Stress and gradient weight in Portuguese

Guilherme D. Garcia
guilhermegarcia.github.io

McGill University

WCCFL 33 | March 2015

Portuguese stress: previous analyses

- Weight effects on stress: only **word-finally**
- - Falling diphthongs or coda segments → heavy
- Weight-sensitivity is categorical
  - Word-final syllables: either H or L
- **Syllable-based** generalizations (rules and constraints)

See Araújo (2007) for a comprehensive review

Portuguese stress: previous analyses

- Weight effects on stress: only **word-finally** ✗
  - Falling diphthongs or coda segments → heavy
- Weight-sensitivity is categorical ✗
  - Word-final syllables: either H or L
- **Syllable-based** generalizations (rules and constraints) ✗

See Araújo (2007) for a comprehensive review

Today: A novel analysis

- Weight effects → across the whole stress domain ✓
- Weight-sensitivity is gradual ✓
- Probabilistic **interval**[1]**-based** account ✓
  - Economy, accuracy, empirical motivation

[1] Steriade 2012

Portuguese main stress

- **Non-verbs** ⇒ phonological factors + exceptional cases
  - Trisyllabic window: $\sigma\sigma\sigma]_{PWd}$
    ma.rí.ti.mo, *ca.vá.lo*, *fa.ról* 'maritime, horse, lighthouse'
  - **final** and **penult** stress: (mostly) **regular**
  - **antepenult** stress: **irregular/unpredictable**
  - Weight effects only word-finally (See Araújo (2007) for a review)
- Verbs ⇒ morphological factors (Wetzels 2007)

Non-verbs

Regular patterns: 72% of the Portuguese Stress Corpus (Garcia 2014; Houaiss et al. 2001, n=154,083)

- Final stress (≈15%): *papél* 'paper'
- Penult stress (≈57%): *caválo* 'horse'

- Standard view: Heavy σs only attract stress word-finally

Non-verbs

Irregular cases: 28% of the lexicon

$\acute{\sigma}\sigma\sigma$	≈13%	*pérola*	'pearl'
$\sigma\acute{\sigma}_L$	≈11%	*café*	'coffee'
$\acute{\sigma}\sigma_H$	≈4%	*nível*	'level'

- Previous work: Marked stress, extrametricality, catalexis and morphological factors (Roca 1999; Araújo 2007)

Portuguese stress

- Previous approaches:
  - i weight domain $= \sigma_{rhyme}\#$
  - ii weight is categorical, word-final and binary: H or L
  - iii onsets do not contribute to weight
- Possible issues:
  - Antepenult stress restricted to XLL]$_{PWd}$ words. Why? i.e., *X́HL, *X́LH, *X́HH
  - Recent studies on onset effects (Gordon 2005, Topintzi 2010, Ryan 2014)

Do onsets affect stress in Portuguese? If so, **how**?

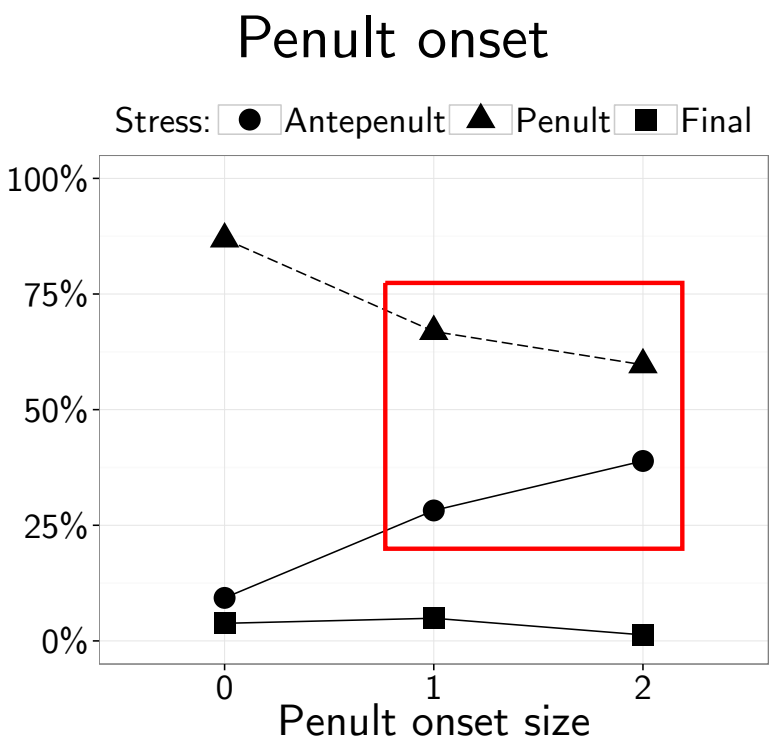
Intervals and syllables

Onsets and weight

- Prediction: no effect or a *positive* correlation CCV́ $\geq$ CV́
- Unexpected: CV́ $>$ CCV́, i.e., *negative* correlation
- Preview: This is found in the Portuguese lexicon (Houaiss et al. 2001)

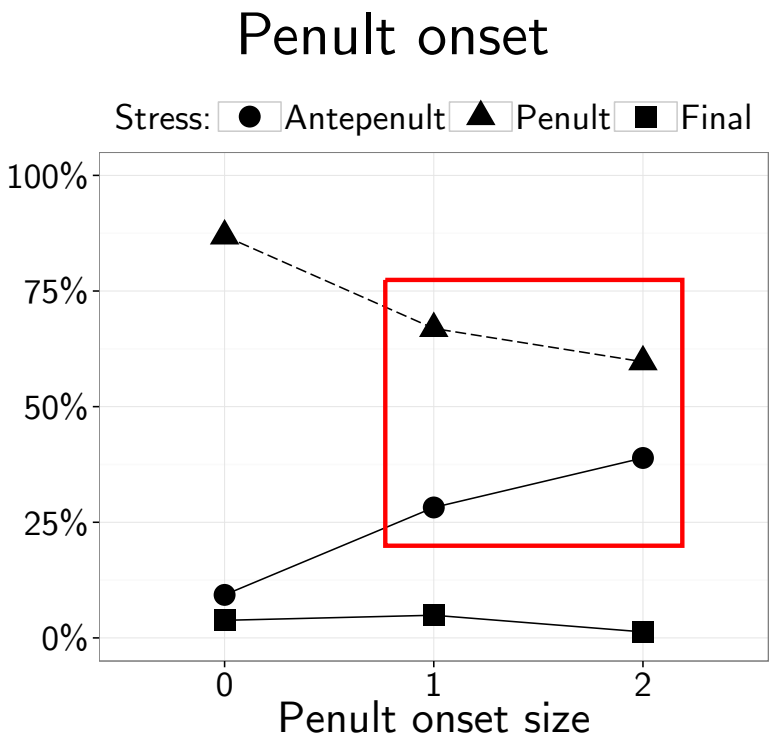
Intervals and syllables

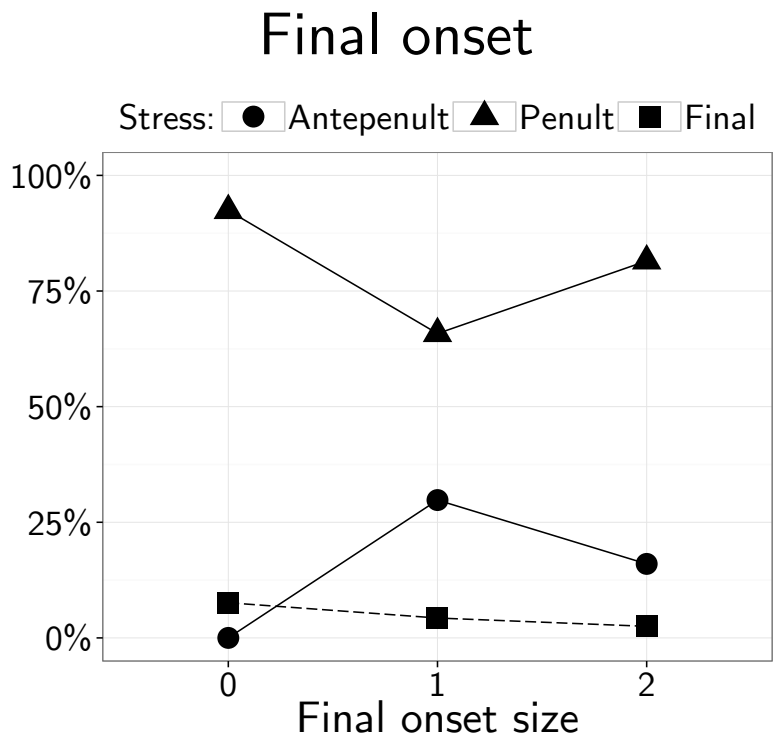
(LLL words)



Intervals and syllables

(LLL words)





Intervals and syllables

Defining intervals (Steriade 2012)

Facts are consistent with *interval* theory

- Interval: rhythmic unit that spans from one V up to the next V: no constituency assumed

  $\text{CVC}_{\sigma_2}\boxed{\text{CCVC}_{\sigma_1}} \Rightarrow \langle\text{C}\rangle\boxed{\text{VCC}_{\iota_2}}\text{VC}_{\iota_1}$

- More segments in one ι $\Rightarrow$ longer duration $\Rightarrow$ more weight

- Intervals argued to be the weight domain in unrelated languages: Finnish, Norwegian, Greek, Latin, Bhojpuri and Estonian

How about Portuguese?

Questions

i How does weight-sensitivity affect stress in Portuguese?
   categorical or gradual?
   word-final or broader in the stress domain?

ii How do syllables and intervals differ in answering question i?

Methods

The Portuguese lexicon (Houaiss et al. 2001)

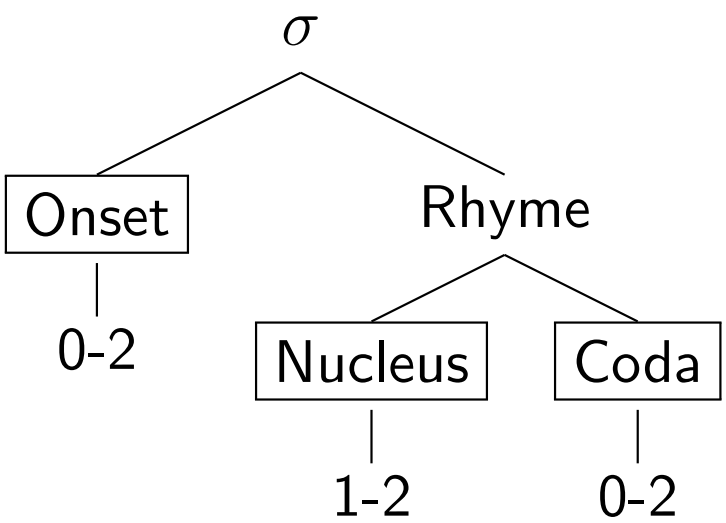
- Houaiss Dictionary ($n = 442{,}000$)
- Non-verbs: $\boxed{n = 154{,}083}$
- Representativeness: spoken vs. written Portuguese
  two frequency corpora: similar proportions of stress patterns
  (penult $>$ final $>$ antepenult)

- **What factors best predict stress position?**

Methods

Factors: stress domain ($\sigma\sigma\sigma$ | $\iota\iota\iota$)

Syllable



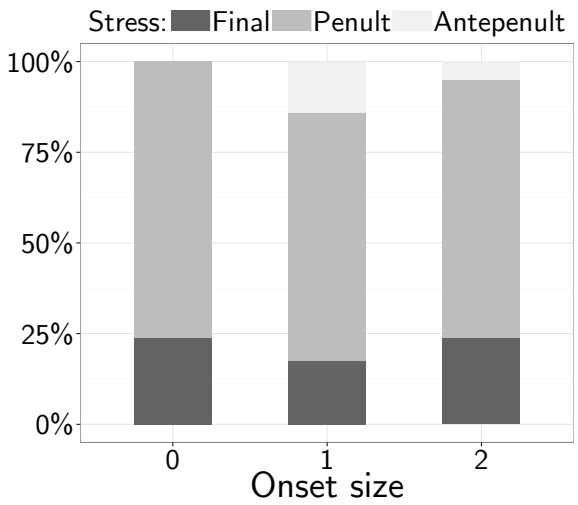
Interval

VCCC 4 VCC 3 $]_{PWd}$

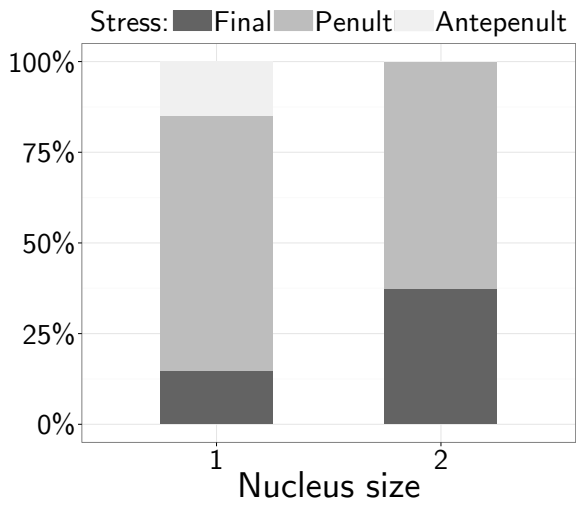
Methods

Data: **final** syllable by stress position

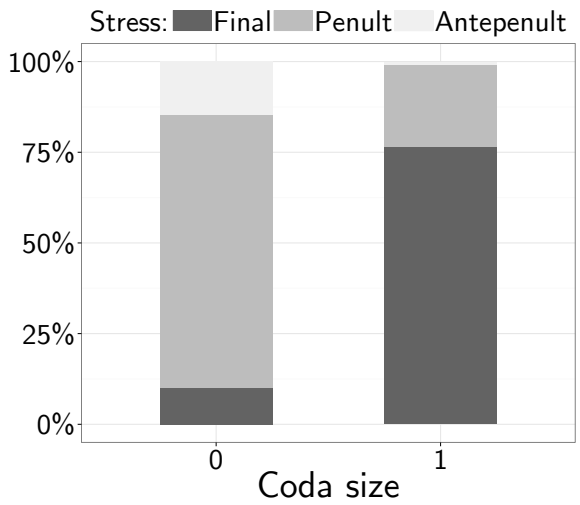
Onset



Nucleus



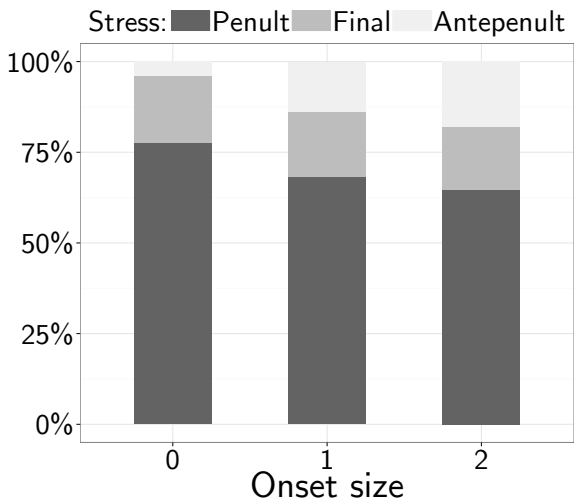
Coda



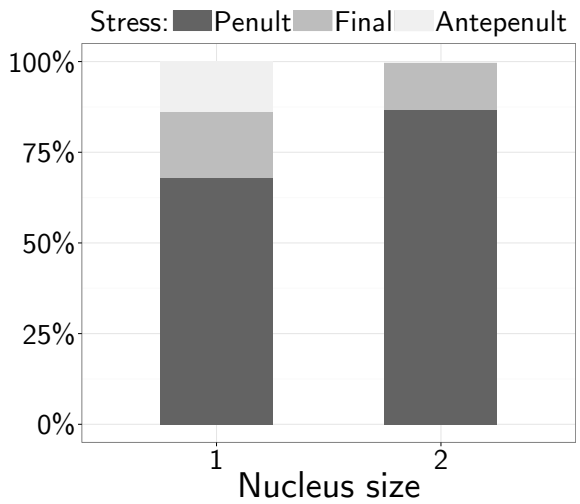
Methods

Data: **penult** syllable by stress position

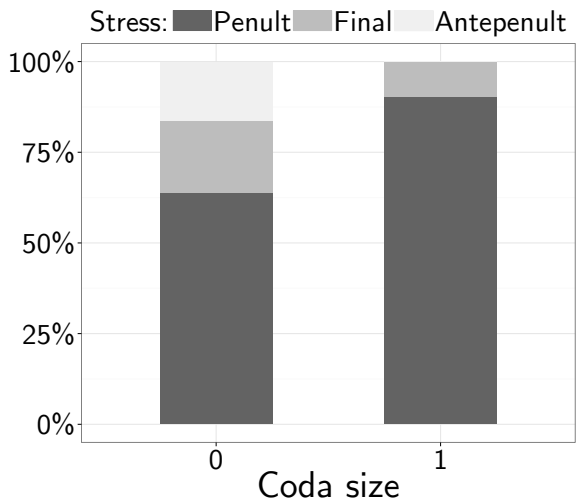
Onset



Nucleus



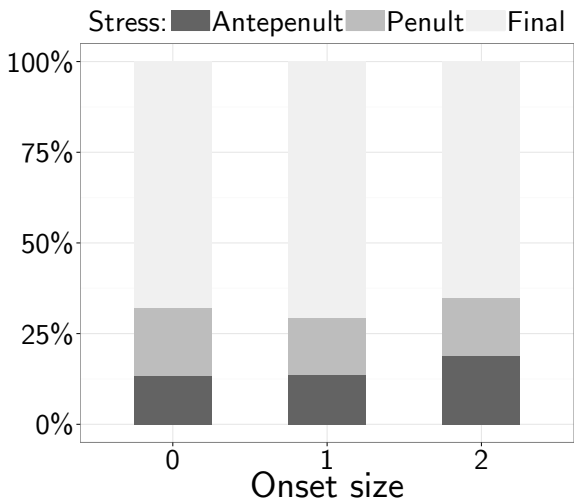
Coda



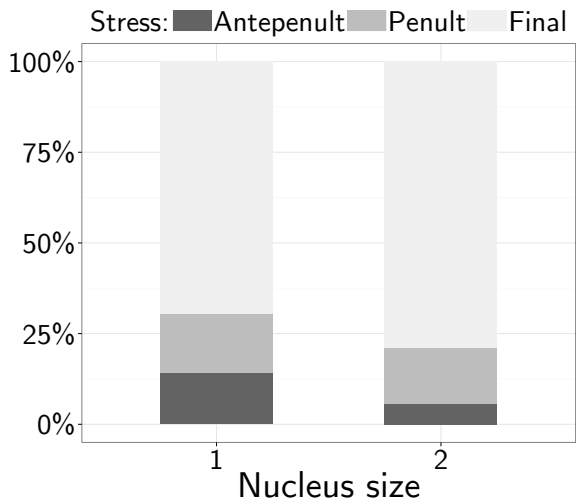
Methods

Data: **antepenult** syllable by stress position

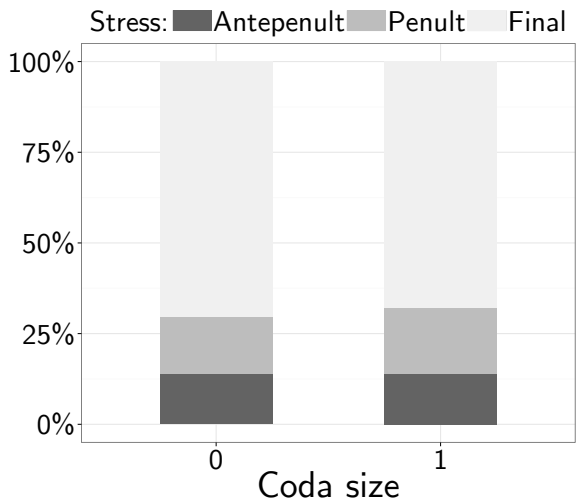
Onset



Nucleus



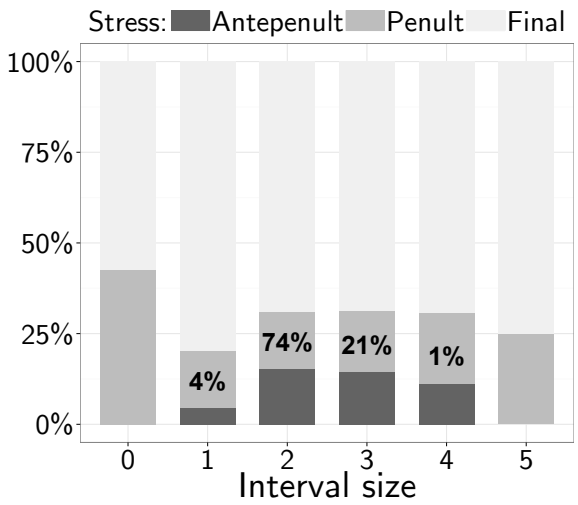
Coda



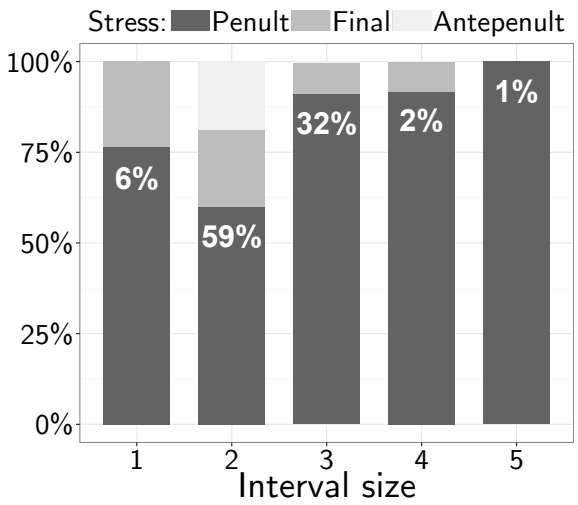
Methods

Data: stress pattern by interval

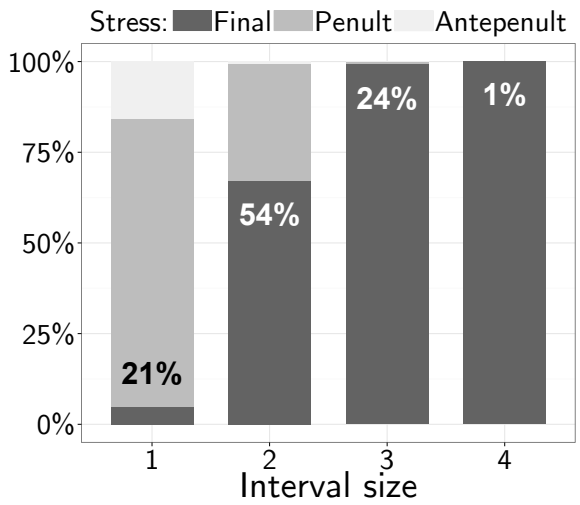
3 (ant)



2 (pen)



1 (fin)



Methods

Variables predicting `stress`: Ordinal Regression* (`clm()` in R)

Syllable model
9 predictors ($3/\sigma$)

`onset` (y/n)
`nucleus` (monoph-/diphthong)
`coda` (y/n)
(binary)

Interval model
3 predictors ($1/\iota$)

`int.0`
`int.1`
`int.2`

Response:
$3 > 2 > 1\ ]_{Pwd}$

Methods: Model evaluation

- **Accuracy:** What % of the lexicon is accurately predicted?
- **Consistency:** Which model is more theoretically consistent?
- **Fit:** Which model has a better fit (AIC, fit *vs.* df)?

Analysis: Syllable model

In what positions does weight affect stress?

What has been argued: CVC.CVC.C[VC]#

Actual patterns: [CV]C. [CVC]. [CVC]#

Analysis: Syllable model

- **Onset behaviour: Inconsistent with syllable theory**

$$\boxed{\text{CVC—}\mathbf{C}_2\text{VC—}\mathbf{C}_1\text{VC\#}}$$

1. Final onsets → penult stress $\quad \hat{\beta} = 1.52 \quad p < 0.00001$
2. Penult onsets → antepenult stress $\quad \hat{\beta} = 0.50 \quad p < 0.00001$

Analysis: Interval model

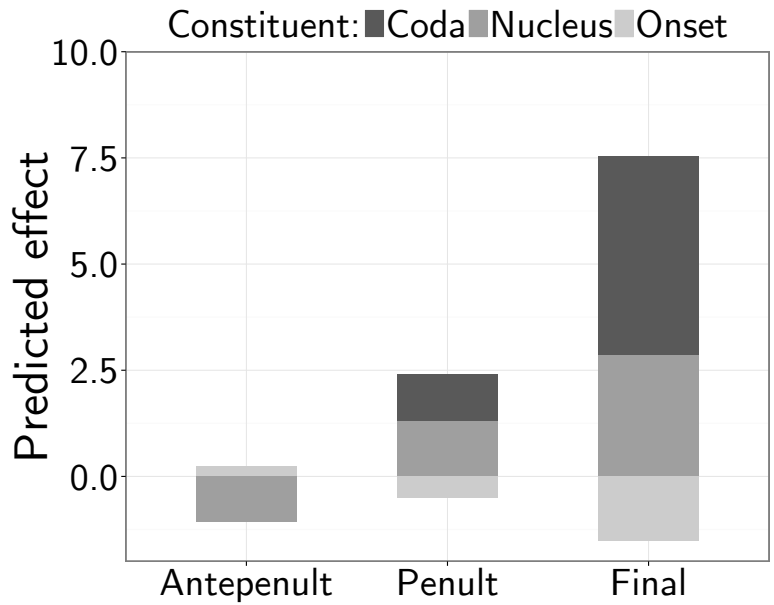
- Longer intervals → higher stress likelihood
- Weight gradience: **between** and **within** interval

CVCC—VCC—VC#

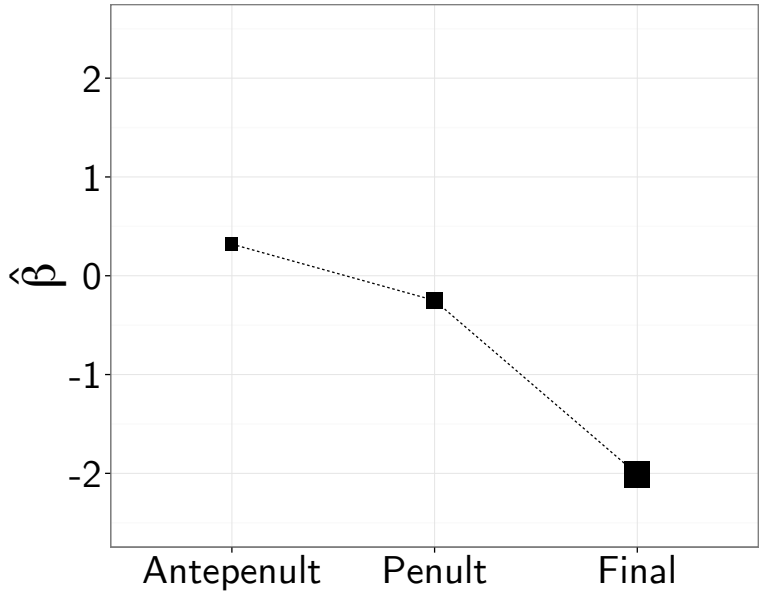
`int1` → final stress	$\hat{\beta} = -2.01$	$p < 0.00001$
`int2` → penult stress	$\hat{\beta} = -0.25$	$p < 0.00001$
`int3` → antepenult stress	$\hat{\beta} = 0.32$	$p < 0.00001$

Question i: Weight gradience (regression weights by model)

Syllables



Intervals



Question ii: Syllables *vs.* intervals

Domain	**Predictors**	AIC	κ	**Accuracy**
Syllable	9	186433.21	40.79	74.75%
Interval	3	180035.58	15.38	78.29%

Intervals:

- More accurate
- More empirically motivated
- More parsimonious

Summary of proposal

- Weight effects on stress are gradual, **not** categorical
- Intervals offer a better analysis of stress in Portuguese
- Onset effects are accurately captured

- Crucially: only **one** predictor, namely, **weight**
- **Probabilistic:** compatible with constraint-based grammars

Final remarks

- Weight effects in Portuguese are *gradient*
- Onset effects suggest that weight is computed in intervals
- Advantages of a probabilistic interval-based approach:
  - Economy, empirical motivation and accuracy
  - Nearly 80% accuracy *vs.* 72% regular cases
  - Only **one** positionally-defined predictor/constraint
- Probabilistic constraint-based implementation (E.g., MaxEnt)

Main references

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

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

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

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

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

Main references

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

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

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

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

Many thanks to Heather Goad, Morgan Sonderegger, Kie Zuraw, Michael Wagner and Colin Wilson.

Thank you | **Obrigado**

`guilherme.ca`

Appendix I: Syllable model

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%		$\kappa = 40.79$

Appendix II: Interval model

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%		$\kappa = 15.38$

Appendix III: Edge effects

[⟨C⟩VC *vs.* [CVC → No onset extrametricality in the lexicon

- Onsets at the left edge have a *positive* effect on stress. Why?
- Hypothesis: Parsing is *exhaustive*
- Test: words with 2, 3 and 4 syllables

If onsets have to be parsed...

Overall, penult onsets have a **negative** effect on penult stress.

How about disyllabic words? Do we find a null effect?

2σ ([**C**V...)	$\hat{\beta} =$ 0.18	$p < 0.002$
$> 2\sigma$	$\hat{\beta} =$ -1.39	$p < 0.00001$

Onsets have a **positive** effect at the left edge of the domain

Appendix IV: Constraint-based approach

An adaptation of the Weight-to-Stress Principle (Prince 1990)

Let ι_u be an unstressed interval:

WSP_{ι_u} Assign one $*$ to every segment in ι_u

Example: MaxEnt Grammar (Hayes & Wilson 2008)

MaxEnt: Lexical modelling

- Lexicon = {different combinations of interval sequences}
- So, *amostragem* 'sampling' $\subseteq$ {4-2-2} (VCCC-VC-VC)
- Assume inputs = unique sequences
- Each input includes a non-empty set of words
- Sets of words (inputs) have different stress patterns
- Problem: sets $\rightarrow$ *very* skewed lexical distribution

MaxEnt

Learned weights for all words in Portuguese

Constraint	WSP_3	WSP_2	WSP_1
Weight	0.00	0.27	0.83

- Weight gradience across the stress domain
- Antepenult interval: No weight effect learned
  - Lexical skewness: 37 out of 86 inputs $\rightarrow$ 99% of the lexicon
  - Model treats each input *equally*

MaxEnt

Example

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