Big data mining and comparative analyses across lexica on the relationship between syllable complexity and word stress

Amanda POST DA SILVEIRA
https://orcid.org/0000-0002-9451-7005; psyphon.ap@gmail.com
Federal University of Jataí (UFJ) (BRAZIL)

Suggested citation: POST, A., (2023), Big data mining and comparative analyses across lexica on the relationship between syllable complexity and word stress, *Langue(s) & Parole,* 8, 149-177, https://doi.org/10.5565/rev/languesparole.132

Abstract

For about four decades, phonological theories have claimed that word stress assignment depends on the word's syllabic structure complexity in relation to syllabic position. This study analyzes the syllabic structure implications for word stress in three languages with weight-sensitive lexical stress, namely Brazilian Portuguese, British English, and American English. After creating three corpora and applying Random Forest modeling, syllabic structure distributions for word stress were found to be bound to stress pattern and word length in number of syllables. To account for these observations, models of word naming must be extended with aspects of word stress.

Keywords: corpus linguistics; syllable structure; word stress; big data; Random Decision Forests

Résumé

Pendant environ quatre décennies, les théories phonologiques ont affirmé que l'affectation de l'accentuation des mots dépend de la complexité de la structure syllabique du mot par rapport à la position syllabique. Cette étude analyse les implications de la structure syllabique pour l'accentuation des mots dans trois langues à accent lexical sensible au poids, à savoir le portugais brésilien, l'anglais britannique et l'anglais américain. Après avoir créé trois corpus et appliqué la modélisation Random Forest, les distributions de structure syllabique pour l'accentuation des mots se sont avérées liées au modèle d'accentuation et à la longueur du mot en nombre de syllabes. Pour tenir compte de ces observations, les modèles de dénomination des mots doivent être étendus aux aspects de l'accentuation des mots.

Mots-clés : linguistique de corpus ; structure syllabique ; accentuation des mots ; big data ; Forêts de décision aléatoire

Resumen

Desde hace unas cuatro décadas, las teorías fonológicas han sostenido que la asignación del acento depende de la complejidad de la estructura silábica de la palabra en relación con la posición silábica. Este estudio analiza las implicaciones de la estructura silábica para la acentuación de palabras en tres idiomas en los que la posición del acento viene condicionada por el peso silábico, a saber, portugués brasileño, inglés británico e inglés americano. Después de crear tres corpus y aplicar modelos de bosques de árboles de decisión (*random forest*), se observó que las distribuciones de la estructura silábica en relación con el acento estaban ligadas al patrón acentual y a la longitud de la palabra en número de sílabas. Para dar cuenta de estas observaciones, los modelos de denominación de palabras deben ampliarse con aspectos relacionados con el acento léxico.

Palabras clave: lingüística de corpus; estructura silábica; acento léxico; datos masivos; bosques de árboles de decisión; Random forest

Resum

Durant unes quatre dècades les teories fonològiques han postulat que l'assignació de l'accent depèn de la complexitat de l'estructura sil·làbica de la paraula en relació amb la posició sil·làbica. Aquest estudi analitza les implicacions de l'estructura sil·làbica pel que fa a l'assignació de l'accent en tres llengües en què la posició de l'accent en la paraula ve condicionada pel pes de la síl·laba: portuguès brasiler, anglès britànic i anglès americà. Després de crear tres corpus i d'aplicar models de boscos d'arbres de decisió (*random forest*), es va observar que les distribucions de l'estructura sil·làbica en relació a l'accent estaven lligades al patró accentual i la longitud de la paraula en nombre de síl·labes. Per explicar aquestes observacions els models de denominació de paraules s'han d'ampliar amb aspectes relacionat amb l'accent.

Paraules clau: lingüística de corpus; estructura sil·làbica; accent lèxic; dades massives; boscos d'arbres de decisió; Random forest

Introduction

How acoustic properties of vowels and suprasegmental acoustic features affect word stress assignment has been increasingly debated and theoretically explored in phonetic and psycholinguistic studies (e.g., Cutler, 1986; Cooper *et al.*, 2002; van Heuven, van Leyden 1996; Braun *et al.*, 2014; Post da Silveira *et al.*, 2014, 2015). At the same time, relatively few empirical studies have considered how the segment sequence within a syllable and the syllabic distributions in words of different lengths affect word stress assignment. Although many studies have been dedicated to this topic in theoretical linguistics,

available models are controversial and lack systematic empirical testing. For instance, studies have sometimes used restricted samples of convenience that meet proposed theoretical generalizations, but leave the empirical coverage of the addressed phenomena unexplored and unresolved (Domahs *et al.*, 2014). This paper is a revised and modified version of Post da Silveira (2016), in which we used corpora of three languages hypothesized to be "weight-sensitive" to word stress, and then subjected these corpora to a 'big data' analysis to investigate the relationship between syllabic structure distributions and word stress.

Phonological claims about weight sensitivity

Sequences of speech sounds in spoken words are considered to be highly language specific. The distribution of sounds (e.g., vowels and consonants) within syllables and words, including the abstract allocation of vowels and consonants to syllables, is called *phonotactics* (Vitevitch *et al.*, 1997). Depending on their syllabic structure, syllables vary in 'weight': They can be light, heavy, and even superheavy given language-specific requirements (e.g., Hayes, 1981; Hyman, 1985; Kager, 1989). Formal phonological approaches consider the initial consonants of syllables as irrelevant to weight. In some languages, there is an opposition between short and long vowels, making structures such as CV and V light compared to VV and VC, which are considered heavy. Sequences such as VVC and VCC are called superheavy syllables in languages that call for the distinction heavy versus superheavy. Syllable weight is considered an important factor in word stress assignment in languages that are referred to as weight or quantity sensitive stress systems. The notion of syllable quantity comes from the counting of syllabic moras - which are metrical units of time. If a syllable contains a long vowel or a diphthong (VV) or a short vowel followed by a consonant in coda (VC), it is assumed to have two moras. In terms of syllable quantity, a syllable is heavy when it has more than one mora (e.g., Kiparsky 1982; Hayes, 1995; Trommelen, Zonneveld 1999).

In weight or quantity sensitive languages, heavy syllables generally attract stress, and heavy syllable stress attraction can depend on their position within a word. For instance, in the English word *employee,* the last syllable attracts stress, because it constitutes a bimoraic, thus, heavy VV syllable (Hayes, 1982; Jensen, 1993; Hammond, 1999). Light syllables are stressed

only according to their position in the word. For instance, the Portuguese word *cabelo* is stressed on the pre-final default syllable of Portuguese, because, given the structure of the word (CVCVCV), there is no syllabic complexity motivation for word stress attraction to any of the three syllables other than based on syllabic position (Mattoso Câmara Jr., 1975).

In Brazilian Portuguese, Mattoso Câmara Jr. (1969) claimed that open (light) syllables, such as (V) and (CV), are the most frequent. Close (heavy) syllables, such as (VC) and (CVC) or (VV / CVV) (diphthongs) are the least frequent, and few consonants are accepted in coda position. Weight-sensitivity is supposed to be conditioned by syllabic position: Heavy syllables are stressed if placed in antepenultimate or final position, such as in *LÂMpada* [light bulb] and *chaPÉU* [hat] (Bisol, 2004)[1].

In English, there is no theoretical agreement on the role syllabic weight plays in determining word stress. Most views claim that English is guided both by stress rules and by weight sensitivity. Heavy syllables are (VV) (long vowels) and VC (vowels + codas), such as in *senSAtion* and *perFECtion*. In these views, the default stress position is antepenultimate, and syllable weight attracts stress to the pre-final syllable (e.g., Chomsky, Halle, 1968; Liberman, Prince, 1977; Hayes, 1982). Other patterns are defined as having lexical stress, meaning that word stress in the language is not assigned based on phonological rules or it is not always on the same syllable – such as in French and Polish (Kijak, 2009). The only way to predict lexical stress is based on word stress pattern frequencies in the lexicon. There is only one theoretical approach in which it is stated that the default stress of English is pre-final for monomorphemic words, while all other cases are explained by lexical stress (Kiparsky, 1982, 1985).

In sum, the syllabic position of word stress must be taken into account when specifying the relationship between syllable complexity and word stress. Importantly, syllable patterns are assumed to relate to word stress patterns in a bottom-up fashion, i.e. from syllable structure to word stress. In the next section, we will review phonological claims for weight sensitivity. This will allow us to check the assumption that syllable complexity motivates word stress and indicate that the reverse may be true.

[1] Antepenultimate and final stress patterns are considered as exceptional in Portuguese and have diacritics on the stressed vowel to mark stress in orthography.

Method

The use of corpora of different languages

Post da Silveira (2016; submitted) used three corpora for this study, involving Brazilian Portuguese (BP), British (UK) English (BE), and American (US) English (AE). They have in common the same variables, which were necessary for the performance of the statistics for the aims of this study. The corpora were briefly described as follows.

(a) Brazilian Portuguese Corpus with Phonetics and Psycholinguistic Factors

The Brazilian Portuguese Corpus with Phonetics and Psycholinguistic Factors (Post da Silveira *et al.*, 2018; Post da Silveira *et al.*, submitted) provides information on the number of graphemes and phonemes in a word, as well as on its number of syllables, word stress pattern, syllable pattern distribution per Syllabic Position, and the word's frequency of occurrence in the corpus. The included words vary in length from 1 to 10 syllables. In total, the corpus contains 123,826 lexical transcriptions (lexical types) and 228,766,402 tokens. For instance, in a search for the word *abacaxi* [pineapple], the corpus provides the following information: It has 7 graphemes, 7 phonemes, 4 syllables, stress is assigned on the last syllable, its SAMPA transcription corresponds to /a.ba.ka'Si/, its syllable structure transcription corresponds to V-CV-CV-CV, and its lexical frequency in the corpus is 874 occurrences.

(b) British English Corpus with Phonetics and Psycholinguistic Factors

The British English Corpus with Phonetics and Psycholinguistic Factors (Post da Silveira *et al.*, submitted), exactly as the Brazilian Portuguese Corpus with Phonetics and Psycholinguistic Factors, provides information on a word's number of graphemes and phonemes, and on its number of syllables, word stress pattern, syllable pattern distribution per syllabic position, and frequency of occurrence in the corpus. Included words vary in length from 1 to 7 syllables. In total, the corpus contains 160,596 lexical transcriptions (lexical types) and a computation of 209,445,212 token frequencies. For instance, in a search for the word *marmalade,* the corpus provides the following information: It has 9 graphemes, 7 phonemes, 3 syllables, stress is assigned on the first syllable, its SAMPA transcription corresponds to / 'mA-m@-leId/, its syllable

structure transcription corresponds to CV-CV-CVVC, and its lexical frequency in the corpus is 44 occurrences.

(c) American English Corpus with Phonetics and Psycholinguistic Factors

The American English Corpus with Phonetics and Psycholinguistic Factors provides information on a word's number of graphemes and phonemes, number of syllables, word stress pattern, syllable pattern distribution per Syllabic Position, and frequency of occurrence in the corpus. Included words vary in length from 1 to 8 syllables. In total, the corpus contains 40,411 types and it comprises 436,166,899 tokens. For instance, in a search for the word *marmalade,* the corpus provides the following information: It has 9 graphemes, 8 phonemes, 3 syllables, stress is assigned on the first syllable, its SAMPA transcription corresponds to /'mAr-m@-led/, the syllable structure transcription corresponds to CVC-CV-CVC, and its lexical frequency in the corpus is 236 tokens.

Data Analysis

To organize our three language corpora, we chose specific factors, such as the number of syllables in a lexical item, its number of graphemes, its number of phonemes, and its SAMPA transcription. The dimensions for analysis were the same as those that structured the corpora, but we selected large data samples from the corpora as a whole dependent on theoretical considerations. The semi-factorial aspects of our analyses allowed an evaluation of novel variables that are hypothesized to be a part of the word recognition process (Balota *et al.*, 2012), but that cannot be tested in experiments with factorial designs, for instance, because they are multi-level factors, such as Syllable Pattern and Word Length (see Table 1 below). A big data approach also allowed the comparison of the prediction power of different statistical models (Balota *et al.*, 2012), such as Multiple Regression Analysis, Confidence Inference Trees, and Random Forest.

Results

Inventory of syllable structure patterns per language

For comparison purposes, we started by extracting the 20 most frequent Syllable patterns of the disyllabic and trisyllabic words from the three corpora we created. All variables we considered for generating the inventory lists from each corpus are listed in Table 1.

Response variable	Levels	Description
Log token frequencies	Logarithmic scale	Syllable patterns
Predictor variables	**Levels**	**Description**
Language Group	3	Brazilian Portuguese, British English, American English
Word Length	2	2- or 3-syllable words
Syllable Pattern	20	CC/ CCC/ CCCVC/ CCD/ CCV/ CCVC/ CCVCC/ CD/ CDC/ CDCC/ CDG/ CV/ CVC/ CVCC/ CVCCC/ D/ DC/ V/ VC/ VCC
Stress Status	2	Stressed or unstressed
Syllabic Position	3	1st , 2nd or 3rd syllable
Stress Pattern	3	1st, 2nd or 3rd syllable stress

Note: "D" means "diphthong"; "G" means "glide".

Table 1: Variables used to generate the inventory of syllable pattern distributions from the three corpora

	Brazilian Portuguese Disyllabic Words				British English Disyllabic Words				American English Disyllabic Words			
	Stressed syllables		Unstressed syllables		Stressed Syllables		Unstressed syllables		Stressed syllables		Unstressed syllables	
Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.	Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.	Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.
1	CV	13.71	CV	14.31	1	D	15.29	CV	11.04	1	CDC	18.05	CV	14.15
2	CD	13.21	CVC	12.77	2	CV	12.93	CVC	10.94	2	CVC	13.91	VC	13.50
3	CVC	13.20	V	12.46	3	V	12.01	VC	10.14	3	CV	13.19	CVC	13.30
4	V	11.72	CCV	12.30	4	CVC	11.72	D	8.89	4	CVCC	13.07	V	13.18
5	VC	11.40	CD	12.07	5	VC	11.71	CD	8.75	5	CCVC	12.43	DC	10.68
6	D	11.30	VC	11.70	6	CCV	11.02	CDC	8.40	6	CD	11.61	CC	6.91
7	CCV	11.14	D	10.29	7	CD	10.78	CC	5.40	7	VC	6.84	CVCC	6.61
8	CDC	10.97	CCVC	10.21	8	DC	10.68	V	5.35	8	V	6.82	CCVC	6.40
9	CCD	10.54	CCD	9.01	9	CCD	4.92	CVCC	5.18	9	CCV	6.60	CCC	6.22
10	CCVC	10.13	CDC	5.14	10	CDC	4.76	CCVC	5.05	10	DC	6.32	CCVCC	6.13
11	CC		CDG	4.94	11	VCC	4.60	CCC	5.01	11	CDCC	5.95	CDC	6.10
12	CCC		CC		12	CC		CCVCC	4.61	12	CCVCC	5.90	CCV	5.84
13	CCCVC		CCC		13	CCC		CCV	4.38	13	CVCCC	5.84	D	5.36
14	CCVCC		CCCVC		14	CCCVC		DC	3.74	14	CC		CD	5.27

15	CDCC	CCVCC	15	CCVC	CCCVC	15	CCC	VCC	4.60
16	CDG	CDCC	16	CCVCC	CCD	16	CCCVC	CCCVC	
17	CVCC	CVCC	17	CDCC	CDCC	17	CCD	CCD	
18	CVCCC	CVCCC	18	CDG	CDG	18	CDG	CDCC	
19	DC	DC	19	CVCC	CVCCC	19	D	CDG	
20	VCC	VCC	20	CVCCC	VCC	20	VCC	CVCCC	

Note: "D" means "diphthong"; "G" means "glide".

Table 2: Syllable pattern frequency distributions in stressed and unstressed syllables of disyllabic words from the three corpora

	Brazilian Portuguese Trisyllabic Words					British English Trisyllabic Words					American English Trisyllabic Words			
	Stressed syllables		Unstressed syllables			Stressed syllables		Unstressed syllables			Stressed syllables		Unstressed syllables	
Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.	Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.	Rank freq.	Phonot.	Log freq.	Phonot.	Log freq.
1	CV	19.42	CV	40.81	1	CDC	20.55	CVC	29.03	1	CVC	18.26	CV	38.01
2	CVC	18.48	CVC	34.04	2	CVC	14.18	CVCC	27.06	2	CV	18.12	CVC	35.60
3	CD	17.64	CCV	33.93	3	CVCC	13.90	VCC	24.67	3	CVCC	16.72	VC	34.87
4	V	16.52	V	33.51	4	CCVCC	11.82	CDC	21.40	4	CCV	16.47	V	33.91
5	D	15.57	VC	28.45	5	VC	9.09	CCVC	19.72	5	CCVC	15.90	CDC	32.07
6	CCVC	14.96	CD	26.75	6	CCVC	9.07	CDCC	18.86	6	CDC	15.72	CC	23.22
7	VC	14.58	D	23.78	7	DC	7.88	CCC	17.42	7	V	12.21	CVCC	22.38
8	CCD	14.02	CCVC	22.17	8	CDCC	7.81	VC	15.96	8	VC	11.70	CD	20.98
9	CCV	12.01	CCD	17.23	9	CVCCC	7.73	CCVCC	15.40	9	CD	10.81	CCV	20.66
10	CDC	9.83	CDC	12.49	10	CDG	6.03	DC	12.72	10	CCVCC	10.61	CCVC	13.88
11	DC	8.73	VCC	5.13	11	VCC	4.66	CVCCC	12.14	11	D	5.48	CCC	11.60
12	CC		CVCC	3.40	12	CCV	4.27	CCV	6.01	12	CCCVC	5.46	VCC	11.00

13	CCC	DC	1.30	13	CD	3.60	CCCVC	4.56	13	VCC	4.52	D	9.64
14	CCCVC	CC		14	D	3.30	D	4.03	14	DC	4.52	DC	9.14
15	CCVCC	CCC		15	CC		CDG	4.01	15	CC		CDCC	5.48
16	CDCC	CCCVC		16	CCC		CD	3.88	16	CCC		CCCVC	
17	CDG	CCVCC		17	CCCVC		V	3.67	17	CCD		CCD	
18	CVCC	CDCC		18	CCD		CC		18	CDCC		CCVCC	
19	CVCCC	CDG		19	CV		CCD		19	CDG		CDG	
20	VCC	CVCCC		20	V		CV		20	CVCCC		CVCCC	

Note: “D” means “diphthong”; “G” means “glide”.

Table 3: Stress pattern frequency distributions in stressed and unstressed syllables of trisyllabic words from the three corpora

A first observation based on Tables 2 and 3 is that the 20 most frequent syllable patterns that emerged from the Brazilian Portuguese corpus mining are less complex than those mined from the British and American English corpora. The syllable patterns that emerge in disyllabic and trisyllabic words also vary in complexity. Syllable patterns become more complex as the number of syllables increase. Across the three languages, more variation in syllable structure complexity is observed in unstressed syllables than in stressed syllables. Syllable structure complexity also appears to be more variable in British and American English than in Brazilian Portuguese, because both varieties of English have more syllable patterns in stressed and unstressed syllables than Brazilian Portuguese. Certain syllable structure sequences were found to be language specific, such as CC and CCC sequences that appeared only in English (British and American) in unstressed syllables.

In Brazilian Portuguese, CV emerged as the most frequent pattern in stressed and unstressed syllables of disyllabic and trisyllabic words. Examples are words like *vida* ["life" in Portuguese] (phonological transcription /'vi.da/, syllable structure transcription CV-CV and log token frequency 5.05), *babá* ["nanny" in Portuguese] (phonological transcription /ba'ba/, syllable structure transcription CV-CV, and log token frequency 3.11), and *estado* ["state' in Portuguese] (phonetic transcription /is'ta.dU/, syllable structure transcription VC-CV-CV, and log token frequency 5.24). This result challenges weight sensitivity theorizations, because complex syllables such as CVC and CD would be expected to emerge as the most frequent ones in stressed position. If Brazilian Portuguese were a language in which vowel length weighs for stress assignment, this result would still be in line with weight sensitivity, because our notation does not include the notation VV - "long" (VV) vowel according to the formal phonological notations -, only V, because phonetic evidence does not support vowel length as a distinguishing phonological feature. Thus, vowel length is not considered weight-sensitive for word stress in Brazilian Portuguese.

In British English, the most frequent syllable structure sequences in stressed and unstressed syllables were D and CV in disyllabic words, and CDC and CVC in trisyllabic words. Examples are words such as *over* (phonological transcription / '@U-v@ /, syllable structure transcription

D-CV, and log token frequency 3.83) and *delightful* (phonological transcription /dI-'laIt-fUl/, syllable structure transcription CV-CDC – CVC, and log token frequency 2.37). In American English, the most frequent syllable structure sequences in stressed and unstressed syllables were CDC and CV, in disyllabic words; and CVC and CV, in trisyllabic words. Examples are words such as *likely* (phonological transcription / 'laIk-li/, syllable structure transcription CDC-CV, and log token frequency 4.85) and *century* (phonological transcription /'sEn-tS@`-ri/, syllable structure transcription CVC-CV-CV, and log token frequency 4.5). It is interesting to note that the patterns that emerged from the corpora of the two varieties of English in stressed and unstressed syllables were considerably different in syllable structure sequence, while common patterns differed in their frequency.

In Tables 2 and 3, we found indications that syllable structure complexity is a factor in word stress assignment for the varieties of English, but not for Brazilian Portuguese. It is likely that other factors, such as the syllabic position of stressed and unstressed syllables (Stress Status) within words of different lengths (Word Length), provide additional information on how the syllable pattern distributions act on word stress assignment across the three lexicons. This hypothesis will now be explored in this study.

Statistical analyses of the syllable pattern properties of the three languages

Next, we applied three different methods to analyze the syllable pattern inventories of our three target languages: Multiple Regression Analyses, Conditional Inference Trees, and Random Forest modeling. We used Multiple Regression Analyses to test to what extent token frequencies of syllable structures have a significant effect on stress assignment with respect to a particular syllable in words that differ in number of syllables. An advantage of Multiple Regression Analysis is that it clearly indicates significant syllable patterns. However, a disadvantage is that it does not consider hierarchical dependencies among the included variables.

Following innovations in the statistical modeling of linguistic database analysis by Tagliamonte and Baayen (2012), we therefore also applied a Random Forest technique available in the 'party' package R (Strobl, Zeileis, 2007; Strobl *et al.*, 2008; Hothorn *et al.*, 2006a; R Development Core Team), which implements forests of conditional

inference trees. Conditional Inference Trees and Random Forest models vary in their explanatory power of factor interactions.

Conditional Inference Trees are constructed based on series of binary decisions that are made with respect to the values of the predictor variables (in our study, Language Group, Syllable Pattern, Stressed Syllable, Syllabic Position, and Word Stress Pattern). The model provides likelihood estimates for predictor variables based on response variable values (in our study, Log token frequency). For instance, for the factor Language Group, it considers whether splitting the data into one of the three languages affects the frequency of use of certain syllable patterns. The Conditional Inference Tree model represents it as a first significant split (or node) based on Language Group and a second significant split (or node) based on Syllable Pattern. Thus, the Conditional Inference Tree algorithm considers all predictors in the analysis and splits the data into subsets whenever the data likelihood allows it. This algorithm is applied in recurrent loops over all the subsets of the model, until no further partitioning is needed, providing an exhaustive and homogenous analysis of predictor interactions based on the data.

The Random Forest approach constructs a large number of these conditional inference trees. Each of these trees contributes a vote based on what it proposes as the most likely outcome response variable (the 'importance measure', implemented in the Random Forest function of the 'party' package (Strobl *et al.*, 2008; Tagliamonte, Baayen, 2012). The advantage of the Random Forest approach is that all predictor variables are exhaustively analyzed with respect to their importance for the response variable, as multiple trees are created by exclusion and inclusion of predictors via the generation of multiple possible trees. A disadvantage is that the hundreds of conditional inference trees created by the Random Forest algorithm are difficult to describe and display in research papers. The sum of the multiple trees may be accessed via a Variable Importance graph, but the multiple variable interactions that the Random Forest model provides are impossible to be visualized. For this reason, Tagliamonte, Baayen (2012) suggested to combine the Confidence Inference Trees with the variable Importance from Random Forest models in the analysis of linguistic corpora.

Next, we conducted a Random Forest analysis on Log Token Frequency to assess the relative importance of six predictors: Language Group, Syllable Pattern, Stressed Syllable, Syllabic Position, and Stress Pattern (see Figure 1).

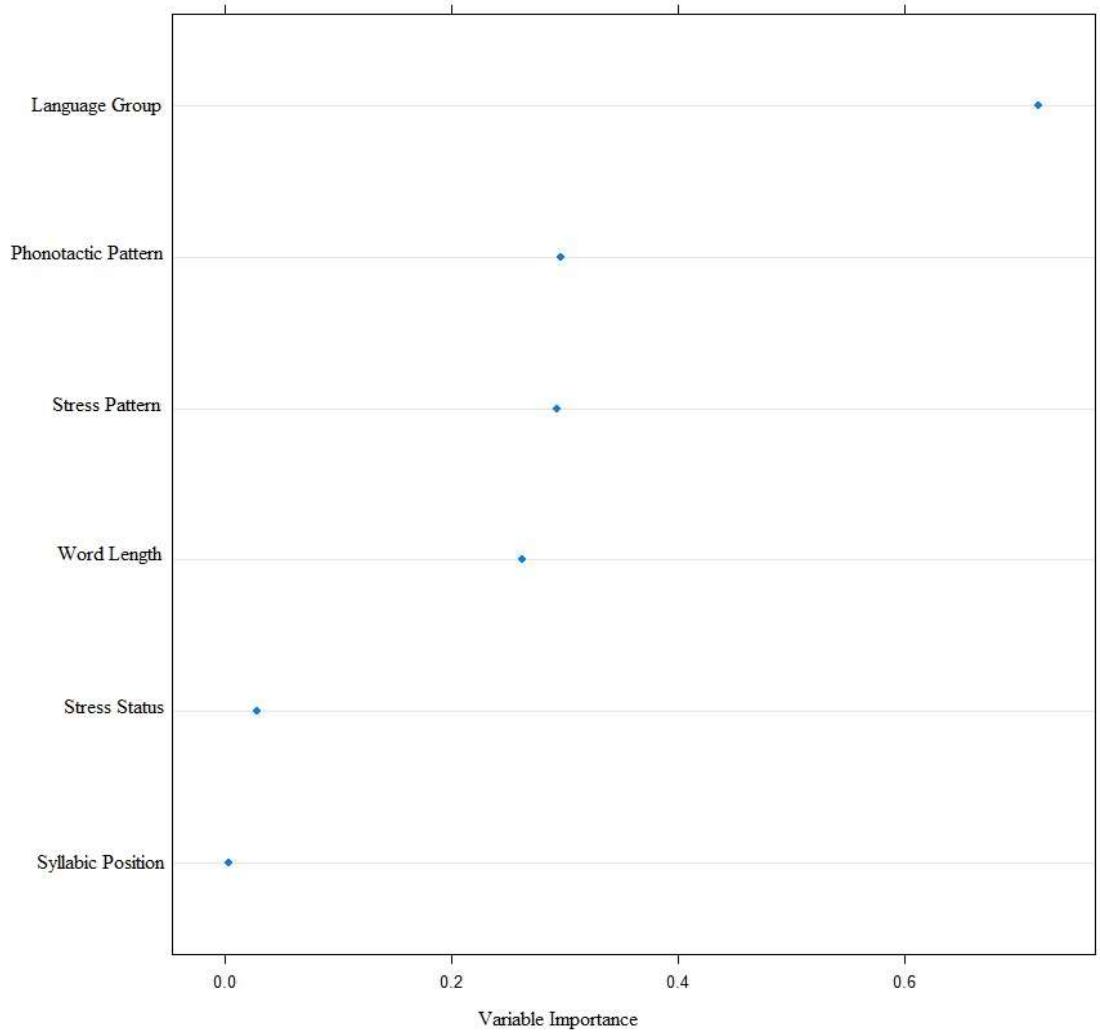


Figure 1: Conditional permutation variable importance for the random forest with all predictors.

The Random Forest model created 386 trees and had an index of concordance *C*=0.85. This index is equal to Pearson r. Figure 1 orders the importance of the conditional permutation-based predictors. According to Figure 1, Language Group is the most important predictor. Other predictors were listed in order of importance: Stress Pattern, Syllable Pattern, and Word Length. However, Stress Pattern was only slightly more important than Syllable Pattern, as indicated by the model. The predictor Stress Status showed weak predictability and the least important predictor was Syllabic Position.

As the importance of Language Group indicates, the three languages analyzed here as a whole constitute the most important predictors of the differences in the interactions among the other five predictors. It is therefore necessary to run separate analyses for each language to better understand the relationship of the predictors in each of them. These analyses are shown in Figures 2(a) and 2(b), 3(a) and 3(b), and 4(a) and 4(b) below.

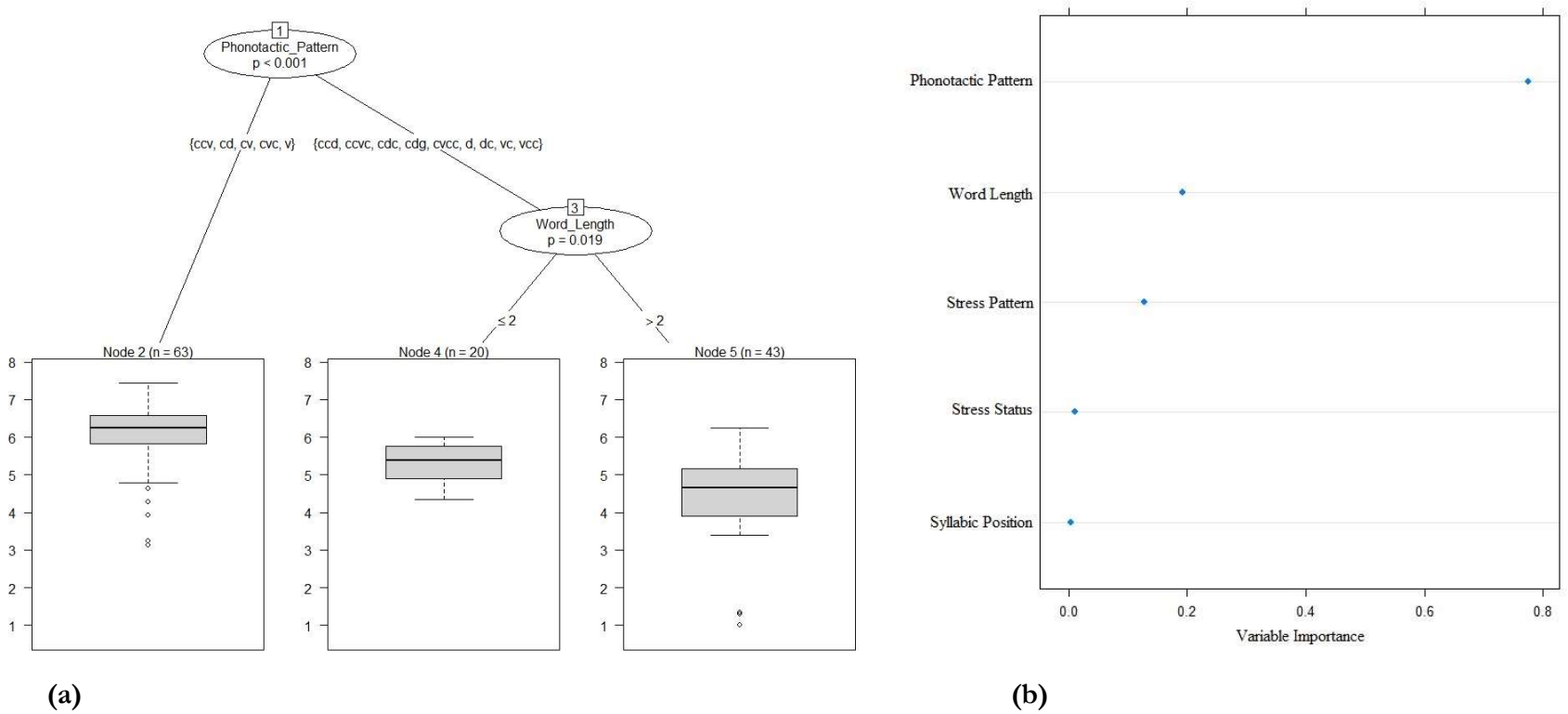
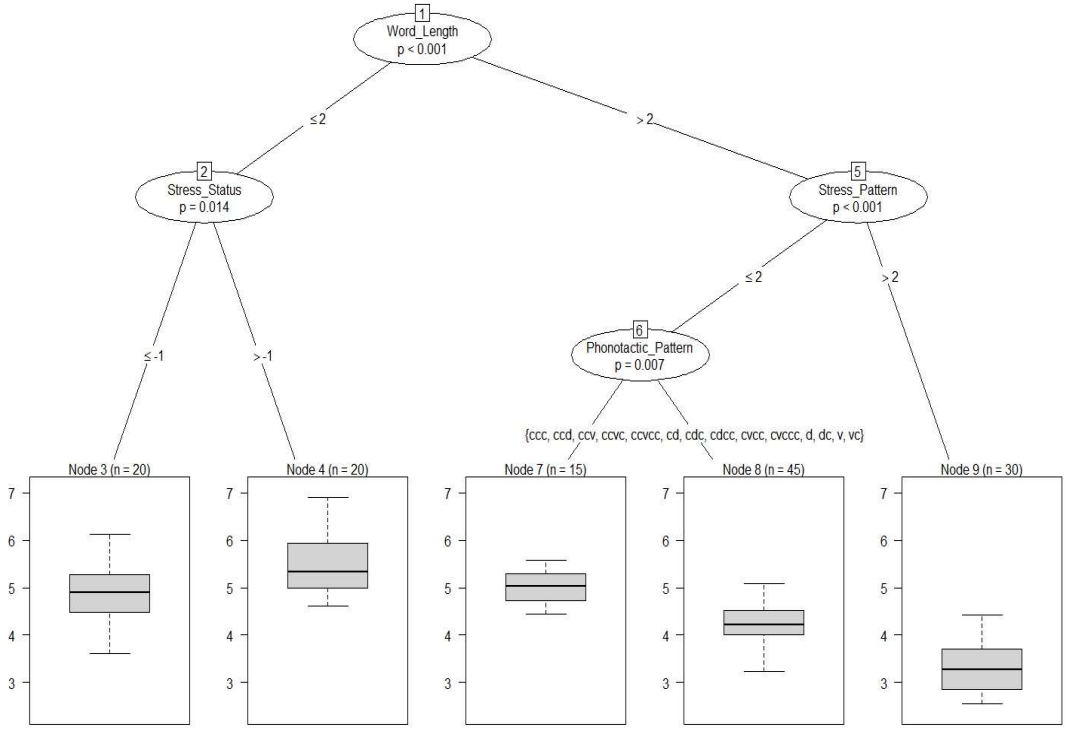
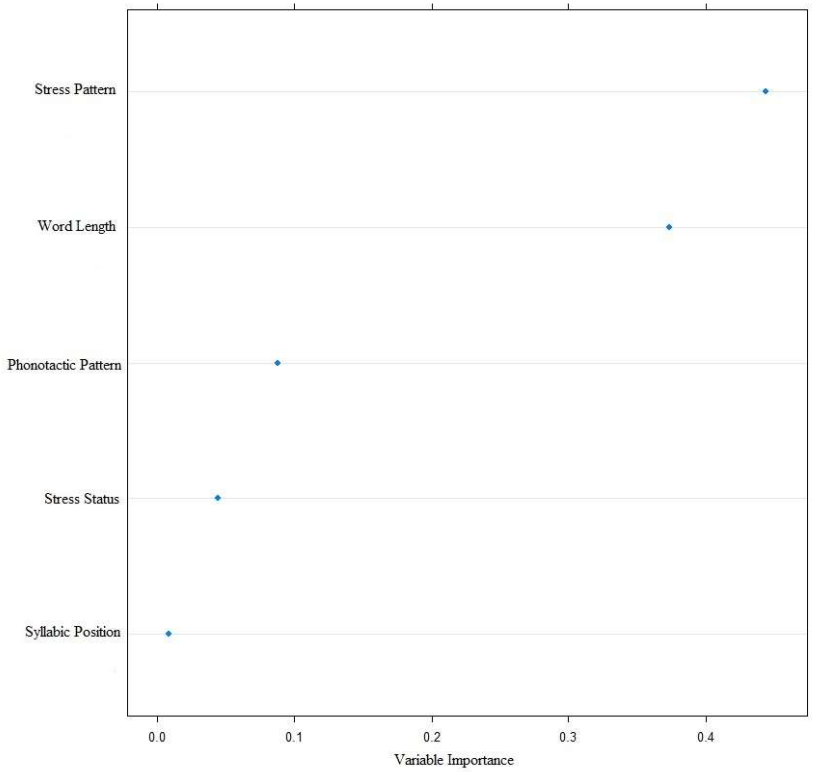


Figure 2: (a) Conditional Inference Tree for Log Frequency of Syllable Patterns to Word Stress of Brazilian Portuguese
(b) Conditional permutation variable importance for the Random Forest for Brazilian Portuguese with 5 predictors.

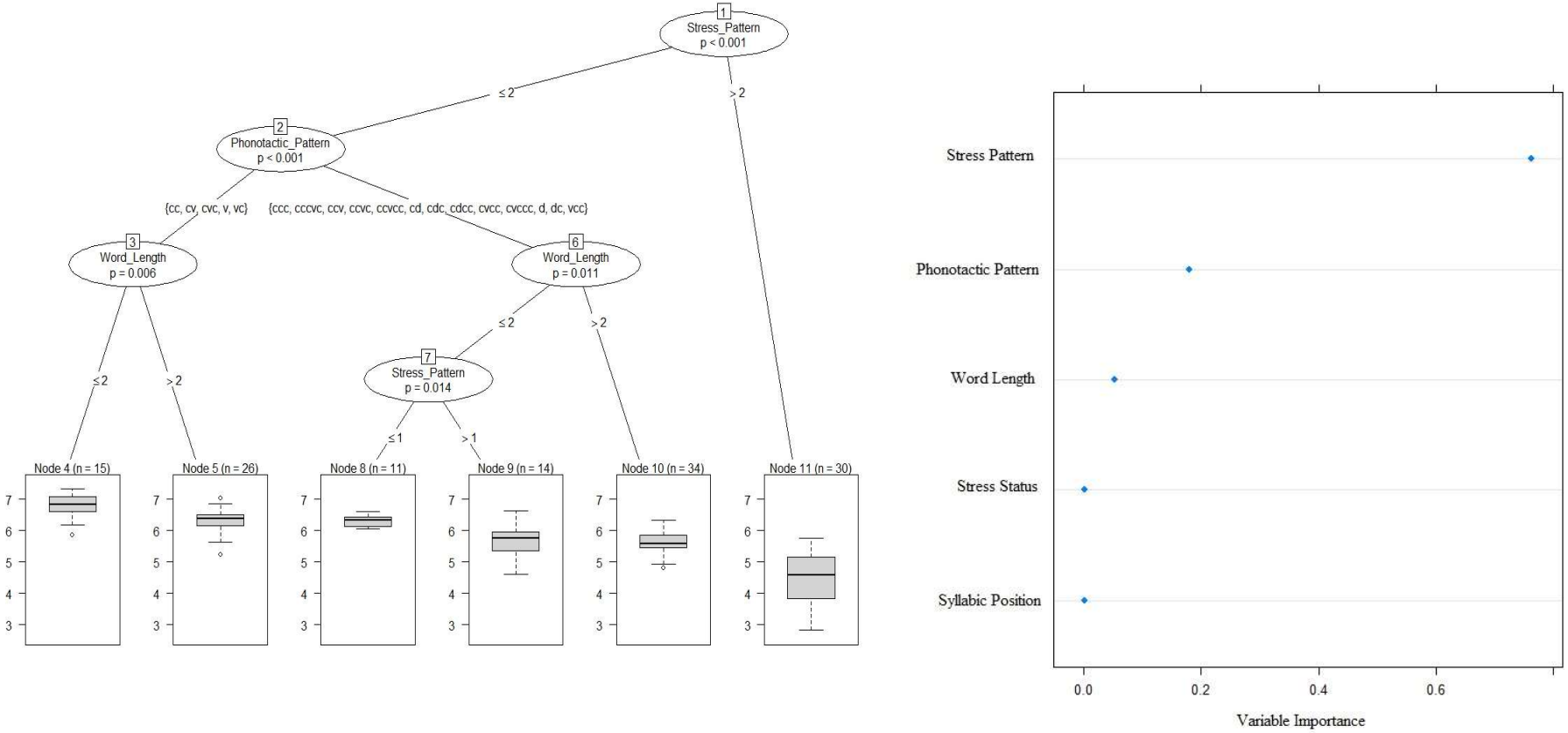


(a)



(b)

Figure 3: (a) Conditional Inference Tree for Log Frequency of Syllable Patterns to Word Stress of British English
(b) Conditional permutation variable importance for the Random Forest for British English with 5 predictors.



(a) (b)

Figure 4: (a) Conditional Inference Tree for Log Frequency of Syllable Patterns to Word Stress of American English (b) Conditional permutation variable importance for the Random Forest for American English with 5 predictors.

The Conditional Inference Tree in Figure 2(a) shows relatively simple interactions in the data. The index of concordance of this model is C = 0.60, which accounts for approximately 36% of the prediction accuracy. The predictors related to word stress (Stressed Syllable, Syllabic Position, and Word Stress Pattern) did not make it into the tree, which indicates that syllable patterns in Brazilian Portuguese are weak predictors of Word Stress. An alternative explanation is that the Word Stress factor is strongly correlated to Syllable Pattern and/or Word Length (Domahs *et al.*, 2014). In this model, the most important subset is SyllablePpattern (node 1). In general, for disyllabic and trisyllabic words the following patterns are preferred (node 2): CCV, CD, CV, CVC, V. Word Length is affected by Syllable Pattern (node 3) and the CCD, CCVC, CDC, CDG, CVCC, D, DC, VC, VCC syllable patterns are the second favorites for disyllabic words and trisyllabic words (node 5). Figure 2(b) shows the variable importance graph based on the evidence from 126 conditional inference trees generated by the random forest model of Brazilian Portuguese. It has a concordance of C=0.76, which corresponds to a prediction accuracy of 58% and to 18% improvement over the performance of the single conditional inference tree. The most important variable according to the Random Forest analysis is Syllable Pattern, followed by Word Length and Stress Pattern in decreasing order of importance. The two least important predictors are Stress Status and Syllabic Position.

In Figure 3, we show the Conditional Inference Tree of British English, which has an index of concordance of C=0.83, corresponding to 68% accuracy of predictions. The tree shows Word Length as a first split (node 1). If Word Length is disyllabic (node 2), Stress Status (-1= unstressed syllable; 1 = stressed syllable) is an important predictor affecting Log frequency of syllable patterns (nodes 3 and 4). If Word Length is trisyllabic (node 5), Stress Pattern is an important predictor. If Stress Pattern of trisyllabic words are first or second syllable stress, it affects Syllable Pattern - CC, CV, CVC are preferred (node 7) and CCC, CCD, CCV, CCVC, CCVCC, CD, CDC, CDCC, CVCC, CVCCC, D, DC, V, VC are preferred (node 8). The final syllable of trisyllabic words is not affected by Syllable Pattern. Figure 3(b) shows the variable importance measure resulting from the Random Forest modeling on Log Frequencies

of Syllable patterns in British English based on evidence from 130 trees. It has an index of concordance *C*=0.86, corresponding to an accuracy of 74% in prediction. The Random Forest analysis shows that the most important predictor of the Log Frequencies is Stress Pattern, followed by Word Length. Less important are Syllable Pattern and Stress Status, while the least important predictor is Syllabic Position.

Finally, in Figure 4(a), the Conditional Inference Tree of US English is shown and its index of concordance is *C*=0.83, which corresponds to a prediction accuracy of 69%. The most important subset of the US English tree is Stress Pattern (node 1). If stress patterns are either first or second syllable stress, Syllable Pattern is an important predictor (node 2). Especially if Syllable patterns are CC, CV, CVC, V, or VC, the predictor Word Length is important (node 3). In other words, disyllabic and trisyllabic words are differently influenced by these syllable patterns (nodes 4 and 5). The CCC, CCCVC, CCV, CCVC, CCVCC, CD, CDC, CDCC, CVCC, CVCCC, D, DC, and VCC syllable patterns are important predictors of the Stress Pattern of disyllabic words (node 6, 7 and 8), but do not affect the Stress Pattern of trisyllabic words (nodes 6 and 9). Stress Pattern is the most important predictor of final syllable stress Log Frequencies (nodes 1 and 10). In Figure 4(b) we show the variable importance graph based on 130 trees generated by the Random Forest modeling of US English Log Frequencies. Its index of concordance is *C*=0.87, which corresponds to 76% accuracy in predictions. The most important predictor here is Stress Pattern. Syllable Pattern and Word Length are considerably less important predictors of Log Frequencies. The least important predictors are Syllabic Position and Stress Status.

Comparing the findings for different approaches, we can state that the agreement between the Conditional Inference Trees and Random Forest Variable Importance was lower for Brazilian Portuguese than for English. Said differently, the variables we used as predictors covered less of the Syllable Complexity-to-Stress phenomena for Brazilian Portuguese. The current analyses provide no indications that there is a straightforward relationship between word stress and syllable pattern distributions in this language. On the other hand, the agreement between the analyses involving Conditional Inference Trees and Random Forest Variable Importance were reasonably high for British English and American

English and less robust but still high for Brazilian Portuguese. In sum, for both British English and American English, the models indicate that the relationship between syllable structure and word stress was top-down in the sense that word stress and word length determined the syllable pattern distributions and not the reverse.

Discussion

To investigate whether syllable complexity motivates word stress assignment or vice-versa, we decided to use corpora analyses, because the factors involved were too many and too novel (such as the factor Word Length in number of syllables) to be tested via experimental methods. We created three phonetic corpora for words of Brazilian Portuguese, British English, and American English, based on three existing corpora, but including the syllable structure transcription of words. We then analyzed the syllable pattern properties of these corpora in detail by means of a number of statistical approaches.

Our question was motivated by two observations, indicating omissions and contradictory evidence. First, we noticed that psycholinguistic models that account for word stress in speech production do not explore the relationship of word stress to syllable complexity, and when they do, their parsing formulations are based on phonological rules (which are rational) and not on corpora mining (which would be the empirically appropriate method). In addition, word length in number of syllables does not seem to be a factor that is considered in formal or empirical studies on the relationship between syllable complexity and word stress. Second, rational-logical phonological theories have argued that the relationship between syllable complexity and word stress is bottom-up, meaning that syllable complexity motivates stressed syllables relative to syllable position within a word in languages classified as weight-sensitive. However, these formulations are not empirically based and are inconsistent with evidence from phonetic experiments testing the value of the duration of syllabic nuclei and syllabic complexity in elicited and spontaneous speech. While phonetic data indicate that differences in phrasal stress and word stress underlie different syllable structures, there is no empirical evidence that syllable structure also motivates stress assignment. If this were the case, word pairs such as *forbear – forbear* (Cutler, 1986) could not have any word stress, because both syllables are

potentially "heavy", in the sense that the nuclei of both syllables consist of full vowels (VV) for phonological theories on weight sensitivity and have one (in this example, the same) consonant in coda.

When we analyzed the especially adapted corpora in our study, we found that, according to the Random Forest analysis, Language was by far the most important predictor of Log Frequencies of Syllable Pattern. This motivated us to conduct separate analyses for each of the three languages. For a language to be considered weight sensitive in empirical terms, we must establish clear relationships between syllable complexity and word stress. Syllable complexity should be the most important predictor affecting predictors related to word stress, such as Stress Pattern, and Stress Status of syllables (stressed or unstressed). As heavy syllables are hypothesized to influence the edges of the trisyllabic window (antepenultimate and final syllables), we should obtain indications that Syllable Pattern influences word stress specifically in those Syllabic Positions. Overall, we found that the most important predictor of Log frequency of syllable patterns by far was Language Group. The next most important predictor was Syllable Pattern, followed closely by Stress Pattern and Word Length. An easy interpretation of this result by the Random Forest model for the three languages would have been that the languages are weight-sensitive. However, as the model ranked the factor Languages above all the others, it also indicates that the three languages vary considerably in their relationships with the other predictors. Therefore, each of them had to be analyzed individually, because big differences between languages might emerge in relation to word stress and syllable pattern interactions.

All statistical methods we applied confirmed differences between the three Language Groups with respect to syllable patterns and word stress. No influence of syllable patterns on word stress was found for any language. Furthermore, there was evidence that the direction of the effects was from word stress patterns and word length in number of syllables to syllable patterns, but not the reverse (in the British and American English varieties). This general result *per se* contradicts the weight-to-stress principle hypothesized to characterize word stress assignment in these languages. We will now discuss the results of the language-specific analyses.

In Brazilian Portuguese, the Syllable Pattern was an important predictor of Word Length, but they were not very important for word stress related predictors such as word Stress Pattern and Stress Status. This suggests that for this language syllable complexity does not motivate stress patterns or play only a marginal role in stress assignment. In Brazilian Portuguese, syllable patterns do not seem to play a direct role in the representation of word stress patterns, but seem to change as a consequence of word length in number of syllables. If we take the example of the disyllabic word *quadril* [body part = hip] and the pentasyllabic word *dificuldade* [difficulty], the latter is more likely to suffer syllable structure reduction /dZi.fi.kuw ˈda.dzi/ - /dZif.kuw ˈdadZ/ than the former (SAMPA transcription = /kwa'd4iw/. The finding can be explained primarily by word length – short words reduce less than long words - and only secondarily by word stress. Note that in long words unstressed syllables are deleted; segments remaining after syllable deletions then cluster as codas of stressed neighbouring syllables. The Confidence Inference Tree and Random Forest of Brazilian Portuguese suggest that predictor probabilities inflated the importance of the Syllable Pattern predictor when we modelled a Random Forest including the three languages.

Although in British English, word stress and word length were the most important aspects of the language, the system as a whole was rather different from Brazilian Portuguese. Syllable patterns were affected mainly when the stress pattern was on the first or second syllable of trisyllabic words, but the expected effect on final syllable was not found, contradicting one of the specific weight-sensitivity claims. Short words were not affected by Syllable patterns. The finding that the frequency of syllable patterns was not affected by the syllable patterns themselves, but primarily by word length and secondarily by word stress patterns and stress status of syllables, was also not in line with the weight sensitivity claim to stress assignment either.

In contrast, the American English relationship between syllable patterns and word stress was similar to the British English system in many ways. For both languages, word stress was the most important predictor of Log frequency of syllable patterns. However, in American English, the Syllable Pattern factor played a role with respect to the stress assignment

of disyllabic words. Certain syllable structure sequences were more likely in stressed syllable position than in unstressed syllable position. Among them are very reduced syllable patterns that appear only in unstressed position, such as CCC; and very complex syllables such as CDC and CDCC, which are more frequently or exclusively found in stressed position. The syllable complexity importance drops considerably in trisyllabic words to a point that was not represented in the Confidence Inference tree, just as in British English. The direction of predictability in American English, as in British English, was from word stress to syllable complexity, instead of in the opposite direction.

In all, these analyses indicate that the relationship between syllable structure complexity and word stress is neither clear, nor is language independent. Our analyses show that all factors related to the frequency of syllable structure patterns have a role to play, and their mutual interaction depends on the language under analysis. When we compare the mutual influence of syllable structure complexity and word stress at a general level, word stress is a more important determiner of syllabic complex patterns than complex syllables are for word stress.

Final remarks

The findings of the present study have implications for psycholinguistic models for word stress assignment in word production and word recognition. First, the relationship between syllable complexity and word stress is not linear or unidirectional. Second, word stress seems to come first in the process of encoding segmental sequences into syllabic nodes prior to speech production. Third, word length may change the relationships that other predictors hold with word stress, such as that between syllable patterns and word stress patterns. Thus, word length should be included as a factor of analysis in psycholinguistic models, for instance.

Because relationships between syllable complexity and word stress vary across lexica, language users of different languages will develop different strategies for word recognition and word production (both concept-to-speech and orthography-to-speech). Most likely, bilinguals will use language processing strategies that are a mix of the linguistic systems they experience. Thus, given the properties of Portuguese, their L1, late

Brazilian-English bilinguals might tend to rely on syllable structures to identify English (L2) words. They might also tend to produce English words with more syllable pattern integrity even though word stress accuracy is less important for word representation in Brazilian Portuguese. The use of syllable patterns would result in misleading clues in English, which favours word stress for word representation. These predictions may be tested in future research.

We hope to have shown that mining big corpora of data can result in sensitive results about linguistic processing that cannot be obtained in empirical studies involving the orthogonal manipulation of only a few variables. We argue that a complete understanding of the relationship of syllable complexity and word stress within a lexicon and across lexicons demands the analysis of large datasets involving various multilevel variables. This sort of analysis is facilitated by the increasing development of statistical analysis techniques, such as the Random Forest approach (Tagliamonte, Baayen, 2012), that can be applied to the analysis of big data (and) linguistic corpora.

Acknowledgments

I would like to thank Erasmus Mundus – Monesia Program for a full PhD grant, and CNPq partial PhD abroad grant received for my PhD project from which this paper originated. I would also like to thank my supervisor, Prof. Dr. Ton Dijkstra (Donders Institute for Brain, Cognition and Behaviour at Radboud University, Nijmegen, the Netherlands) for his support and supervision, as well as I thank Dr. Eric Sanders and MSc. Gustavo Mendonça for helping me build the corpora used in this study. Finally, I thank Prof. Dr. Harald Baayen for his suggestions for the statistical analyses of this study.

References

BALOTA, D. A., YAP, M. J., HUTCHISON, K. A., CORTESE, M. J., Megastudies: What do millions (or so) of trials tell us about lexical processing? In ADELMAN, J. S. (ed.), *Visual Word Recognition: Vol. 1. Models and Methods, Orthography and Phonology,* New York, Psychology Press, 2012, 90-115.

BISOL, L., Mattoso Câmara Jr. e a palavra prosódica, *DELTA*, 2004, **20**, 59-70. https://doi.org/10.1590/S0102-44502004000300006

BRAUN, B., GALTS, T., KABAK, B., Lexical encoding of L2 tones: the role of L1 stress, pitch accent and intonation, *SECOND LANGUAGE RESEARCH,* 2014, **30 (3)**, 323-350. https://doi.org/10.1177/0267658313510926

CHOMSKY, N., HALLE, M., *The Sound Pattern of English,* New York, Harper and Row, 1968.

COOPER, N., CUTLER, A., WALES, R., Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners, *LANGUAGE AND SPEECH*, 2002, **45(3)**, 207-228. https://doi.org/10.1177/00238309020450030101

CUTLER, A., Forbear is a homophone: Lexical prosody does not constrain lexical access, *LANGUAGE AND SPEECH,* 1986, **29**, 201-220. https://doi.org/10.1177 /002383098602900302

DOMAHS, U., PLAG, I., CARROLL, R., Word stress assignment in German, English and Dutch: quantity-sensitivity and extrametricality revisited, *JOURNAL OF COMPARATIVE GERMANIC LINGUISTICS,* 2014, **17(1)**, 59-96 https:// doi.org/10.1007/s10828-014-9063-9

HAMMOND, M., *The Phonology of English: A Prosodic Optimality-Theoretic Approach*, London, OUP, 1999.

HAYES, B., *A Metrical Theory of Stress Rules,* [Doctoral Thesis MIT, US], 1981. Revised version distributed by IULC, published by Garland Press, New York, 1985.

HAYES, B., Extrametricality and English Stress, *LINGUISTIC INQUIRY*, 1982, **13**, 227-276.

HAYES, B., *Metrical Stress Theory: Principles and Case Studies,* Chicago, University of Chicago Press, 1995.

HOTHORN, T., HORNIK, K., ZEILEIS, A., *Party: A Laboratory for Recursive Part(y)itioning,* 2006. [http://CRAN.R-project.org/package=party]

HYMAN, L., *A Theory of Phonological Weight,* Dordrecht, Foris, 1985 https://doi.org /10.1515/9783110854794

JENSEN, J. T., *English Phonology*, Amsterdam, John Benjamins, 1993 https://doi.org /10.1075/cilt.99

KAGER, R., *A Metrical Theory of Stress and Distressing in English and Dutch* [Doctoral Thesis, Utrecht University, NL], 1989

KIJAK, A.M., *How Stressful is L2 Stress? A Cross-linguistic Study of L2 Perception and Production of Metrical Systems* [Doctoral Thesis] Utrecht, LOT, 2009.

KIPARSKY, P., From cyclic phonology to lexical phonology, *THE STRUCTURE OF PHONOLOGICAL REPRESENTATIONS*, 1982, **1**, 131-175 https://doi.org/10.1515/9783112328088-008

KIPARSKY, P., Some consequences of lexical phonology, *PHONOLOGY*, 1985, **2 (01)**, 85-138 https://doi.org/10.1017/S0952675700000397

LIBERMAN, M., PRINCE, A., On stress and linguistic rhythm, *LINGUISTIC INQUIRY*, 1977, **8**, 249–336.

MATTOSO CÂMARA, JR, J., *Problemas de linguística descritiva*, Petrópolis, Vozes, 1969.

MATTOSO CÂMARA, JR, J., *História e Estrutura da Língua Portuguesa,* Rio de Janeiro, Padrão, 1975.

POST DA SILVEIRA, A., VAN HEUVEN, V., CASPERS, J., SCHILLER, N.O., Dual activation of word stress from orthography: The effect of the cognate status of words on the production of L2 stress, *DUTCH JOURNAL OF APPLIED LINGUISTICS*, 2014, **3**, 2, 170-196. DOI: https://doi.org/10.1075/dujal.3.2.05sil

POST DA SILVEIRA, A., VAN LEUSSEN, J. W., Generating a bilingual lexical corpus using interlanguage normalized Levenshtein distances, *Proceeding of the 18th International Conference of Phonetic Sciences (XVII ICPhS)*, Glasgow, UK, 2015.

POST DA SILVEIRA, A., *Word stress in second language word recognition and production*. 1. ed. Enschede: Ipskamp, 2016 Available in: https://www.researchgate.net /publication /312219967_Word_stress_in_second_language_word_recognition_and_production

POST DA SILVEIRA, A., SANDERS, E., MENDONÇA, G., DIJKSTRA, T., What Weighs for Word Stress? Big Data Mining and Analyses of Phonotactic Distributions in Brazilian Portuguese, in VILLAVICENCIO, A. *et al.* (eds) *Computational Processing of the Portuguese Language,* PROPOR 2018. Lecture Notes in Computer Science, Dordrecht, Springer, vol 11122, 2018, 399-408 https://doi.org/10.1007/978-3-319-99722-3_40

POST DA SILVEIRA, A., SANDERS, E., MENDONÇA, G., DIJKSTRA, T. (submitted), Corpora para investigação fonética, fonológica e psicolinguística: corpora do português brasileiro, inglês americano e inglês britânico com variáveis fonológicas e psicolinguísticas, Preprint DOI: 10.13140/RG.2.2.14814.00323

R DEVELOPMENT CORE TEAM: R, *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, 2008. [http://www.R-project.org/]

STROBL, C, BOULESTEIX, AL, ZEILEIS, A, HOTHORN, T., Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution, *BMC BIOINFORMATICS*, **8**, 25., 2007. https://doi.org/10.1186/1471-2105-8-25

STROBL, C., ZEILEIS, A., Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance, *Proceedings of the 18th International Conference on Computational Statistics,* Porto, Portugal, 2008 https://doi.org/10.1186/1471-2105-8-25

TAGLIAMONTE, S., BAAYEN, H., Models, forests and trees of York English: Was/were variation as a case study for statistical practice, *LANGUAGE VARIATION AND CHANGE*, 2012, **24(2)**, 135-178 https://doi.org/10.1017/S0954394512000129

TROMMELEN, M., ZONNEVELD, W., Dutch, in VAN DER HULST, H. (ed.), *Word Prosodic Systems in the Languages of Europe*, Berlin, Mouton de Gruyter, 1999, 492-514.

VAN HEUVEN, V.J.J.P., VAN LEYDEN, K., Lexical stress and spoken word recognition, Dutch versus English, in DIKKEN, M. DEN, CREMERS, C. (eds.) *Linguistics in the Netherlands 1996*, Amsterdam, John Benjamins, 1996. 159-170 https://doi.org/10.1075/avt.13.16ley

VITEVITCH, M. S., LUCE, P. A., CHARLES-LUCE, J., KEMMERER, D., Phonotactics and syllable stress: Implications for the processing of spoken nonsense words, *LANGUAGE AND SPEECH*, 1997, **40**, 47–62 https://doi.org/10.1177/002383099704000103

Amanda POST DA SILVEIRA is graduated in English Letters and obtained her master's degree in Linguistic Studies both from the Federal University of Santa Maria, Brazil. She obtained a PhD in Psycholinguistics and Phonetics from the Donders Institute (DCC), Radboud University Nijmegen, the Netherlands. In her research, she uses experimental methods, such as eye-tracking and word naming reaction times, to investigate bilingual phonological representations and their role in lexical retrieval from orthography and from the acoustic signal. She developed a bilingual measure called *interlanguage normalized Levenshtein distances* (inLd). She is currently working at UFJ as an Assistant Professor.