Probabilistic grammar and the Portuguese Stress Corpus

This paper proposes a weight-based probabilistic approach to stress in Portuguese. Previous analyses have argued that weight-sensitivity in the language is categorical and restricted to word-final syllables. I show that weight effects in Portuguese are *gradient* and can be found across *all* three positions in the stress domain (three-syllable window). I also compare two domains of weight computation, namely, the syllable and the interval [5], and show that interval-based statistical models are more internally consistent.

A thorough phonological analysis of stress in a given language requires a comprehensive and detailed corpus. Unlike frequency corpora, such a corpus must contain quantitative and qualitative segmental information as well as stress marking, syllabification, syllable shapes and weight profiles, among other variables. The present study introduces the *Portuguese Stress Corpus* (PSC, author). PSC was developed to provide researchers with a large and reliable annotated lexicon of Portuguese. The corpus contains nearly all non-verbs in the language (n = 154,611), and is based on the wordlist found in the Houaiss Dictionary [4], the most comprehensive dictionary in the Portuguese language.

Background

Weight, Syllables and Intervals: In languages where stress is weight-sensitive, syllables with greater weight are more likely to be prominent, i.e., to attract stress [2]. In interval theory [5], greater weight entails greater duration in a given interval, defined as the rhythmic unit that spans from a vowel up to (but not including) the following vowel; i.e., V-to-(V). Segments preceding the leftmost vowel are not included in any interval. Intervals (ι) have no *a priori* constituency, and predict different rhythmic units when compared to syllables (σ): onset segments in a given syllable are computed as part of the preceding interval. For example, the string $\mathrm{CVC}_\sigma\mathrm{CCVC}_\sigma$ is equivalent to $\langle\mathrm{C}\rangle\mathrm{VCCC}_\iota\mathrm{VC}_\iota$ in interval theory. The onset effects found in the Portuguese Stress Corpus motivate intervals, as they are *negatively* correlated with stress.

Portuguese: Previous analyses ([1], [6], among others) have argued that weight effects on stress in Portuguese non-verbs are restricted to the word-final syllable (stress in verbs is not phonologically conditioned): stress is final if the word-final syllable is heavy. Otherwise, stress falls on the penult syllable (regardless of weight). Both final and penult stress patterns are (mostly) *regular*. Antepenult stress is *irregular*, and no pre-antepenult stress is allowed. Approximately 72% of the non-verbs in the language (N=163,626) have regular/predictable stress. Researchers have employed different factors to account for stress regularities (e.g., foot binarity, foot type, metrical alignment) and irregularities (e.g., extrametricality, catalexis, theme vowel influence) in the language. Under previous analyses, irregular and regular words are by definition treated differently.

Methodology: In the present study, stress in the Portuguese Stress Corpus was modelled using Binomial Logistic Regressions (`glm()` in R). Given that three stress positions are possible, two binomial models were required to predict stress: one model predicts antepenult stress *vs.* penult or final stress ((1-a), (2-a)). Because all words with antepenult stress are considered to be irregular, this model predicts the following: given a word, *how likely is it to bear antepenult stress as opposed to penult or final stress?* Another model predicts penult *vs.* final stress ((1-b) and (2-b)); i.e., the two regular positions in the language. Because syllables and intervals are compared, a total of four models are used.

Syllables have three constituents (onset, nucleus and coda). Intervals, on the other hand, are a single string of segments (no constituency). As a result, 9 predictors are used in (1-a), 6 in (1-b); 3 in (2-a), and 2 in (2-b).

(1) Syllable-based models (∼ = as a function of)

a. `antPenFin model: stress (ant vs {pen/fin})` ∼ `[onset + nucleus + coda] sizes (×3)`

b. `penFin model: stress (pen vs fin)` ∼ `[onset + nucleus + coda] sizes (×2)`

(2) Interval-based models (∼ = as a function of)

a. `antPenFin model: stress (ant vs {pen/fin})` ∼ `interval size (×3)`

b. `penFin model: stress (pen vs fin)` ∼ `interval size (×2)`

Results: Traditional approaches predict that (i) only the size of *word-final* constituents should have a significant effect on stress, and that (ii) a syllable is either heavy (H) or light (L). In other words, vowel+glide$]_\sigma$ syllables (heavy in Portuguese) should behave as heavy as vowel+consonant$]_\sigma$ syllables (also heavy in the language). However, both syllable- and interval-based models confirm that weight effects are *gradient* and found across *all three* positions in the stress domain (all effects at $p < 0.00001$, including antepenult syllables/intervals; interval-based effects are shown in Fig. 1). These results show that weight-sensitivity is much more intricate than previously thought. Importantly, the syllable-based models also show effects that are inconsistent with the representational assumptions in syllable theory. For example, antepenult nuclei are negatively correlated with antepenult stress ($\hat{\beta} = -0.219, p < 0.00001$); i.e., monophthongs are preferred to diphthongs in the antepenult syllable of words with antepenult stress. The interval-based models are more internally consistent; i.e., every interval has a positive impact on stress (the longer the interval, the more likely it is to bear stress).

Crucially, both intervals and syllables provide a significantly more accurate characterization of stress in Portuguese. Fig. 2 exemplifies the mean predicted probabilities of penult (*vs.* final) stress. A LH word, for example, is predicted by previous analyses to always bear final stress (dotted line). However, nearly 25% of such words have penult stress (◆). The syllable- and interval-based models (σ and ι, respectively) presented here accurately approximate the actual proportion.

The probabilistic grammar implied in the present analysis assumes that words are assigned stress based on the probability distribution of stress patterns already in the language. This entails that stress is lexically marked once assigned, which means all words are assigned stress based on the same principle. In other words, we no longer need to differentiate regular and irregular patterns. Rather, patterns are more or less likely. Finally, this predicts that newly-coined words may also be assigned antepenult stress, unlike previous analyses, which treat such a pattern as completely unpredictable. This proposal can be mapped into a constraint-based approach such as MaxEnt Grammar [3].

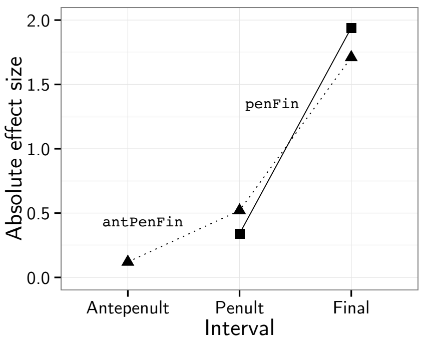

Figure 1: Absolute effect size by interval. `antPenFin` model predicts antepenult *vs.* penult/final stress; `penFin` model predicts penult *vs.* final stress.

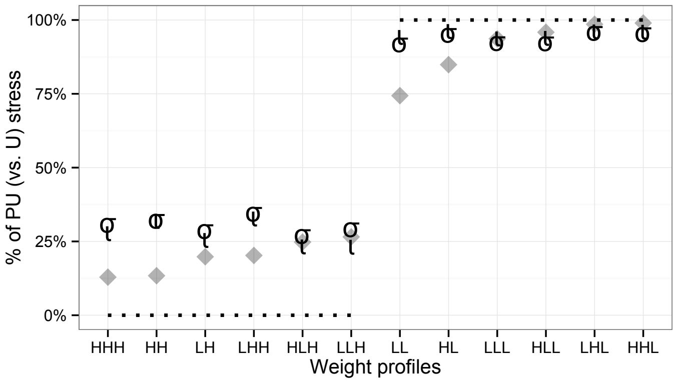

Figure 2: Mean predicted probability of penult (*vs.* final) stress by weight profiles based on syllable- and interval based models. Dotted lines represent predicted probabilities based on previous (categorical) approaches. ◆ represents actual proportions in PSC.

References

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

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

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

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

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

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