 present different weight hierarchies: $j, j-2 > j-1$ (syllables) and $j-2 > j > j-1$ (intervals). Importantly, any segment preceding a vowel is parsed into the previous interval. Word-initial consonants are extrametrical ($\langle$C$\rangle$ in Fig. 2).

Figure 2: Syllables (.) and intervals (•)

$$\text{CVC}_{j-2}.\text{CCV}_{j-1}.\text{VC}_{j}]_{PWd} \qquad\qquad \langle\text{C}\rangle\text{VCCC}_{j-2}\bullet\text{V}_{j-1}\bullet\text{VC}_{j}]_{PWd}$$

The weight of an interval is a function of the overall duration of $\pi$, and so is affected by different phonological/phonetic properties, such as the length of post-vocalic segments and inherent properties of vowels (height, frontness etc.). Interval size is therefore correlated with duration, rather than segment count. In this paper, for convenience, I adopt Hirsch's (2014) less fine-grained metric where duration is assumed to positively correlate with the number of segments in a given interval.

Because intervals assume different groupings from syllables, the predictions that follow, too, are different: since adjacent consonants are now merged into a single interval, this entails that every segment that contributes to duration (i.e., all of them) also contributes to weight. Table 4 compares both theories by parsing BP words with different stress patterns.

Table 4: Syllables and intervals: Portuguese words

| | **Stress pattern** | **Word** | $\pi = \sigma$ | $\pi = \iota$ |
|---|---|---|---|---|
| (a) | Final | *casal* [kaˈzaw] 'couple' | ka.zaw | ⟨k⟩az • aw |
| (b) | | *café* [kaˈfɛ] 'coffee' | ka.fɛ | ⟨k⟩af • ɛ |
| (c) | Penult | *pato* [ˈpatʊ] 'duck' | pa.to | ⟨p⟩at • o |
| (d) | | *nível* [ˈnivew] 'level' | ni.vew | ⟨n⟩iv • ew |
| (e) | Antepenult | *parábola* [paˈɾabʊla] 'parabola' | pa.ɾa.bu.la | ⟨p⟩ar • ab • ol • a |

At first glance, by looking at number of segments (i.e., intervals) rather than syllables, weight could motivate the location of stress in cases (c) and (d) in Table 4: a word like 'pato', for example, is traditionally accounted for by proposing that, given the absence of a word-final coda, stress should fall onto the penultimate syllable. With intervals, this pattern arises due to the longer duration (i.e., the greater number of segments) in that interval. Words like 'nível', which are deemed irregular in syllable theory, would not be accounted for in interval theory, given that the two intervals have the same size. Here, as in syllable-based approaches, other mechanisms would have to be employed to define which interval will bear stress (still assuming a categorical calculation of weight, where two intervals with the same number of segments are treated as equally heavy). Such mechanisms may include, for example, directionality.

Interval theory does worse than syllable theory in cases such as (a), which is regular in a syllabic approach (§2). In addition, exceptional cases like (b) go against intervals, and cases like (e) are problematic under both views. Of course, this oversimplified comparison tells us very little about how each theory would account for the patterns in the language as a whole, since coda and onset sizes (or interval sizes) naturally vary in the lexicon. Given that Table 4 only contains singleton onsets and codas, this should not be taken to be a representative analysis of intervals for one simple reason: increasing the number of onset or coda segments will not affect a syllabic view of stress, where weight is treated as categorical (CVC and CVCC are equally heavy in syllable-based analyses of Portuguese stress). On the other hand, intervals will be affected by such a change. Therefore, we need to look at a large enough corpus in order to undertake a more realistic and comprehensive comparison of the two approaches.

In this section, I briefly reviewed an alternative view of weight computation, namely, intervals and p-centers—both of which consider onsets to be relevant. However, these alternatives need not be mutually exclusive, and the interval-based analysis provided in this paper (§5) could incorporate p-center theory. Before moving to an examination of the data in §4, let us briefly consider how both theories could work together.

### 3.3 Integrating intervals and p-center theory

Since p-center theory is not based on a syllabic representation, the duration of segments in an interval could also cause the interval boundary to shift leftwards. Recall that this paper assumes for convenience that intervals count segments, and not duration *per se*. However, in order to integrate p-center theory and interval theory, information about duration would be necessary, given that a boundary shift may be only a fraction

of a segment. Based on Ryan (2014), we can hypothesize that the shift ($S$) of an interval ($\iota$) boundary would be equal to 35%[12] of the sum of the mean added duration ($D$) of the segments ($C$) included in that interval (in ms): $S(\iota) = -\sum_{i=1}^{n} D_{C_{i_\iota}} \cdot 0.35$.

Under syllable theory, onsets are pre-nuclear elements, and therefore the p-center is shifted leftwards as a function of the number of onset segments added to a given syllable. Under interval theory, however, onsets in syllable $j-1$ are parsed as post-nuclear segments of interval $j-2$ (Fig. 2). If codas and onsets have different effects on the p-center, intervals need to distinguish such units. One alternative would be to include the differences between onsets and codas in intervals. In this case, intervals and syllables would simply be two different algorithms for parsing onsets, nuclei and codas. Because the presence of constituents and a hierarchical relation between such constituents are two separate dimensions, syllables and intervals would still be structurally different, but would share the same elements.

Another possibility would be to assume two levels of parsing: firstly, (i) sequences of segments are syllabified according to universal and language-specific phonotactic constraints—no stress is assigned at this level. Then, (ii) such sequences are 'reparsed' as intervals, after which stress is assigned. Crucially, onsets and codas, determined by (i), would now be two distinct units in (ii). As a result, we would have interval rhythmic units whose internal constituents are differentiated as per syllable theory.

This possibility would not be parsimonious, as it would require that strings be scanned twice. Instead, moras could be assigned to segments on the basis of relative sonority, regardless of any structural assumptions. Let us take two quantitatively identical strings: $\text{VCC}^i.\text{V}$ and $\text{VCC}^i_\mu.\text{V}$. In the second string, $\text{C}^i$ is more sonorous than in the first string, and therefore contributes more to weight.

In this subsection, I have shown how p-centers and intervals could be implemented together. Such an integration is relevant in Portuguese, where complex onsets have a weaker effect than coda segments (§5). Recall that the crucial issue to be examined in this paper is whether weight is gradient and positionally-defined. In the next section, I show that weight gradience is found under both syllable theory and interval theory, but that onsets seem to be correlated with stress in an interval fashion. In §5 I compare both syllables and intervals, and discuss which domain better fits the lexicon examined here.

# 4 Data

This section explores the Portuguese lexicon in an attempt to answer the three questions posed in §1, repeated in (10) for convenience.

[12] This value is based on the English lexicon (Ryan 2014), and therefore only serves as an example here.

(10) 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 (10) are clearly connected, since (10a) examines where weight-sensitivity is found and (10b) examines how it affects stress. Likewise, question (10c) also affects the answer to question (10b).

The data examined in this paper is largely based on the most comprehensive lexicon available in the Portuguese language: the Houaiss Dictionary (Houaiss et al. 2001). The Houaiss corpus contains 442,000 entries (unlemmatized types), of which 164,291 are non-verbs, including monosyllables. Thus, the lexicon used in this paper contains nearly all non-verbs in Portuguese, and includes a list of orthographic words with pronunciations, syllabifications and parts of speech.

Given its large size, the 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 [t͡ʃn] 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 lexicon (≈ 5.6%). Monosyllables were also removed from the lexicon (≈ 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 diphthongization in Portuguese:. . . ˈCV.CV.V ⇒ . . . ˈCV.CGV. For example, *terráqueo* /te.ˈxa.ke.o/ is realized as [te.ˈxa.kjʊ] ‘earthling’. Diphthongization 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 diphthongization yields penult stress. As a result, these data could potentially bias the analysis.[13] Thus, this subset of words was removed from the data. The final corpus (Garcia 2014)

[13]In fact, statistical models were run with and without such words, and the predicted negative correlation was confirmed. No other effects were influenced by these forms.

contains 154,611 unlemmatized types of non-verbs (see Table 5).

Grapheme-phoneme conversion was done by different scripts and regular expression substitutions. Some cases, however, are ambiguous. For example, the grapheme *x* can be realized 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 realizations involve 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.

Among the rare words in the Houaiss 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 corpus, 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. The IL corpus and the LaPS corpus are used here to ensure that all three stress patterns are balanced in the corpus used in this study, i.e., that the proportions of each stress pattern in the Houaiss lexicon are mirrored in the spoken language.

Table 5: Portuguese/BP* corpora

| **Stress pattern** | **Houaiss** | **IL*** | **LaPS*** |
|---|---|---|---|
| Final | 18% | 21% | 27% |
| Penult | 69% | 71% | 62% |
| Antepenult | 13% | 8% | 11% |
| | $n$=154,611 | $n$=39,705 | $n$=8,468 |

## 4.1 Weight-sensitivity: the Portuguese lexicon

In this subsection, I examine how weight-sensitivity affects stress placement in the Portuguese lexicon (Houaiss et al. 2001). 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 (§4.1.1) or interval (§4.1.2) may affect stress: both subtle and strong effects are found in all three syllabic positions, namely, onset, nucleus and coda, as well as in all intervals in the stress domain. In section 5, I present statistical models that capture such trends in the lexicon as well as empirical differences regarding syllables and intervals.

### 4.1.1 Weight and syllables

The Houaiss 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, [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 6). 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 6). 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 6) or consonantal correlations with stress.

#### 4.1.1.1 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. 3

Table 6: Most frequent onset and coda segments by stress pattern

| | **Final $\sigma$** | | **Penult $\sigma$** | | **Antepenult $\sigma$** | |
|---|---|---|---|---|---|---|
| **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/ |

(Antepenult $\sigma$). The primary focus of the data analysis that follows is to visualize 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. 3 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 $\sigma$, Penult $\sigma$ and Final $\sigma$, respectively).

Figure 3: Onset size effects by syllable and stress pattern

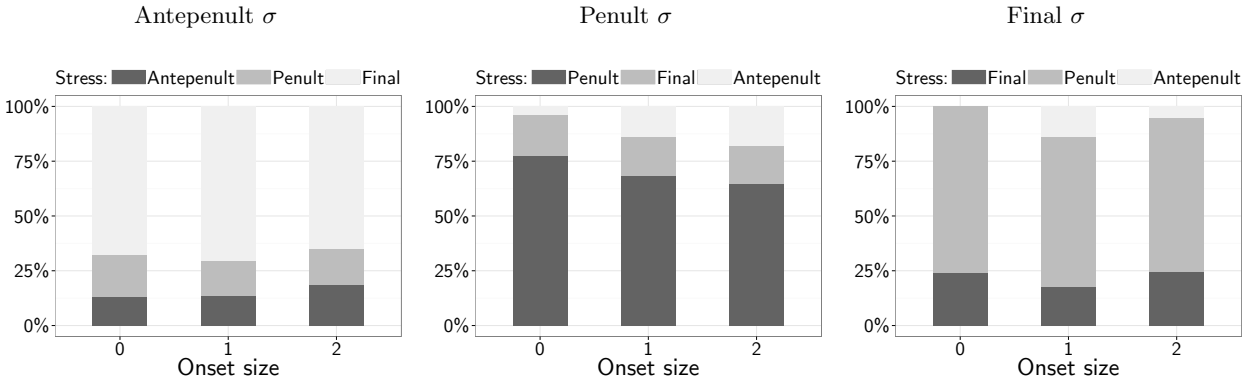



Fig. 3 suggests that onsets are positively correlated with stress in the antepenult and final syllables. The number of words with antepenult stress does not seem to be affected in different ways when the antepenult onset size is either 0 or 1. Rather, the difference in the Antepenult $\sigma$ plot lies between {0,1} and 2 segments. For both antepenult and final syllables, onset effects on stress are not clear in the figures.

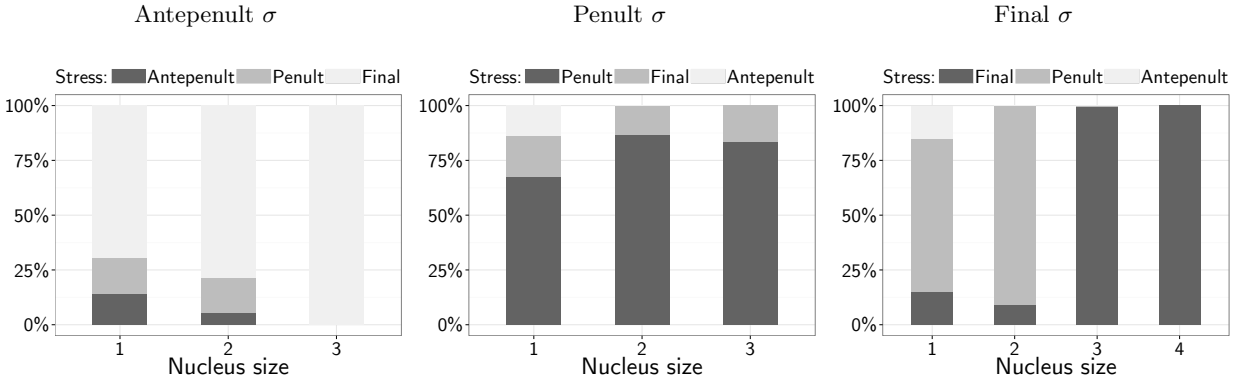
We can see in Fig. 3 that onset size is negatively correlated with stress in penult syllables. In other words, as we increase the number of onset segments in the penult syllable, we observe a decrease in the number of words with penult stress. Interestingly, it is the number of words with *antepenult* 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

(§4.1.2). The importance of these effects will be examined in §5.

#### 4.1.1.2 Nucleus size effects

Nuclei and codas are expected to have stronger effects on stress than onsets. This is the case in traditional views of syllable theory (discussed in §3) as well as in p-center theory (§3.1). In Fig. 4, we can see that words with penult and final stress seem to be affected by penult and final nucleus size, respectively. Diphthongs (nucleus size `2`) have a stronger effect on stress than monophthongs (`1`), consistent with typological weight distinctions, where complex nuclei are heavier than V nuclei. Note, however, that the distinction 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). Surprisingly, antepenult nuclei seem to have a *negative* effect on antepenult stress, which is clearly unexpected.

Figure 4: Nucleus size effects by syllable and stress pattern

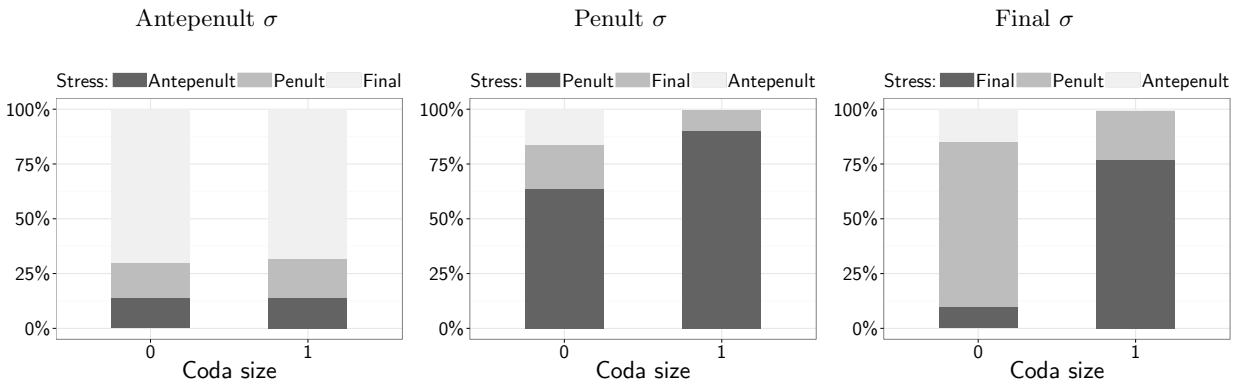



#### 4.1.1.3 Coda size effects

Let us now examine the effect of coda size on stress placement. Fig. 5 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 likely 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 show 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, on the other hand, 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 has a very strong (negative) effect on antepenult stress.

Figure 5: Coda size effects by syllable and stress pattern




The trends observed above suggest that the effect of syllable weight is gradient, not categorical: coda effects are 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, and such cases count as complex nuclei in Fig. 4. 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 §5 below.

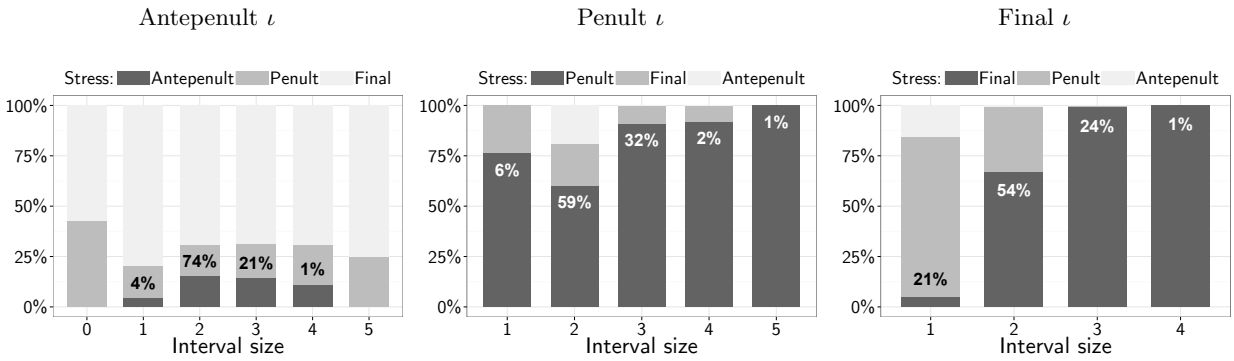
### 4.1.2 Weight and intervals

Thus far, the data presented in Figures 3, 4 and 5 assume standard syllabic representations. Recall, however, that the onset patterns in Fig. 3 suggested a negative correlation between onsets and stress, which is inconsistent with syllable theory. Likewise, the antepenult nucleus effects in Fig. 4 also pose problems for an approach based on syllables. Let us now examine the data according to interval theory, the alternative domain of weight computation discussed in §3.2.

In Fig. 6, the proportion of words with each stress location in the lexicon is plotted as a function of the size of each interval. Recall that the penult interval includes all segments between the penult and final vowels

(as per Fig. 2). Likewise, the antepenult interval includes all segments between the antepenult and penult vowels. The $x$-axis represents the size of each interval ($\iota$), where *1* represents no intervening consonantal segments in a V-to-(V) interval (i.e., an interval consisting of a single vocalic segment)—rising and falling diphthongs each count as two interval segments. A comparison of Fig. 6 with Figs. 3, 4 and 5 allows us to contrast what the data (and the stress patterns) look like under the two different representational approaches.

Figure 6: Intervals and stress patterns




To interpret the trend in Fig. 6, we need to take into consideration the number of words in the lexicon that fall into each category, i.e., interval size. For example, although antepenult intervals vary from 1 to 5, less than 1% of the data actually contain a 5-segment interval. What can be seen here is that longer intervals tend to be positively correlated with stress in all three cases, although antepenult intervals present a less clear picture—we will see in §5 that the trend in the penult and final intervals also applies to the antepenult interval.

Given that intervals conflate onsets, nuclei and codas, it might be difficult to tell exactly where the independent effects within each interval originate (assuming one wishes to directly compare interval- and syllable-based empirical patterns). However, as we already know the effects of each of these constituents from §4.1.1 above, it should be possible to unpack which parts of different syllables trigger the effects in Fig. 6. Crucially, the antepenult interval may have a more straight-forward answer to that question: firstly, recall that antepenult onsets are not counted in the antepenult interval (§3.2). Rather, the antepenult interval is composed of the antepenult nucleus and coda, and the penult onset. Secondly, we have seen that neither nuclei nor codas in the antepenult syllable seem to have a positive effect on stress (Figs. 4 and 5). Therefore, the effect in the antepenult interval seen in Fig. 6 must be largely due to the penult onsets. We can find support for this hypothesis by removing codas and both types of diphthongs from the data, and considering

only LLL words, since only onsets would vary in that particular subset.

Before we proceed, it should be noted that, because syllables and intervals are independent theories, a direct comparison between the representation under each theory is perhaps inappropriate. In other words, one should be cautious when interpreting (and evaluating) the interval patterns observed in the data solely based on syllables.

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. 5, $\ldots$LL$]_{PWd}$ words will most likely have pre-final stress regardless of onset size. Even if p-center theory is supported in Portuguese, 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 7: Stress patterns by final onset size in $\ldots$LLL words

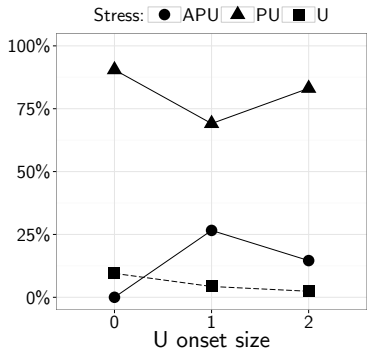


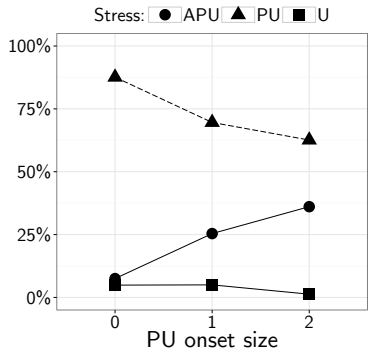

Fig. 7 suggests that larger final onset sizes (specifically from 1 to 2) are more highly correlated with penult stress than final stress. It should be noted that singleton onsets are much more frequent in the Houaiss corpus than complex onsets: 84.3% *vs.* 4.7% in words with final stress, 89.1% *vs.* 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. 7 could indicate that intervals play some role in the patterns observed in these data, given that final stress seems to be negatively correlated with final onset size—a result we would not expect if the domain of weight computation in Portuguese is the syllable. Let us now examine how penult onset size

affects stress.

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. 8 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. Interval theory, on the other hand, predicts an increase in the likelihood of antepenult stress.

Figure 8: Stress patterns by penult onset size in ...LLL words




We see in Fig. 8 that the likelihood of antepenult stress increases when the penult syllable contains onset segments. Figs. 7 and 8 show a clear pattern, which is consistent with interval theory. Antepenult onset size (not shown here), 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. Antepenult syllables are at the edge of the stress domain, and their high degree of unpredictability might explain (at least in part) the unexpected patterns that we find. As we will see in the next section, antepenult syllables show a pattern that is not accounted for under syllables nor under intervals.

# 5 Statistical analysis

In the previous section, we observed that both syllables and intervals seem to show gradient weight effects on 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 syllables and intervals. In §5.1 and §5.2, I describe each statistical model proposed, analyse the results, and examine how they relate to

the main questions in this paper, stated in (10). In §6, these models are compared to previous approaches, which serve as the baseline for the present analysis.

The factors examined in §4 are listed in Table 7—note that the independent variables (i.e., predictors) are separated into two groups, namely, syllables and intervals. Given the representational differences between these two domains, statistical models based on syllables naturally have more predictors (3 syllables × 3 constituents = 9). Intervals, on the other hand, may be characterized with only three predictors (i.e., as many predictors as the number of positions in the stress domain).[14] Antepenult factors are coded as `NA` in disyllabic words.

Table 7: Predictors and response

| | | |
|---|---|---|
| Syllables | `onset.fin` | Number of onset segments in the **final** $\sigma$ (`0-2`) |
| | `nucleus.fin` | Number of segments in the nucleus of the **final** $\sigma$ (`1-4`) |
| | `coda.fin` | Number of coda segments in the **final** $\sigma$ (`0,1`) |
| | `onset.pen` | Number of onset segments in the **penult** $\sigma$ (`0-2`) |
| | `nucleus.pen` | Number of segments in the nucleus of the **penult** $\sigma$ (`1-3`) |
| | `coda.pen` | Number of coda segments in the **penult** $\sigma$ (`0,1`) |
| | `onset.ant` | Number of onset segments in the **antepenult** $\sigma$ (`0-2`) |
| | `nucleus.ant` | Number of segments in the nucleus of the **antepenult** $\sigma$ (`1-3`) |
| | `coda.ant` | Number of coda segments in the **antepenult** $\sigma$ (`0,1`) |
| Intervals | `int1` | Number of segments in $\iota$ `0` (**final**): `1-4` |
| | `int2` | Number of segments in $\iota$ `1` (**penult**): `1-5` |
| | `int3` | Number of segments in $\iota$ `2` (**antepenult**): `0-5` |
| | Response | `antepenult`, `penult`, `final` |

The analysis presented in this section employs multiple 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

[14] As pointed out by an anonymous reviewer, the absence of internal constituency is not a necessary assumption in interval theory. Rather, this is the position adopted in this paper, following Steriade (2012) and Hirsch (2014).

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. 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 2014)). As mentioned, 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 7). This division is aligned with traditional analyses, which classify antepenult stress as irregular, and penult/final stress as (mostly) regular (§2). Both models will be interpreted separately, as each predictor will yield two different effects (coefficients).

Logistic Regressions predict the log-odds of $y = 1/0$ based on a set of predictors. In this case, $y = antepenult/\{penult, final\}$ in one model and $y = penult/final$ in another model. The fitted model is given in (11), where $Pr(y_i = 1)$ denotes the probability that response $y = 1$; $\beta^0$ 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. For example, assume we have a CVCCVCV word such as *martelo* 'hammer', which would be parsed into intervals as ⟨C⟩VCC.VC.V (⟨m⟩art.el.o). In the `antPenFin` model, $Pr(y_i = APU/\{PU, U\})$ would be $logit^{-1}(\beta^0 + 3 \cdot \beta^1 + 2 \cdot \beta^2 + 1 \cdot \beta^3)$, given that we have 3 segments in the antepenult interval (`int3`), 2 segments in the penult interval (`int2`) and 1 segment in the final interval (`int1`)—assuming our $\beta^1$ represents `int3`. In this case, we are interested in how much `int1`, `int2` and `int3` affect such a probability (i.e., the $\beta^{1-3}$ values).

(11) **Logistic Regression**

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

Because both syllables and intervals will be compared, a total of four statistical models will be presented. Crucially, as we will see, *both* pairs of models capture gradient weight in Portuguese. Besides, given the probabilistic nature of the approach proposed in this paper, both syllable- and interval-based 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 all four models and discuss their results and predictions. In §6, I contrast these predictions with the actual patterns in the lexicon, and show how the present analysis differs from previous approaches.

## 5.1 Syllable models

### 5.1.1 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 8, 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 only present word-finally. 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 8 have been scaled, and are therefore directly comparable to one another.

The results in Table 8 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* affect antepenult stress.

The onset effects we observe in Table 8 are consistent with the data trends discussed in §4.1.1, i.e.,

Table 8: Scaled 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** | $\hat{\beta}$ | **OR** | **SE** | ***z* value** | ***p* value** |
|---|---|---|---|---|---|
| `onset.ant` | 0.109 | 1.115 | 0.009 | 12.503 | < 0.00001 |
| `nucleus.ant` | -0.219 | 1.245 | 0.012 | -18.331 | < 0.00001 |
| `coda.ant` | -0.051 | 1.052 | 0.008 | -5.989 | < 0.00001 |
| `onset.pen` | 0.334 | 1.396 | 0.009 | 36.756 | < 0.00001 |
| `nucleus.pen` | -1.107 | 3.025 | 0.047 | -23.456 | < 0.00001 |
| `coda.pen` | -2.725 | 15.256 | 0.119 | -22.843 | < 0.00001 |
| `onset.fin` | 0.626 | 1.87 | 0.014 | 45.807 | < 0.00001 |
| `nucleus.fin` | -2.774 | 16.023 | 0.132 | -20.970 | < 0.00001 |
| `coda.fin` | -1.169 | 3.219 | 0.026 | -44.155 | < 0.00001 |
| AIC: 84402 | | | | | $\kappa$ = 28.19 |

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 assigned 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 3.2, whereas the presence of a penult coda reduces the odds of antepenult stress by a factor of 15.8 (see §5.1.2 for a discussion on nucleus-coda effects). These observed differences in effect size also capture a 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*[15] (Belsley et al. 1980). The predictors included in the model in Table 8 have high collinearity ($\kappa = 28.19$).

[15]Represented here by $\kappa$. 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).

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, 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 $\rho^2$ 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 major problems for the analysis.

Finally, Table 8 also lists the AIC value (Akaike Information Criterion, Akaike 1974) of the syllable model, which is a measure of the relative quality of the regression. By adding parameters (i.e., variables/predictors) to a model, we expect the model's fit to improve. However, the model may overfit the data, thus losing relevant information regarding the process under examination. Therefore, there is a trade-off between the number of parameters in a model and the resulting increased error in that model. The Akaike information criterion tells us that the lower the AIC value, the better the fit. Naturally, the AIC value for a given model is only meaningful in comparison to the AIC value of another model.

#### 5.1.2 Model B: `penFin`

The `penFin` model in Table 9 shows that only penult onsets have no significant effect on penult (*vs.* final) stress—all other five 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 can see that most of the trends discussed in §4.1.1 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 1.15. This effect is inconsistent with syllables. We also see that both `nucleus.fin` ($\hat{\beta} = -1.108, p < 0.00001$) 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 in the `antPenFin` model discussed above for both final and penult positions. This contradicts a strong typological tendency, whereby VV is heavier than VC (Gordon 2011). However, recall that Portuguese has no long vowels. Instead, complex nuclei consist of a single vowel and a glide. Importantly, not all complex nuclei in the language are assumed to affect stress, as rising diphthongs are traditionally treated as light (§2.2). The model presented

in Table 9 makes no distinction between rising and falling diphthongs, since `nucleus.fin` and `nucleus.pen` simply count the number of segments in the domain.[16] This could explain why the effect of nuclei is smaller than that of codas in this case. 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 exact same pattern is found in the `antPenFin*` model.

Table 9: Scaled 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

| $\sigma$ **predictor** | $\hat{\beta}$ | **OR** | **se**($\hat{\beta}$) | $z$ **value** | $p$ **value** |
|---|---|---|---|---|---|
| `onset.pen` | 0.011 | 1.01 | 0.01 | 1.13 | 0.315 |
| `nucleus.pen` | −0.084 | 1.09 | 0.01 | −7.99 | < 0.00001 |
| `coda.pen` | 0.139 | 1.15 | 0.01 | 12.06 | < 0.00001 |
| `onset.fin` | 0.142 | 1.15 | 0.01 | 15.50 | < 0.00001 |
| `nucleus.fin` | −1.108 | 3.03 | 0.01 | −136.00 | < 0.00001 |
| `coda.fin` | −1.493 | 4.45 | 0.01 | −181.18 | < 0.00001 |
| AIC: 78307 | | | | | $\kappa = 18.23$ |

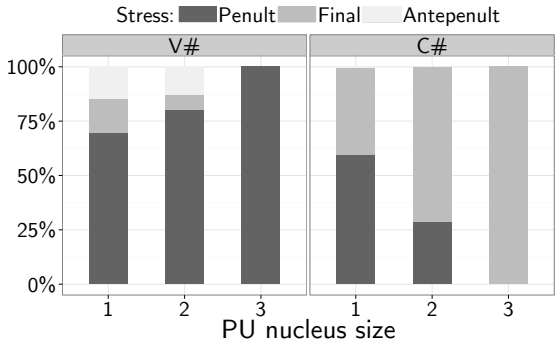
Let us now examine the results of `nucleus.pen` and `coda.pen`. First, `nucleus.pen` shows a negative effect on penult stress, which is unexpected. This, again, could be connected to the distinction between rising and falling diphthongs discussed above: if most diphthongs in `nucleus.pen` happen to be *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 alone 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. 4. The interaction between these two variables, however, is not captured in Fig. 4, since nuclei are plotted independently. Once we visually inspect these two variables (Fig. 9), 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. 9). A word such as *fácil* 'easy' (ĹH) is, in fact, more frequent in

[16] This is further discussed in §5.1.3.

the Portuguese lexicon than a word such as *álbum* 'album' (ĤH). In other words, if we only look at these two weight profiles in the lexicon, the proportion of words with penult stress in HH words is lower than the proportion of words with penult stress in LH words (13% *vs.* 20%).

Figure 9: 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 9 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 §5.1.3 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 capture the weight gradience in Portuguese.

The fact that penult nuclei and word-final codas do interact and that such an interaction is significant ($p = 0.02$) is puzzling, and deserves further investigation. In fact, once we examine the entire lexicon, it becomes clear that CVG.CVC words, for example, are much less likely to have penult stress (9%) when compared to CV.CVC words (23%). Interestingly, almost all CV́G.CVC words are in fact borrowings (e.g., *cáiser*, *lêucon* 'kaiser', 'leucon'), and are rarely used in Portuguese.

Let us now turn to `coda.pen`, which had a significant effect in the `penFin` model. The positive coefficient value of this predictor ($\hat{\beta} = 0.139$) indicates that adding a coda segment to the penult syllable increases the odds of penult stress by a factor of 1.15. This is naturally a much smaller effect than, for example, `coda.fin`, but it is highly significant. The effect sizes listed in Table 9 clearly show a gradient effect, whereby predictors in the final syllable have a greater absolute effect than predictors in the penult syllable.

In Table 9, `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. P-center theory (§3.1) would predict that word-initial

onset size in such cases should favour penult stress. Indeed, if we restrict the `penFin` model to disyllables only ($n$=11,356), we do find that `onset.pen` has a positive effect on penult stress ($\hat{\beta} = 0.16, p < 0.00001$).

### 5.1.3 Syllable models: 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.

The reason why the models above do not differentiate rising and falling diphthongs is as follows. How rich a model is has to do with the types of theoretical and representational assumptions said model should encode. 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? We are interested in a model that predicts stress based on segmental (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. 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 syllable 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) or (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 ĤL trochees, (ĹL)⟨X⟩ is better than (ĤL)⟨X⟩, and therefore the former is more likely than the latter—all else being equal. This could indicate that light antepenult syllables are more stress-attracting for footing reasons, and not weight *per se*. Needless to say, this explanation relies on extrametricality.[17] A third footing option, namely, (Ĥ)L⟨X⟩, is preferred to (ĤL)⟨X⟩. However, this leaves a medial syllable unfooted, which contradicts traditional foot-based analyses of Portuguese.

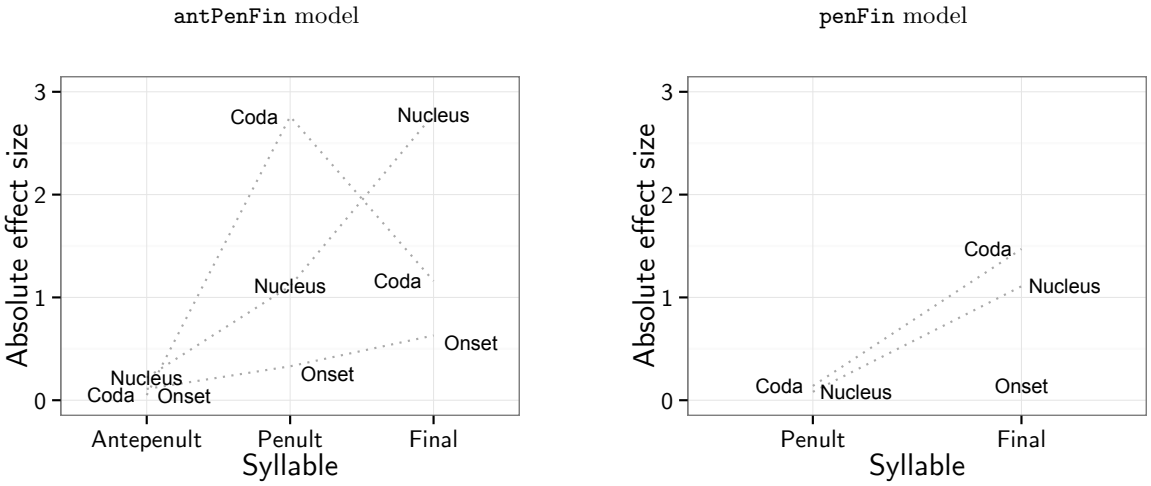
Not all facts are accounted for by extrametricality and footing patterns, however. For example, whereas the negative effect of `nucleus.pen` would be explained, the positive effect of `coda.pen` would not. Furthermore, the onset effects found in both models would require an additional explanation, as 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 syllable-based models discussed above show that weight is not a categorical phenomenon in the language. In Fig. 10, 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 (`penAnt` model).

As can be seen in Fig. 10, 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 (`penFin` model) and penult syllables (`antPenFin` model), but Nucleus > Coda > Onset for final syllables in the `antPenFin` model.

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

Figure 10: Absolute effect sizes in the syllable models




## 5.2 Interval models

Let us now turn to the interval models. Given the representational assumptions in this paper, such models will, by definition, have fewer predictors (one per interval). Importantly, all phonemes in the language count as *one* interval segment, including palatals ([ʎ, ɲ]). Separate models were run assuming such phonemes count as two segments (§2.2), but the effects found were not different from the models presented in this paper. Below, I present and discuss the results of the `antPenFin` and `penFin` models. Then, in §6, I compare both syllable and interval models against the actual patterns found in the data.

### 5.2.1 Model A: `antPenFin`

The `antPenFin` model based on intervals is presented in Table 10. In this model, the probability of antepenult stress (as opposed to penult or final stress) is predicted based on `int1`, `int2` and `int3`. We can see that both `int1` and `int2` negatively affect antepenult stress ($\hat{\beta} = -1.70, p < 0.00001$ and $\hat{\beta} = -0.52, p < 0.00001$, respectively). This indicates that adding a segment to the final interval decreases the odds of antepenult stress by a factor of 5.47. The fact that the antepenult interval (`int3`) also has a highly significant effect on antepenult stress contradicts the assumption that this stress position is not affected by weight.

The level of collinearity between the predictors in this model is lower when compared to the `antPenFin` predictors in the syllable model in Table 8 (14.59 *vs.* 28.19). The present model, however, has a worse fit, i.e., a higher AIC value (104146 *vs.* 84402).

Table 10: Scaled 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

| $\iota$ **predictor** | $\hat{\beta}$ | **OR** | **se**$(\hat{\beta})$ | $z$ **value** | $p$ **value** |
|---|---|---|---|---|---|
| `int1` | −1.70 | 5.47 | 0.01 | −42.80 | < 0.00001 |
| `int2` | −0.52 | 1.68 | 0.01 | −61.39 | < 0.00001 |
| `int3` | 0.11 | 1.12 | 0.01 | 14.67 | < 0.00001 |
| AIC: 104146 | | | | | $\kappa = 14.59$ |

### 5.2.2 Model B: `penFin`

The `penFin` model based on intervals is presented in Table 11. Positive coefficient values indicate that a given predictor favours penult stress (as opposed to final stress). As expected, penult stress is positively affected by `int2` ($\hat{\beta} = 0.34, p < 0.00001$) but negatively affected by `int1` ($\hat{\beta} = -1.92, p < 0.00001$). In other words, each segment added to `int1` decreases the odds of penult stress by a factor of 6.82—recall that this discrepancy is also captured in the syllable models discussed in §5.1.

Table 11: Scaled 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

| $\iota$ **predictor** | $\hat{\beta}$ | **OR** | **se**$(\hat{\beta})$ | $z$ **value** | $p$ **value** |
|---|---|---|---|---|---|
| `int1` | −1.92 | 6.82 | 0.01 | −185.12 | < 0.00001 |
| `int2` | 0.34 | 1.40 | 0.01 | 30.94 | < 0.00001 |
| AIC: 72924 | | | | | $\kappa = 11.82$ |

Note that the set of variables used in this model (`int1` and `int2`) presents relatively low collinearity when compared to the syllabic variables. As well, compared to its syllabic counterpart (Table 9), the present model has a slightly better fit, i.e., a lower AIC value.

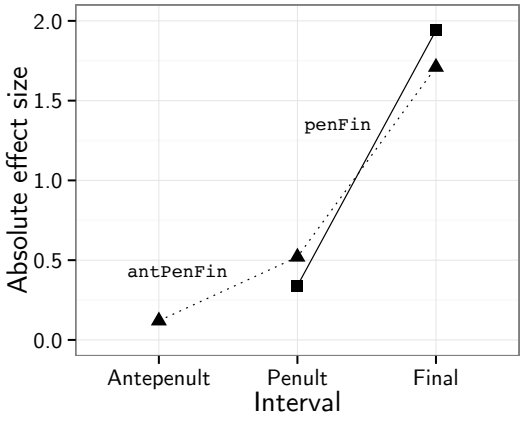
### 5.2.3 Interval models: assessment

Both interval models are consistent with the trends observed in Fig. 6, where more segments in an interval correlated with higher stress rates in the vast majority of words in the lexicon. The models show that interval size is highly significant across all positions in the stress domain (`int1`, `int2` and `int3`). This argues against previous analyses, which assume that weight effects are only found word-finally.

Crucially, the models clearly show a gradient weight effect, whereby weight-sensitivity weakens as we

move away from the right edge of the word. This gradient pattern is summarised in Fig. 11, which plots the absolute effect size of each interval for both interval models.

Figure 11: Absolute effect sizes in the interval models by interval




# 6 Discussion

In this section, I summarise and discuss the main results presented in this paper. §6.1 evaluates the accuracy of the syllable- and interval-based models. §6.2 briefly explores the implications of the approach adopted here for the grammar of Portuguese.

The models discussed in §5 clearly answer the questions in (4). 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 all four statistical models discussed above. Third, onsets do contribute to weight, but in a way which is consistent with intervals rather than syllables being the domain over which weight is computed (at least for penult and final positions).

The syllable-based models discussed 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 in predicting antepenult stress; (ii) final codas have a stronger effect than final nuclei in predicting penult stress; (iii) heavy antepenult rhymes disfavour antepenult stress. Indeed, these facts are potentially more problematic than the onset effects observed, given that neither syllables nor intervals are consistent with them.

The crucial aspect of the analysis proposed here does not depend on the results of the comparison between intervals and syllables. Rather, the most important characteristic of the present approach is its probabilistic

nature. A categorical approach cannot explain why a certain irregular pattern exists (e.g., ĹLL), given that it deviates from traditional generalizations about the language (XXH́ else XX́L). The present proposal, however, predicts that all licit stress patterns are possible (including so-called irregular cases), but some are more likely than others. Consequently, it is no longer the case that all irregular forms are *equally* unlikely, an implication of traditional analyses. As we will see below, the probabilistic nature of the present approach results in a more accurate characterization of stress in Portuguese.

## 6.1 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 both `antPenFin` models in Fig. 12, 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., sequence of light (L) and heavy (H) syllables) in the language.[18] These weight profiles conflate different word shapes; e.g., CV.CV.CV and CV.CCV.CV are both treated as LLL, since the baseline in this case is the syllable. 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 size corresponds to lexical representativeness). In some cases (e.g., HHH, HHL, LHL), categorical predictions accurately match the actual lexical proportions (Fig. 12). However, a clear mismatch is observed for HLL and LLL words. Finally, $\sigma$ and $\iota$ represent the mean predicted probability of antepenult stress based on the syllable and interval models, respectively.

We can see that both models approximate the actual lexical values—although intervals are less accurate for HHL and LHL words when compared to the syllable-based model. 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 of final. Traditional approaches, on the other hand, would not predict antepenult stress in this particular case.

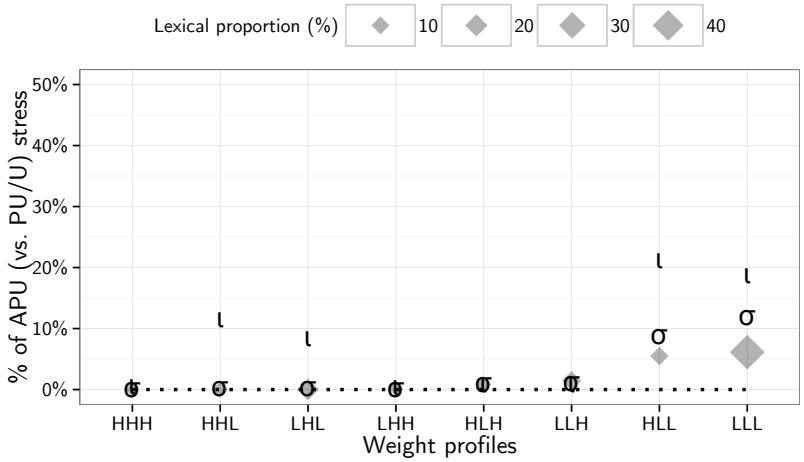
Assuming that a word is not assigned antepenult stress, we now need to consider penult *vs.* final stress, which account for the vast majority of words in the lexicon (Table 1). Fig. 13 plots the proportion (or probability) of words with penult stress (*vs.* final stress) across the different weight profiles in the language. Recall that traditional approaches predict final stress for all words with a heavy final syllable (XXH́) and penult stress elsewhere (X́XL)—this is represented by the dotted lines in Fig. 13. Clearly, these predictions

[18] Predicted probabilities in each model are averaged across all words with a given weight profile.

deviate considerably from the actual lexical proportions of penult stress (◆).

Like Fig. 12, Fig. 13 shows that both syllable- and interval-based 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, ĤL words are more frequent than ĹL words—and this difference is mirrored in the models' mean predicted probabilities.

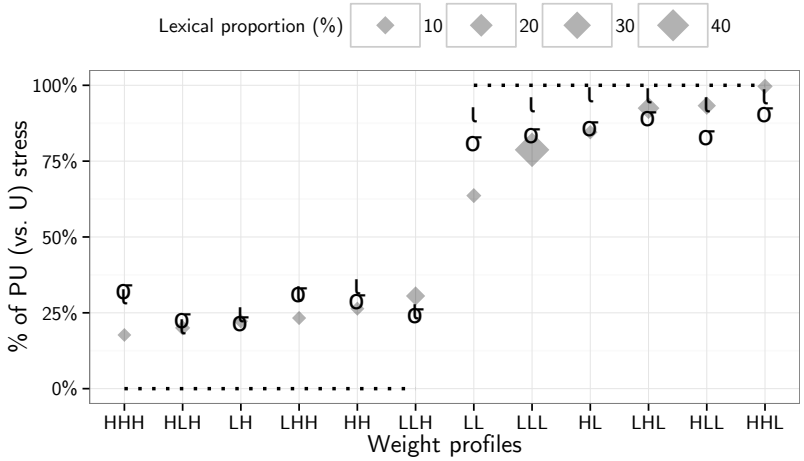
Figure 12: `antPenFin` models' accuracy: Mean predicted probabilities of antepenult (*vs.* penult/final) stress by weight profile. Predictions of interval model ($\iota$), syllable model ($\sigma$), and actual lexical frequencies (◆) are plotted. Dotted lines indicate predicted stress based on a standard categorical analysis




Figs. 12 and 13 both provide a means to compare the present proposal to traditional analyses of Portuguese stress. In these cases, categorical weight was used to define sets of words, so that traditional weight-based approaches serve as the baseline. The use of H/L, however, does not fully allow us to compare syllable and interval models, or visualize how well they do for different cases, since H/L profiles collapse many cases where the models make different predictions. Thus, a more appropriate representation of accuracy would consist of CV strings instead of weight profiles. This is shown in Figs. 14 and 15, where the different parsings of intervals and syllables can be contrasted.

One difficulty with graphically representing predictions for CV strings is that too many such templates exist: 592 for all words included in the `antPenFin` model, and 735 for all words included in the `penFin` model. For that reason, Figs. 14 and 15 only consider CV strings that comprise at least 1000 words each. These representative templates ($n$=16 in Fig. 14 and $n$=23 in Fig. 15) account for approximately 80% and 75% of the data modelled, respectively. In both figures, syllable boundaries are added for clarity, specifically to

Figure 13: `penFin` models' accuracy: Mean predicted probabilities of penult (*vs.* final) stress by weight profile. Predictions of interval model ($\iota$), syllable model ($\sigma$), and actual lexical frequencies (◆) are plotted. Dotted lines indicate predicted stress based on a categorical analysis




disambiguate alternative syllabifications for medial CC strings; in interval theory, such strings would always be parsed with the preceding interval.

Figure 14: `antPenFin` models' accuracy: Mean predicted probabilities of antepenult (*vs.* penult or final) stress by CV string. Interval model ($\iota$), Syllable model ($\sigma$) and actual lexical frequencies (◆) are plotted. Dotted lines indicate predicted stress based on a categorical analysis

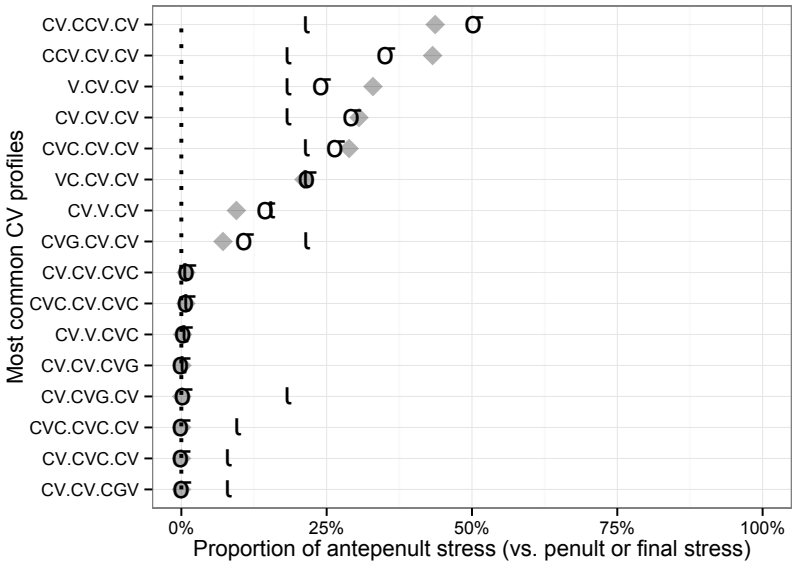


Figure 15: `penFin` models' accuracy: Mean predicted probabilities of penult (*vs.* final) stress by CV string (syllabification is provided). Interval model ($\iota$), Syllable model ($\sigma$) and actual lexical frequencies (◆). Dotted lines indicate predicted stress based on a trochaic categorical analysis

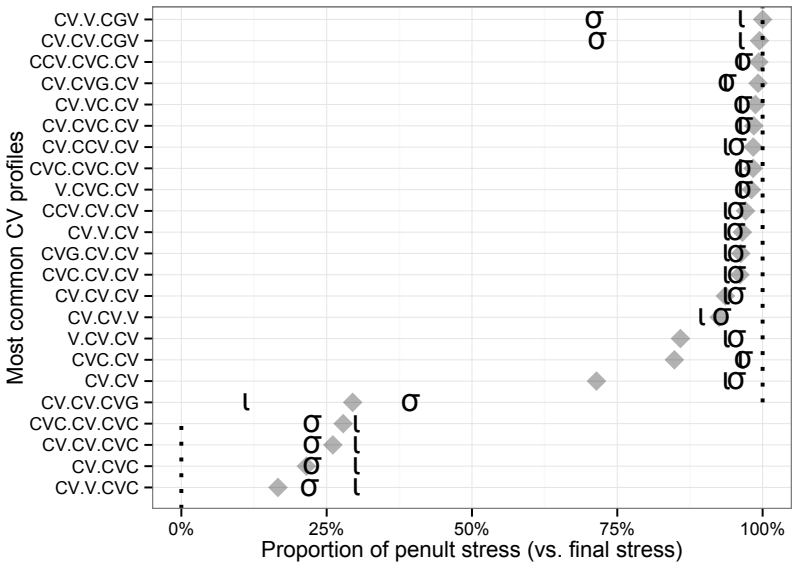



Fig. 14 suggests that syllable-based predictions (`antPenFin` model) are overall more accurate than interval-based predictions, which perform poorly in CV.CVG.CV words such as *coveiro* 'gravedigger', for example. This overall pattern is confirmed if we calculate the weighted mean[19] deviation of predicted probabilities from actual lexical proportions (considering *all* CV strings in the lexicon). On average, the deviation of interval-based predictions is 11% (SD = 9%), whereas the mean deviation of syllable-based predictions is 2% (SD = 5%). For example, the interval-based model predicts a ≈25% probability of antepenult stress, which clearly deviates from the actual proportion (0%). Thus, we can conclude that the syllable-based `antPenFin` model is generally more accurate than its interval counterpart—even though the model itself is not internally consistent (§5).

In Fig. 15, syllables and intervals perform more similarly (`penFin` model). We can see that CV.CV words (⟨C⟩VC•V under interval theory) are especially difficult to predict for both syllable- and interval-based models. CV.V.CGV and CV.CV.CGV words, however, have stress placement better predicted by an interval-based model (e.g., *vestuário, vigário* 'clothing', 'vicar', respectively). Therefore, once we examine the `penFin` models, intervals are slightly more accurate than syllables in certain cases. On average, the deviation of interval-based predictions is 5% (SD = 8%), whereas the mean deviation of syllable-based predictions is

[19]Different CV strings contain different numbers of words in the lexicon. Here, means are weighted on the basis of the representativeness of each string.

6% (SD = 9%).

Overall, the present models approximate the actual lexical proportions more accurately than a categorical approach—regardless of the weight domain one chooses. The weighted mean deviation of predictions based on a categorical analysis is 91% (SD = 11%) for the set of words modelled by the `penFin` model, and 14% (SD = 16%) for the set of words modelled by the `antPenFin` model (see Table 12). As well, the patterns we observe are very similar whether we use weight profiles (Figs. 12 and 13), which conflate segmental information, or CV strings (Fig. 14 and 15), which may make a comparison with traditional approaches less straight-forward.

Table 12: Weighted mean deviation of mean predicted probabilities from actual lexical proportions: syllable model, interval model, and categorical analyses

| | **$\sigma$ model** | | **$\iota$ model** | | **Cat. analyses** | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| `antPenFin` | 2% | 5% | 11% | 9% | 14% | 16% |
| `penFin` | 6% | 9% | 5% | 8% | 91% | 11% |

## 6.2 A probabilistic grammar

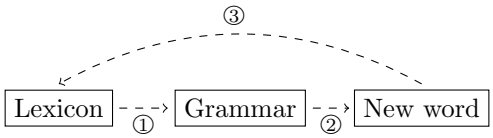
Thus far, we have investigated the stress patterns in the Portuguese lexicon by employing different statistical models. Little has been said, however, about what these patterns mean for the *grammar* of Portuguese speakers. If the lexical patterns explored in this paper are psychologically real, an important question is (i) how such patterns could be implemented in a phonological grammar and (ii) how lexicon and grammar interact. We will not construct such a grammar here, but will sketch what it would look like.

The grammar implied by the present proposal can be captured within probabilistic versions of Optimality Theory (Prince & Smolensky 1993) where constraints are weighted (Pater 2009), such as MaxEnt Grammar (Hayes & Wilson 2008) or Harmonic Grammar (Legendre et al. 1990, Potts et al. 2010). A MaxEnt Grammar would make particular sense, given that constraints correspond to different predictors (Goldwater & Johnson 2003). In other words, 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 $\text{WSP}_n$ (Weight-to-Stress Principle, Prince (1990)) would assign one violation mark to every non-nuclear segment in an unstressed syllable/interval in position $n$ (where $n$ represents the possible positions in the stress domain). The weight of each such constraint would be determined based on

the distribution of stress patterns in the lexicon.

As a new word enters the language, it is assigned stress based on a probability distribution (② in Fig. 16). Once an output is selected, stress remains lexically marked on the word, which is now part of the lexicon (③ in Fig. 16). In other words, the analysis proposed here entails that stress is assigned probabilistically, and that, once assigned, stress information is retained in the lexical entry. As a result, patterns are no longer treated as regular or irregular, but rather *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 inviolable constraints 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 realized in the output). This relationship between lexicon and grammar is represented in Fig. 16.

Figure 16: Relationship between lexicon and grammar 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 were derived based on the generalizations already discussed. This entails that speakers would memorize only the 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. Finally, because stress is lexically marked, mechanisms such as extrametricality are no longer needed. Instead, the grammatical approach assumed here includes weight-based constraints as well as constraints banning illicit stress (e.g., in pre-antepenultimate position).

# 7 Conclusion

In this paper, I argued for a gradient notion of weight-sensitivity in Portuguese. The lexicon examined clearly shows that weight affects stress in all three positions in the stress domain, contra previous analyses (Bisol 1994, Lee 1994, 2007, Wetzels 2007, Mateus & d'Andrade 2000). These effects, however, reveal unexpected results if we assume that syllables are the domain of weight computation. For example, onset size showed a negative correlation with stress, a fact which is consistent with Interval Theory. Interval-based models are more internally consistent, but are less accurate when predicting antepenult (*vs.* penult or final) stress.

The probabilistic grammar implied in this paper assumes that stress is assigned based on a probability distribution derived from the patterns already present in the lexicon of the language. By definition, stress remains lexically marked once assigned. This approach is substantially different from traditional analyses. First, the notion of regular and irregular patterns no longer exists. Rather, a given stress location is more or less likely. Likewise, the weight of rhythmic units (syllables or intervals) 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. The second important difference between the present analysis and previous approaches is the fact that *all* words need to be learned with their respective stress position. Traditional analyses assume (i) that specific mechanisms (e.g., extrametricality) have to be learned by speakers and associated with particular words, given the so-called unpredictable patterns, or (ii) that a subset of words (namely, the irregular cases) needs to be learned with stress (e.g., Lee (2007) assumes the input is stressed in such cases). In other words, regular and irregular words are treated as belonging to formally different classes.

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 generalizable. Future research is needed to test whether these predictions are confirmed, and whether the weight effects in the Portuguese lexicon are also 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 generalized) by speakers—and whether intervals are indeed more internally consistent based on speakers' intuitions. As the relationship between the Portuguese lexicon and speakers' grammars become clear, the probabilistic approach to grammar assumed here can be further developed.

## References


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

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Amaral, M. P. d. (1999). *As proparoxítonas: teoria e variação*. Ph.D. thesis, PUC-RS.

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

Araújo, G. A., Zwinglio, O. G.-F., Oliveira, L., & Viaro, M. (2007). As proparoxítonas e o sistema acentual do português. In G. A. Araújo (Ed.) *O Acento em Português: abordagens fonológicas*, (pp. 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., & 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).

Belsley, D. A., Kuh, E., & 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 Lingüística 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.

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

Chomsky, N., & 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 L. Bisol (Ed.) *Introdução a estudos de fonologia do português brasileiro*, (pp. 132–165). Porto Alegre: EDIPUCRS, 5th ed.

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*, vol. 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., & 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., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, (pp. 1–38).

Frota, S., & 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., & 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. (2012). *Aquisição de acento primário em inglês por falantes de português: uma análise de derivações com sufixos não neutros via algoritmo de aprendizagem gradual—GLA*. Master's thesis, Universidade Federal do Rio Grande do Sul (UFRGS).

Garcia, G. D. (2014). *Portuguese Stress Corpus*. GitHub repository available at `https://github.com/guilhermegarcia/portuguese_corpus/wiki`.

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

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

Gordon, M. (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 J. A. Goldsmith, J. Riggle, & C. L. Alan (Eds.) *The handbook of phonological theory*, vol. 75. Hoboken: John Wiley & Sons.

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

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

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

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

Harris, J. (2011). Deletion. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.) *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., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

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

Hirsch, A. (2014). What is the domain for weight computation: the syllable or the interval? In *Proceedings of Phonology 2013*.

Houaiss, A., Villar, M., & 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*, vol. 19. Dordrecht: Foris Publications.

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*. Ph.D. thesis, Unicamp.

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

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, (pp. 884–891). Mahwah, NJ: Lawrence Erlbaum Associates.

Liberman, M., & 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 lingüístico no ritmo dos trovadores: três momentos da história do acento*. FCL, Laboratório Editorial, UNESP.

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

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

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, *83*(5), 405.

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

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 G. A. Araújo (Ed.) *O acento em português: abordagens fonológicas*, (pp. 61–83). São Paulo: Parábola.

Potts, C., Pater, J., Jesney, K., Bhatt, R., & Becker, M. (2010). Harmonic grammar with linear programming: from linear systems to linguistic typology. *Phonology*, *27*(01), 77–117.

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

Prince, A., & Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. Published in 2004 by Oxford: Blackwell.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org/`

Roca, I. M. (1999). Stress in the Romance languages. In H. van der Hulst (Ed.) *Word Prosodic Systems in the Languages of Europe*, (pp. 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.

Shukla, S. (1981). *Bhojpuri grammar*. Washington, DC: Georgetown University Press.

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.

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

Vigário, M. (2003). *The prosodic word in European Portuguese*, vol. 6. Berlin: Walter de Gruyter.

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.