LANGUAGE TRANSFER AND POSITIONAL BIAS IN ENGLISH STRESS*

Guilherme D. Garcia (Ball State University)

To appear in *Second Language Research*

ABSTRACT

This paper shows that L1 transfer may not be effectively maintained in the interlanguage due to confounding factors in the L2. When two factors, $\mathcal{A}$ and $\mathcal{B}$, are correlated in the L2, second language learners may only acquire $\mathcal{B}$, even if $\mathcal{A}$ is present in the L1. Transfer may not be effective because $\mathcal{B}$, being more robust in the input, conceals $\mathcal{A}$. Native speakers, on the other hand, generalize $\mathcal{A}$ in spite of $\mathcal{B}$. The variables in question are weight-sensitivity ($\mathcal{A}$) and positional bias ($\mathcal{B}$) in English, both of which can predict the location of stress in the language. I show that two seemingly target-like groups of second language learners of English (speakers of Mandarin and speakers Portuguese) fail to accurately generalize weight-sensitivity in the language, and instead display response patterns which are predictable given the existing positional bias in English stress.

Keywords: transfer, stress, weight, positional bias, English, Mandarin, Portuguese, Bayes

1 Introduction

A common assumption in second language acquisition is that structures in the first language (L1) are transferred to the second language (L2)—White 1989, Schwartz and Sprouse 1996. As a result, the patterns we observe in second language learners' (L2ers) interlanguage (Selinker 1972) are overall systematic, and can be predicted if we analyze relevant patterns in the L1. Needless to say, the interlanguage is gradually restructured in response to L2 input, i.e., transfer is stronger at initial stages, and slowly weakens as L2ers become more proficient in the L2 (e.g., Broselow and Park 1995, Schwartz and Sprouse 1996).

Transferring a pattern that exists in the L1 may not be as straightforward as it sounds, however. Suppose a pattern $\mathcal{P}$ exists in the L2 which can be predicted on the basis of *two* separate but highly correlated variables, $\mathcal{A}$ and $\mathcal{B}$. Now, assume that while variable $\mathcal{A}$ can be easily transferred from

*Thanks to Heather Goad, Natália Brambatti Guzzo, Jeffrey Lamontagne, and Jiajia Su. Thanks also to the audience at GALANA 8, and to the anonymous reviewers at *Second Language Research*. This research was supported by grant 435-2015-049 from the Social Sciences and Humanities Research Council of Canada, awarded to Heather Goad and Lydia White, and grant 19-0214 from Ball State University, awarded to Guilherme D. Garcia.

the L1, variable $\mathcal{B}$ is more robust in the input to which L2ers are exposed (i.e., variable $\mathcal{B}$ is more frequent and/or more phonetically salient in the input). L2ers may fail to effectively maintain $\mathcal{A}$ in their interlanguage simply because $\mathcal{B}$, being more robust, conceals the effects of $\mathcal{A}$. Indeed, given that $\mathcal{B}$ predicts $\mathcal{P}$ most of the time, we may even ask ourselves if the grammars of native speakers of the L2 accommodate $\mathcal{A}$ in the first place.

The scenario described above characterizes the topic of the present study: the pattern in question is stress in English; variable $\mathcal{A}$ refers to phonological weight effects (weight-sensitivity), and variable $\mathcal{B}$ refers to a positional bias favoring word-initial stress. L2ers in the present study are represented by two groups with typologically distinct L1s: speakers of Mandarin, and speakers of Brazilian Portuguese (henceforth Portuguese).

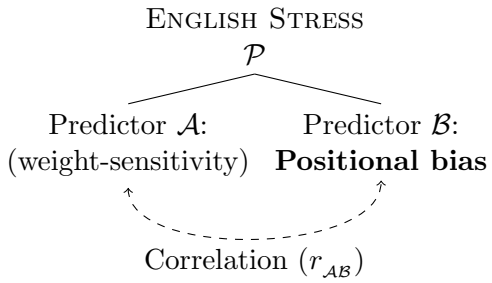
English is a weight-sensitive language, which means the location of stress is at least in part determined by whether a given syllable is heavy or light (Hayes 1995). In most nouns and adjectives (non-verbs), stress falls on the penultimate (PU) syllable if the syllable is heavy (*agénda*), and on the antepenultimate (APU) syllable otherwise (*Cánada*; §2). Given that most common words in English are short, a substantial percentage of such words have initial stress (Cutler 2012). As a result, weight-sensitivity and position are correlated predictors of stress location in the language—this confounding factor[1] is graphically illustrated in Fig. 1.

Like English, both Mandarin and Portuguese are sensitive to weight (Qu 2013, Garcia 2017a). In Mandarin, the notion of phonological weight is tied to lexical tones (Qu 2013). As a result, tones can be classified as weightless, light, heavy, or superheavy—see §2.2. In Portuguese, on the other hand, phonological weight is tied to syllable structure (§2). Crucially, in all three languages phonologically prominent syllables are phonetically longer (§2), which reflects the idea that "weight can be thought of as a property of the time dimension: a syllable is heavy because it is long" (Hayes 1995, p. 271). Therefore, given that both Mandarin and Portuguese associate prominence with phonological weight, and phonological weight with syllable duration (length), speakers of either language could transfer phonological weight from their L1 to the L2 (English).

The objectives of this paper are two-fold. First, I will show that the two groups of L2ers mentioned above fail to successfully generalize weight-sensitivity to English, even though their L2 grammars seem target-like on the surface (i.e., their stress assignment in existing English words is accurate). Native speakers' grammars, in contrast, generalize weight-sensitivity in spite of the presence of another predictor, namely, a positional bias favoring word-initial stress. This positional bias is more robust than weight-sensitivity as it can be observed not only for primary stress, but also for secondary stress (Cutler 2012). Weight-sensitivity, on the other hand, is mostly constrained to primary stress, and thus can only be fully active within the stress domain. Furthermore, word edges tend to be phonetically more salient relative to word-internal positions (Cutler 2012), which

[1]This could be a case of collinearity in the input: in mono- and disyllabic words in English, stress can be completely predictable based on either weight or position. Crucially, because the vast majority of common words (non-verbs) in English fit the length in question, it is possible that during the initial stages of acquisition of English stress, learners' lexica contain truly collinear variables. As learners' lexica increase in size and complexity (longer, less common words are acquired), the correlation between the two variables should decrease.

Figure 1: English stress position can be largely predicted by both weight-sensitivity (heavy syllables are more likely to be stressed) and positional bias (word-initial stress happens to be more likely overall). Both variables are correlated (dashed arrow).



in turn means that word-initial stress may be more easily perceived relative to word-internal stress.

The second objective of the present study is methodological. I will provide a statistical analysis using Bayesian estimation with uncertainty, which allows us to incorporate theoretical assumptions into our statistical models through mildly informed priors. As a result, the statistical models in §5 will not only be used to analyze the data, but also to simulate distinct grammars which are directly related to the hypothesis of the study (§3). I will show that native speakers' behavior is best captured by a statistical model that simulates a grammar *with* weight-sensitivity. L2ers' behavior, on the other hand, does not support such a model. Comparing models with different theoretical assumptions not only further strengthens the analysis of the phenomenon in question, but also demonstrates how a Bayesian approach to data analysis can be more informative and meaningful than traditional (Frequentist) approaches in second language research.

The paper is organized as follows. §2 briefly reviews stress in English, Mandarin, and Portuguese. §3 details the hypothesis of the present study—alluded to above. In §4, I describe in detail the experimental design and the statistical analysis employed in this study. In §5, I examine the experimental results and provide a statistical analysis for the data. Finally, §6 discusses the findings.

2 Background

Weight-sensitivity is a frequent topic of interest in phonological studies, and has played a crucial role in the development of phonological theory (e.g., Liberman and Prince 1977, Hyman 1985, Halle and Vergnaud 1987, Kenstowicz 1994, Hayes 1995, Gordon 2006, Ryan 2019). Nevertheless, while different studies have examined the second language acquisition of stress (e.g., Archibald 1993, Broselow and Park 1995, Pater 1997, Kijak 2006, Tremblay 2008, Özçelik 2014), the role of syllable weight *per se* has not received equivalent attention (though see Pater 1997). Face (2005), for example, is one of the only studies which directly examine whether (and how) L2ers (English speakers) acquire weight-sensitivity in the L2 (Spanish). Yet, the subtlety of weight-sensitivity can

help us better estimate the limits of second language grammars: can L2ers see past the competing positional bias in English stress and effectively transfer weight-sensitivity from their L1s? In other words, can they transfer *and maintain* weight-sensitivity active in the L2?

This section briefly reviews the stress and weight patterns in English, Mandarin, and Portuguese. The main objective is to establish that similar weight/prominence patterns can be found in all three languages.

2.1 Stress in English

English stress has been studied extensively (Chomsky and Halle 1968, Liberman and Prince 1977, Selkirk 1980, Halle and Vergnaud 1987, Hayes 1982, 1995, Moore-Cantwell 2016, among others). Overall, stress in the language is not completely predictable, but clear patterns can certainly be observed. For example, Liberman and Prince (1977) point out that a nonce word such as *pódectal* not only cannot be found in the English lexicon, but is also considered to be unnatural by native speakers. This has led researchers to conclude that English stress is both listed in the lexicon and derived by rule (Hayes 1982, p. 237).

In English, verbs and non-verbs (nouns and adjectives) display different stress patterns. For example, whereas final stress is not uncommon in verbs, it is typically avoided in non-verbs (see below). This distinction can be observed in minimal pairs such as *presént* vs. *présent*, and *impórt* vs. *ímport*—note that vowel quality is often affected by the location of stress in English; I return to this below. In spite of this difference in final stress, both verbs and non-verbs can be accounted for by similar metrical patterns (e.g., extrametricality; see Hayes 1982).

2.1.1 Weight-sensitivity in English stress

The general stress pattern in English is tied to the notion of weight, i.e., that heavy syllables, which are typically longer in duration, attract stress (Hayes 1995). In English, heavy syllables contain a coda consonant (e.g., *agénda*), or a diphthong (e.g., *Arizóna*, *recítal*). Primary stress in non-verbs typically falls on the PU syllable if a heavy syllable is present (*ve.rán.da*), and on the APU syllable otherwise (*Cá.na.da*, *dí.sci.pline*)—a pattern which resembles that of Latin stress. Final stress is typically avoided in non-verbs, but can be found in words which have a heavy final syllable (e.g., *typhóon*, *sardíne*, *shampóo*)—note that both *typhóon* and *sardíne* have a tense vowel, which is longer than its lax counterpart, as well as a coda consonant in their final syllables. As we can see, even though final stress deviates from the patterns in (1), it is also affected by weight.

The avoidance of final stress in English non-verbs has been captured in the literature with the concept of *extrametricality*: word-final syllables that do not contain a long vowel in non-verbs[2] are extrametrical, and are therefore skipped when stress is assigned. Simply put, what counts as a

[2]In verbs, word-final *segments* are assumed to be extrametrical (Hayes 1982). As a result, in verbs with two word-final codas such as *presén⟨t⟩*, the final syllable is still heavy and is therefore predicted to attract stress. Verbs with a single word-final coda, on the other hand, have a light final syllable, and are predicted to have non-final stress: *lítte⟨r⟩*.

heavy syllable varies depending on the position of said syllable (i.e., final *vs.* non-final).

(1) Stress and weight in English non-verbs

a. Heavy PU syllable → PU stress
agénda, *recítal*

b. Light PU syllable → APU stress
Cánada, *díscipline*

The relationship between weight and stress in the English lexicon (already established in the literature, e.g., Hayes (1995)) can be observed even in relatively small (but representative) subsets of the lexicon. Table 1 lists the percentage of APU stress for different weight profiles—both adjectives and nouns are listed. The data come from a sample of the CMU Dictionary (CMUdict 2014). The sample is based on the filtered wordlist used in Moore-Cantwell (2016) and Garcia (2017a), and contains only trisyllables with at most one heavy syllable to avoid conflicting effects of multiple heavy syllables in a single word (n = 4,573). If we compare HLL (heavy-light-light) and LHL (light-heavy-light) words to LLL (light-light-light) words, we can see that the percentage of APU stress is very similar between HLL and LLL words, but drops to nearly 0% in LHL words—a result which reflects the patterns listed in (1).

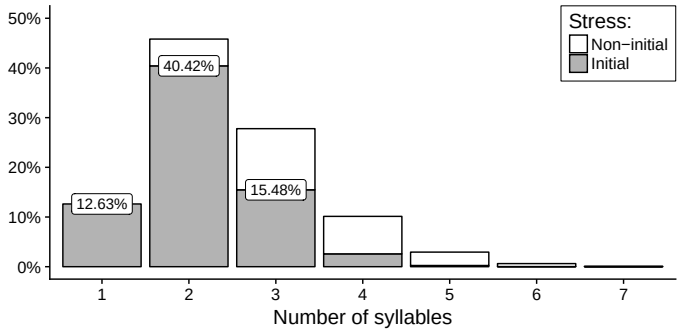
Table 1: Percentage of APU stress (as opposed to PU stress) by weight profile and part of speech (PoS) in English non-verbs in a sample (n = 4,573) of the CMU Dictionary (CMUdict 2014). "H" and "L" represent heavy and light syllables, respectively. More than 98% of the words in the sample have either APU or PU stress. For HLL and LLL words, both weight and position correctly predict the location of stress in most words.

Weight profile	PoS	%	Example	
HLL	Adj	69.54	*ábsolute*	ˈæb.sə.⟨lut⟩
HLL	N	74.17	*ábstinence*	ˈæb.stə.⟨nəns⟩
LHL	Adj	0	–	
LHL	N	2.49	*gálaxy*	ˈgæ.lək.⟨si⟩
LLL	Adj	68.65	*géneral*	ˈd͡ʒɛ.nə.⟨ɹəl⟩
LLL	N	75.05	*précedence*	ˈpɹɛ.sə.⟨dəns⟩

2.1.2 Positional bias in English stress

The patterns in (1) entail that most words will have either PU or APU stress. For disyllables or trisyllables, which are especially common in English, that means that stress will often be word-initial. This positional bias is shown in Fig. 2, which plots the percentage of words with initial and non-initial stress (y-axis) against word length in number of syllables (x-axis). The set of words with one, two, or three syllables accounts for over 85% of the CMU Dictionary (CMUdict 2014). In that set, words with initial stress account for nearly 70% of the entire lexicon being plotted.

Figure 2: Percentage of words with initial and non-initial stress by number of syllables in the CMU Dictionary (n = 133,557). Specific percentages are provided for words with one, two, or three syllables (relative to all words; y-axis) which have initial stress, a subset that accounts for nearly 70% of the words plotted.



The observation that most words in English have initial stress is not new. Cutler and Carter (1987), for example, show that about half of all words in English are polysyllabic and have initial stress—see Cutler (2012) for a comprehensive review of the literature on word-initial stress in English. If initial secondary stress is taken into account, then about two thirds of the language have initial stress. This word-initial bias is even stronger if we consider common words, which are typically what we observe in everyday speech: Cutler and Carter's study shows that about 60% of the London-Lund Corpus consists of monosyllabic words. Polysyllabic words with a weak initial syllable made up less than 10% of the corpus.

The fact that most words in English have initial stress can be useful to listeners, who can use stress to identify where words end and begin (Cutler 2012)—stress in English is phonetically signaled through higher pitch, higher intensity, longer duration, as well as vowel quality (Bolinger 1958, Beckman 1986, Lieberman 1960), which facilitates perception. Second language learners could naturally benefit as well: given that most words that they will hear and use will be relatively short, assuming that stress will be initial is statistically reasonable, and, crucially, simpler than considering the correlation between weight and stress in (1). Indeed, the correlation in question can be easily confirmed if we consider that most common English words are mono- or disyllabic. For non-verbs in such a sample, weight and position will be completely correlated, by definition.

In summary, English stress has (at least) two main reliable predictors: weight and position. As we saw above, because the size of the stress domain in English (1) is a superset of the typical length of common words (Fig. 2), weight effects and the positional bias discussed above are correlated, and when two variables are correlated, it becomes more difficult to determine which variable is driving a particular effect. More importantly, it is safe to conclude that the positional bias in question is more robust than the weight effects in the language: not only is it lexically more frequent in the input, it is also phonetically more salient (Cutler 2012), and computationally simpler (e.g., unlike weight-driven stress, no if-statements are needed to characterize word-initial stress). As a result, the positional bias in English could potentially conceal the weight effects observed in the language:

monosyllables and disyllables are extremely frequent in English, as shown in Fig. 2, but they offer little evidence for the effect of weight on stress; they do, however, reinforce the positional bias discussed above.

2.2 Stress in Mandarin

Unlike in English, stress and weight effects in Mandarin are disputed in the literature—in part due to the typological differences between these two languages. Hyman (1977), for example, classifies Mandarin as a language without stress. Duanmu (2007), on the other hand, not only argues that Mandarin has stress, but also shows that the realization of stress is similar between English and Mandarin. Likewise, some researches have argued that Mandarin is sensitive to weight (Yip 1980, Duanmu 1990, Qu 2013), while some have argued that the language is *in*sensitive to weight (e.g., Feng 1995). In the present paper, I follow Qu (2013), who proposes a four-way weight distinction based on durational differences across all four phonemic tones in the language.

Table 2: Mandarin tones and syllable weight (Qu 2013, p. 71).

Tone		**Weight**	**Pitch**
$\mathbf{T_{1/2/3/4}}$ in isolation		Super-heavy	
$\mathbf{T_1}$: *mā*	'mother'		High level
$\mathbf{T_2}$: *má*	'helm'	Heavy	High rising
$\mathbf{T_4}$: *mà*	'scold'		High falling
$\mathbf{T_3}$: *mǎ*	'horse'	Light	Low falling
$\mathbf{T_0}$: *ma*	'question marker'	Weightless	Low level

Mandarin words are typically one, two, or three syllables long. The structure of stressed syllables in the language can be represented by (C)VX, where (C) may contain a glide (C^j), as in *n^{j}au* 'bird' (Duanmu 2007, p. 82), and where X can be a coda consonant or a long vowel. Contrary to popular belief, most common words in the language are disyllabic, not monosyllabic (Duanmu 2007)—most polysyllabic words are compounds. Monosyllables in Mandarin can be stressed ([ʈ͡ʂau] 'contact', T_4) or unstressed ([ʈ͡ʂə] 'be doing', T_0). In disyllabic words, we find different weight patterns: heavy-weightless (H0), where stress is initial (2a), and heavy-heavy (HH), where stress can be initial (2b) or final (2c), depending on the word (i.e., stress is lexically determined). Finally, even though the main phonetic correlate of lexical tones in Mandarin is pitch, studies have shown that different tones have distinct duration and intensity contours (Zhang et al. 2008).

(2) Stress in Mandarin disyllables (adapted from Duanmu (2007))

a. H0: [ˈpaa.pa] 'dad' — initial stress (T_4-T_0)

b. HH: [ˈtɕii.xʷaa] 'plan' — initial stress (T_4-T_4)

c. HH: [sʷuu.ˈʂɤɤ] 'dorm' — final stress (T_4-T_4)

In contrast to English (§2.1), weight in Mandarin is not sensitive to syllable shape. As a result, a Mandarin speaker acquiring English must learn that weight in the L2 is correlated with inherently longer vowels (e.g., diphthongs) as well as coda consonants, both of which are absent in Mandarin (Qu 2013)—Mandarin codas are restricted to [n], [ŋ], and the [ɚ] suffix. On the other hand, heavy syllables in both languages are correlated with longer duration (Hayes 1995, Qu 2013). Thus, even if one disagrees that phonological weight plays a role in Mandarin, the existing correlation in the language between syllable duration and the perception of stress (Qu 2013) could still be successfully transferred to English.

2.3 Stress in Portuguese

If the existence of stress and weight-sensitivity is debatable in Mandarin, it certainly is not for Portuguese (Câmara Jr. 1970, Bisol 1992, Araújo 2007, Wetzels 2007, Garcia 2017b). Indeed, Portuguese and English are surprisingly similar when we observe the stress patterns in both languages. Like English, Portuguese also differentiates stress in verbs and non-verbs: whereas stress in verbs is affected by morphological factors, stress in most non-verbs (>70%) can be accurately captured with weight-sensitivity (Garcia 2017b).

Portuguese non-verbs typically have final stress if a heavy final syllable is present (3a)—like in English, heavy syllables contain diphthongs or coda consonants. If the final syllable is light, stress is typically PU (3b). APU stress is the least common pattern in the language, and is much less predictable than PU and final stress—though weight effects are found in all three positions (Garcia 2017a). All three stress patterns are summarized in (3). (3a) and (3b) together account for over 70% of the Portuguese lexicon (Garcia 2014). (3c) captures the observation that whereas most LLL words have PU stress, most words with APU stress are LLL—the lower predictability of APU stress is represented by "⇝".

Unlike in English, word-final syllables in Portuguese are *not* assumed to be extrametrical (given that final stress is relatively common in the language). However, once extrametricality is taken into account, the overall pattern in both languages is exactly the same: stress the rightmost syllable if heavy, else stress the next syllable to the left. Another similarity between Portuguese and English is that stress in Portuguese is also correlated with duration (Major 1985, Massini-Cagliari 1992, Vogel et al. 2018). Heavy syllables are therefore longer, mirroring what we observe in both English and Mandarin.

(3) Stress and weight in Portuguese non-verbs

a. Heavy final syllable → final stress
jornál 'newspaper', *carnavál* 'carnival'

b. Light final syllable → PU stress
caválo 'horse', *planálto* 'plateau'

c. Light final and PU syllables ⇝ APU stress
prático 'practical', *tópico* 'topic'

One important difference between Portuguese and English is that Portuguese words are longer on average. As a result, the language has comparatively fewer monosyllabic and disyllabic words, which means that no strong correlation exists between word-initial position and weight-sensitivity. Indeed, excluding monosyllables, nearly 95% of all words in the Portuguese lexicon have non-initial stress (Garcia 2014).

In summary, once Portuguese speakers learning English understand that word-final syllables in the L2 are often skipped in non-verbs (extrametricality), most of what remains can be accurately captured with the weight-sensitivity patterns already present in Portuguese. Ultimately, both Mandarin and Portuguese speakers can benefit from transferring the correlation between syllable length (weight) and stress from their native languages.

3 Hypothesis

As seen above, in spite of important typological differences, both Mandarin and Portuguese display a correlation between duration, weight, and stress. Consequently, it is fair to assume that speakers of these languages can make use of such a correlation when learning English. For example, speakers of Mandarin have been shown to use the phonetic correlates present in Mandarin to signal stress in English as a second language (Zhang et al. 2008). This is certainly not surprising: one's native grammar contains a number of correlations. Importantly, whereas Mandarin speakers can be argued to only *indirectly* transfer weight (i.e., transfer duration as a phonetic correlate of prominence), given the phonotactic differences between Mandarin and English, Portuguese speakers can *directly* apply the weight patterns present in their L1 to the L2—as long as they learn that word-final syllables tend to be skipped when stress is assigned to non-verbs, a pattern that is considerably robust in English, given its frequency.

The question that arises is whether L2ers are sensitive to the weight effects in English in spite of the positional bias in the language. We saw above that word-initial stress can certainly conceal weight effects in English—a possibility that is even more likely when we consider the fact that common words in English tend to be short. Given these two correlated predictors of stress location in English, the hypothesis of the present study is given in (4).

(4) HYPOTHESIS
Whereas the grammars of native speakers of English will generalize weight-sensitivity to novel words, the grammars of second language learners will not. Weight-sensitivity will not be actively maintained in the interlanguage (transferred from the L1) because the positional bias in English is more robust, and thus conceals any weight effects present in the L2.

The hypothesis in (4) is agnostic as to whether (prosodic) parameters (e.g., Dresher and Kaye 1990, Hayes 1995, Snyder and Lillo-Martin 2011) are needed to formalize second language acquisition. Naturally, the same hypothesis stated in (4) could be phrased in terms of parameter resetting. For example, we could hypothesize that WEIGHT-SENSITIVITY, which is set to YES in learners' na-

tive grammars, would be reset to No in the absence of robust positive evidence for weight effects in the L2 input. The question, then, is whether the researcher can empirically differentiate parameter resetting from parameter deactivation in cases such as the one at hand. The Prosodic Acquisition Path Hypothesis (PAPH; Özçelik 2016), for example, assumes that parameters can never be completely deactivated—but that they can be reset. Thus, for PAPH, a predictor of stress that is not actively maintained in the L2 grammar (e.g., weight-sensitivity) would be equivalent to a parameter that has been reset to No.

As will be clear below, a core assumption of the present study is that acquisition is probabilistic, which in turn means that we should not expect categorical patterns in our data (parameters values are fundamentally categorical, even though they can be implemented probabilistically). I deliberately choose to use the theory-neutral term "predictor" instead of "rule", "parameter", or "constraint"—even in a non-probabilistic parametric approach, parameters will *predict* what we should observe in language and in language acquisition. Note, however, that the statistical approach presented in this paper can be easily mapped onto a constraint-based approach where constraints are weighted, such as MaxEnt (Hayes and Wilson 2008).

4 Methods

4.1 Experimental design

To investigate the hypothesis in (4), a forced-choice experiment was designed and implemented using Praat (Boersma and Weenink 2020). Participants were auditorily presented with trisyllabic English nonce words, each of which was recorded with APU and PU stress by a male native speaker of Canadian English with training in linguistics. For every word in the experiment, participants were asked which version of the word sounded more natural to them. The stimuli ($N = 180$) were created in R (R Core Team 2020), and were later manually inspected by a native speaker of English who is also a linguist to make sure that they were all phonotactically well-formed. The script in question generates nonce words on the basis of (i) a phonemic inventory, and (ii) a set of phonotactic rules. Crucially, the stimuli were divided into three weight profiles, namely, LLL, HLL, and LHL. Representative examples are provided in Table 3.

Table 3: Examples of stimuli used in the experiment.

LLL	HLL	LHL
[prɪ.tɑ.rək]	[nɑr.pɛ.lət]	[dɑ.sɛŋ.kəl]
[lɑ.prɛ.sən]	[prɑŋ.kɛ.mət]	[pɛ.trɑŋ.kəp]
[sɑ.pɪ.nər]	[krɪm.pɛ.dən]	[tɪ.prɛs.dəl]

As can be seen in Table 3, word-final syllables in the stimuli consisted of [CəC], a syllable shape that patterns as light (i.e., is extrametrical), as discussed above. This ensured that all stimuli

were phonotactically natural, given that open final syllables are not common in English.[3] Identical heterosyllabic consonantal sequences were manually removed to avoid OCP (Obligatory Contour Principle) effects (Leben 1973). All stimuli, as well as the order of presentation of stress patterns, were pseudo-randomized.

Participants were asked which version of each stimulus sounded more natural in English in their opinion, and were instructed to choose their answers using specific keys in the keyboard. They were also asked to rate their level of certainty using a 6-point scale. Crucially, all participants were told that only nouns were used in the experiment—they were instructed to consider all words as possible object names. Finally, the reaction time of participants' responses was also recorded.

4.2 Participants

Participants consisted of three groups, namely, native speakers of English ($n = 13$), and L2ers whose L1 is either Mandarin ($n = 24$) or Portuguese ($n = 25$). All participants were living in Canada at the time of the experiment. They were tested in a sound-attenuated booth, and wore headphones during the experiment. Overall, participants took 20-40 minutes to complete the experiment.

L2ers' proficiency level ranged from upper-intermediate to proficient, as determined by self-reported proficiency in a pre-experiment survey, TOEFL scores (if available), as well as a global accent task (Jesney 2004), in which L2ers showed highly accurate stress assignment. The control group consisted of native speakers of North American English, nearly all of whom were university students (undergraduate and graduate levels) at the time of the experiment.

The global accent task consisted of a paragraph with 138 words, which learners were asked to read aloud. The paragraph contained words with final, penultimate, and antepenultimate stress, and thus provided a representative sample with regard to stress patterns in the language. The recordings were then judged using a 6-point scale by a native speaker of English, who was provided with two training samples representing the end-points of said scale: 1 = least native-like; 6 = most native-like. The rater was linguistically trained, and was instructed to rate each learner's overall pronunciation on the basis of the training samples. Because nearly all stressed syllables in the learners' samples were sufficiently salient, the location of stress was rarely ambiguous: most learners signalled stress with longer duration and higher pitch relative to unstressed syllables. The results of the global accent task were consistent with learners' self-reported proficiency levels,

[3]It should be noted that a correlation exists between tense and lax vowels and stress location in English, as alluded to earlier in the paper. While tense-lax vowel pairs are phonetically similar in quality (e.g., /i/-/ɪ/), in most dialects of English they form a long-short opposition. Crucially, tense and lax vowels have a semi-complementary distribution: in monosyllabic words, tense vowels can occur in open and closed syllables (*bee* /bi/, *beat* /bit/), but lax vowels can only occur in closed syllables (*bit* /bɪt/ but */bɪ/). To capture the asymmetry above, tense vowels are traditionally assumed to be phonologically heavier than lax vowels—see Giegerich (2005) for a comprehensive review. The problem, however, is that once a syllable becomes unstressed, it will often undergo reduction, which in turn affects the quality of its nucleus. As a result, if a stimulus such as [pɹɪ.tɑ.ɹək] has initial stress, its second syllable will be affected, and so will the quality of the vowel in question (/ɑ/). Not reducing the syllable may either introduce secondary stress or may result in a less natural stimulus. More research is needed to determine to what extent learners of English are aware of the relationship between tense and lax vowels and weight-sensitivity and their effects on stress location in English. The stimuli in the present study include a variety of vowels, and the statistical models presented include by-word random intercepts.

which is not surprising given the literature on the reliability of self-reported measures of language proficiency (e.g., Marian et al. 2007).

4.3 Statistical analysis

The data discussed in the present paper were modeled in R (R Core Team 2020) using Bayesian hierarchical logistic regressions run with Stan (Carpenter et al. 2017), a platform for statistical modeling and high-performance statistical computation. All models included by-participant random effects and random intercepts, represented by `(1 + weight | part)` in (5), as well as by-item random intercepts, represented by `(1 | item)` in (5). The model computes the effects of weight on whether participants will choose APU stress (as opposed to PU stress). As discussed above, the three weight profiles (levels of the variable `weight`) are LLL, HLL, and LHL—LLL, underlined in (5), serves as the reference level.

(5) Modeling weight effects on English stress
`APU ~ weight + (1 + weight | part) + (1 | item)`
`weight = {LLL, HLL, LHL}`

Because Bayesian data analysis is considerably different from traditional (Frequentist) statistics, I briefly discuss some important concepts below. As will be shown, the concept of priors is of particular relevance to the present study.

4.3.1 Bayesian data analysis

Traditional (Frequentist) statistics provides the probability of observing the data at hand given a particular parameter value—assuming that the null hypothesis (H_0) is true. We can represent that probability as $p(D|\theta)$, where D represents the data, and θ represents any parameter value (e.g., $\hat{\beta}$ in regression models), and p represents the well-known p-value. In a typical logistic regression, we normally extract the coefficient values ($\hat{\beta}$), given in log-odds, which are single point estimates (i.e., the effect sizes). We also extract Confidence Intervals, which help us quantify the magnitude of effect sizes, and, of course, p-values.

In contrast to Frequentist statistics, Bayesian statistics provides the probability of a parameter value given the data, or $p(\theta|D)$—which is both more intuitive and more meaningful. Crucially, Bayesian estimation with uncertainty is *not* conditioned on a (null) hypothesis (cf. Bayes factors). By definition, then, no p-values or Confidence Intervals are provided—for well-known issues related to p-values, see Nuzzo (2014), or see Kruschke (2015) and McElreath (2016) for a comprehensive discussion. A brief introduction to Bayesian data analysis can be found in Kruschke and Liddell (2018).

When examining the output of a Bayesian model, one important advantage over outputs of Frequentist models is that, instead of a single point estimate for effect sizes, we have entire distributions of statistically credible parameter values given the data (i.e., *posterior distributions*). The model shown in (5), for example, will provide entire posterior distributions of credible effects of

`weight` *given the experimental data*—the same applies to all other parameters in the model (e.g., random effects). Because three weight profiles are included in the factor `weight`, three posterior distributions will be provided. We can then summarize each individual distribution by extracting its mean or median, and by establishing different credible intervals (see below). We can also summarize *comparisons* of distributions, which allows us to examine whether the probability of choosing APU stress in HLL words is *higher* than that in LLL or LHL words, for example. By definition, the comparison itself will result in a (posterior) distribution of credible differences between the two original distributions.

One common option to interpret posterior distributions is to extract 95% Highest Density Intervals (HDIs), which offer a straightforward interpretation: values that are closer to the peak of a given distribution are more credible given the data. Indeed, this is often how Frequentist Confidence Intervals are misinterpreted: unlike Bayesian Credible Intervals, however, Frequentist Confidence Intervals are *not* distributions.

Perhaps the most relevant advantage of Bayesian data analysis for the present study is that statistical models can incorporate different hypotheses in the form of informed priors. The intuition is simple: if we have empirical evidence that supports a particular assumption, our models should be informed with such an assumption. For example, suppose that dozens of studies point to a particular effect. A new experiment is run, and its findings contradict all previous studies. Should we change our existing assumptions? Clearly, a single study cannot be that powerful. In a Bayesian framework, we can modulate the impact of the data on the posterior distribution of effect sizes by using informed priors, which incorporate into our statistical model the existing body of knowledge of a particular effect.

It should be noted that while Bayesian methods can be used as models of the mind, for example, their use in the present paper is data-analytic. One does not need to subscribe to the idea that the human mind is best captured with Bayesian models to appreciate the advantages of such models to analyze data more generally. Simply put, the methodological question examined in this paper is not whether second language acquisition is a Bayesian process, but rather how Bayesian models can be used to analyze second language acquisition data.

Being able to use informed priors also allows us to bridge the gap between our theoretical/representational assumptions and our statistical methods. In the context of second language acquisition, for instance, it is possible for our statistical models to simulate the effects of learners' L1 grammars, and compare how different assumptions about transfer and learning can account for the data we observe. For example, if we assume that the initial state in the interlanguage is characterized by full transfer, then we necessarily assume an initial bias in favor of the patterns found in the L1. This bias can be encoded/emulated in our statistical models by adjusting the prior distributions of the parameters of interest. If, on the other hand, we assume that no concrete bias exists (i.e., no transfer at all), our models' priors can be adjusted accordingly. Naturally, these are two extreme hypotheses along a continuum of possibilities. Demonstrating how such an approach can be advantageous and informative is the methodological goal of this paper (§5).

In the next section, I will examine the hypothesis in (4) by comparing three different types of models for each of the three groups of participants. The first type will be a "naïve" model. The conclusions of such a model depend essentially on the experimental data it models. This is a naïve model insofar as it completely ignores that learners may have natural biases that stem from their L1 grammars—this model approximates what is assumed in typical Frequentist models (even though studies using such models rarely assume that learners are a blank slate). For that reason, the priors in such a model are weakly informative, and thus provide little information relative to the more informed priors of the second and third types of models described below. In other words, a naïve model simply models the data, without informed prior expectations—the posterior maxima in such a model will be similar to single point estimates in equivalent Frequentist models (the interpretation of such values, of course, will be different). Simply put, a naïve model is equivalent to assuming that no language transfer occurs in second language acquisition. Naturally, this model can be used as our statistical baseline, much like LLL is used as our weight baseline.

The second type of model will include the weight-sensitivity effects present in English: PU stress in LHL words and APU stress otherwise (LLL and HLL words). As a result, this model ignores the positional bias in the language, and assumes that weight will be effectively transferred from Mandarin and Portuguese to the L2. Finally, the third type of model incorporates the positional bias discussed above, and ignores/overrides weight-sensitivity. In other words, it simulates a learner who does not have weight-sensitivity in his or her L2 grammar (i.e., no effective transfer).[4] All three models are summarized in Table 4, where the different prior distributions (all Gaussian) simulate different biases in learners' grammars.

Table 4: Three different models to assess the effects of different weight profiles on APU stress. Effects of HLL and LHL are interpreted relative to LLL (baseline). Weight models' assumptions are based on weight-sensitivity in English. *Effective transfer* characterizes a model which assumes weight-sensitivity is effectively transferred from the L1.

	MODELS' ASSUMPTIONS AND ASSOCIATED PRIORS		
	1. Naïve	**2. Weight**	**3. Positional**
LLL	*Effect:*–	Positive	Positive
	Prior: $\mathcal{N} \sim (0, 10^6)$	$\mathcal{N} \sim (1, 1)$	$\mathcal{N} \sim (1, 1)$
HLL	*Effect:*–	Neutral	Neutral
	Prior: $\mathcal{N} \sim (0, 10^6)$	$\mathcal{N} \sim (0, 1)$	$\mathcal{N} \sim (0, 1)$
LHL	*Effect:*–	Negative	Neutral
	Prior: $\mathcal{N} \sim (0, 10^6)$	$\mathcal{N} \sim (-1, 1)$	$\mathcal{N} \sim (0, 1)$
		Effective transfer	*No effective transfer*

In Table 4, LLL (in bold) serves as the baseline (intercept) against which we can assess weight effects in HLL and LHL words. LLL has a prior that is normally distributed around a *positive* mean

[4]By "no effective transfer", what is meant is that learners transfer weight-sensitivity but do not actually use it effectively in the L2 grammar due to the lack of sufficient evidence (i.e., the confounding factor in question).

for the weight and positional models: $\hat{\beta}(\text{LLL}) = \mathcal{N} \sim (1, 1)$.[5] This merely captures the expectation that APU stress should be more likely than PU stress in LLL words, regardless of whether we assume weight or position as the main predictor of stress location.

We can see in Table 4 that the crucial difference between the weight and positional models is the predicted effect of LHL: while in a positional model we predict APU stress, in a weight model we predict PU stress. Thus, in weight models, LHL has a prior that is normally distributed around a *negative* mean: $\hat{\beta}(\text{LHL}) = \mathcal{N} \sim (-1, 1)$. Both the weight and positional models predict APU stress is equally likely in LLL and HLL words—hence their neutral effects in Table 4, which are represented with a prior that is normally distributed around zero: $\hat{\beta}(\text{LLL}) = \hat{\beta}(\text{HLL}) = \mathcal{N} \sim (0, 1)$. In other words, if stress is word-initial and ignores weight-sensitivity, HLL words should not present different response patterns relative to LLL words. Likewise, given the weight patterns in (1), APU stress in HLL and LLL words should be equally likely.

Once all three models are run, we can compare which models best fit participants' response patterns. This model comparison will be done using two information criterion methods, namely, LOO (Leave-One-Out cross validation; Vehtari et al. 2017) and WAIC (Watanabe-Akaike Information Criterion; Watanabe 2010). Both methods allow us to choose the model(s) with the best fit.

5 Results and analysis

In this section, I first present and discuss the main results of the study, which are then complemented with a brief discussion about participants' certainty levels and reaction times. As we will see, both metrics are consistent with the weight effects observed in the data. After discussing the results, I provide a statistical analysis of participants' responses that first focuses on naïve models. I then evaluate such models to show that they accurately capture participants' response patterns, and thus form a reliable statistical baseline. After establishing that naïve models are reliable and accurate, I compare such models to positional and weight models, which incorporate different assumptions about L2ers' grammars, as discussed above.

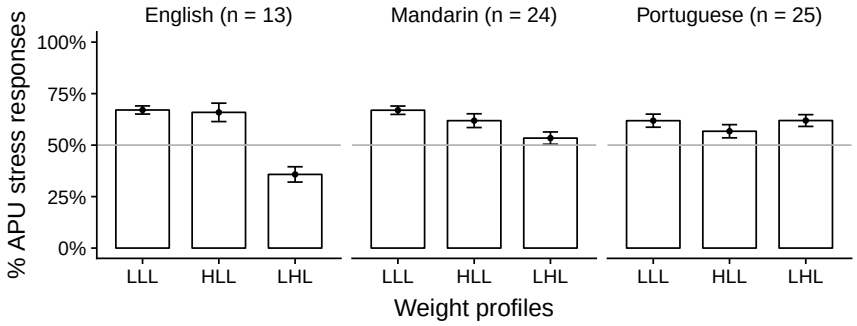
5.1 Main results

Fig. 3 plots the experimental results for all three groups. The gray horizontal line at 50% represents a categorical cut-off point: English controls' mean responses (and associated standard errors) are above 50% for LLL and HLL words, but below 50% for LHL words. That is exactly what we would predict given the weight and stress patterns in English shown in (1). This is consistent with the literature (e.g., Liberman and Prince 1977), and confirms once more that native speakers are indeed aware of weight-sensitivity in English *despite* the strong positional bias in the language.

In contrast to native speakers, second language learners' mean responses (and associated standard errors) are all *above* 50%. In other words, L2ers favor APU stress across the board. Mandarin

[5] Note that the distribution is considerably wide, given its standard deviation, which means the model is relatively free to adjust its conclusions based on the data it models.

Figure 3: Mean percentage of APU stress responses and associated standard errors by weight profile across all three groups of participants. Means and standard errors are provided.



speakers are more target-like, given that their preference for APU stress in LHL is lower relative to LLL words. On the other hand, a similar trend can be seen in HLL words, which is what we would expect given their L1. This shows that these learners do not favor stress on heavy syllables overall.

One possible metrical explanation for the HLL pattern found in the Mandarin data could be that Mandarin speakers disprefer APU stress in HLL words (relative to LLL words) because ĹLL words result is a metrically marked foot, namely, an uneven trochee (Hayes 1995). Optimal trochees in weight-sensitive languages have two light syllables, (LL), or a single heavy syllable, (H). An uneven trochee has a heavy syllable and a light syllable, (HL).[6]

The marked metrical configuration of uneven trochees, however, is an unlikely explanation for the Mandarin data, given that different studies have argued that uneven trochees are exactly the type of foot Mandarin builds (Goad et al. 2003, Goad and White 2006, Qu 2013). Thus, if Mandarin speakers were to transfer and actively apply the metrical pattern in their L1 to English, we should see a higher percentage of APU stress responses in HLL words in Fig. 3.

The overall pattern observed above is consistent with the assumption that the positional bias in English is the main driving factor behind Mandarin speakers' responses. The lack of consistent weight effects is even more apparent when we examine Portuguese speakers' response patterns: not only are APU stress responses above 50% for all weight profiles in question, these learners actually judge ĹHL and ĹLL words as equally natural in the L2.

Recall that both groups of L2ers are fluent in the L2, and had no problems assigning stress to existing words in English. Thus, given their response patterns for nonce words, it is safe to conclude that weight is *not* the main variable responsible for their stress accuracy in existing words.[7] If

[6]In weight-sensitive languages that build trochees, *moras* are counted instead of syllables (Hyman 1985). One light syllable contains one mora, whereas one heavy syllable contains two moras. Uneven trochees therefore contain three moras.

[7]Naturally, we cannot conclude that weight-sensitivity plays absolutely no role in the L2 grammar. It is not the objective of this paper to probe a null effect—which could be done if we established a Region Of Practical Equivalence (ROPE) for the posterior distributions examined. The objective is to show that position overrides weight, which does not entail that weight has zero effect.

learners' stress accuracy in existing English words were due to weight-sensitivity alone, then they should mirror native speakers when generalizing stress to nonce words. Learners' response patterns for LHL words, coupled with the clear preference for APU stress overall, are consistent with the hypothesis in (4): most of Mandarin and Portuguese learners' response patterns can be explained by the positional bias in English stress. Such a bias is so strong that, when faced with cases where weight and position are *not* correlated, namely, LHL words, learners favor position over weight. In other words, even though learners could effectively employ weight effects from their L1 in their interlanguage, the positional bias in the L2 appears to be the main factor behind learners' response patterns.

The overall results shown above can be contrasted with learners' production of existing words in the global accent task discussed above, which was practically error-free with regard to stress location. This should not be surprising, however: being able to accurately produce stress in existing words does not mean that the underlying patterns that govern stress are actually internalized by the learners—hence the importance of nonce words, where learners' generalizations can be observed.

5.2 Beyond stress: certainty and reaction time

As previously mentioned, participants were asked to rate their certainty level for each response on a 6-point scale. Fig. 4 plots participants' mean certainty (and associated error bars) for APU and PU stress responses. Note that L2ers are overall *more* certain than controls. More importantly, however, the level of certainty exhibited by controls is aligned with weight-sensitivity: they are more certain about their APU responses in LLL and HLL words, and more certain about their PU responses in LHL words. This further strengthens the evidence that native speakers' grammars are indeed sensitive to weight effects.

If we now look at certainty patterns for Mandarin and Portuguese speakers, we see that Portuguese speakers' certainty levels do not seem to be affected by weight: they are more certain about their APU responses than their PU responses *regardless* of the weight profile of the stimuli. Mandarin speakers, on the other hand, do show a slightly different pattern for LHL words, and are again more similar to controls. Overall, the certainty response patterns in Fig. 4 are consistent with participants' responses in Fig. 3—for both controls and L2ers.

Participants' responses are also consistent with their reaction times. As expected, native speakers are overall faster. More importantly, their responses are faster when stress is consistent with weight-sensitivity. In other words, they are faster when choosing APU stress in LLL and HLL words, but faster when choosing PU stress in LHL words. This once again confirms that weight-sensitivity is active in native speakers' grammars. L2ers' reaction times, on the other hand, seem to ignore weight completely: both Mandarin and Portuguese speakers are faster when choosing APU stress over PU stress, a pattern which does not seem to depend on the weight profile of the stimuli, but which does follow from the existing positional bias in the L2.

Taken together, the results discussed above are consistent: whereas native speakers' response patterns are clearly aligned with weight-sensitivity in English, L2ers' response patterns are not.

Figure 4: Participants' mean certainty levels and associated standard errors by response (APU and PU stress) and weight profile across groups: English (En), Mandarin (Ma), and Portuguese (Pt).

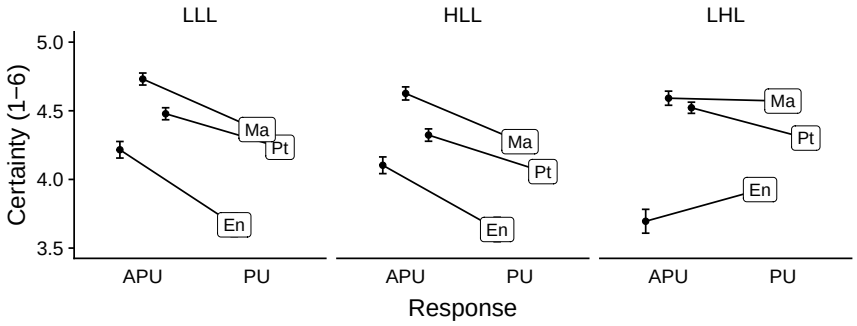
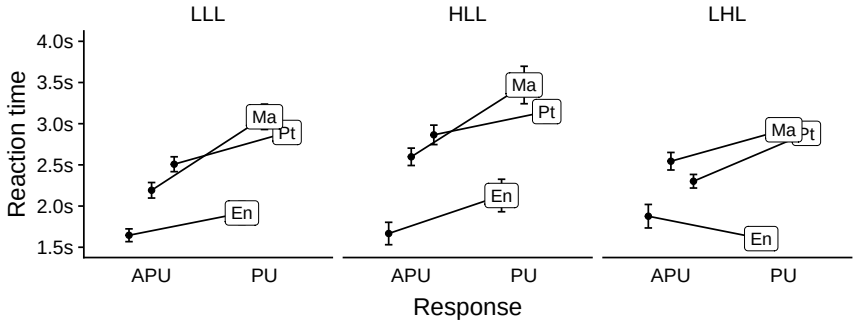


Figure 5: Participants' mean reaction time (in seconds) and associated standard errors by response (APU and PU stress) and weight profile across groups: English (En), Mandarin (Ma), and Portuguese (Pt).

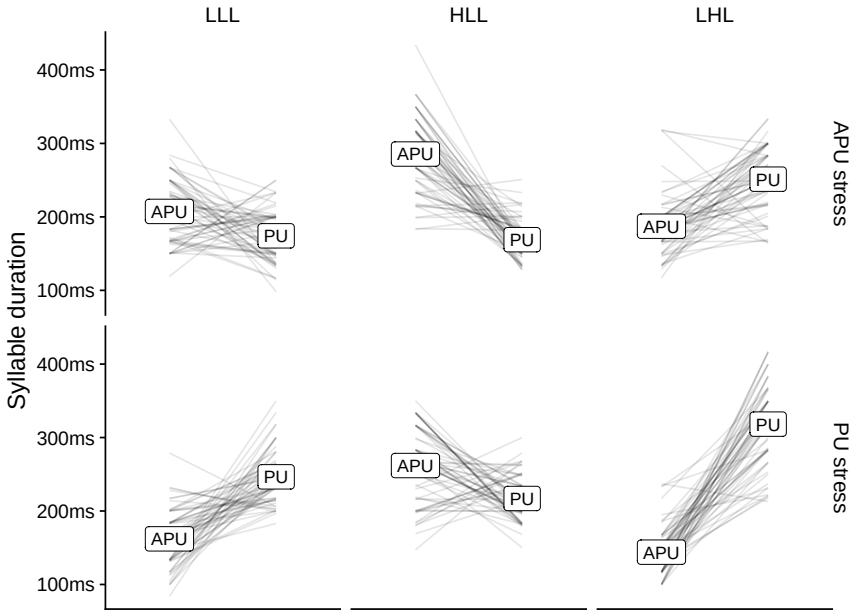


We observe that in the main results, shown in Fig. 3, but also when we examine participants' certainty levels and reaction times. These results are consistent with the hypothesis in (4) that the positional bias in English is sufficiently robust to conceal weight-sensitivity in the language.

Before we proceed, it is worth considering one possible alternative explanation for the absence of weight-sensitivity in participants' response patterns: perhaps the correlation between weight, stress, and syllable duration is not sufficiently robust in the stimuli. In such a scenario, transfer would not be captured in the data because heavy syllables in the stimuli are not long enough relative to light syllables—and not because the positional bias discussed in this paper favors APU stress in English.

Fig. 6 plots syllable duration by weight profile and stress position—each line represents one stimulus in the experiment. APU and PU labels inside the plot correspond to antepenultimate and penultimate syllables, respectively, and represent the mean duration values across all stimuli in each of the six facets of the plot. As we can see, heavy syllables are considerably longer than light syllables. For example, APU syllables are longer in HLL words than in LLL words—in LLL words, where no heavy syllable is present, stressed syllables are longer. As a result, the stimuli mirror the

Figure 6: Syllable duration of stimuli ($N = 180$) by weight profile and stress location. APU and PU labels correspond to the different syllables in the stimuli, and represent the mean duration in each group of stimuli. Heavy syllables are considerably longer than light syllables. Within LLL words, stressed syllables are longer.



exact patterns we would expect given what we know about English stress (e.g., Hayes 1995). We can therefore safely conclude that the lack of phonological weight effects in participants' response patterns cannot be attributed to a lack of acoustic cues in the stimuli.

Given participants' responses discussed in this section, it is expected that a model that incorporates in its priors the positional bias in English should have a better fit for the L2ers' data when compared to a model that includes weight-sensitivity in its priors. That is the focus of the next section, which provides the statistical analysis of the patterns in Fig. 3.

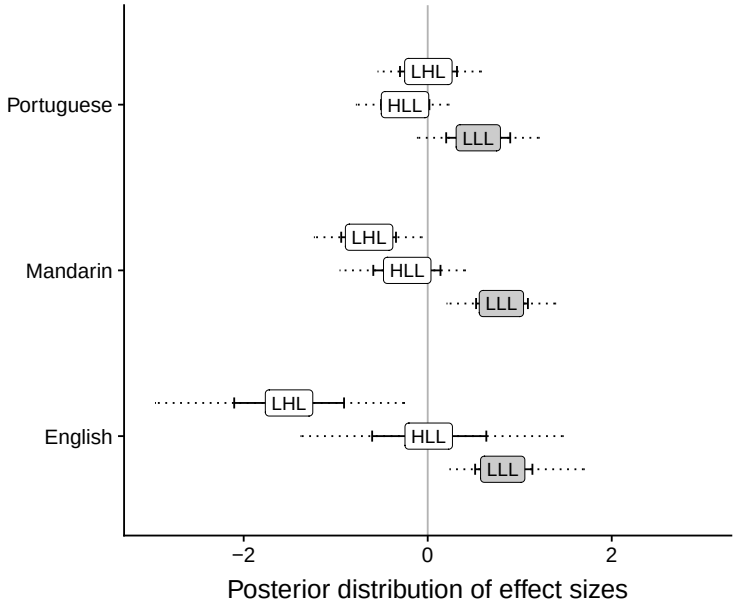
5.3 Analysis

Let us start by examining the results of our naïve models,[8] which have weakly informative priors and thus do not assume positional or weight effects on stress in English. Fig. 7 provides the posterior distributions of weight effects ($\hat{\beta}$) for all three groups of participants—as usual, effect sizes are provided in log-odds. Error bars indicate the 95% HDIs, and dotted lines represent the entire posterior distribution, i.e., 100% of all credible values. Recall that, because HDIs are actual distributions, values closer to the center of the distribution are more plausible given the data. Positive values indicate that APU stress is favored—for HLL and LHL, effects are interpreted

[8]The models reported in the present study were all diagnosed for chain convergence and Effective Sample Size (ESS; Kass et al. 1998). The Gelman-Rubin statistic ($\hat{R}$) was also checked to ensure that between- and within-chain variance were the same (Brooks and Gelman 1998).

relative to LLL (baseline), highlighted in gray in the figure. For example, if we examine LLL for native speakers, we see that the mean of its posterior distribution is $\hat{\beta} = 0.82$ (i.e., its most probable effect size given the data; see Table 5). Because effect sizes are given in log-odds, this is equivalent to stating that the odds of APU stress increase by a factor of 2.27 ($e^{|0.82|}$) in LLL words when we consider the mean effect size of the distribution. If we now examine LHL for native speakers, where the mean posterior distribution is $\hat{\beta} = -1.51$, we note that the odds of APU stress *decrease* by a factor of 4.5 relative to LLL words.

Figure 7: Posterior distributions of effect sizes ($\hat{\beta}$) for weight effects across all three groups of participants (naïve models). Error bars indicate the 95% HDIs for each distribution. Dotted lines indicate the entire posterior distributions. Effects must be interpreted relative to reference levels (LLL), highlighted in gray—LLL effects are interpreted relative to zero. Positive distributions indicate that APU stress is favored.



The naïve models' estimates in Fig. 7 are consistent with the patterns observed in Fig. 3: for native speakers, HLL is centered around zero, which means that controls' response patterns in HLL words are not credibly different from those in LLL words. Note that the distribution of LLL is entirely positive, which reflects the observation that APU stress is expected in those words. The distribution of LHL, on the other hand, is entirely *negative*, which is again expected, since APU stress should be disfavored in LHL words.

If we now inspect the Mandarin model, we can see that the posterior distributions of weight effects are closer to each other when compared to the distributions in the English model. Here, the model also reflects the empirical patterns discussed above: both HLL and LHL disfavor APU stress relative to LLL. Finally, the posterior distributions in the Portuguese model are clustered together, and show that LHL is not credibly different from LLL, the most striking pattern in the data. Like

Mandarin speakers, Portuguese speakers do not favor APU stress in HLL words (relative to LLL words), which is conveyed by the almost entirely negative HDI in HLL for both groups.

When examining Fig. 7, it is important to note that if an HDI includes zero, we still need to examine *where* in the distribution zero is located. For example, if zero is right in the middle of the HDI, we conclude that zero is indeed a very likely parameter value given the data. That is very different from an HDI where zero is in the tail of the interval.[9]

In summary, even though Mandarin speakers are more similar to controls than Portuguese speakers are when we use LLL as our baseline, this does not change the fact that both groups of L2ers still favor APU stress across all weight profiles under examination. Table 5 lists the means and 95% HDIs for all posterior distributions plotted in Fig. 7.

Table 5: Summary of weight estimates ($\hat{\beta}$) for naïve models. Means and 95% HDIs are provided for each posterior distribution—see Fig. 7.

	GROUPS		
	English	*Mandarin*	*Portuguese*
LLL	0.82	0.80	0.55
95% HDI	[0.52, 1.14]	[0.53, 1.09]	[0.20, 0.90]
HLL	0.01	-0.22	-0.25
95% HDI	[-0.60, 0.64]	[-0.59, 0.14]	[-0.51, 0.02]
LHL	-1.51	-0.64	0.01
95% HDI	[-2.10, -0.91]	[-0.94, -0.34]	[-0.30, 0.32]

5.3.1 Model adequacy

One way to evaluate the naïve models in Fig. 7 and Table 5 is to compare the models' predictive distribution (often referred to as y^{rep}) to the observed data (y). If the models are appropriate, we should not observe systematic differences between y^{rep} and y. Posterior predictive checks allow us to assess whether a given model fits the data adequately—see Gelman et al. (2014, Ch. 6).

Fig. 8 compares (a) English speakers' actual responses (bars) with (b) replicated responses ($n = 100$)—note that each replication is a distribution, not a point estimate. We can see that (a) and (b) are clustered together, which shows that the naïve model in question accurately captures speakers' behaviors. The same level of accuracy is observed in the Mandarin and Portuguese models, shown in Figs. 9 and 10, respectively.

In summary, we can safely conclude from Figs. 8-10 that the naïve models in question fit the data appropriately and thus form a reliable baseline. The question is whether positional and weight models can offer a *better* fit for the data. To answer that question, let us now compare all three

[9] In Frequentist Confidence Intervals, on the other hand, we do not care where exactly zero is within the interval: if the 95% Confidence Interval includes zero, that is sufficient to consider our result non-significant (assuming $\alpha = 0.05$).

Figure 8: Posterior predictive check for naïve model (English speakers). Mean predicted replicated responses (black dots) and associated error bars (based on 100 samples; not visible) match participants' actual response patterns (white bars).

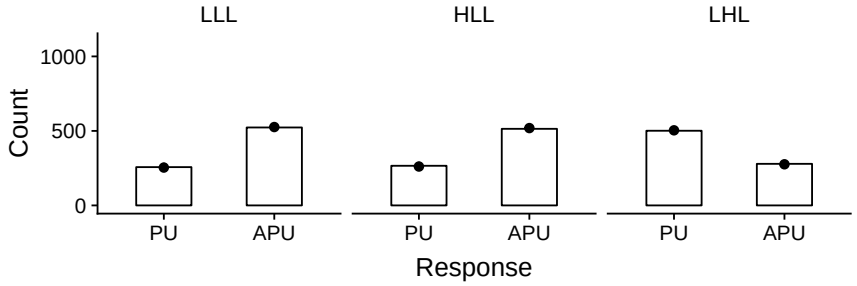
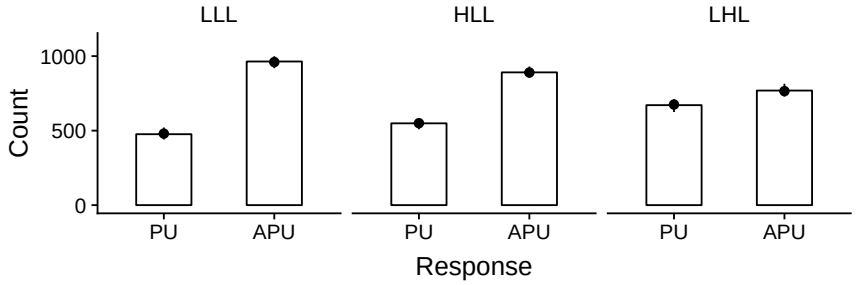


Figure 9: Posterior predictive check for naïve model (Mandarin speakers). Mean predicted replicated responses (black dots) and associated error bars (based on 100 samples; barely visible) match participants' actual response patterns (white bars).



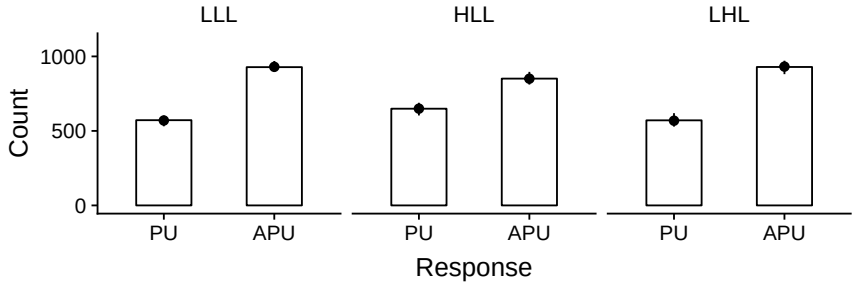
types of models.

5.3.2 Model comparison

Recall that three models were run for each group of participants. The naïve models described above can be used as a baseline against which we can compare positional models and a weight models. Fig. 11 compares all three models using two information criteria, namely, LOO and WAIC, as mentioned above. Information criteria are used to estimate the performance and accuracy of different models, an essential step of the analysis, given that multiple models can provide good fits of the data. In both methods employed, lower values indicate a better fit.

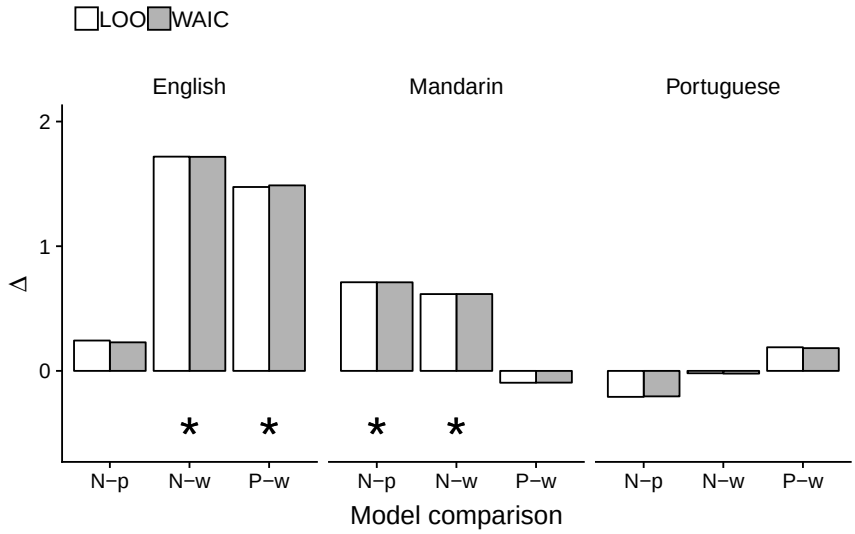
If we examine the comparisons for English models in Fig. 11, we note that no credible difference exists between a naïve model and a positional model (no asterisk above **N**(aïve)-**p**(ositional) in the figure). We do, however, observe credible differences when we compare the naïve and positional models with the weight model (**N**(aïve)-**w**(weight) and **P**(ositional)-**w**(weight), respectively): in both cases, a model that includes weight-sensitivity in its priors offers a better fit of the data—this can be seen in the positive values above N-w and P-w in Fig. 11, which indicate that the naïve and

Figure 10: Posterior predictive check for naïve model (Portuguese speakers). Mean predicted replicated responses (black dots) and associated error bars (based on 100 samples; barely visible) match participants' actual response patterns (white bars).



positional models have higher LOO and WAIC values when compared to the weight model, thus yielding the positive differences we observe.

Figure 11: Model comparison based on LOO and WAIC: N = Naïve model; P/p = Positional model; W/w = Weight model. An asterisk signals that the difference in question (**Δ**) is greater than its associated standard error. Positive differences indicate that the model in capitals (x-axis) has a *worse* fit.



Let us now examine the L2ers' models. In the Mandarin models, both N-p and N-w show a positive difference in Fig. 11, which indicates that a naïve model has a *worse* fit than both positional and weight models. The fact that N-p has a slightly higher value than N-w is consistent with the hypothesis that the positional bias in English stress is more robust than the weight patterns in the language—a trend which can also be seen in P-w, even though the differences in LOO and WAIC values between models are not greater than their associated standard errors.

Finally, the Portuguese models reveal that no credible difference in fit can be found across all three models, which is again consistent with the data discussed so far—consider, for example, the almost identical LOO and WAIC values for N-w.

In summary, a model that incorporates weight-sensitivity better fits native speakers' response patterns when compared to both a positional and a naïve model, as indicated in Fig. 11. In contrast, L2ers' models show that including weight-sensitivity in our models does not result in a better fit of response patterns found in the data.

The different fits in Fig. 11 are naturally affected by the priors assumed by each informed model. For example, if the informed priors used in the present paper had a narrower distribution around the same mean (i.e., a smaller standard deviation, which would emulate a higher degree of certainty), that could impact our comparison, potentially intensifying the differences observed.

Table 6: Model comparisons: differences in LOO and WAIC values are provided along with associated standard errors (SE). N = Naïve model; P/p = Positional model; W/w = Weight model. Asterisks indicate differences (**Δ**) which are greater than their associated SEs. Positive differences indicate that the model in capitals has a *worse* fit.

Group	Method	Comparison	Δ	SE
English	LOO	N-p	0.25	0.66
		N-w*	1.71	0.61
		P-w*	1.47	0.61
	WAIC	N-p	0.23	0.66
		N-w*	1.72	0.61
		P-w*	1.49	0.61
Mandarin	LOO	N-p*	0.71	0.52
		N-w*	0.62	0.45
		P-w	-0.10	0.53
	WAIC	N-p*	0.71	0.52
		N-w*	0.62	0.45
		P-w	-0.09	0.53
Portuguese	LOO	N-p	-0.20	0.55
		N-w	-0.01	0.52
		P-w	0.18	0.55
	WAIC	N-p	-0.20	0.55
		N-w	-0.02	0.52
		P-w	0.18	0.55

6 Discussion and conclusions

The analysis above shows that weight-sensitivity is not properly maintained in the interlanguage of Mandarin and Portuguese speakers. The empirical patterns observed for both groups of L2ers deviate from those exhibited by native speakers of English, who clearly disfavor APU stress if a

heavy penultimate syllable is present (LHL). These results are consistent with the hypothesis stated in (4), i.e., that the strong positional bias in English can conceal weight-sensitivity in the language.

It is important to note that the prevalence of position in the data examined above does *not* mean that weight is not used at all by learners (e.g., recall that Mandarin speakers do show some effects of phonological weight). What the results indicate is that, if weight-sensitivity is partially active in L2ers' grammars, it is not sufficiently strong to override the existing positional bias in the language and lead learners to the correct generalizations.

The overall empirical patterns in the data were further corroborated by participants' certainty levels and reaction times. Both metrics show that L2ers' responses were not affected by weight patterns, and clearly favored word-initial (APU) stress overall. These results are again consistent with the hypothesis in (4), and imply that even though transfer may not be input-driven, its effective maintenance in the interlanguage is at least in part input-dependent.

The statistical analysis proposed in this paper had two objectives. First, to further strengthen the argument that leaners use position rather than weight to generalize stress in English, unlike native speakers. Second, to demonstrate how Bayesian data analysis can help us model our data with priors which are informed by different theoretical or representational assumptions. To my knowledge, this is the first paper that employs such a statistical approach in the context of stress acquisition in second language research (see Wilson and Davidson 2013 on the acquisition of phonotactics).

The comparison between naïve and informed models in §5 simulated different assumptions about L2ers' grammars. The baseline assumption (naïve models) essentially mirrored what is assumed by traditional (Frequentist) statistical models. We know that such models are not linguistically appropriate, insofar as L2ers never "start from zero", but they are statistically relevant in the model comparison in question.

The model comparison in §5 demonstrated that, for native speakers, a statistical model that simulates weight-sensitivity *a priori* has a better fit than a positional model and a naïve model. This provides further statistical evidence that native speakers are aware of weight-sensitivity *despite* its high correlation with the robust positional bias in English. The model comparison for L2ers, on the other hand, shows the opposite pattern. For Portuguese speakers, no credible difference in fit was found between the models examined. For Mandarin speakers, even though the weight and positional models had no difference greater than one standard error, the positional model did show lower LOO and WAIC values. In summary, the model comparison above is consistent with the hypothesis in (4), which stated that L2ers would fail to generalize the weight-sensitivity in English in spite of their L1s—unlike native speakers.

We have seen above that when two variables are correlated in the L2 (positional bias and weight), second language learners may fail to employ one of the variables (weight) even when such a variable can be transferred from the L1. One important question then is whether and how learners realize that they have the wrong grammar. Presumably, learners can adjust their grammars when their lexica are large enough that longer words become more representative. Simply put, learners need

more evidence that simply assuming word-initial stress will not be sufficient, and such evidence can only be found in LHL words or words that are longer than the stress domain in the L2. The question is how many words it would take for weight-sensitivity to override the existing positional bias. The Tolerance Principle (Yang 2016) can help us answer that question.

(6) Tolerance Principle (Yang 2016)
A rule R, defined over a set of N lexical items, will only be productive if e, the number of items not supporting R, does not exceed θ_N.

$$e \leq \theta_N = \frac{N}{lnN}$$

We could hypothesize that $R_{\mathcal{A}}$ is defined as "stress falls on heavy syllables". $R_{\mathcal{A}}$ is transferred from the L1, but is not granted productive status in the L2 because it is opaque in the vast majority of items (i.e., the learner does not find positive evidence in support of $R_{\mathcal{A}}$ in the input). An alternative $R_{\mathcal{B}}$, namely, "stress is word-initial" is then considered. In such case, $R_{\mathcal{B}}$ would be productive until the number of items not supporting $R_{\mathcal{B}}$ reached θ, our threshold.

To illustrate the Tolerance Principle, let us assume a learner's lexicon that contains 1,000 words, which would give us $\theta_{1000} = 145$. In other words, 14.5% of said learner's lexicon would have to contain words with non-initial stress which were consistent with weight-sensitivity—recall that stress in English is not completely predictable, so even when position fails as a predictor we cannot be sure that weight will not fail as well. However, we already know that the vast majority of English words have initial stress (primary or secondary). Therefore, it is plausible to assume that learners will only be able to conclude that $R_{\mathcal{B}}$ is unproductive after their lexica are considerably large, at which point $R_{\mathcal{A}}$ may be "promoted" or activated in the interlanguage.

The assumption that more data will lead learners' grammars to adjust to weight-sensitivity is perfectly compatible with the probabilistic approach presented in this paper: as the data change (i.e., larger lexica emerge), so do parameter values (i.e., the posterior distributions). Importantly, the impact of new data decreases as speakers' lexica increase: the effect that 100 words that follow a certain pattern have on a given lexicon is inversely proportional to the size of said lexicon.

The relationship between the size of learners' lexica and their grammars is an important one. Learners' lexica are, by definition, a subset of the lexicon of a typical native speaker. Furthermore, the present paper does not assume that learners have access to the full corpus of L2 lexical items. Such an assumption would be unrealistic, and is certainly not necessary for the present analysis. Quite the contrary: the smaller the lexica, the more biased the stress pattern input will be towards position (not weight)—recall that the vast majority of common words are short, and have initial stress. Thus, the more constrained one assumes learners' lexica are, the more correlated position and weight will be, and the more consistent the results discussed above will become.

A question for future research is whether L2ers acquire weight when it is not pervasively correlated with another variable in the L2. Interestingly, as we saw above, Portuguese is a language where no strong correlation exists between word-initial stress and weight-sensitivity. In a recent study,

Garcia (2016) shows that English speakers learning Portuguese clearly generalize weight-sensitivity to nonce words in the language—a pattern which can be observed even among intermediate learners in the sample. Similar results are reported in Face (2005), where native speakers of English are shown to generalize weight-sensitivity to nonce words in Spanish. Such findings make sense: when no correlated variable exists, weight-sensitivity is not concealed and can thus be transferred from the L1 grammar and, crucially, actively maintained in the L2 grammar.

Finally, I have shown that the use of Bayesian data analysis in second language research can be valuable not only because it offers more intuitive, meaningful, and comprehensive answers to our questions, but also because it allows us to approximate theory and method by employing priors which are theoretically informed. In the case in question, such priors allowed us to arbitrate between different linguistically-informed models, which in turn provided a more thorough examination of the hypothesis under consideration.

References

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

Archibald, J. (1993). *Language learnability and L2 phonology: the acquisition of metrical parameters*, volume 19. Springer, New York.

Beckman, M. E. (1986). *Stress and non-stress accent*. Walter de Gruyter, Berlin.

Bisol, L. (1992). O acento e o pé métrico binário. *Cadernos de Estudos Linguísticos*, 22:69–80.

Boersma, P. and Weenink, D. (2020). Praat: doing phonetics by computer [Computer program].

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, 14(2-3):109–149.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Broselow, E. and Park, H.-B. (1995). Mora conservation in second language prosody. In Archibald, J., editor, *Phonological acquisition and phonological theory*, pages 151–168. Erlbaum, Hillsdale, NJ.

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

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

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

CMUdict (2014). Carnegie Mellon Pronouncing Dictionary.

Cutler, A. (2012). *Native listening: language experience and the recognition of spoken words*. MIT Press, Cambridge, MA.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, 2(3-4):133–142.

Dresher, B. E. and Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, 34(2):137–195.

Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages*. PhD thesis, Massachusetts Institute of Technology.

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University Press.

Face, T. L. (2005). Syllable weight and the perception of Spanish stress placement by second language learners. *Journal of Language and Learning*, 3(1):90–103.

Feng, S. (1995). *Prosodic structure and prosodically constrained syntax in Chinese*. PhD thesis, University of Pennsylvania.

Garcia, G. D. (2014). *Portuguese Stress Lexicon*. Comprehensive list of non-verbs in Portuguese. Available at http://guilhermegarcia.github.io/psl.html.

Garcia, G. D. (2016). Extrametricality and second language acquisition. In Hansson, O. G., Farris-Trimble, A., McMullin, K., and Pulleyblank, D., editors, *Proceedings of the Annual Meetings on Phonology*, pages 1–12.

Garcia, G. D. (2017a). *Weight effects on stress: lexicon and grammar*. PhD thesis, McGill University.

Garcia, G. D. (2017b). Weight gradience and stress in Portuguese. *Phonology*, 34(1):41–79.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC, Boca Raton, 3rd edition.

Giegerich, H. J. (2005). *English phonology: an introduction*. Cambridge University Press, New York, 8th edition. First published in 1992.

Goad, H. and White, L. (2006). Ultimate attainment in interlanguage grammars: A prosodic approach. *Second Language Research*, 22(3):243–268.

Goad, H., White, L., and Steele, J. (2003). Missing Inflection in L2 Acquisition: Defective Syntax or L1-Constrained Prosodic Representations? *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 48(3-4):243–263.

Gordon, M. (2006). *Syllable weight: phonetics, phonology, typology*. Routledge, New York.

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

Hayes, B. (1982). Extrametricality and English stress. *Linguistic Inquiry*, 13(2):227–276.

Hayes, B. (1995). *Metrical stress theory: principles and case studies*. University of Chicago Press, Chicago.

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

Hyman, L. (1977). On the nature of linguistic stress. In Hyman, L., editor, *Studies in Stress and Accent*, pages 37–82. Los Angeles: University of Southern California, Los Angeles.

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

Jesney, K. (2004). The use of global foreign accent rating in studies of L2 acquisition. *Calgary, AB: University of Calgary Language Research Centre Reports*, pages 1–44.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100.

Kenstowicz, M. (1994). *Phonology in generative grammar*. Blackwell, Oxford.

Kijak, A. (2006). Native intuitions of speakers of a lexical accent system in L2 acquisition of stress. The case of Russian learners of Polish. In *Speech Prosody, 3rd International Conference*, Dresden, Germany.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan*. Academic Press, London, 2nd edition.

Kruschke, J. K. and Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1):155–177.

Leben, W. R. (1973). *Suprasegmental phonology*. PhD thesis, Massachusetts Institute of Technology.

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

Lieberman, P. (1960). Some acoustic correlates of word stress in american english. *The Journal of the Acoustical Society of America*, 32(4):451–454.

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

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4):940–967.

Massini-Cagliari, G. (1992). *Acento e ritmo*. Editora Contexto, São Paulo.

McElreath, R. (2016). *Statistical rethinking: a Bayesian course with examples in R and Stan*, volume 122. Chapman & Hall/CRC, Boca Raton.

Moore-Cantwell, C. (2016). *The representation of probabilistic phonological patterns: neurological, behavioral, and computational evidence from the English stress system*. PhD thesis, University of Massachusetts.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature News*, 506(7487):150.

Özçelik, Ö. (2014). Prosodic faithfulness to foot edges: the case of Turkish stress. *Phonology*, 31(2):229–269.

Özçelik, Ö. (2016). The prosodic acquisition path hypothesis: Towards explaining variability in l2 acquisition of phonology. *Glossa*, 1(1):1.

Pater, J. (1997). Metrical parameter missetting in second language acquisition. *Language Aquisition and Language Disorders*, 16:235–262.

Qu, C. (2013). *Representation and acquisition of the tonal system of Mandarin Chinese*. PhD thesis, McGill University.

R Core Team (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ryan, K. M. (2019). *Prosodic weight: categories and continua*. Oxford University Press, Oxford.

Schwartz, B. D. and Sprouse, R. (1996). L2 cognitive states and the full transfer/full access model. *Second Language Research*, 12(1):40–72.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Selkirk, E. (1980). The role of prosodic categories in English word stress. *Linguistic Inquiry*, 11(3):563–605.

Snyder, W. and Lillo-Martin, D. (2011). Principles and parameters theory and language acquisition. In Hogan, P., editor, *The Cambridge encyclopedia of language sciences*, pages 670–673. Cambridge University Press, Cambridge, UK.

Tremblay, A. (2008). Is second language lexical access prosodically constrained? processing of word stress by French Canadian second language learners of English. *Applied Psycholinguistics*, 29(04):553–584.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432.

Vogel, I., Athanasopoulou, A., and Guzzo, N. B. (2018). Timing properties of (Brazilian) Portuguese and (European) Spanish. In Repetti, L. and Ordóñez, F., editors, *Romance Languages and Linguistic Theory 14. Selected papers from the 46th Linguistic Symposium on Romance Languages (LSRL)*, pages 325–340.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

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

White, L. (1989). *Universal grammar and second language acquisition*, volume 1. John Benjamins Publishing, Amsterdam.

Wilson, C. and Davidson, L. (2013). Bayesian analysis of non-native cluster production. In *Proceedings of the Annual Meeting of the North East Linguistic Society*, volume 40.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press, Cambridge, MA.

Yip, M. (1980). *The tonal phonology of Chinese*. PhD thesis, Massachusetts Institute of Technology.

Zhang, Y., Nissen, S. L., and Francis, A. L. (2008). Acoustic characteristics of english lexical stress produced by native mandarin speakers. *The Journal of the Acoustical Society of America*, 123(6):4498–4513.