Intervals and weight gradience in Portuguese*

Guilherme D. Garcia
McGill University

Abstract

This paper investigates how the location of primary stress in Portuguese is influenced by weight. Traditionally, the effect of weight in Portuguese is seen as categorical, and only word-final syllables are considered to be weight-sensitive—i.e., the weight of penult and antepenult syllables is not relevant for stress. I examine stress placement across a comprehensive lexicon of Portuguese non-verbs (Houaiss et al. 2001), and show that weight effects (i) are also present word-internally and (ii) gradually weaken as we move away from the right edge of the word—this gradience is supported by different Ordinal Regression models. I compare two theories of weight domain, namely, the *syllable* and the *interval* (Steriade 2012)—defined as a rhythmic unit that spans from one vowel up to (but not including) the next vowel. Under both theories, weight has a gradient effect on stress, and is not limited to the right edge of the word. The data analysed suggest that intervals may be more accurate representations of the weight domain in Portuguese, in that they appropriately capture onset effects found in the lexicon that are not accounted for under syllable theory. In addition, interval-based models are more accurate than syllable-based models. Finally, I briefly demonstrate how the analysis proposed here can be mapped into a MaxEnt model (Goldwater & Johnson 2003, Wilson 2006, Hayes & Wilson 2008).

Keywords: stress, weight, interval theory, MaxEnt

1 Introduction

This paper examines Brazilian Portuguese (BP) primary stress in non-verbs,[1] and argues for a gradient notion of weight-sensitivity 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, that is, 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 (i.e., a syllable is either heavy or light according to the shape of its rhyme, unlike ternary systems).

*Thanks to Heather Goad (McGill) and Morgan Sonderegger (McGill), for crucial comments on this paper. Thanks also to Kie Zuraw (UCLA), for valuable suggestions on MaxEnt models, Michael Wagner (McGill), and to the GMHP (Group on Historical Morphology of Portuguese) at the University of São Paulo (USP).

[1] BP and European Portuguese (EP) are practically identical vis-à-vis primary stress, thus most of what follows could 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 all transcriptions are in BP, I use 'BP' and 'Portuguese' interchangeably in this paper, as the lexicon examined here is not limited to Brazilian Portuguese.

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 if the word-final syllable is heavy—where *heavy* is defined as containing a falling diphthong,[2] a nasal vowel or a coda consonant (1a). Otherwise, stress falls on the penult syllable (1b). This is the regular stress pattern in the language, which is 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 cases: final stress when the word-final syllable is light (2a); penult stress when the word-final syllable is heavy (2b); and antepenult stress (2c)[3].

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

a. *jacaré* [ʒakaˈɾɛ] 'alligator'

b. *nível* [ˈnivew] 'level'

c. *fósforo* [ˈfɔsforʊ] 'match' *n* (often repaired as [ˈfɔsfrʊ])

Scholars 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 scholars to propose that morphological factors govern the location of stress as well—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 (2a) is an exception to that pattern, in that no theme vowel is present—i.e., words in that category are monomorphemic (3a). By positing that regular stress in Portuguese falls on the stem-final vowel, such forms are no longer irregular. Both theme vowels and

[2] (Oral) falling diphthongs (VG) in Portuguese are heavy. (Oral) rising diphthongs (GV) are light—an example is provided in (1b). Portuguese has no long vowels.

[3] Throughout this paper, I use 'oxytones', 'paroxytones' and 'proparoxytones' interchangeably with 'words that have final', 'penult' and 'antepenult stress', respectively.

catalexis formally account for the cases in (2a).

(3) a. *jacaré* $[\text{ʒakaˈɾɛ}]_{stem}$ ‘alligator’

b. *boca* $[\text{ˈbok}]_{stem}[\text{-a}]_{\text{TV}}$ ‘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 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 also seem to affect stress location in the lexicon, which indicates that weight computation in Portuguese may not be restricted to the rhyme.

A better notion of how weight is computed in Portuguese is naturally important if one wishes to have a more comprehensive and accurate understanding of how stress and weight interact in said language. In this paper, I present an 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[4] and quantitatively determined, i.e., weight effects are gradient across and within each position in the stress domain.

The analysis is developed in this paper by addressing four 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?

d. Which theory of weight computation best captures the stress patterns in Portuguese?

Statistical models (§5) indicate that weight-sensitivity gradiently declines/weakens as we move away from the

[4] For positional weight, see Gordon (2004).

right edge of the word. The fact 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.

Previous research in BP is based on the assumption the 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 that 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 some languages (Gordon 2005, Topintzi 2010, Ryan 2014). In Portuguese, however, such effects have a very particular profile (§4.1.1.1).

In answering question (4c), one would anticipate that onsets have either a positive or a null effect on stress. The lexicon examined here, however, showed that neither seems to be the case: onsets in Portuguese appear to be 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. However, an alternative theory of weight domain seems to capture such effects. Interval theory, proposed by Steriade (2012), proposes that 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. Unlike syllables (σ), intervals (ι) have no internal constituency. In addition, onsets of a given syllable are grouped with the previous interval (VC$_\sigma$CCVC$_\sigma$ *vs.* VCCC$_\iota$VC$_\iota$).

Question (4d) investigates which theory of weight computation, namely, *syllables* or *intervals*, best captures the stress patterns in Portuguese. Although both domains capture the weight gradience in the language, I show that an interval-based analysis is (i) empirically better motivated, (ii) more economical, and (iii) more accurate than syllables. This is confirmed by two different statistical models,[5] presented in §5.

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. I show that there is no compelling argument for a morphological role in Portuguese non-verb stress. 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,

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

results and implications of each theory in different statistical models. Both models based on intervals and syllables show effects that are consistent with a gradient notion of weight-sensitivity in Portuguese. Furthermore, both models are capable of predicting a substantial portion of irregular patterns in the language. In section 6, I map the statistical models proposed in this paper into a Maximum Entropy (MaxEnt) grammar, where one positionally-defined weighted constraint is capable of capturing the gradient effects in Portuguese (see Ryan (2011) for a similar approach). I discuss important differences between a purely statistical analysis and a MaxEnt model, given the methodology employed. Finally, section 7 summarises the findings of this paper, and discusses directions for future work.

2 Stress in Brazilian 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 cases 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[6]. 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 the trisyllabic window (Domahs et al. 2014).

Several studies on stress in BP (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 words 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.[7] Proparoxytones have always existed in BP, 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), but some of them are repaired via syncope and resyllabification (2c), as long as the resulting form obeys the phonotactic patterns in the language (see Amaral 1999). This is the case for most dialects of BP, though in some northeastern varieties 'this pattern

[6] 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.

[7] Houaiss contains, in its entirety, approximately 442,000 entries, including synonyms, antonyms, and historical words.

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 by far the most common and productive patterns in spoken Portuguese (≈18% and ≈68% in the Houaiss corpus, respectively). It is highly unlikely that a new word in the language will have antepenultimate stress (Hermans & Wetzels 2012). Rather, new words tend to have either final or penultimate stress, aside from some borrowings.[8]

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,[9] stress in non-verbs is heavily influenced by weight. As mentioned in §1, some scholars have suggested that stress in non-verbs is also influenced by morphological factors, namely, the presence of a theme vowel (TV) (Pereira 2007 and Lee 2007, among others). These scholars assume stress in non-verbs is sensitive to both morphological and phonological factors. Whether or not morphology influences stress in non-verbs, the systematic phonological patterns in BP stress make it very difficult to assume Portuguese has lexically-marked stress, as suggested by Camara Jr. (1979), Martins (1982) and others. Table 1 summarizes the stress patterns in non-verbs—'H' stands for a heavy syllable. Heavy syllables may have a nasal vowel, a coda consonant, and/or a falling diphthong: *pagã* [paˈgã] 'pagan'; *valor* [vaˈloɾ] 'value'; *funil* [fuˈniw] 'funnel'. *Light* syllables ('L') are open and contain a short vowel or rising diphthong: *abacaxi* [abakaˈʃi] 'pineapple'; *ópio* [ˈɔpju] 'opium'. 'X' stands for either 'H' or 'L'. Note that very few words have antepenult stress and a heavy penult or final syllable (also noted in Wetzels (2007) for 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* [ˈpɾɔ.pʊ.lis]. Some of these words undergo syncope in spoken BP: *óculos* [ˈɔ.kʊ.lʊs] ⇒ [ˈɔ.klʊs] 'glasses'.

2.1 Morphological approaches to stress in Portuguese non-verbs

In this section, I review the arguments for morphological influence in non-verb stress. 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 compelling evidence suggesting that such vowels actually influence stress: effects often attributed to theme vowels can be accounted for by phonological factors alone.

[8] The words 'penalty' [ˈpenaltʃi] and 'performance' [peɾˈfɔrmãnsi], for example, are present in Portuguese dictionaries with the original stressed syllable, even though this may result in a different stress position (as in 'performance') when compared to the source language (penult stress in English, antepenult in Portuguese). This is respected in the spoken language as well, despite following a very rare pattern in Portuguese.

[9] For a comprehensive analysis of BP verb stress, see Wetzels (2007).

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

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

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]
STEM PL
Alligators

b. boc -a -s [ˈbokas]
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*[10] (5a), for example, stress falls onto the stem-final vowel (/ɛ/)—this approach entails that all words with irregular final stress are monomorphemic. A word like *boca* (5b), on the other hand, has a theme vowel (/a/), and therefore stress falls on /o/, the only vowel in the stem.

The main argument for this proposal lies in derived forms. If we add a suffix to both words above, the theme vowel is typically deleted, whereas the stem-final vowel cannot be. In (6), the diminutive suffix *-inho* [-iɲʊ] is attached to *pato* and *sofá*. In (6a), the theme vowel is deleted, yielding ‘patinho’; in (6b),

[10] 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 oxytones ending in /u/ or /i/, as these vowels cannot be thematic.

since the word-final vowel is part of the stem, an epenthetic consonant (/z/) is inserted to avoid hiatus (Bachrach & Wagner 2007).

(6) a. *patinho* (cf. _*patoinho_) [paˈtʃiɲu]

pat	-o	-inh	-o
STEM	MASC.TV	DIM	MASC

‘Small duck’

b. *sofazinho* (cf. _*sofinho_) [ˌsofaˈziɲu]

sofá	-inh	-o
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. *livrinho* or *livrozinho* [liˈvɾiɲu] ∼ [ˌlivɾʊˈziɲu]

livr	-o	-inh	-o
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: $\ldots$XˈL$]_{PWd}$ 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.[11] 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

The scholar 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).[12] 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 case, stress is shifted to avoid ungrammaticality. (9b), on the other hand, simply follows the stress patterns listed in (1).

[11] Vowel deletion across word boundaries.

[12] The suffix in example (6), 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]).

(9) a. [ˈatomo -iko] ⇒ [aˈtomiko] (σ [σ σ σ)
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)). That is, 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, from Table 1, 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. The claim that they affect stress, on the other hand, is essential here, and solid evidence is indispensable to support such a claim. I conclude that at present there is no definitive reason to propose a morphological component to Portuguese non-verb stress. The analysis proposed in this paper is therefore based only on phonological factors, discussed in the next section.

2.2 Phonological approaches to Portuguese stress in non-verbs

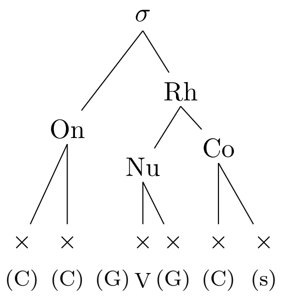
Even if we assumed that morphological factors did impact stress in Portuguese, we would still need to consider phonological factors, which heavily influence stress in the language. In this section, I examine such factors in more detail, focusing on weight and how it affects the stress patterns found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

In Portuguese, only two segments can occupy the onset position, and up to four segments (including the nucleus) can occupy the rhyme[13] (see Fig. 1). In the very few cases where two coda consonants are present, the second element is an /s/, and is almost always found in word-internal syllables, mostly words with the

[13] I assume all glides are nuclear. Rhymes with five segments (or more) are unattested.

prefix *trans-*. Therefore, only one coda consonant is commonly found in Portuguese. 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.: *crisp* ⇒ [ˈkrispi]).

Figure 1: Syllabic structure in Portuguese



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. Recent studies, however, show that onsets also have an impact on stress in several languages, suggesting at least some contribution to the calculation of weight (Gordon 2005, Topintzi 2010, Ryan 2011).

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 simplified in unstressed syllables (Harris 1997): *prato* [ˈpɾatʊ] *-inho* [iɲʊ] ⇒ [paˈt͡ʃiɲʊ] ‘plate’, ‘small plate’. In other words, complex onsets are preferred in more prominent positions. This simplification is relatively common in some spoken BP varieties: words such as *próprio* [ˈpɾɔpɾjʊ] are often produced as [ˈpɾɔpjʊ] ‘proper’. In addition, onset metathesis is observed in words such as *obstetra* [obsˈtɛtɾɐ] ⇒ [obsˈtɾɛtɐ] ‘obstetrician’. Despite the apparent correlation between onset clusters and stressed syllables in such processes, Cristófaro-Silva (2002) argues that what triggers onset cluster simplification and onset metathesis is the segmental structure of liquids, not stress.[14] Therefore, as no study has directly examined the impact of onsets on BP 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.

[14] Whether or not the arguments are compelling is important to consider, but investigating onset cluster simplification is not within the scope of this paper. Rather, the more relevant fact here is that no study has examined how onsets may influence weight computation in the language.

As seen in §2, Portuguese stress in non-verbs is weight-sensitive (Wetzels 2007), but most studies have argued 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 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.[15] 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 among proparoxytones (≈ 3.8% of the corpus contain such onsets). Wetzels (2007, p. 25) analyses such consonants as geminated, which therefore occupy both onset and (previous) coda slots: *baralho* ⇒ [ba.ˈɾaʎ.ʎo] (*[ˈba.ɾaʎ.ʎo]) 'deck of cards' (see Fig. 1). This would be consistent with the fact that very few proparoxytones 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 patterns in the lexicon, such as the distribution of different syllable shapes 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 σ		**Closed σ**	
Pattern	n	%	n	%
Final stress	5780	19.02%	24608	80.98%
Penult stress	72531	64.60%	39730	35.40%
Antepenult stress	17242	79.67%	4400	20.33%

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

[15] 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).

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. We will test whether this is the case in §4.

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 domain 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 paroxytones' stressed syllables. 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'	[ˈmɔvew] *vs.* ∅	'furniture'
[ˈsɛd͡ʒi] *vs.* [ˈsed͡ʒi]	'head office', 'thirst'	[eˈlɛtɾoŋ] *vs.* ∅	'electron'
[ˈbɔxa] *vs.* [ˈboxa]	'bird species', 'sediment'	[ˈdɔɾis] *vs.* ∅	'Doris'
[ˈmɔʎʊ] *vs.* [ˈmoʎʊ]	'bundle', 'sauce'	[ˈfɛzis] *vs.* ∅	'feces'

Syllabic 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.

In phonological approaches to stress in Portuguese, exceptions are dealt with by 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: (ˈσ σ) $\langle\sigma\rangle$. 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: *cafe*$\underline{t}$*eira* 'coffee pot' *vs.* *cafe*$\underline{z}$*al* 'coffee plantation'— note that the quality of the catalectic consonant varies 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 predicted to be just as likely or unlikely to bear final stress, i.e., 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 previous analyses to account for the stress patterns found in Portuguese. Most scholars 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 is focused on phonological factors only, in particular weight.

Examining the effects of onsets on weight is relatively recent in the literature. 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, to appear). 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 onsets positively correlate with stress—this would be consistent with what is discussed in §2.2. For instance, onset cluster simplification in unstressed syllables, discussed in §2.2, would advocate for a positive correlation. 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. 22).

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. This predicts that, all else being equal, onset size positively correlates with weight, which in turn correlates with stress likelihood (i.e., CCVC is heavier than CVC; CVC is heavier than VC). In fact, Ryan (2014, p. 24) 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'.

P-center theory could help explain onset cluster simplification in BP (§2.2), where complex onsets tend to be simplified in unstressed positions. One could hypothesize that such a simplification takes place to favour the most prominent position in the word. For example, in a ˈCCV.C_iCV word, deleting C_i delays the p-center of the final syllable, reducing its percept of heaviness. In addition, the difference in duration between both syllables increases even more—recall that stress in Portuguese strongly correlates with duration (§1).

The bigger question, however, is whether the Portuguese lexicon actually contains the patterns predicted by the 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 stress 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 according to p-center theory. In what follows, I review an alternative domain for weight computation, namely, intervals.

3.2 Intervals

P-center theory does not assume any particular domain of application. As a result, it is an open question how the boundaries of rhythmic units are defined. In a word such as $\mathrm{CC}\overset{a}{V}\mathrel{¦}\mathrm{C}\mathrel{¦}\mathrm{C}\mathrel{¦}\overset{b}{V}\mathrm{C}$, a boundary (¦) between units a and b has to be established before one calculates the p-center of unit b, since that calculation is based on which Cs count as post-nuclear segments of a, and which Cs count as pre-nuclear segments of b. Different representational assumptions may result in different perceptual centres.

Given how syllabic constituency is understood, it would be phonologically surprising if onset size *negatively* correlated with stress in a given language (where 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). If this is in fact the case, onset effects in the language would contradict the representational assumptions of syllable theory. 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 'π',[16] is the syllable (σ) or the interval (ι) 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. Thus, segments preceding the leftmost vowel in a word are not included in any interval.[17] All intervals begin with a vowel, which means that onset segments in σ_{j-1} in syllable theory are part of the interval ι_{j-2} (see Fig. 2). Unlike standard syllable theory, intervals have

[16] π is intended to be theory-neutral notation (see Steriade (2012)).

[17] This, as we will see in §5.1.1, will have to be revised for Portuguese.

no 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][18] ‘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•àː•ìt•î] and [⟨m⟩àht•áːr•ì], respectively. Because intervals include segments that would otherwise be part of the following syllable as onsets, 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.[19] 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.[20] Because VCC is heavier than VV but lighter than VVC, both ‘páncàːìtî’ and ‘màhtáːrì’ are accounted for.

Fig. 2 exemplifies how π differs in both theories in a CVCCCVVC word (syllables/intervals are counted from right to left). 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).

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

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

The weight of an interval is affected by different phonological/phonetic properties, such as (a) the overall duration of π, (b) the length of post-vocalic segments and (c) inherent properties of vowels (height, frontness etc.). Intervals are therefore closely related to duration—in this paper, for convenience, I adopt Hirsch's 2014 less fine-grained metric where duration positively correlates with the number of segments in a given interval.

[18] Where ‘ˆ’ stands for *no stress* (Steriade 2012, p. 8).

[19] On the weight difference between VV and VC, see below.

[20] Crucially, the different structural assumptions in syllables and intervals result in different alignments between the rhythmic unit and segmental material. This, in turn, yields differences in the computation of weight.

Because intervals assume different groupings from syllables, the predictions that follow, too, are different: since adjacent ‘onsets’ and ‘codas’ are now merged into one given 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) of 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, 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. Crucially, the number of segments in each interval will vary considerably across all BP words, and this will have a direct impact on the theory of weight domain one chooses, given the structural differences explained above between syllables and intervals.

In this section, I briefly reviewed two alternatives for 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 could incorporate p-center theory. Before moving to §4, let us briefly hypothesize 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 (see Hirsch (2014) for a brief discussion, where word-initial Cs are not excluded from the word-initial interval). 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 (ι) boundary would be equal to 35%[21] of the sum of the mean added duration (D) of the segments (C) included in that interval (in ms): $S(\iota) = -\sum\limits_{i=1}^{n} D_{C_{i_\iota}} \cdot 0.35$.

Under syllable theory, onsets are pre-nuclear elements, and therefore the p-center is shifted leftwards as a function of the number of onsets 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 possibility is 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), are now 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: VCC^i.V and VCC$^i_\mu$.V. In the second string, C^i is more sonorous than in the first string, and therefore contributes more to weight.

[21] Ryan (2014, p. 23). This value is based on the English lexicon, and therefore only serves as an example here.

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. Crucially, we will see that onset behaviour in Portuguese is best accounted for if one integrates intervals and p-center theory: intervals provide a more accurate picture of where weight is computed (i.e., in which domain), and p-center theory explains onset effects at the left edge of the word, i.e., *how* weight is computed. 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 four questions posed in §1, repeated in (10) for convenience.

(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?
d. Which theory of weight computation best captures the stress patterns in Portuguese?

The questions in (10) are clearly connected, in that the answer to question (10b) is highly dependent on the answer to question (10a). Likewise, question (10d) depends (at least in part) on the answer to question (10c).

The data examined in this paper come from the most comprehensive lexicon available in the Portuguese language: the Houaiss Dictionary (Houaiss et al. 2001). The Houaiss corpus contains 442,000 entries, of which 164,291 are non-verbs, including monosyllables. The lexicon contains a list of orthographic words with pronunciations, syllabifications and parts of speech.

The corpus also includes many words that are rarely used in spoken Portuguese. Some words are also borrowings whose phonotactic patterns do not match those found in the language—e.g., German words such as *schnitzel* and *Bretschneidera* (the sequence [t͡ʃn] is not allowed in Portuguese, and undergoes [i] epenthesis). Words with more than two onset or coda segments were excluded from the lexicon, as were monosyllables

($\approx 0.4\%$). No constraints were imposed on word length.[22]

One further adaptation is 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.[23] This directly affects stress, since the diphthongization yields penult stress. These data could potentially bias the analysis as follows: word-final rising diphthongs are not underlying diphthongs, and are light (§2.2); falling diphthongs, on the other hand, are heavy, and attract stress. A superficial analysis that amalgamates both groups could therefore find a negative correlation between word-final diphthongs and final stress, for example.[24] Thus, this subset of words was removed from the data. The final corpus (Garcia 2014) contains 154,083 non-verbs (see Table 5).

Grapheme-phoneme conversion was done by different scripts and regular expression substitutions. Some cases, however, are idiosyncratic. For example, the grapheme *x* can be realized as [s], [z], [k.s] and[ʃ]: *máximo* 'maximum', *exato* 'exact', *oxigênio* 'oxygen', *coxa* 'thigh'—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 the corpus used here is not representative of spoken Portuguese vis-à-vis stress patterns. Although the analysis in this paper is concerned with the lexicon *per se*, and not the spoken language, it would be ideal if both lexicon and spoken Portuguese shared similar proportions of stress patterns. To verify this representativeness, two frequency corpora were examined—both of which contain only the most frequent words in the language: the OPUS corpus (Tiedemann & Nygaard 2004) and the LaPS corpus (Klautau 2013), from the Federal University of Pará, in Brazil.[25] In all three corpora,[26] the proportions of each pattern are relatively similar. More importantly, the order *penult* > *final* > *antepenult* is observed in all three cases. The OPUS corpus and the LaPS corpus are used here to ensure that all three stress patterns are balanced in the

[22] 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 (Frota et al. 2010), for example, more than 90% of the words in the spontaneous speech sample available ($n = 188,269$) contain fewer than four syllables. Thus, a separate analysis was implemented where only words with fewer than six syllables were considered. However, 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.

[23] E.g., *terráqueo* [te.ˈxa.ke.ʊ] ⇒ [te.ˈxa.kjʊ] 'earthling'. Diphthongization is not categorical when the second V in a VV sequence is stressed: *piada* [pi.ˈada] ∼ [ˈpja.da] 'joke'.

[24] 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.

[25] Unlike the Houaiss corpus, the OPUS and LaPS corpora are based solely on Brazilian Portuguese.

[26] Verbs were excluded from both frequency corpora.

corpus used in this study, i.e., that the proportions in each stress pattern in the Houaiss lexicon are mirrored in the spoken language.

Table 5: Portuguese/BP* corpora

Stress pattern	**Houaiss**	**OPUS***	**LaPS***
Final	18%	22%	27%
Penult	69%	70%	62%
Antepenult	13%	8%	11%
	n=154,083	*n*=29,901	*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. 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 (i.e., onset, nucleus and coda) as well as segmental quality for all three possible positions, namely, final, penult and antepenult syllables. No clear qualitative differences were observed, that is, stress likelihood does not seem to be consistently predicted by segment quality in onset position (this was also the case for onset clusters such as /kɾ, gɾ, tɾ, dɾ/). The most predictive onset segments of final stress are /d,s,t/, but /t/ is also the most predictive segment of penult stress. Likewise, the most predictive coda segments word-finally in oxytones are /ɾ,l,m/, of which /l/ is also the most predictive word-final coda in paroxytones. Most onset and coda segments are limited to a few segments, but whether or not such segments are predictors of stress is unclear, given that the sets of segments in stressed and unstressed positions are almost equivalent in terms of sonority. If a clear correlation existed, stressed positions (shaded cells in Table 6) should contain more sonorous segments. Onset segments are particularly informative, given the wider range of sounds allowed in such positions (sonority-wise), namely, liquids, nasals and stops. Coda segments, on the other hand, are restricted to liquids, nasals and /s/. Given the lack of clear patterns indicating segmental quality relevance to stress, this paper focuses on the size of onsets, nuclei and codas, which does have a clear correlation with stress.

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

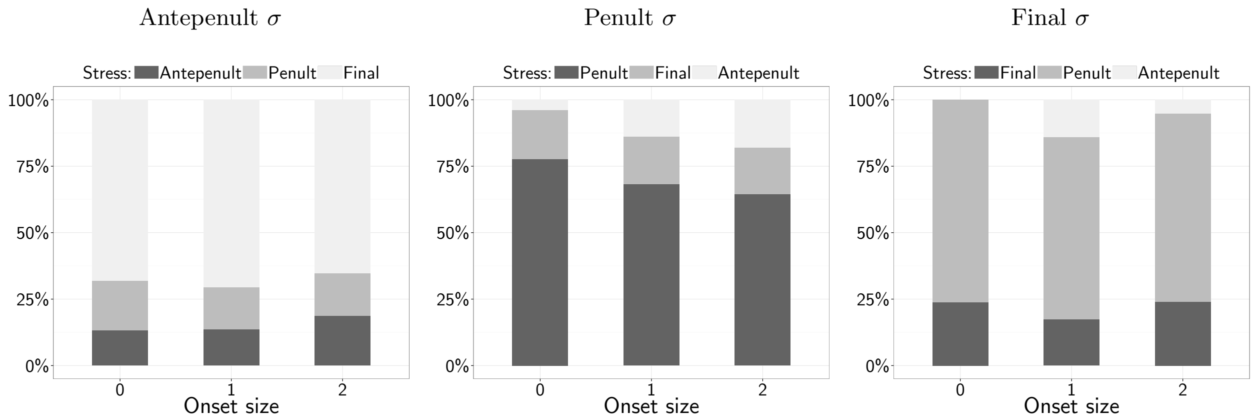
	Final σ		**Penult** σ		**Antepenult** σ	
Stress pattern	Onset	Coda	Onset	Coda	Onset	Coda
Final	/d,s,r/	/ɾ,l,s/	/k,t,ɾ/	/n,ɾ,m/	/k,t,s/	/n,ɾ,s/
Penult	/t,d,s/	/l,m,s/	/t,d,n/	/n,s,ɾ/	/l,k,m/	/n,ɾ,s/
Antepenult	/k,l,ɾ/	/s,n,ɾ/	/t,f,n/	/n,ɾ,l/	/t,l,n/	/s,n,ɾ/

4.1.1.1 Onset size effects

Let us now explore the data by examining the impact of onset size on stress. 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 σ, penult σ and final σ, respectively).

The first figure in Fig. 3 shows that onsets might be positively correlated with stress in the antepenult and final syllables. However, note that penult stress also seems to be affected by final onset size. In the penult syllable, the plot suggests a clear negative correlation—these percentages are analysed in more detail in §4.1.2. 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 (though penult and final onset sizes do have a clear negative effect on antepenult stress, as we can see in the plots for Penult σ and Final σ). Rather, the difference in the Antepenult σ plot lies between {0,1} and 2 segments. Except for the penult syllable, onset effects on stress are not clear in the figures. Still, the observed monotonic trends should not be neglected, as they suggest some relation between the presence/absence of (complex) onsets and stress. As we will see below, these effects

Figure 3: Onset size effects by syllable and stress pattern



become clearer once we control for coda size (§4.1.2). The significance of these effects will be examined in §5.

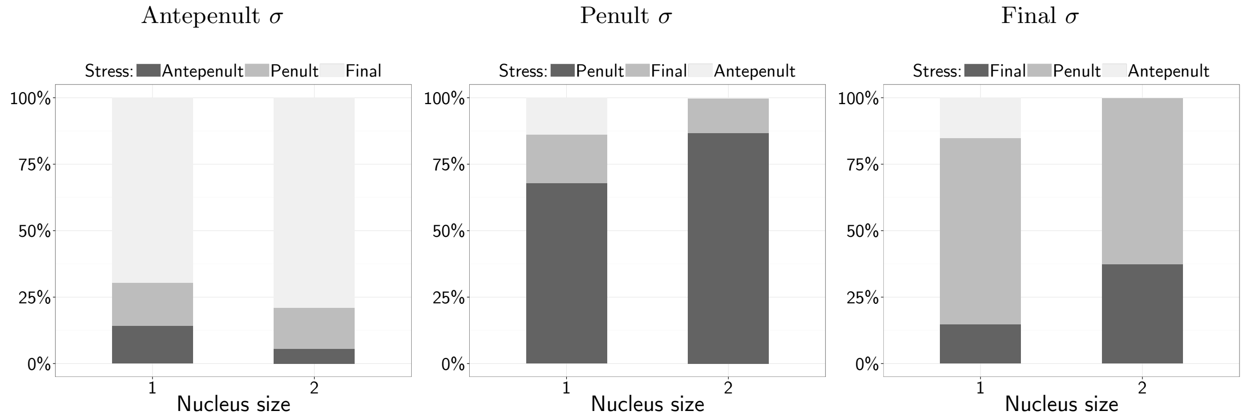
4.1.1.2 Nucleus size effects

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 VV/VG nuclei are heavier than V nuclei. Note, however, that the distinction is visible not only word-finally, but also on 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 under syllable theory.

4.1.1.3 Coda size effects

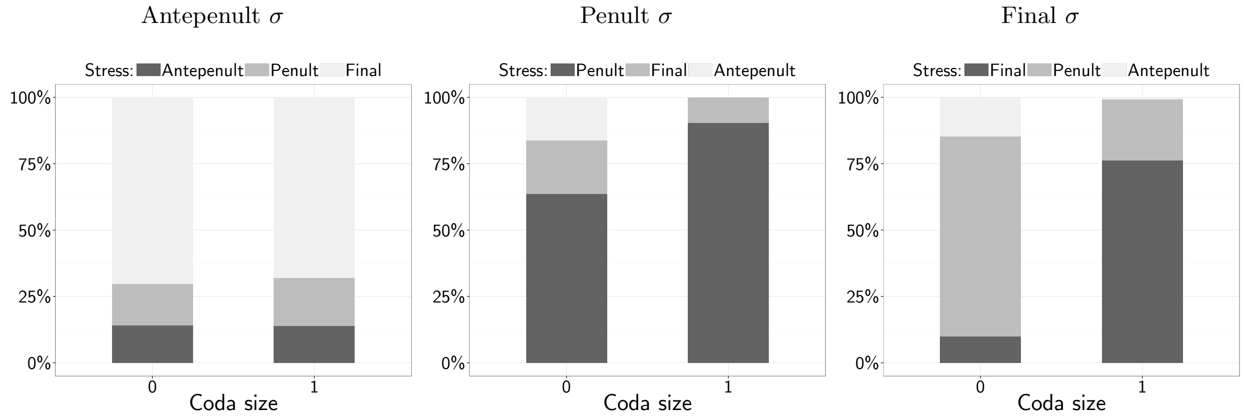
Let us now examine the effect of coda size on stress placement. Fig. 5 shows a very strong effect of the presence of a final coda on stress placement, consistent with the standard approaches to stress in Portuguese discussed in §2: final stress is far more likely when the final syllable has a coda. On 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,

Figure 4: Nucleus size effects by syllable and stress pattern



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

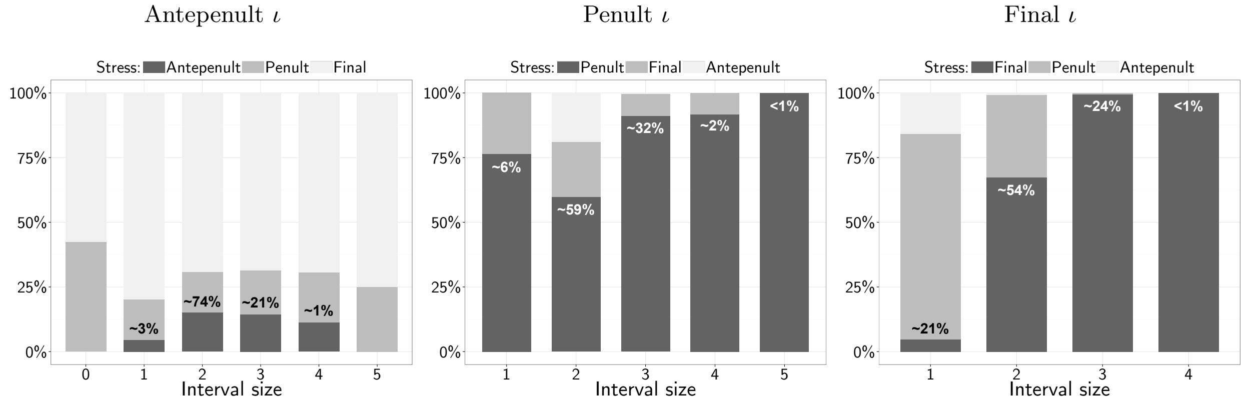


The trends observed above suggest that the effect of syllable weight is gradient, not categorical: coda effects are stronger than nucleus effects, but both seem to have an impact on stress. How much weight influences stress also depends on which syllable one examines: final stress is more strongly affected 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 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 the standard syllabic representations. This is also the assumption in §3.1, where we discussed p-center theory. Recall, though, that p-center theory is not tied to syllable theory *a priori*. In §3, I also discussed interval theory, which has a non-hierarchical representational schema. Therefore, I now turn to Fig. 6, where the proportion of words with each stress location in the lexicon is plotted as a function of the size of each interval. The x-axis represents the size of each interval (ι), where *1* represents no intervening consonantal segments in a V-to-V interval (i.e., an interval consisting of a single vocalic segment). 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



Antepenult ι

Penult ι

Final ι

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 contains 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. Because intervals conflate onsets, nuclei and codas, it is difficult to tell exactly where the independent effects within each interval originate. However, as we already know the effects of each of these constituents from §4.1.1 above, it should be possible to unpack what 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 diphthongs from the data, and considering only LLL words, since only onsets would vary in that particular subset.

Given the trisyllabic window, we can verify the onset-stress relation in the two final syllables. Considering the coda effects in Fig. 5, $\ldots\text{LL}]_{PWd}$ 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 more onset segments are present. Thus, let us examine whether final onset size affects penult/final stress.

Figure 7: Stress patterns by final onset size in ...LLL words

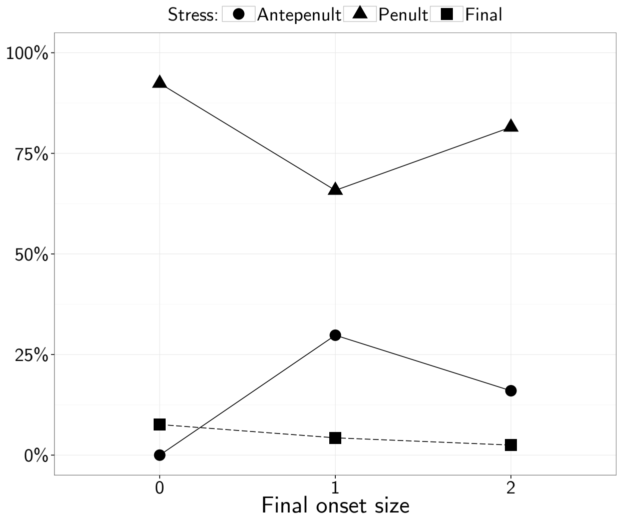


Fig. 7 shows that larger final onset sizes are more correlated with penult stress than final stress. It should be noted that singleton onsets are much more frequent in the Houaiss corpus than complex onsets: 84.3% *vs.* 4.7% in oxytones, 89.1% *vs.* 2.1% in paroxytones, and 81% *vs.* 10.2% in proparoxytones.[27]

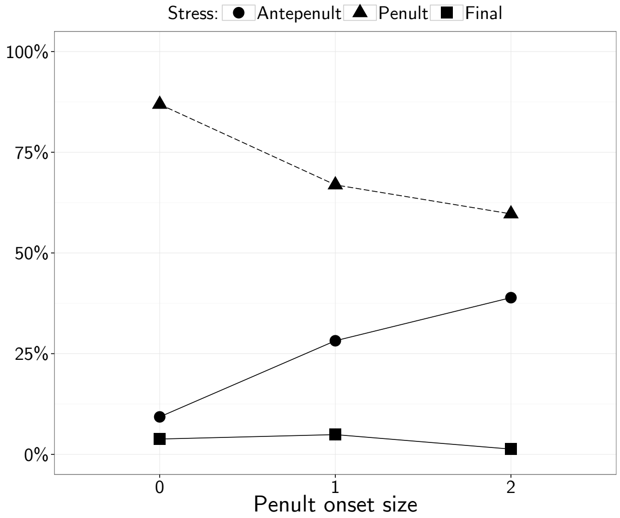
The trend in Fig. 7 could indicate that intervals play some role in these data, given that final stress seems to be more negatively correlated with final onset size—a result we would not expect if the domain of weight computation in Portuguese is the syllable. Let us now examine how penult onset size affects stress.

Penult and antepenult syllables are locations where coda effects are less apparent (standard analyses

[27] 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, as can be seen in Fig. 4.

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. As seen in Fig. 2, syllable theory predicts that an increase in onset size in the penult syllable would increase the amount of material in that constituent, positively impacting its duration, which would 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 favours interval theory for Portuguese stress. Antepenult onset size (not shown here), on the other hand, presents a less clear pattern: onset clusters seem 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 intervals.

5 Statistical analysis

In the previous section, we examined patterns found in the Portuguese lexicon. Both syllable structure 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 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).

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, only need three predictors (i.e., as many predictors as the number of positions in the stress domain).

Given that the stress patterns found in the Portuguese lexicon involve non-continuous responses, the data in this study could be modelled using Logistic Regressions (`glm()` in R[28]). However, because standard logistic models involve binary responses (i.e., binomial), two such models would be necessary to accommodate the stress domain in Portuguese. As a result, each predictor would yield two different effects. An alternative is to employ a Multinomial Logistic Regression (e.g., `mlogit()` in R), where more outcomes can be accounted for in a single model (three, in this case). 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. Here, again, each predictor would yield two effects. Furthermore, the literature on multinomial models applied to linguistic data is scarce when compared to binomial models.

A more parsimonious alternative—the one employed in this paper—is to model the data using *Ordinal Regression* (see Agresti 2010), also known as *Cumulative Link Model* or *Ordered Logit Model* (`clm()` in R).[29] In this case, the stress domain in the data is 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.[30] Another advantage of ordinal regressions is that predictors in such models tend to have lower standard errors when compared to equivalent binomial regressions (Christensen 2013a, p. 6).

In a Cumulative Link Model, an ordinal response (Y_i) can be classified into different categories $j_1, j_2, ..., j_n$ (J). The response (Y_i) follows a multinomial distribution with parameter π, where π_{ij} indicates the proba-

[28] R is an open-source statistical programming language (R Core Team 2014).

[29] {`ordinal`} package (Christensen 2013b) in R.

[30] For the purposes of comparison, the lexicon examined in this paper was also modelled with two Binomial Logistic Regressions (final *vs.* penult stress; antepenult *vs.* final/penult stress). The effects observed in such regressions are consistent with those present in the Ordinal models analysed in this section.

bility that the ith observation falls in category j,[31] i.e., one of the three stress positions in the lexicon. The model predicts the log-odds of response Y_i based on a set of predictors. The fitted model is given in (11), where $Pr(Y_i \leqslant j)$ denotes the probability that response Y_i is $\leqslant$ category j; $\{\theta_j\}$ parameters (*thresholds*) represent the intercept of each cumulative logit (i.e., each j; in this case, *two* such intercepts are necessary, given the three possible stress categories); $x_i^T\beta$ represents the vector of regression coefficients for each predictor. Note that the coefficients are independent of j, which means effect size of a given predictor does not rely on a specific category—this is a key difference between Ordinal and Binomial/Multinomial Logistic models, where multiple responses yield different coefficients. In other words, standard multinomial/binomial models would result in multiple coefficient values.

(11) **Ordinal Regression**

$$\text{logit}(Pr(Y_i \leqslant j)) = \text{logit}(\gamma_{ij}) = \theta_j - x_i^T\beta$$

In our discussion, this means the models predict the probability of a given stress pattern for each data point in the lexicon (i.e., word). Two intercepts (thresholds, j) are necessary: (1) final | penult-antepenult, and (2) final-penult | antepenult (i.e., $1|2,3$ and $1,2|3$). For example, we know that $Pr(Y_i \leqslant 3) = 1$, given the trisyllabic window in Portuguese. Thus, an interval-based model (see Table 7) would be defined as logit($Pr(Y_i \leqslant 1)$)= $\theta_{j_1} - \beta_1(\texttt{int1}_i) - \beta_2(\texttt{int2}_i) - \beta_3(\texttt{int3}_i)$ for final (1) *vs.* penult (2) or antepenult (3) stress, and as logit($Pr(Y_i \leqslant 2)$)= $\theta_{j_2} - \beta_1(\texttt{int1}_i) - \beta_2\ (\texttt{int2}_i) - \beta_3(\texttt{int3}_i)$ for final (1) or penult (2) *vs.* antepenult (3) stress. In order to contrast the two theories of weight, two models are necessary, namely, a syllable model and an interval model. Syllable predictors in Table 7 are binary, unlike interval predictors. In §5.1 and §5.2, the two models are discussed.

In each of the two models presented in §5.1 and §5.2, positive $\hat{\beta}$ values (i.e., regression coefficients) indicate higher likelihood of antepenult stress, whereas negative $\hat{\beta}$ values indicate higher likelihood of final stress—which reflects the end-points of the ordinal stress domain defined above (i.e., $3 > 2 > 1]_{PWd}$). In addition to regression coefficients, standard errors, Wald z values and significances are also reported and explained. The bottom row of each model lists the model's thresholds (θ), accuracy level and collinearity (κ)—predictors in both models have been checked for linearity.

[31] Where the number of response categories $(J) \geqslant 2$.

Table 7: Predictors and response

Syllables	onset.fin	Whether the **final** σ contains an onset segment (0/1)
	nucleus.fin	Whether the **final** nucleus is monopositional/light (0) or bipositional/heavy (1)
	coda.fin	Whether the **final** σ contains an coda segment (0/1)
	onset.pen	Whether the **penult** σ contains an onset segment (0/1)
	nucleus.pen	Whether the **penult** nucleus is monopositional/light (0) or bipositional/heavy
	coda.pen	Whether the **penult** σ contains an coda segment (0/1)
	onset.ant	Whether the **antepenult** σ contains an onset segment (0/1)
	nucleus.ant	Whether the **antepenult** nucleus is monopositional/light (0) or bipositional/heavy
	coda.ant	Whether the **antepenult** σ contains an coda segment (0/1)
Intervals	int1	Number of segments in ι 0 (**final**): 1-4
	int2	Number of segments in ι 1 (**penult**): 1-5
	int3	Number of segments in ι 2 (**antepenult**): 0-5
	stress	**response**: final (1), penult (2) or antepenult (3)

5.1 Syllable model

In this model (Table 8), final coda and final nucleus both have a positive effect on final stress ($\hat{\beta} = -4.68$ and $\hat{\beta} = -2.86$, respectively; $p < 0.0001$). Both effects are highly significant, and final codas have a stronger effect than final nuclei: whereas the presence of a final coda raises the odds of final stress by a factor of 108.10 ($e^{|\hat{\beta}|}$),[32] the presence of a final heavy nucleus (i.e., diphthong) raises the odds of final stress by a factor of 17.55. Final onsets have a negative impact on final stress ($\hat{\beta} = 1.52$, $p < 0.00001$), lowering the odds of final stress by a factor of 4.59. These results mirror the trends in Figs. 3, 4 and 5.

As for the set of predictors that refer to the penult syllable, a similar pattern is found: penult onsets ($\hat{\beta} = 0.50$, $p = 0.00001$) increase the likelihood of antepenult stress by a factor of 1.65. On the other hand, a heavy nucleus ($\hat{\beta} = -1.32$, $p = 0.00001$) or a coda segment ($\hat{\beta} = -1.09$, $p < 0.00001$) in the penult syllable all lower the odds of antepenult stress.

Antepenult onsets and nuclei have significant effects: while onsets positively affect antepenult stress ($\hat{\beta} = 0.24$, $p < 0.00001$), heavy nuclei *negatively* affect antepenult stress ($\hat{\beta} = -1.06$, $p < 0.00001$). The latter effect is particularly difficult to explain, given that the weight effect found is the opposite of what one would expect. Given these effects, the pattern observed in the antepenult syllable does not seem to be

[32] Gelman & Hill (2006, p. 60).

consistent with syllables or intervals.

Not only is weight significantly affecting all three positions in the stress domain, but it does so in a gradient way, in which final syllables are more weight-sensitive than penult syllables, which in turn are more weight-sensitive than antepenult syllables. This is an important finding, given that standard analyses of Portuguese stress assume no weight-sensitivity outside of the final syllable. Another important finding is that codas and nuclei have stronger absolute effects than onsets (i.e., $\hat{\beta}$ final $\sigma > \hat{\beta}$ penult $\sigma > \hat{\beta}$ antepenult σ, and $\hat{\beta}$ {codas, nuclei} $> \hat{\beta}$ onsets)—even though the negative onset effects are, in fact, unexpected under syllable theory. For example, given the coefficient values in Table 8, this model predicts that a CVG.CV.CV word will have the following stress probabilities: final = 5.4%; penult = 84%; and antepenult = 10.5%. On the other hand, a CVG.CV.CV$\boxed{\text{C}}$ word will have have very different predicted probabilities, as word-final syllable coefficients have a very strong impact on stress likelihood: final = 86.2%; penult = 13.6%; and antepenult = 0.1%.

Table 8: Coefficient values for σ model ($\hat{\beta} > 0 \Rightarrow$ higher likelihood of antepenult stress), with associated odds ratio ($e^{|\hat{\beta}|}$), standard errors, Wald z values and significances

σ **predictor**	$\hat{\beta}$	$e^{\lvert\hat{\beta}\rvert}$	**se**($\hat{\beta}$)	z **value**	p **value**
`onset.fin`	1.52	4.59	0.02	53.97	< 0.00001
`nucleus.fin`	−2.86	17.55	0.02	−137.99	< 0.00001
`coda.fin`	−4.68	108.10	0.02	−194.62	< 0.00001
`onset.pen`	0.50	1.65	0.02	23.75	< 0.00001
`nucleus.pen`	−1.32	3.75	0.02	−46.18	< 0.00001
`coda.pen`	−1.09	3.00	0.01	−63.85	< 0.00001
`onset.ant`	0.24	1.27	0.02	11.64	< 0.00001
`nucleus.ant`	−1.06	2.90	0.02	−40.50	< 0.00001
`coda.ant`	−0.02	1.02	0.01	−1.646	0.0999
$\theta = \{-1.64, 3.34\}$	AIC: 186433.21		Accuracy: 74.75%		$\kappa = 40.79$

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*[33] (Belsley et al. 1980). The

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

predictors included in the model in Table 8 have high collinearity ($\kappa = 40.79$[34]). The syllabic shapes found in Portuguese explain why collinearity is not low between onsets, nuclei and codas: although both heavy nuclei and complex codas are allowed, VGC syllables are rare in the language—i.e., syllabic predictors are not completely orthogonal. Furthermore, words with coda segments in different syllables are uncommon in the Portuguese lexicon. A Spearman ρ^2 test reveals that the most collinear pair of predictors is `onset.ant` and `nucleus.ant` ($\rho^2 = -0.55$). 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, note that, except for antepenult codas, which are non-significant, all the effects in question are highly significant ($p < 0.00001$), and therefore high collinearity should not pose major problems to the analysis.

In order to assess the goodness of fit of the model in question, we can compare it to a null, baseline model (using a likelihood ratio test), wherein no predictors are considered (intercept-only), and all words are predicted to have penult stress, which is the most common pattern in the data. The model proposed here does significantly better than such a baseline model ($p < 0.0001$[35]). A more meaningful evaluation is to examine how accurately the model in Table 8 predicts stress location. This classification allows us to know the percentage of words that have a predicted stress position that matches the actual data. The baseline model has an accuracy rate of 69.06%, whereas the model proposed here achieves 74.75% accuracy. More importantly, this proportion exceeds the proportion of regular cases in the lexicon modelled (72%, n=154,083), that is, the syllable-based model is capable of predicting at least part of the irregular cases.

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 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 values, the better the fit. Naturally, the AIC value for the syllable model is only meaningful in comparison to the AIC value of the interval model, which is discussed in §5.2 below.

5.1.1 Analysis

The model above has both clear and unclear results. Final and penult nuclei and codas, for example, show a clear (and expected) effect. The strong effect of final nuclei and codas could possibly explain why previous analyses of Portuguese stress have constrained weight effects to the right edge of the word: such analyses

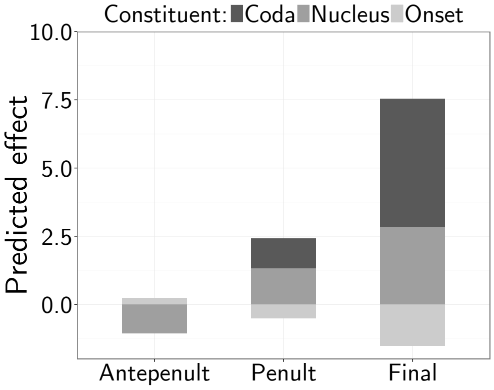
[34] `collin.fnc()` in R.

[35] Likelihood-ratio test comparing two models—χ^2 test (`anova()` in R).

have concentrated on word-final syllables only most likely because of the considerably different coefficient values between final and penult syllables ($\frac{\hat{\beta}_{\texttt{coda.fin}}}{\hat{\beta}_{\texttt{coda.pen}}} \approx 4.3$). Therefore, though the structure of earlier syllables does affect stress placement, these effects are small compared to the structure of the final syllable, and thus may not be noticed if the whole lexicon is not considered. The different coefficient values in (i) final and penult syllables, and (ii) codas and nuclei argue for a clear *gradient* notion of weight-sensitivity in Portuguese. Contrary to what previous analyses assume, the model shows that weight is not a categorical phenomenon in the language (see Fig. 9).

Final and penult onset effects, on the other hand, are negatively correlated with word-final stress (i.e., both predictors increase the likelihood of antepenult stress). Under syllable theory (and p-center theory), the fact that such an effect is significant poses a problem. Although the coefficient values of onsets is small, it is highly significant. In addition, within each syllable, onset effects are smaller than nucleus or coda effects, arguing again for a gradient notion of weight within syllables.

Figure 9: Weight gradience in the σ model:
different predicted effect sizes within and across syllables



Antepenult onsets have a positive effect, which is consistent with syllable theory. One could therefore argue that this particular effect is by definition unexpected under interval theory: assuming that onsets at the left edge of the word are extrametrical (§3), and given that only three syllables/intervals fall within the stress domain, antepenult onsets should have a null effect on stress, not the positive effect we see in Table 8. Such a positive effect indicates that extrametricality may not be the right explanation here. An alternative hypothesis is that onsets are always parsed. This would account for the positive onset effect that we observe in antepenult syllables. This hypothesis also predicts that, in words with two syllables, penult onsets would have to be parsed within the penult interval, therefore making that interval heavier, and more likely to attract stress. This would violate the *a priori* assumption that intervals begin with a vowel (V-to-(V))—but it would be compatible with the view that onsets are expected to perturb the perceptual centre of a given rhythmic unit (see §2.2).

In order to test these predictions, let us model two sets of words: disyllables and longer words.[36] If we only include `onset.pen` as our stress predictor, we can find out if penult onsets in disyllables have a null effect (thus supporting extrametricality) or a positive effect (thus supporting the alternative hypothesis discussed above). As it turns out, penult onsets actually have a positive effect on penult stress in disyllables ($\hat{\beta} = 0.18, p < 0.002$), but a negative effect on penult stress in longer words ($\hat{\beta} = -1.39, p < 0.00001$).[37] This behaviour is exactly what we would expect if onsets were not extrametrical—consistent with the antepenult onset effect in Table 8.

The same predictions can be tested in antepenult syllables. If we compare trisyllables with longer words, and examine the effects of the antepenult onsets on antepenult stress, we expect trisyllables to have a positive onset effect on stress, and longer words to have a null effect. While we do find a positive onset effect in trisyllables ($\hat{\beta} = 0.007, p = 0.01$), thus mirroring Table 8, we find a negative effect in longer words ($\hat{\beta} = -0.03, p < 0.00001$). This particular effect is unexpected and cannot be accounted for under syllables or intervals. In sum, antepenult onsets have a positive effect on antepenult stress, but once we only examine long words, the effect is negative. By examining word-edge effects of onsets, we can conclude that extrametricality under intervals is clearly not empirically supported in the Portuguese lexicon.

Antepenult heavy nuclei also have an unexpected negative effect: relative to monophthongs, diphthongs lower the odds of antepenult stress. This effect is particularly difficult to interpret, given that nuclei play a crucial role in quantity-sensitive languages. One possible explanation is that, since antepenult stress is dispreferred in the language, antepenult syllables tend to avoid the features that would normally attract stress. We can safely conclude that the antepenult position in the stress domain presents distinct patterns.

The syllable-based model captures subtle (and gradient) weight effects and is more accurate than a baseline model—recall that its accuracy level is in fact higher than the proportion of words with predictable/regular stress in the lexicon. On the other hand, antepenult nucleus (and partially onset) effects are surprising and require further investigation.

In sum, the syllable-based model shows that weight-sensitivity is present in all three syllables in the stress domain—consistent with Figs. 3, 4 and 5. Effect sizes vary within and across syllables, showing that weight-sensitivity should not be understood as categorical, i.e., only present or absent in a given syllable. Finally,

[36] In disyllables, a binomial model predicts penult *vs.* final stress; in longer words, penult *vs.* antepenult stress.

[37] Given that most words in the lexicon have more than two syllables, this is what we see in the syllable model in Table 8.

some effects do not have a clear interpretation, as they seem to contradict representational assumptions that underlie the weight domain in question. Recall that three questions in (10) were related to (i) word-internal weight-sensitivity, (ii) categorical *vs.* gradient weight, and (iii) onset effects. Thus far, the syllable-based model has a clear answer to (i) and (ii): weight-sensitivity is not limited to the word-final syllable, and weight effects are *gradient* rather than categorical (see Fig. 9). Onsets also have an effect on weight, which answers (iii), but results contradict what one would expect under syllable theory. As we will see in the next section, models based on intervals are not only more empirically motivated, but also fare better than when we consider the results in the syllable-based model discussed here.

5.2 Interval model

The interval model (Table 9) has all three predictors in the domain.[38] All three variables have a significant effect on stress: having more segments in `int1` positively affects word-final stress ($\hat{\beta} = -1.99$, $p < 0.0001$), and having more segments in `int3` positively affects antepenult stress ($\hat{\beta} = 0.31$, $p < 0.0001$). In other words, each segment added to the final interval raises the odds of final stress by a factor of 7.33, whereas each segment added to the antepenult interval increases the odds of antepenult stress by a factor of 1.37. `int2` shows a negative coefficient value, which indicates that having more segments in that interval increases the likelihood of penult stress (as opposed to antepenult stress, provided that `int1` is sufficiently low—if the final interval is not low enough, stress is final in the vast majority of cases). All three effects are highly significant, and have very different $\hat{\beta}$ values. This model enjoys medium collinearity ($\kappa = 13.57$), and has higher accuracy when compared to the baseline model (78.26% *vs.* 67.60%). In addition, a likelihood-ratio test reveals that the model fits the data significantly better than the intercept-only model (χ^2 test: $p < 0.00001$).

Table 9: Coefficient values for ι model ($\hat{\beta} > 0 \Rightarrow$ higher likelihood of antepenult stress), with associated odds ratio ($e^{|\hat{\beta}|}$), standard errors, Wald z values and significances

ι **predictor**	$\hat{\beta}$	$e^{\vert\hat{\beta}\vert}$	**se**($\hat{\beta}$)	z **value**	p **value**
`int1`	-1.99	7.33	0.01	-197.70	< 0.00001
`int2`	-0.21	1.23	0.01	-33.40	< 0.00001
`int3`	0.31	1.37	0.01	48.04	< 0.00001
$\theta = \{-2.18, 2.60\}$	AIC: 181389.03		Accuracy: 78.26%		$\kappa = 13.57$

Here, again, we see that weight effects are present in all rhythmic units (i.e., intervals) in the domain.

[38] Predictors have been scaled and centred (`scale()` in R).

Furthermore, the different effect sizes also argue for gradient weight-sensitivity. For example, a VC•VC•V word has the following predicted stress pattern probabilities: final = 4%; penult = 79.4%; and antepenult = 16.5%. If we now add one segment to the final interval (VC•VC•V$\boxed{\text{C}}$), the model predicts very different probabilities: final = 66.4%; penult = 33.1%; and antepenult = 0.4%. Finally, note that the AIC value of the interval model is (slightly) lower than the AIC value of the syllable model, which indicates that the interval-based model has a better fit.

One important assumption of the interval model presented above is that onsets at the (left) edge of the word are extrametrical (Steriade 2012), and therefore do not affect stress likelihood. However, we saw in §5.1 that this assumption is not met in Portuguese: in reality, whether penult or antepenult onsets have a positive or negative impact on stress depends on whether they are at the edge of the word. The implications are relevant: the antepenult interval in a word like *francesa* 'French' (fem.) will have three segments if onsets at the left edge of the word do not matter, and five segments otherwise. The model in Table 10 adjusts the assumption of onset behaviour: now, penult onsets in disyllables are included in the penult interval; in longer words, such onsets are included in the antepenult intervals. Antepenult onsets are now always included in the antepenult interval—this conflates the distinct effects discussed above, but this simplification accurately mirrors the positive effect found in the syllable-based model. Given what we have seen thus far, this model is more empirically motivated, since it reflects the exact patterns we see in the lexicon and observed in Table 8.

The adjusted model in Table 10 has a better fit (i.e., lower AIC) and is slightly more accurate than the model in Table 9. That is, besides being more faithful to the actual patterns observed, this model is also capable of predicting more items correctly. With regard to coefficient values, the adjusted model is essentially the same as the model in Table 9.

Table 10: Interval model adjusted for onset effects at word edges

ι **predictor**	$\hat{\beta}$	$e^{\lvert\hat{\beta}\rvert}$	**se**($\hat{\beta}$)	z **value**	p **value**
`int1`	−2.01	7.46	0.01	−196.60	< 0.00001
`int2`	−0.25	1.28	0.01	−35.33	< 0.00001
`int3`	0.32	1.38	0.01	43.82	< 0.00001
$\theta = \{-2.21, 2.63\}$	AIC: 180035.58		Accuracy: 78.29%		$\kappa = 15.38$

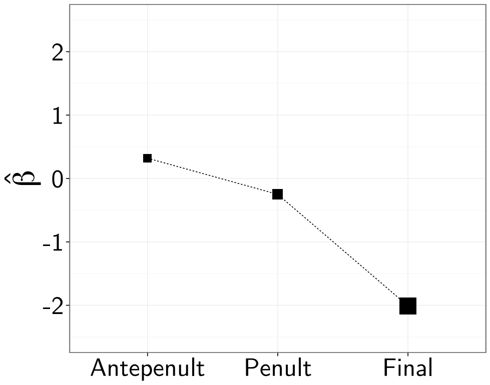
5.2.1 Analysis

Both versions of the interval model have higher accuracy when compared to their syllabic counterpart (78.26% and 78.29% *vs.* 74.75%), even though an interval-based model contains only three predictors (in Portuguese). Importantly, this model predicts ≈ 6% more data than the proportion of regular patterns in the lexicon modelled, which means that at least 16% of the irregular forms are accounted for under the interval-based model (*vs.* at least 7.7% under the syllable-based model). Both interval- and syllable-based models fit the data significantly better than baseline models. In Fig. 10, the coefficient value of each interval (y-axis, $\hat{\beta}$) increases as we approach the right edge of the word (x-axis, interval)—recall that negative effects are associated with the lower (i.e., right) end of the response scale (i.e., stress domain).

The main motivation for considering intervals in Portuguese stress lay on the role of final and penult onsets, which have a significantly negative effect on stress in the final and penult syllables (§4), and on the role of antepenult onsets discussed above (§5.1.1). This explains most of the unexpected results in the syllable model. In interval theory, such effects are accounted for (§3.2).

All onsets in Table 8 are as expected under interval theory. Antepenult nuclei, on the other hand, have an unexpected effect in both weight domains, namely, they negatively impact antepenult stress. In other words, having a diphthong in the antepenult syllable reduces the likelihood of antepenult stress. This is indeed a surprising pattern in the lexicon, and further investigation is necessary to explore this particular position in the stress domain. For the purposes of the present analysis, the essential findings here are: (i) weight effects are found outside the word-final position, including the antepenult syllable/interval; (ii) more importantly, such effects are gradient within and between syllables/intervals; and (iii) intervals account for more patterns observed in the lexicon than syllables do. Fig. 10 shows the coefficient values (in absolute terms) of each interval in the adjusted model.

Figure 10: Weight gradience in the adjusted interval model: $\hat{\beta}$ value by interval



6 A constraint-based implementation of weight gradience

In previous sections, we saw how regression models accounted for weight effects on stress in Portuguese based on segmental predictors (syllabic constituents and interval size). These models provide a comprehensive statistical analysis, with precise coefficient values, standard errors and significances. In this section, I briefly extend my analysis to a phonological grammar, and show that the weight gradience in Portuguese can be captured in such a grammar via positionally-defined weight-based constraints. More specifically, a Maximum Entropy Grammar (Goldwater & Johnson 2003, Wilson 2006, Hayes & Wilson 2008). As we will see, the advantage of such a grammar is that effects are more easily interpretable (given the multinomial nature of the stress domain in Portuguese). In such a grammar, constraints are numerically weighted and the well-formedness of an output is defined by a probability, which is derived from the sum of the violations incurred by each candidate under evaluation.

In the Maximum Entropy (MaxEnt) Grammar proposed by Hayes & Wilson (2008), each constraint has a non-negative weight. The higher the weight of constraint C, the stronger its effect in lowering the probability of every candidate that violates C. In order to calculate the probability of a given candidate, we first need to calculate the *score* of each output candidate (Hayes & Wilson 2008, p.383). The score of candidate x (denoted by $h(x)$) is the sum of all constraint violations incurred by x multiplied by the weight of each violated constraint.

(12) **Definition: Score**

The score of a given candidate x, denoted by $h(x)$, is

$$h(x) = \sum_{i=1}^{n} w_i C_i(x),$$

where

w_i is the weight of the ith constraint,

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

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

Next, a MaxEnt value $P^*(x)$ is calculated by taking the exponential of the negated score $-h(x)$, thus candidates with more violations receive lower MaxEnt values $P^*(x)$. The probability of x is then calculated according to its respective $P^*(x)$ as well as the total sum of other MaxEnt values under evaluation.

(13) **Definition: MaxEnt value**

Given a form x and its score $h(x)$, the MaxEnt value of x, denoted by $P^*(x)$, is

$$P^*(x) = exp(-h(x))$$

Let us now take the interval-based model as an example and map it into a MaxEnt approach, which requires (a) inputs, (b) possible candidates, (c) candidates' observed frequencies, and, naturally, (d) constraints and (e) weights. In this case, our constraints could be based on the Weight-to-Stress Principle, proposed by Prince (1990, p. 3) and defined as a constraint in Prince & Smolensky (1993, p. 63).

(14) **Weight-to-Stress Principle**

WSP Heavy syllables are stressed

Given the gradient nature of weight proposed in this paper, (14) needs to capture three important facts: (i) the domain of weight computation can also be the interval; (ii) weight varies across the three intervals (or syllables) in the stress domain and (iii) weight effects are cumulative within each position. One solution to (ii) is to propose a positionally-defined version of WSP, which entails that each position in the domain needs to be evaluated by a different WSP constraint. Weighted constraints appropriately capture (iii) by definition (15).

(15) **Gradient Weight-to-Stress Principle**

Let n be an unstressed position in the stress domain:

$\textsc{wsp}_n$ Assign one violation mark to every segment in n

Where the number of highly-ranked $\textsc{wsp}_n$ constraints in a grammar is defined by the number of positions in the stress domain in that grammar.

Thus, an unstressed VCCC interval in final position would violate $\textsc{wsp}_1$ four times. Next, we need to define the weights for $\textsc{wsp}_{1,2,3}$. Recall that the coefficients provided by the regression in Table 9 showed a clear gradient pattern in the stress domain, whereby if a given interval has a larger effect on stress, not stressing that interval is costlier.

The weights for $\textsc{wsp}_{1,2,3}$ were defined using the MaxEnt Grammar Tool (Wilson 2006). The lexicon was modelled as follows: each input represented a unique combination of all intervals in the stress domain. In total, 86 combinations (i.e., input forms) were attested in the lexicon—assuming the adjusted count of onsets discussed in §5.2. Because inputs represent sets of lexical items, they each have different frequencies for each stress pattern (candidates). For instance, the input with `int3=3`, `int2=2` and `int1=2` (or '`3-2-2]`' for short) represents the set of words that contain that interval combination (i.e., VCCVCVC words). In that set, the stress patterns are: ≈ 62% final, ≈ 36% penult and ≈ 0.7% antepenult. This allows us to set the frequency of each possible candidate in the language (n=3, given the stress domain).

This is a simpler method to define the inputs of the model, which differs from the analysis in §5: here, the sets of items (under unique interval sequences) being modelled are given *equal* weight, since no information regarding lexical representativeness of each input is provided. However, interval combinations represent very distinct proportions of the lexicon: for example, of the 86 combinations/inputs, 37 such combinations account for over 99% of all the words in Portuguese. Given that inputs do not specify lexical proportions, they conflate all words that contain a specific interval sequence.

In a MaxEnt grammar, two parameters are specified for each constraint, namely, μ, which is the 'preferred' value of constraint C (1 in this case), and σ, which, when small, forces the value of C to be close to μ—here, $\sigma = 10^7$, so μ is not forced to be close to 1 (Goldwater & Johnson 2003). As the data are learned, each weight is adapted so that the model is able to maximize the probability of the observed forms. The weights learned by the MaxEnt model are given in Table 11. Note that, like the $\hat{\beta}$ values in the analysis proposed in §5, the weights of all three constraints decrease as we move away from the right edge of the word. However, `int3` ($\textsc{wsp}_3$) has no effect.

Table 11: Constraint weights learned in a MaxEnt model

Constraint	WSP$_1$	WSP$_2$	WSP$_3$
Weight	0.83	0.27	0.00

Assuming the weights in Table 11, the ranking for Portuguese is WSP$_1$ $\gg$ WSP$_2$ $\gg$ WSP$_3$, as expected. In Tableau 1, we see the evaluation of an underspecified candidate of the shape /VCCCVCVC/ (e.g., *amostragem* [amosˈtraʒem] 'sampling'). Scores and MaxEnt values are provided for each candidate (a-c), along with predicted probabilities ($P(x)$) and actual lexical frequencies ($Freq(x)$). Note that $P(x)$ does not deviate much from $Freq(x)$. Once we take all 86 candidates into consideration, $P(x)$ and $Freq(x)$ are positively correlated, $r(256) = 0.52, p < 0.00001$.

Tableau 1: MaxEnt evaluation of /VCCCVCVC/

	/VCCCVCVC/	WSP$_1$	WSP$_2$	WSP$_3$	$H(x)$	MaxEnt	$P(x)$	$Freq(x)$
a	[VCCC•VC•ˈVC]	0	2	4	0.54	0.58	0.66	0.63
b	[VCCC•ˈVC•VC]	2	0	4	1.66	0.19	0.22	0.36
c	[ˈVCCC•VC•VC]	2	2	0	2.2	0.11	0.13	0.01

As expected, the MaxEnt approach briefly examined here also captures the weight gradience proposed in this paper. A constraint-based approach such as the one sketched in this section is advantageous, in that it formally characterizes a phonological grammar. In addition, constraint interactions tend to be more easily interpreted than multinomial/ordinal regressions. On the other hand, because of the methodology employed and the fact that input representativeness is not considered by the model, a MaxEnt approach may not capture subtle effects observed in the Portuguese lexicon, which involve different data proportions when we examine interval combinations (inputs). Small/subtle effects may also shrink due to the regularization term used in Maximum Entropy models to avoid overfitting. In the present analysis, the fact that all inputs are treated as equivalent meant that the effect of the antepenult interval (WSP$_3$) was *null*, even though the statistical model in Table 9 was able to capture a highly significant weight effect for that position (i.e., `int3`). In fact, two other models (binomial and multinomial logistic regressions, not presented in the paper) were also able to capture the very same patterns presented in §5. Presumably, if inputs encoded one more variable (lexical proportion), a MaxEnt grammar would also capture the effects of the antepenult interval.[39]

The analysis proposed in this paper and mapped into a MaxEnt model in this section predicts that other quantity-sensitive systems should present variations in the ranking proposed for Portuguese. For example, one should find a system where penult syllables/intervals, followed by antepenult syllables/intervals, affect stress more strongly than final syllables ($\textsc{wsp}_2 \gg \textsc{wsp}_3 \gg \textsc{wsp}_1$).

7 Conclusion

Questions (4a) and (4b) raised in §1 and repeated in §4 were, respectively, *Is weight-sensitivity only found word-finally in Portuguese?* and *Is weight-sensitivity* categorical *or* gradient*?*. In sections 4 and 5, I showed that the Portuguese lexicon clearly does not limit weight-sensitivity to the word-final syllable. In addition, weight-sensitivity is not categorical under either set of theoretical assumptions examined in this paper (i.e., syllables or intervals). Rather, weight has different effects depending on the position we investigate. Standard approaches to stress in Portuguese did not take such effects into consideration, probably because word-internal effects are substantially weaker than word-final effects.

Crucially, the answer to both questions is essentially the same regardless of which weight domain we choose: weight-sensitivity is not limited to word-final position, and it is not categorical. Now, let us move on to question (4c), namely, *How do onsets affect stress likelihood?*

In the data and models analysed in §4 and §5, onsets had a surprising effect: in final and penult syllables, having an onset in σ_j negatively correlated with stress on σ_j. These effects are unaccounted for in syllable theory, as one would expect either a null or a positive effect of onsets. The monotonic negative correlation between penult onsets and antepenult stress in Fig. 3 motivated an alternative weight domain, namely, interval theory. In sum, onsets did have a significant effect, but it is not clear how we can explain such an effect under syllable theory. Importantly, once we assume a revised version of intervals (one where extrametricality plays no role), onsets behave as we would expect in almost all cases.

The additional question examined here involved different weight domains, namely, syllables and intervals (4d). The interval model outperforms a baseline (intercept-only) model as well as the syllable model examined. Additionally, interval theory accounts for the patterns observed in the data more comprehensively than syllable theory does; it also has simpler assumptions, which makes it more parsimonious—this is reflected in the statistical model (Tables 9 and 10), which had fewer predictors (3 *vs.* 9). The fact that intervals miss the

[39] A different model was run with approximately half the inputs (thus only considering representative sets), and $\textsc{wsp}_3$ was no longer null.

different effects of codas and nuclei observed in §4 and §5 did not affect the models' accuracy when compared to the syllable-based model. This may seem surprising, but is likely due to the low frequency of onset clusters in the language. As a result, the vast majority of CC sequences involve a coda segment followed by an onset segment.

In the interval approach to stress proposed in this paper, I showed that the notion of extrametricality is not applicable, given the empirical patterns observed. In other words, onsets in all cases behave exactly as we would expect if they were not extrametrical in the language.

By examining weight effects on stress in Portuguese, this paper proposes a statistical/probabilistic analysis to stress in non-verbs in the language. This analysis has three main advantages: firstly, it is more parsimonious, since weight alone is able to account for the vast majority of cases. In fact, the interval-based model, which is more empirically motivated, is based on the very simple assumption that more segments increase stress likelihood; secondly, it is not stipulative, given that no abstract notions are employed to account for irregular cases (e.g., segmental/syllabic or word-edge extrametricality, catalexis); thirdly, both the syllable- and interval-based approaches are more accurate than analyses that assume all irregular forms are lexically marked (the accuracy percentage of these models goes beyond the proportion of regular forms in the lexicon modelled). I have shown that at least part of the irregularities are accounted for by a single factor under interval theory (i.e., weight), which suggests that this analysis could be even more accurate if more factors were considered (e.g., segmental duration). The paper also shed some light on the role of extrametricality in interval theory: in Portuguese, such a notion is not empirically motivated.

The analysis of stress proposed here entails that all elements in the stress domain (syllables or intervals) are evaluated in *parallel*: each rhythmic unit competes to bear stress, and the impact that segmental material has on each unit differs according to its position in the domain. In §6, I showed that this scenario can be mapped into a constraint-based approach where constraints have different weights, such as a MaxEnt grammar (Goldwater & Johnson 2003, Wilson 2006, Hayes & Wilson 2008). In such an approach, the weight gradience in Portuguese is also captured, as expected. However, if we assume inputs are interval combinations, the MaxEnt analysis in §6 is not successful in reproducing the effects of antepenult intervals found in §5, which are highly significant.

This paper examined the Portuguese lexicon, thus, variable phenomena regarding spoken Portuguese were outside the scope of this research. However, an important question is *How different is the lexicon from speakers' grammars with regard to stress and weight?* Onset cluster simplification, for instance, is relatively common in some dialects of Brazilian Portuguese (e.g., Southeastern Brazilian Portuguese (Harris

1997)). This phonological phenomenon is possibly related to stress, as we saw in §3. One important question regarding cluster simplification has to do with direction: is it the case that only unstressed syllables undergo this process (e.g., *próprio* [ˈpɾɔ.pɾjʊ] ⇒ [ˈpɾɔ.pjʊ] ‘proper’) or can we also find simplified *stressed* syllables, as in [ˈpɾɔ.pɾjʊ] ⇒ [ˈpɔ.pɾjʊ]? Answering these kinds of questions will clarify how the lexicon and speakers’ grammars may differ—which is essential for a comprehensive understanding of the language as a whole. Experimental research could also show whether speakers treat antepenult syllables differently, thus mirroring the surprising patterns in the lexicon. If that turns out to be the case, what could explain such distinct patterns?

Segmental quality is another important factor, since it is expected to significantly affect duration, which in turn causes intervals to lengthen or shorten. Segment duration is difficult to control for without production experiments where speakers’ dialects are carefully observed (e.g., coda /s/ and /r/ vary greatly in Brazil; vowel reduction/deletion is widespread in European Portuguese, including Faialense Portuguese (Silva 1997)).

The present analysis suggests that as far as the Portuguese lexicon is concerned, intervals are more empirically motivated. Let us assume that, in Portuguese, both the lexicon and speakers’ grammars motivate intervals as the domain of weight computation. In that case, one question we should consider is *How can we differentiate interval-internal segments, given that nuclei and codas have different effects in the lexicon?* Intervals do not preclude constituency *a priori*, and a unified view of syllables and intervals might be a more comprehensive answer to that question: V-to-V rhythmic units that still differentiate onsets from codas and nuclei. Such an analysis could imply different levels of representation, where syllabification takes precedence, determining which segments bear more or less weight, and intervals apply at a later stage, ‘re-parsing’ syllables—only then would (primary) stress be assigned. Perhaps a better solution lies on differentiating weight by (relative) sonority, as discussed in §3.3.

The analysis presented here may also be applicable to other Romance languages with similar stress patterns, such as Spanish and Italian—this would allow for a more comprehensive understanding of stress and weight computation across different but related languages. Do these languages also present gradient weight effects in their stress domains? In Portuguese, intervals seem to account for more patterns present in the lexicon; would that apply to other Romance languages? Answering such questions is crucial if we wish to have a broad understanding of how weight is computed (and how it affects stress) cross-linguistically.

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*.

Bisol, L. (2013). O Acento: Duas Alternativas de Análise. *Organon*, *28*(54).

Camara Jr., J. M. (1979). *The Portuguese language*. Chicago: University of Chicago Press.

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

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

Christensen, R. H. B. (2013b). ordinal—regression models for ordinal data. R package version 2013.9-30 http://www.cran.r-project.org/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. (2002). Branching onsets in Brazilian Portuguese. *Revista de Estudos da Linguagem*, *10*(1), 91–107.

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

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: a database with all non-verbs in Portuguese.
URL `https://github.com/guilhermegarcia/portuguese_corpus`

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Goedemans, R., & van der Hulst, H. (2013). *Weight Factors in Weight-Sensitive Stress Systems*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
URL `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.

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

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

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

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

Harris, J. (2011). Deletion. In 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*. 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 14 September, 2013.
URL `http://www.laps.ufpa.br/falabrasil`

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 *O Acento em Português: abordagens fonológicas*, (pp. 120–143). São Paulo: Parábola Editorial.

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.

Martins, M. R. D. (1982). *Aspects de l'accent en Portugais*. Hamburg: Hamburg: Buske Verlag.

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. (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.

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.

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.

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.

Silva, D. J. (1997). The variable deletion of unstressed vowels in Faialense Portuguese. *Language Variation and Change*, *9*(03), 295–308.

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

Tiedemann, J., & Nygaard, L. (2004). The opus corpus-parallel and free. In *LREC*. Retrieved 23 February, 2014.
URL `http://opus.lingfil.uu.se`

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.

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