Syllabic Size Restrictions on Verb Reduplication in Brazilian Portuguese

Conference Paper · November 2018

1 author:

Jamilläh Rodriguez
University at Albany, The State University of New York

Syllabic Size Restrictions on Verb Reduplication in Brazilian Portuguese

JAMILLÄH S. RODRIGUEZ
*University at Albany, State University of New York**

1 Introduction

Brazilian Portuguese incorporates a process of nominalization in which a verb root undergoes total reduplication to form a noun. For example, *pega* 'catch' reduplicates to create *pega-pega* 'a game of tag'. Reduplicated verbs differ from the larger set of verbs with respect to disyllable preference and initial consonants. In this paper, I analyze this difference with the Null Parse theory (McCarthy & Wolf 2009) and compare the Brazilian Portuguese data to an analogous pattern in Cuban Spanish (Lederer 2003).

Using the SUBTLEX corpus (Tang 2012) and past literature (Araújo 2002; Gonçalves 2004; Sempere 2006), a total of 49 reduplicants were compiled. The properties of the reduplicated verbs in this list were compared to the properties of Brazilian Portuguese verbs overall in the SUBTLEX corpus, and there was a difference in size and shape between the two sets of data (henceforth referred to as the reduplicant corpus and SUBTLEX corpus). In the SUBTLEX corpus, disyllabic verbs are common, but they are not the most frequently occuring. However, in the reduplicant corpus, disyllabic verbs are the most frequent. I propose that there is a size restriction of two syllables in verbs that are reduplicated to account for this difference between the characteristics of verbs overall and verbs that are reduplicated. I use the Maxent Grammar Tool (Hayes & Wilson 2006) to examine the predicted outputs using weighted constraints to restrict the size of verbs with different syllable counts. The results show that there is indeed a size restriction on verbs within the reduplication process, and the lack of reduplicated verbs longer than two syllables can be accounted for by a null parse output.

A novel observation of my study on the reduplication pattern is that while monosyllables and disyllables are exclusively consonant-initial, trisyllables are exclusively vowel-initial. Furthermore, a comparison of the two corpora shows that monosyllables are not un-

*My sincere thanks to Michael Becker, Lee Bickmore, Lauren Clemens, and Maria Luisa de Andrade Freitas for their helpful comments, edits, and support. I would also like to thank the audience of BLS 44.

derattested and are not predicted to be worse than disyllables. The forms that are underrepresented are vowel-initial disyllables and both vowel-initial and consonant-initial trisyllables. Contrary to past literature on Portuguese reduplication (Gonçalves 2004) there were not only disyllabic reduplicants, but consonant-initial monosyllables and vowel-initial trisyllables observed in the data as well.

It is common cross-linguistically for phonological phenomena to be restricted to disyllables or at least to preferentially target them. For example, expletive infixation in English is restricted in that the infix appears to the left of a syllabically binary foot (McCarthy 1982).

(1a) abso-bloody-lútely
fan-fuckin-tástic
to-bloody-géther
im-fuckin-pórtant

(McCarthy 1982)

On a related note, there is a disyllabic word minimum requirement in many languages, including several Australian languages like Uradhi and Yidiɲ (Downing 2006).

There are also restrictions that are only applicable to reduplicative processes, such as a reduplication pattern in Tonkawa. Generally in the language, word-initial syllables are usually heavy. In the case of reduplication, word-initial syllables are restricted to light syllables (Gouskova 2007). When a word with a heavy initial syllable is reduplicated, a segment is deleted and the copied reduplicant becomes a light syllable, as in 2.

(1b) naa.toʔs → na-na.toʔs *'I step on it (repeatedly)'*
sal.koʔs → sa-sal.ke.noʔs *'I pull (repeatedly)'*

(Gouskova 2007)

In addition to the factor of word size, the identity of the word-initial segment is a primary determining factor in the reduplication pattern of Brazilian Portuguese. While monosyllables and disyllables are exclusively consonant-initial, trisyllables are exclusively vowel-initial. This is comparable to a reduplication pattern in Timugon Murut. If

the stem begins with a consonant the first syllable will be reduplicated, but if the stem begins with a vowel, the first syllable will be skipped and the second syllable will be copied (McCarthy 2008), as in 3.

(1c) li.mo *'five'* → li.li.mo *'about five'*
a.ba.lan *'bathes'* → a.ba.ba.lan *'often bathes'*

(McCarthy 2008)

To summarize, as we will see in Sections 2 and 3, Brazilian Portuguese reduplication is sensitive to commonly typologically observed properties: dyllabicity and whether the word is consonant or vowel-initial.

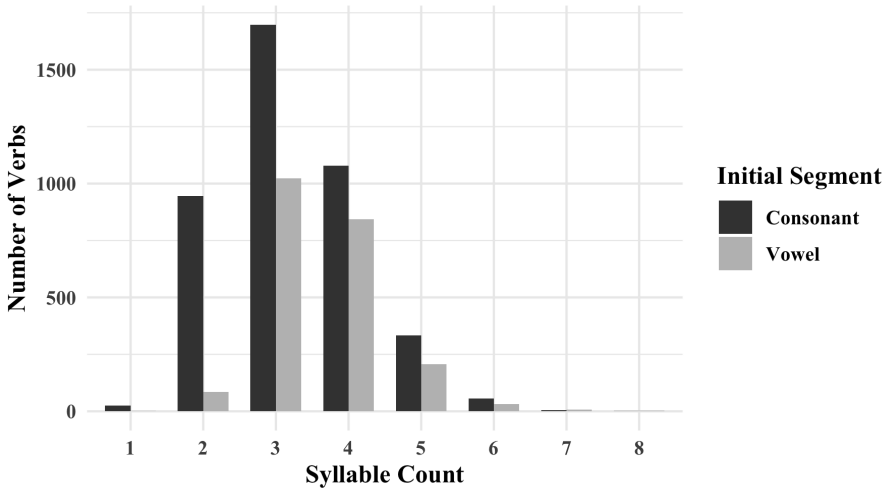
2 Corpus Data

To examine the properties of the Brazilian Portuguese reduplication process, a total of 49 reduplicated verbs were collected from both the past literature and the SUBTLEX corpus. Of the 49 forms, 4 (8%) were monosyllabic, 39 (80%) were disyllabic, and 6 (12%) were trisyllabic. All of the bases were in the third-person singular indicative form. All forms were shared with a native speaker and confirmed to be verbal with a nominal outcome.

In this paper, I use the SUBTLEX corpus to investigate the distribution of syllables and initial segments in Brazilian Portuguese verbs in the overall language. The SUBTLEX corpus consists of approximately 60 million words collected from the subtitles of movies and television shows. Using R (R Development Core Team 2008), I separated the corpus by part of speech, isolated the verbs, and organized them based on syllable count and initial segment. To find the third person verbs, I used the infinitive form, but manually corrected a set of verbs in which the number of syllables in the infinitive differs from the third-person singular indicative form. For example, disyllabic *fazer* 'to have' is conjugated as monosyllabic *faz*.

Figure 1 displays the type frequencies of the third-person singular indicative verbs in the SUBTLEX corpus, separated into consonant-initial and vowel-initial groups. There was a total of 6,340 unique verbs in the corpus.

Figure 1: Type frequencies of consonant-initial and vowel-initial verbs by syllable count in the SUBTLEX corpus



Consonant-initial verbs are more frequent than vowel-initial verbs across all syllable counts. As the syllable count in a verb increases, the difference between consonant-initial and vowel-initial frequency decreases.

To create a realistic model of the reduplication pattern for the Maxent learner, the observed outputs should reflect the proportions that are presented in Table 1.

Table 1: Proportions of predicted (SUBTLEX corpus) and observed (reduplicants corpus) outputs

σ	**Initial Segment**	**Predicted (%)**	**Observed (%)**
1σ	C	0.6	12.0
	V	0	0
2σ	C	25.0	80.0
	V	2.6	0
3σ	C	26.9	0
	V	44.9	8.0

If we look at the ratio of verbs by syllable in the SUBTLEX corpus, we can give the

Maxent learner a predicted output based on observation. In the set of reduplicants that was collected, the ratio of monosyllables to disyllables to trisyllables is 4:39:6. However, based on the corpus of verbs in the overall language, for every 39 disyllabic verbs with an onset there should be 1 consonant-initial monosyllabic verb, 4 vowel-initial disyllabic verbs, 42 consonant-initial trisyllabic and 70 vowel-initial trisyllabic verbs. To account for the difference in the two groups, I added 4 null outputs to the observed ouput of vowel-initial disyllables in the Maxent learner, 42 observed null outputs for the consonant-initial trisyllables, and 64 for the vowel-initial trisyllables.

Although the SUBTLEX corpus contains verbs with more than three syllables, this analysis will focus primarily on monosyllablic, disyllabic, and trisyllabic verbs because there were no observed occurrences of verbs with more than three syllables in the reduplicant corpus. It can be assumed that all stems with more than three syllables do not have a phonetically realized output.

3 Maxent Predictions

In this section, I use the Maxent Grammar Tool to ascertain whether we can accurately portray the reduplication pattern in the language by accounting for the presence of both null and pronounced outputs for disyllables and trisyllables. Although the language consists of verbs with more than three syllables, I will focus only on a maximum of three syllables in the word because that is the limit of the observed reduplication pattern.

In Maxent theory, each constraint is assigned a weight. The harmony ($\mathscr{H}$) is the combined total of the violation marks multiplied by the constraint weight for each candidate, and the probability (p) is the harmony divided by the sum of the harmony of all the candidates.

The null parse candidate, $\odot$, is an output form that holds no phonological or morphological properties. The only constraint that the null parse violates is MPARSE, and as a result, when the null output is present alongside other potential outputs that violate higher-ranked constraints, the null output will always win (McCarhy & Wolf 2009).

I use the following constraints to restrict the size of bases:

(1) PARSE-σ: All syllables must be parsed into feet. (Prince & Smolensky 2004)

(2) FT-BIN: All feet must be syllabically binary. (McCarthy & Prince 1990)

The FT-BIN constraint assigns a violation to candidates with monosyllablic bases due to the absence of a necessary syllable in the binary foot. Because monosyllables and trisyllables have distributional differences in the corpus, the Size-Restrictor constraint, introduced by Junko and Mester 1996, was separated into two constraints: FT-BIN and PARSE-σ rather than one constraint against any non-disyllabic base to account for this output difference.

3.1 Monosyllabic Verbs

Distributional differences between verbs with and without an initial onset were evident within the corpora. As a result, the Maxent learner was given inputs for vowel-initial and consonant-initial verbs separately.

Tableau 2 shows the results of the Maxent learner for monosyllabic reduplication with a verb that has an initial consonant. FT-BIN is only violated by the fully faithful form, and MPARSE is only violated by the null parse candidate.

Table 2: Monosyllabic with word-initial onset, *Dói* 'hurt'

/ˈdɔj/	FT-BIN w=1.2	MPARSE w=11.1	$\mathscr{H}$	p
(ˈdɔj.)(ˈdɔj)	**		-2.4	>**.99**
⊙		*	-11.1	<.01

The fully faithful reduplicated form will most likely be the output form. The null parse candidate has a nearly impossible chance of being the winning candidate.

In the tableau with the vowel-initial input below, the ONSET constraint is added to account for the difference in acceptability of verbs with an initial consonant and verbs without. MPARSE has a high weight, so the null parse candidate should not win, but the combined weights of the other constraints give the phonetically realized candidate a lower harmony.

Table 3: Monosyllabic without word-initial onset, *Há* 'have'

/ˈa/	FT-BIN w=1.2	MPARSE w=11.1	ONSET w=11.9	$\mathscr{H}$	p
(ˈa.)(ˈa)	**		**	-26.2	<.01
⊙		*		-11.1	>**.99**

Tableaux 2 and 3 have inverse results. With consonant-initial verbs, the reduplicated form is almost always the fully faithful, totally reduplicated candidate. However, with vowel-initial verbs the null parse candidate will be the output.

The reduplication of monosyllabic verbs presents a potential problem with the utilization of Null Parse theory, due to the fact that monosyllables are overrepresented in the data and the null parse can only account for underrepresented forms. The proportions of verbs by syllable count and initial segment were re-examined in the Mac-Morpho corpus (Aluísio et al. 2003; Fonseca & Rosa 2013; Fonseca et al. 2015), a corpus of Brazilian Portuguese texts, and were found to be almost identical to the proportions of the SUBTLEX corpus, so the overrepresentation is a consistent issue across corpora and consequently the language.

For disyllabic and trisyllabic data, the observed occurrence for the null parse output was calculated by subtracting the number of pronounced reduplications of that syllable and initial segment observed in the reduplicant corpus from the expected number of verbs according to the SUBTLEX corpus. However, the number of observed monosyllabic reduplications exceeded the number expected based on verbs in the language overall. There were a total of four monosyllablic reduplicants collected, but only one expected. This problem can not be solved using the null parse candidate or the MPARSE constraint.

3.2 Disyllabic Verbs

In the list of reduplicated verbs, disyllabic bases were most common. However, vowel-initial disyllabic reduplicants were absent and only verbs with word-initial consonants were observed. Similar to monosyllables, the MaxEnt Grammar Tool made the following predictions given a disyllabic input.

The fully faithful candidate does not violate any of the constraints and therefore has a harmony of 0. The null parse candidate only violates MPARSE, but because this constraint has a higher weight than 0, the null parse will lose.

Table 4: Disyllabic with word-initial onset, *pega* 'catch'

/ˈpɛ.gə/	FT-BIN w=1.2	MPARSE w=11.1	$\mathscr{H}$	p
(ˈpɛ.gə.)(ˈpɛ.gə)			0	**1**
⊙		*	-11.1	0

If there is a reduplicated verb and that verb is disyllabic and has a word-initial consonant, we expect that it will always be fully faithful and pronounced.

Unlike the consonant-initial verb in Tableau 4, the fully faithful form in Tableau 5 violates the ONSET constraint twice, giving it a low harmony.

Table 5: Disyllabic without word-initial onset, *achar* 'find'

/ˈa.ʃə/	FT-BIN w=1.2	MPARSE w=11.1	ONSET w=11.9	$\mathscr{H}$	p
(ˈa.ʃə.)(ˈa.ʃə)			**	-23.8	$<$.01
⊙		*		-11.1	$>$**.99**

According to the Maxent learner, the predicted probability of a vowel-initial disyllabic reduplicant having a fully faithful output is less than .01%. This differs from Tableau (5) because of the ONSET constraint. Although the MPARSE constraint has a high weight, the addition of the ONSET constraint gives the pronounced candidate a lower harmony than the phonetically unrealized candidate. As a result, the null parse candidate has an almost absolute chance of winning.

3.3 Trisyllabic Verbs

Trisyllabic verb reduplication differs greatly from monosyllabic and disyllabic verb reduplication in Portuguese because of the observed reduplications, both monosyllables and

disyllables only had observed occurences of consonant-initial reduplicants. Any vowel-initial reduplicated verbs were assumed to be null pronounced. Contrastingly, only vowel-initial verbs were observed to be reduplicated in trisyllabic verbs.

Furthermore, it was observed that although trisyllabic verbs are most frequent in the SUBTLEX corpus and monosyllables are quite rare, trisyllables are more marked than monosyllables when reduplicated. I add an additional constraint here, LAPSE-MEDIAL, based on constraints in Kager 2001.

(3) LAPSE-MEDIAL: No word-medial unstressed syllables.

Table 6: Trisyllabic with word-initial onset, *carimba* 'stamp'

/ka.(ˈɾĩ.bə)/	PARSE-σ w=0.0	LAPSE-MEDIAL w=23.1	MPARSE w=11.1	$\mathscr{H}$	p
ka.(ˈɾĩ.bə.)ka.(ˈɾĩ.bə)	**	*		-23.1	<.01
⊙			*	-11.1	>**.99**

As observed in Tableau 6, the null parse candidate will be the output form. The second candidate, [ka.hi.ka.hi.bə], appears to be an ideal candidate. It follows the size restriction requirement for the base and begins with a consonant, which was observed in all of the disyllabic reduplications; but the requirement to remain faithful to the input means it has a small chance of winning.

For monosyllables and disyllables, the ONSET constraint drastically changed the results of the output. With trisyllables, that is not the case.

Table 7: Trisyllabic without word-initial onset, *empurra* 'shove'

/ĩj.(ˈpu.hə)/	PARSE-σ w=0.0	MPARSE w=11.1	ONSET w=11.9	MAX w=1.8	$\mathscr{H}$	p
ĩj.(ˈpu.hĩj.)(ˈpu.hə)	*		*	*	-13.7	.07
⊙		*			-11.1	**.93**

Although the ONSET constraint is introduced to vowel-initial inputs, the output is still

most likely to be the Null Parse candidate, with a probability of approximately 93%. However, there is now a 7% chance that the reduplicated form would be the candidate that deletes a segment of the base to avoid a vowel cluster.

Forms in which the base ends in a vowel and the reduplicant begins with a vowel are avoided in Brazilian Portuguese. We can look at verbs with a different initial vowel and final vowel, like [ĩj.ˈpu.hə]. If the base vowel is deleted, we should expect [ĩj.ˈpu.hĩj.ˈpu.hə]. If the reduplicant vowel is deleted, we expect [ĩj.ˈpu.həj.ˈpu.hə]. If the vowels are merged, we should expect something similar to [ĩj.ˈpu.hẽj.ˈpu.hə]. The front, central vowel, [ẽ], is the mid-point between [i] and [ə], but since this is not the observed vowel in the reduplicated form, we can assume that the high, front [i] and the central [ə] are not merging. The data presents the reduplicated form as [ĩj.ˈpu.hĩj.ˈpu.hə].

3.4 Verbs with more than three syllables

Although verbs with more than three syllables are not observed in the reduplication pattern of Portuguese, the probabilities of the candidates for an input can still be predicted. If the input contained a word-initial onset, it is most likely that the output would be null. If the input did not contain a word-initial onset, it is most likely that the output would be null as well, but there would be a small chance that the output could be a phonetically represented form with a deleted segment from the base. This is most similar to trisyllables due to the separation of the size restriction constraint into FT-BIN and PARSE-σ.

4 A Comparison to Reduplication in Cuban Spanish

The Portuguese phenomenon discussed here is similar to a pattern found in Cuban Spanish, which also nominalizes verbs. The resulting noun has a meaning that indicates a repetitive action of the verb (Lederer 2003). For example, *come* ‘eat’ becomes *come-come* ‘a lot of eating’.

Compared to the Cuban Spanish pattern, the reduplication process in Brazilian Portuguese is less restricted in regards to syllable constraints on the length of the base.

Tableau 4 shows the ranking of constraints in Spanish.

Table 8: Verb reduplication in Cuban Spanish, *escucha* ‘listen’

/es.ˈku.tʃa/	PARSE-σ	FT-BIN	MPARSE
es.(ˈku.tʃes.)(ˈku.tʃa)	*!		
(es.ˈku.tʃes.)(ˈku.tʃa)		*!	
☞ ⊙			*

The PARSE-σ constraint will reject trisyllabic stems with the condition that all syllables must belong to a foot, and the FT-BIN constraint mandates that there can not be more than two syllables in that foot. PARSE-σ and FT-BIN are ranked higher than the MPARSE constraint, so although the null output, ⊙, violates the MPARSE constraint, it is the winner. The ranking of the constraints in Spanish is PARSE-σ, FT-BIN >> MPARSE.

Table 9: Verb reduplication in Brazilian Portuguese, *agarra* ‘grab’

/ə.ˈga.hə/	MPARSE	FT-BIN	PARSE-σ
☞ ə.(ˈga.hə.)(ˈga.hə)			*
(ə.ˈga.hə.)(ˈga.hə)		*!	
⊙	*!		

The ranking of the two constraints in Brazilian Portuguese is the inverse of Spanish: MPARSE >> FT-BIN >> PARSE-σ. Although the fully faithful output violates the size restriction, it will win over the unpronounced candidate.

Although the variance in ranking can account for the difference of outputs between Cuban Spanish and Brazilian Portuguese, it cannot explain why the output in Portuguese is sometimes null for monosyllables and trisyllables and sometimes pronounced. Therefore, I utilize a Maxent framework to give the constraints weights and portray the probabilities of each possible output.

5 Conclusion

In conclusion, there is a size restriction on reduplicated words in Brazilian Portuguese that can be represented by two constraints: FT-BIN and PARSE-σ, with PARSE-σ having a

lower weight than FT-BIN. Consonant-initial verbs are more acceptable as bases, except in the case of trisyllables when vowel-initial bases are more acceptable.

Our understanding of the process of verb reduplication in both Brazilian Portuguese and Cuban Spanish would benefit from extension beyond corpus-based studies to an investigation into well-formedness judgments by native speakers. Of particular interest is the fact that this pattern is not as productive as other forms of nominalization and is extremely restricted in a predictable way. Since vowel-initial monosyllables and disyllables, and consonant-initial trisyllables were not observed in the corpora, an experiment utilizing words from those categories could provide more insight to reduplication phonology.

REFERENCES

ALUÍSIO, SANDRA; PELIZZONI, JORGE; MARCHI, ANA RAQUEL; DE OLIVEIRA, LUCÉLIA; MANENTI, REGIANA and MARQUIAFÁVEL, VANESSA. 2003. An Account of the Challenge of Tagging a Reference Corpus for Brazilian Portuguese. In *International Workshop on Computational Processing of the Portuguese Language,* 110-117.

ARAÚJO, GABRIEL. 2002. Truncamento e reduplicação no Português Brasileiro. *Revista de Estudos da Linguagem*, *10*(1).

BARBOSA, PLÍNIO A. and ELEONORA C. ALBANO. 2004. Brazilian Portuguese. *Journal of the International Phonetic Association*, *34*(2), 227-232.

DOWNING, LAURA J. 2006. *Canonical Forms in Prosodic Morphology*. Oxford University Press.

FONESCA, ERIK ROCHA and JOÃO LUÍS G. ROSA. 2013. *Mac-morpho revisited: Towards robust part-of-speech tagging.*

FONSECA, ERICK ROCHA; JOÃO LUÍS G ROSA; and SANDRA MARIA ALUÍSIO. 2015. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. *Journal of the Brazilian Computer Society*.

GOUSKOVA, MARIA. 2007. The reduplicative template in Tonkawa. *Phonology*, *24*(3), 367-396.

HAYES, BRUCE and COLIN WILSON. 2006. *Maxent Grammar Tool* (Computer Software). http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/: UCLA.

ITO, JUNKO; YOSHIHISA KITAGAWA; and ARMIN MESTER. 1996. Prosodic faithfulness and correspondence. *Journal of East Asian Linguistics*, *5*, 217-194.

KAGER, RENÉ. 2001. Rhythmic Directionality by Positional Licensing. In *Fifth HIL Phonology Conference (HILP 5).* Handout.

MCCARTHY, JOHN J. 1982. Prosodic Structure and Expletive Infixation. *Language*, *58*(3), 574-590.

MCCARTHY, JOHN J. 2008. *Doing Optimality Theory: Applying Theory to Data*. Wiley-Blackwell.

MCCARTHY, JOHN J. and ALAN PRINCE. 1990. Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language and Linguistic Theory*, *8*(2), 209-283.

MCCARTHY, JOHN J. and MATTHEW WOLF. 2009. Less than Zero: Correspondence and the null output. In (p. 17-66). Equinox Publishing.

PRINCE, ALAN S. 1980. A Metrical Theory for Estonian Quantity. *Linguistic Inquiry*, *11*(3), 511-562.

PRINCE, ALAN and PAUL SMOLENSKY. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Wiley-Blackwell.

R DEVELOPMENT CORE TEAM. 2008. R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`

SEMPERE, ANTONIO GARU. 2006. *Conflicting quantity patterns in Ibero-Romance prosody* (Unpublished doctoral dissertation). University of Texas at Austin.

TANG, KEVIN. 2012. A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research. *UCL Work Pap Linguist*, *24*, 208-214.