Prosodic end-weight effects in Malay echo reduplication: modeling the role of naturalness and lexical attestedness

Background: Echo reduplication involves copying of a word with some minor alternation, such as a change in consonant (e.g., *helter-skelter*) or a change in vowel (e.g., *pitter-patter*). It often respects prosodic end-weight, whereby the prosodically heavier constituent tends to come second. Several factors have been shown to contribute to prosodic end-weight (Ryan, 2019:193). Across languages, (i) more syllables, (ii) less sonorous onsets, (iii) lower vowels, (iv) closed syllables, (v) more sonorous codas and (vi) onsetful syllables induce prosodic end-weight (preferred in final position) more than fewer syllables, more sonorous onsets, higher vowels, open syllables, less sonorous codas and onsetless syllables.

Motivation: Malay echo reduplication is theoretically interesting because both typologically natural and unnatural statistical patterns coexist in the lexicon. Table I below summarizes the lexical statistics by number of forms and some examples in Malay showing the natural and unnatural trends. Figure 1 gives the prosodic end-weight tendencies in percentages calculated using the formula: no. of natural forms / (no. of natural forms + no. of unnatural forms) × 100%. As seen in the figure, the onset sonority and vowel height factors are unnatural (i.e., prosodic end-weight tendency < 50%), whereas the other factors are natural (i.e., prosodic end-weight tendency > 50%). Besides, prosodic factors can differ in terms of their lexical attestedness, i.e., the number of forms instantiating a particular trend in the lexicon. For instance, even though vowel height and onset sonority are both unnatural factors, the strength of evidence in the lexicon is stronger for the former than for the latter (129 forms vs. 44 forms).

Previous findings: Using a binary forced-choice nonce-probe test, XXX found that learners of Malay largely matched the statistical trends in the lexicon but with some interesting divergences (Figure 2). First, natural patterns were learned better than unnatural patterns when lexical attestedness was kept constant. To illustrate, the discrepancy between the lexicon and the experiment for the natural syllable count factor (|92% − 74%| = 18%) is smaller than the one for the unnatural onset sonority factor (|7% − 41%| = 34%), suggesting that the former is learned better than the latter. Second, the equally unnatural (but highly attested) vowel height factor was well-learned, as the subjects' responses closely matched the statistical distribution in the lexicon. Third, not all natural patterns in the lexicon were well-learned. For the poorly attested coda sonority and onset size factors, the subjects' responses clustered around the 50% chance level baseline, which suggests that they were not able to internalize these natural trends.

Modeling: The present paper aims to provide learning simulations for the experimental results couched in Maximum Entropy (MaxEnt) Harmonic Grammar (Goldwater & Johnson, 2003), using the entire corpus of 244 Malay echo-reduplicated forms as training data. In order to test the effects of naturalness and lexical attestedness, four different learning models were constructed (cf. Table II). They differed in whether the input frequencies used in the training data were raw count (the numbers given in Table I) or proportion (the numbers given in Figure 1), and whether a uniform or substantive bias was implemented. Bias was implemented in the models using a Gaussian prior defined over each constraint in terms of a mean (μ) and a standard deviation (σ) (Wilson, 2006; White, 2017). All the constraints employed the $X \rightarrow \varphi_s$ format proposed in Ryan (2019:204), where X can be any phonological element, φ denotes a prosodic node at or above the Prosodic Word, and *s* indicates a strong prosodic node. A natural constraint prefers a prosodically heavier element to be set in a phrasally strong position, in this case final position (e.g., low vowel $\rightarrow \varphi_s$), whereas an unnatural constraint prefers a prosodically lighter element to be found in the same position (e.g., high vowel $\rightarrow \varphi_s$). A substantively biased model had different σ values for each constraint depending on naturalness (natural: $\sigma = 0.6$, a weaker prior; unnatural: $\sigma = 0.2$, a stronger prior). On the other hand, a uniformly biased model had the same σ value across all constraints ($\sigma = 0.4$). μ was kept at 0 for all models. The simulation results are given in Table II. All the models were assessed by how well their predictions fit the experimental results. As seen in Table II, the substantively biased raw count model (Model D) has the highest r^2 and log likelihood value, outperforming the other models that lack either or both of the key components. Thus, the learning simulations suggest that both sensitivity to lexical attestedness and substantive bias are crucial and are implicated in the learning of statistical trends in Malay echo reduplication.

Table I: Lexical statistics in raw counts with examples of natural and unnatural forms for each prosodic factor, where σ: syllable, C: consonant, V: vowel, R: sonorant, T: voiceless obstruent, Hi: high vowel, Lo: low vowel, Ø: empty coda/onset.

Prosodic factor		**Natural**		**Unnatural**	
(i)	Syllable Count	35	gelap-gelita (σσ-σσσ)	3	saudara-mara (σσσ-σσ)
(ii)	Onset Sonority	3	lauk-pauk (RV-TV)	41	piut-miut (TV-RV)
(iii)	Vowel Height	40	mandi-manda (Hi-Lo)	89	warna-warni (Lo-Hi)
(iv)	Coda Size	8	cucu-cicit (CVØ-CVC)	5	gerak-geri (CVC-CVØ)
(v)	Coda Sonority	4	sorak-sorai (CVT-CVR)	1	geliang-geliat (CVR-CVT)
(vi)	Onset Size	3	inca-binca (ØV-CV)	0	-- (CV-ØV)

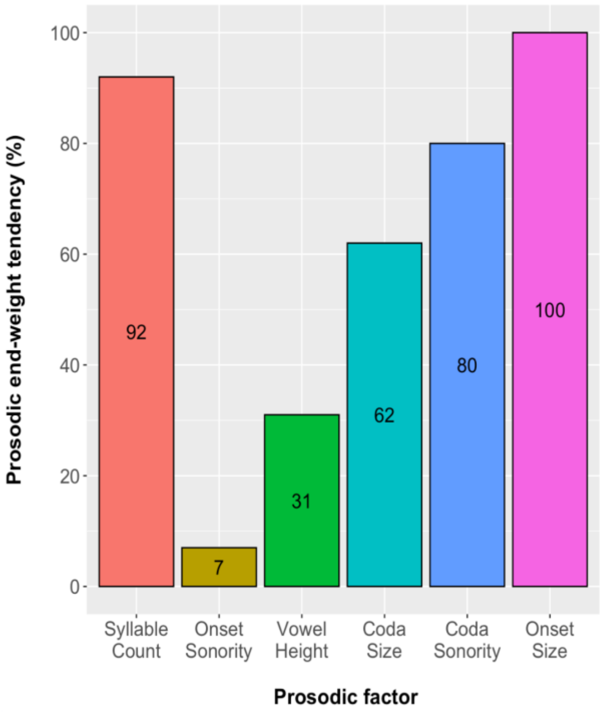

Figure 1: Lexical statistics in percentages

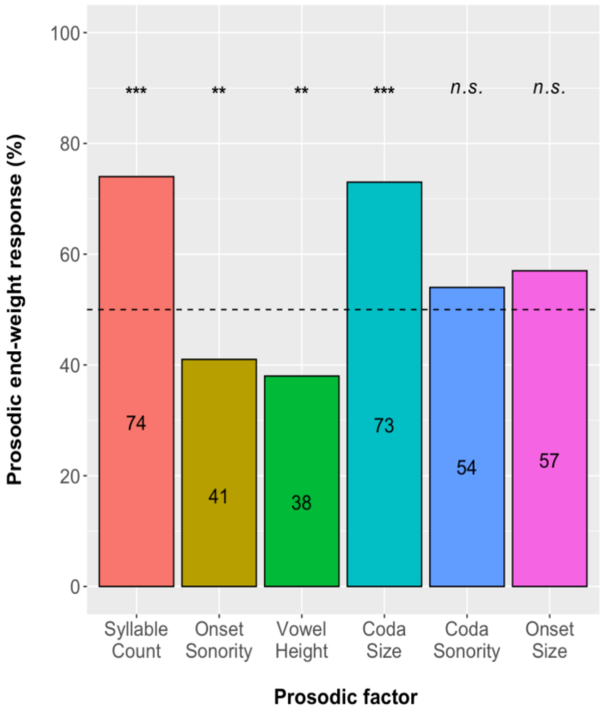

Figure 2: Experimental results from wug testing

Table II. r^2 and log likelihood values for all MaxEnt models.

Model	**Naturalness**	**Lexical Attestedness**	r^2	**Log likelihood**
A	Uniform (–)	Proportion (–)	.51	−488.9
B	Substantive (+)	Proportion (–)	.51	−494.9
C	Uniform (–)	Raw count (+)	.60	−472.3
D	Substantive (+)	Raw count (+)	.76	−464.3