

Markedness in Kaytetye Reduplication: An Information-Theoretic Analysis

Forrest Panther¹, Mark Harvey²

¹The New Zealand Institute of Language, Brain & Behaviour, University of Canterbury, New Zealand

²University of Newcastle, Australia

forrest.panther@canterbury.ac.nz; mark.harvey@newcastle.edu.au

Abstract

Kaytetye is an Arandic language spoken in Central Australia. Kaytetye has a partial reduplication pattern, in which the reduplicant is a disyllable. Surprisingly, the reduplicant in this form is required to be vowel-initial, and consonant-initial forms are ungrammatical. We show through quantitative analysis of the Kaytetye lexicon that the form of the reduplicant results from the distribution of information: vowel-initial forms are less informationally complex than consonant-initial forms, and an informational ‘floor’ effect constrains the minimal size. These results support an analysis of markedness that it results from information optimization within a language, not from universal patterns of markedness.

Index Terms: markedness, phonology, information, phonotactics, Australian languages, reduplication

1. Introduction

Markedness is defined as a factor that makes a construction or form more complex or rare, resulting in it being restricted in distribution, or even ungrammatical [1-4]. Hume [5] identifies two broad approaches to markedness in the literature:

- *Universal Markedness:* Markedness results from universal patterns of complexity and simplicity. Therefore, ‘unmarked’ and ‘marked’ patterns are identical for all languages. Markedness is a property of Universal Grammar, as framed in formal approaches to phonology [1, 3, 6].
- *Language-Specific Markedness:* Markedness results from communication optimization; in other words, it emerges from speaker-specific and language-specific phenomena [4, 7-11]. Markedness in phonology, in part, results from phonotactics and the distribution of information in that language, which affect these patterns [5, 12]

Under *Universal Markedness*, languages will conform to universal patterns of markedness, while under *Language-Specific Markedness*, markedness patterns emerge from the structuring of information in the specific language. This latter approach, of course, predicts recurring patterns across languages, as languages generally share many common properties, and the facts of human psychology and biology will result in similarities in markedness patterns across languages. However, it also predicts that there will be occasional patterns that are consistent with patterns in that language, but not with patterns common in language generally. Evidence of the existence (or non-existence) of these patterns allows these two analyses of markedness to be empirically testable.

2. Reduplication in Kaytetye

2.1. Markedness in Reduplication

Markedness is a factor in the formation of partial reduplicants [6]. For example, in Nookta (Wakashan, Pacific North-West), a reduplication pattern selects the initial consonant of a word and the following vowel: *čičims’ih* [13]. The reduplicant is the initial syllable of the base, apart from the coda of that syllable, i.e. the reduplicant is *čit*, not *čim*. The analysis is that CV syllables are unmarked while CVC syllables are marked, as evidenced by syllable typology [10, 14].

Partial reduplicants also typically show a fixed size. In the Nookta reduplication pattern, the shape is CV; CVC is ungrammatical. These shapes typically correspond to “templates”, such as a syllable, or a prosodic foot [15]. Earlier models of reduplication proposed that the size of the reduplicant is fixed and not reducible to any other factors [16-17]; some recent models have also maintained this general approach [18]. However, more recent approaches, such as Generalized Template Theory, view the size of the reduplicant as being constrained by markedness, just like other constraints on the content of the reduplicant [19-21].

2.2. Kaytetye Verbal Reduplication

Kaytetye verbal reduplication occurs in an Associated Path (henceforth AP) construction, a set of constructions that associate a path to a predicate [22-24]. Examples (1) & (2) illustrate a sentence that occurs with and without an Associated Path construction with the root /kwaṭə-/ ‘drink’. When in an AP construction, as in (2), the verb /kwaṭə-/ receives a ‘participial suffix’ /j(ə)/, that indicates that the action (in this case drinking) occurs after the path. It is then followed by a ‘path auxiliary’ /alpə-/ , which indicates a path back to a location, with the meaning of the overall construction meaning ‘drink when getting back’.

- 1) /acə aŋju kwaṭə-ṗə/
1SG.ERG water drink-PST.PFV
“I drank water.”
- 2) /acə aŋju kwaṭə-j+alpə-ṗə/
1SG.ERG water drink-after+return-PST.PFV
“I drank water when I got back.”

Most Associated Path constructions have Path Auxiliaries that correspond to a lexical verb -- /alpə-/ means ‘go back’ (compare English ‘have’ as both a lexical verb and an auxiliary verb). However, the most frequent Associated Path construction has a reduplicant as the Path Auxiliary. In this construction, the verb

occurs with the participial suffix /-lp(ə)/, which means ‘during’. The reduplicant indicates a path without any particular direction away or from a place, which is translated as ‘(on) the way’, and consequently the meaning of the construction is ‘do X on the way’.

The reduplicant in this Associated Path construction may take two forms. When the verb root in the construction is a consonant-initial monosyllable (henceforth CV), the reduplicant is a total copy of the verb root. A constructed example is shown in (3) involving the verb root /pu-/ ‘cook’.

- 3) /ɬə a:rə pu-lpə+pu-ŋə/
 3SG.NOM kangaroo cook-during+RED-PST.PFV
 “He cooked a kangaroo on the way.”

When the verb root is larger than CV (i.e. at least disyllabic), the reduplicant has the shape of a vowel-initial disyllable (henceforth VCV). Constructed examples based on attested forms, are shown in (4)-(6).

- 4) /ɬə a:rə alarə-lp+arə-ŋə/
 3SG.NOM kangaroo kill-during+RED-PST.PFV
 “He killed a kangaroo on the way.”
- 5) /acə aŋtu kwatə-lp+atə-nə/
 1SG.ERG water drink-during+RED-PST.PFV
 “He drank water on the way.”
- 6) /ərmicinə akuwɪncə-lp+ɪncə-ŋə/
 dust rise-during+RED-PST.PFV
 “The dust rose.”

In these constructions, the content of the reduplicant is copied from the right edge of the verb root base. Note, however, that the reduplicant does not copy the onset of the penultimate syllable, and to do so would produce an ungrammatical form. This requirement is typologically non-standard: if the shape of reduplicants is conditioned by markedness, then while they may show obligatory onsets, they will not typically show vowel-initial conditions. This is because syllables with onsets are analyzed as less marked than syllables without onsets [10, 14]. This then raises the question of whether there is a motivation for this pattern within Kaytetye itself, and by extension, the *Language-Specific Markedness* analysis, as opposed to the *Universal Markedness* analysis. The goal of this paper is to show that, in fact, this pattern is predicted under this *Language-Specific* approach.

2.3. Hypotheses

The *Language-Specific Markedness* analysis of Kaytetye reduplication hypothesizes that the forms are selected due to their lack of complexity. That is, CV and VCV forms are less complex than CVCV forms. We hypothesize that this follows from existing patterns in the Kaytetye lexicon.

The monosyllabic and disyllabic shapes are motivated by the minimal information content of short forms. However, we hypothesize that the predominance of disyllabic forms relates to an information ‘floor’ effect. While the shape of the reduplicant is informationally simple, it shows an informational minimality effect, in which any productive smaller form would be too simple to be informative as a reduplicative auxiliary verb.

The literature on gradient phonotactics agrees that speakers use the statistics of types in their mental lexicon to

develop phonotactic intuitions [25-31]. Consequently, patterns relating to phonological complexity will be inferable from statistics in the Kaytetye lexicon. Based on the shape of the Kaytetye reduplicants, we have three hypotheses about the distribution of roots in the Kaytetye lexicon:

- H1. There will be a positive correlation between: (i) the number of syllables; (ii) the initial phonotactics (i.e. whether the root is consonant or vowel-initial) of lexical items, and their overall complexity.
- H2. The effect of initial phonotactics on the complexity of lexical items will be present independent of the number of segments overall in an item. In other words, the average complexity of the configurations in an item will still show the effect of initial phonotactics.
- H3. There will be distributional evidence that monosyllabic roots are informationally deficient in a way disyllables are not.

3. Method

3.1. *kRoot*, a Database of Kaytetye Word Roots

In order to carry out an analysis of the Kaytetye lexicon, it is necessary to have a representative list of Kaytetye lexemes. Consequently, we developed a set of Kaytetye word roots from headwords in the *Kaytetye-to-English Dictionary* [32].

From the list of all the Kaytetye headwords, we removed any items that had their listed category as an affix or a clitic. We removed multi-word expressions, and forms with a hyphen. Headwords with no assigned part of speech were also removed. From the remaining set, we used replacement procedures using regular expressions to remove any transparent morphology, and the resulting set was hand corrected for any possible errors in the remaining data. From this remaining set, items that were judged to be word roots met one of two criteria: (i) there was no evidence of morphological complexity in the item; (ii) the item appeared complex, but its semantics were not transparently derived from these components. An example of the latter category is the word *akwerrepenhe* /akurəpəŋə/, a word that appears to be transparently ‘from the coolamon’ (i.e. the word /akurə/ ‘coolamon’ in the sequential case), but means ‘small baby’. Finally, semantically transparent roots in complex headwords that were not independent entries in the dictionary were added to the database as independent roots.

This process resulted in a set of 2,762 word roots belonging to seven parts of speech: noun, verb, pronoun, demonstrative, preverb, coverb, and adverb. The dataset was then transformed from Kaytetye orthography into an IPA representation using regular expressions. This IPA representation made use of the four-vowel phonemic analysis of [23].

3.2. Complexity Scores

The *kRoot* dataset was used to train a bigram model. For the modelling, root boundary characters (represented as ‘#’) were added to indicate the beginning and the end of the root. The bigram in each root type was then retrieved. For example, for a root type /aləkə/, boundary characters were added: #aləkə#. This root then contributed six bigrams: #a, al, lə, ək, kə, ə#. This process was conducted for every root type in the database, and the full list of bigrams was retrieved.

The score calculated for each unique bigram is the positive log of its conditional probability, which is used in research to model the phonotactics of word forms [25, 28, 33-35]. The

conditional probability of the bigram was calculated as the frequency of the bigram in the set of roots in the database divided by the frequency of the initial unigram. For example, the conditional probability of *al* is the equivalent of the frequency of *al* divided by the frequency of *a*. When this conditional probability was retrieved, the negative base-2 log probability was derived to produce a “bits per phoneme” measure as an estimate for the complexity of the bigram. The formula used to calculate the score of the bigram *al* is summarized in (7).

$$7) \text{ score}(al) = \log_2\left(\frac{\text{Count}(al)}{\text{Count}(a)}\right)$$

We used this bigram model to calculate two measures of the complexity of word roots in the *kRoot* dataset, both of which were used in the data modelling. A *sum complexity score* is the sum of the complexity scores of the bigrams in a word root. The *average complexity score* is the mean complexity score of the bigrams in a word root.

3.3. Modelling Procedure

In order to determine the effect that root size and root-initial phonotactics have on the sum and mean complexity of Kaytetye word roots, we created two linear mixed effects regression models using the *lme4* package [36] in R [37]. **Model 1** uses the sum complexity of roots as the response variable, while **Model 2** uses the average complexity of roots as the response variable.

In developing these models, we considered two fixed factors: (i) number of syllables in the root (1 - 9 syllables); (ii) whether the root begins with a vowel (True/False). Random intercepts considered were: (i) the part of speech of the root; (ii) the quality of the final vowel (data exploration showed that this has a significant effect on the complexity of the root). Linear modelling utilizes the *kRoot* dataset, excluding 8 and 9 syllable roots due to the absence of consonant-initial forms at these sizes.

4. Results

Descriptive Statistics Table 1 summarizes the distribution of the *kRoot* word roots by number of syllables and whether they are consonant-initial or vowel-initial. Roots of 8 and 9 syllables are excluded. Only 24 roots are monosyllables, which is less than 1% of the set of word roots. All but two of these are consonant-initial, and the two vowel-initial forms are homophones: /a:/, which can mean ‘fight’, or ‘entrance’. The smallest size of root with a significant proportion of the lexicon is the disyllables, with 560 items, or 20.3% of the set of roots.

For every root size apart from monosyllables and disyllables, the number of vowel-initial forms is higher than the number of consonant-initial forms. Vowel-initial forms are quantitatively predominant overall: of the 2,762 roots in the *kRoot* dataset, 1,705 (61.7%) are vowel-initial.

Figure 1 provides the mean sum and average complexity scores for *kRoot* roots, categorized by their syllable count and root-initial phonotactics. The vowel-initial monosyllable category is excluded from this figure, which has only one root type (two roots with the form /a:/). It is, however, included in the modelling.

First, in support of Hypothesis H1, these results show that there is the expected positive correlation between the size of word roots and their sum complexity. Second, in support of

Syllable Count	Vowel-Initial	Consonant-Initial
1	2	22
2	217	343
3	756	458
4	507	182
5	162	47
6	47	2
7	11	3

Table 1: Distribution of Kaytetye Word Roots by Syllable Count and Root-Initial Phonotactics

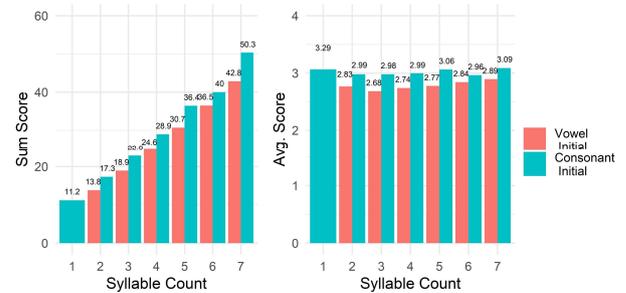


Figure 1: Mean Sum and Average Complexity Scores By Syllable Count and Root-Initial Phonotactics

Hypothesis H2, there is an effect of being vowel-initial on the average complexity in word roots. However, there is not an obvious effect of word root size in the average complexity score. The exception to this is in the monosyllable category, which shows the highest mean average complexity score of all the categories presented here. One possible interpretation of this is that it results from a combination of the small sample size of this category, and the low number of bigrams in comparison to other categories. In this instance, a comparatively small number of infrequent phonotactic configurations can inflate the score in this category, while in categories with more bigrams these configurations will be averaged out. This interpretation is unlikely given the fact that other categories with low numbers of bigrams (i.e. disyllables) and categories with low counts (such as 7-syllables) show numbers consistent with the rest of the average complexity counts. Instead, it is likely these numbers reflect more on-average complex configurations in monosyllables. This possibility is supported by other aspects of the distribution of monosyllabic roots, that point to a preference for contexts with more information. That is, a vast majority of monosyllabic forms either occur in contexts where there is morphology to provide context for their interpretation, or they are function words that are frequent in discourse. Of the 24 monosyllables, 12 (50%) are verbs, which in Kaytetye always occur with a verbal suffix. Only four (20%) are nouns, and two of these nouns are homophones that consist of the rare long vowel /a:/. Four CV roots are pronouns, and two are demonstratives – both are closed classes. The other two are coverbs, which always occur with a following verb. In other words, monosyllables are over-represented in classes where there is context for their interpretation. Compare this distribution with disyllables, in which 335 (59.8%) are nouns and only 112 (20%) are verb roots.

All these observations are consistent with the notion of a type of complexity ‘floor’, in which forms below this floor show patterns that increase their salience. These observations support Hypothesis H3.

Model 1				
Fixed Effect	Estimate	t value	p value	Sig.
Intercept	2.5	1.2	0.42	
Initial Segment – Cons.	3.5	7.9	<0.001	***
Syllables	5.9	70.4	<0.001	***
Init. Segment – Cons. : Syl.	0.3	1.9	0.054	
Model 2				
Fixed Effect	Estimate	t value	p value	Sig.
Intercept	2.9	8.5	0.072	
Initial Segment – Cons.	0.4	8.9	<0.001	***
Syllables	0.05	5.4	<0.001	***
Init. Segment – Cons. : Syl.	-0.03	-2.2	<0.05	*

Table 2: Summary of Fixed Effects for Model 1 (Sum Complexity) & Model 2 (Average Complexity)

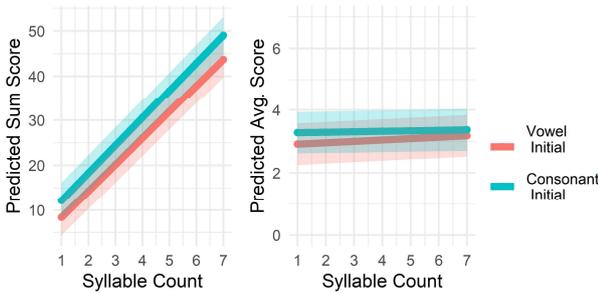


Figure 2: Predicted Sum and Average Complexity Score in Model 1 (left) & Model 2 (right) by Syllable Count and Initial Phonotactics.

4.1. Model Results

In order to confirm the findings relating to Hypothesis H1 and Hypothesis H2 in the distribution of complexity scores, we conducted data modelling on the sum (Model 1) and mean (Model 2) complexity scores of word roots. The resulting models included syllable count and whether the root is vowel-initial as fixed effects, including an interaction term between these two factors.

A summary of the fixed effects of Model 1 & Model 2 are shown in Table 2. In all, Model 1 shows the expected positive effect of vowel-initial forms and syllable count on the sum complexity score, meeting the predictions of Hypothesis H1. Model 2 shows the effect of vowel-initial roots on average complexity scores, confirming that the predictions of Hypothesis H2 are met. Surprisingly, Model 2 also shows an effect of syllable count, as well as an interaction between vowel-initial roots and syllable count. This effect is observable in Figure 2: vowel-initial roots show a positive correlation between average complexity score and syllable count, while this trend is almost non-existent for consonant-initial roots. We propose that this effect relates to the comparatively high complexity of consonant-initial monosyllables, that was described above. In other words, the comparatively high complexity of monosyllables largely negates the otherwise apparent trend of increasing average complexity of roots by their syllable count. These statistical results confirm the predictions of Hypothesis H1 & H2, and are consistent with the prediction of Hypothesis H3.

5. Discussion

The results presented in this paper show that the form of the Kaytetye verbal reduplicant is predictable based on the information content of the Kaytetye lexicon. The vowel-initial

condition is based on the fact that consonant-initial forms are more informationally complex than vowel-initial forms. The disyllabic default of the reduplicant is based on the low information content of monosyllables, showing an apparent ‘floor’ effect. This is consistent with a *Language-Specific Markedness* analysis of markedness.

The focus of this paper has been presenting evidence in favour of a *Language-Specific Markedness* analysis to markedness in Kaytetye reduplication, rather than presenting evidence against the *Universal Markedness* analysis. However, we noted briefly in §2.2 that the vowel-initial condition is inconsistent with the evidence of markedness and syllable typology, which favours consonant-initial forms rather than vowel-initial forms [10, 14]. VC reduplicants appear in other languages. For example, Lushootseed exhibits a reduplication pattern with a VC shape: $\text{ʔaxid} \rightarrow \text{ʔax-ix-əd}$ [20]. However, the reduplicant attaches after C2 in the base, and this C2 provides an onset for the reduplicant when the overall Base + Reduplicant word is syllabified. Consequently, the universal preference for onsets is satisfied.

This might also appear a plausible analysis of the Kaytetye reduplication. For example, in the construction in (4) alarə-lp+arə-ŋə ‘kill-during+RED-PST.PFV’, the participial suffix could appear to be underlyingly consonant-final, i.e. $/-lp/$. If it was underlyingly consonant-final, then this could license a following V-initial reduplicant because the participial suffix would be available to supply an onset. However, all morphemes in Kaytetye are V-final, and the final vowel of the participial suffix surfaces when there is a following C-initial morpheme, as in (8).

- 8) $/acŋu-lpə=lk+acŋu-jə$ $ɪə/$
 descend-during=then+RED-FUT 3SG.NOM
 ‘Then he will go down.’ (Example from [32])

The appearance of the consonant-final form of the participial suffix is due to vowel hiatus resolution, which deletes the first of two adjacent vowels, and occurs in other aspects of Kaytetye phonology [24]. Thus, (4) is underlyingly $/alarə-lpə+arə-ŋə/$, but the hiatus sequence $/ə+a/$ between the participial suffix and the reduplicative auxiliary root is not permitted. Hiatus reduction is a post-lexical, rather than a lexical process, and so will not satisfy onset maximization at a lexical level.

In this paper we showed that the facts of Kaytetye verbal reduplication are explicable under a *Language Specific Markedness* analysis, while there are significant problems with a *Universal Markedness* analysis. These results necessarily raise questions for the motivations for reduplication patterns generally, and whether alternative approaches are necessary.

6. References

- [1] Rice, K. (2007). Markedness in phonology. In P. De Lacy (Ed.). Cambridge University Press.
- [2] Hume, E. (2011). Markedness. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The blackwell companion to phonology*. Wiley-Blackwell. <https://doi.org/10.1002/9781444335262>
- [3] Kager, R. (1999). *Optimality theory*. Cambridge university press.
- [4] Hume, E. (2016). Phonological markedness and its relation to the uncertainty of words. *Phonological Studies. Journal of the Phonological Society of Japan*, 19, 107–116.

- [5] Hume, E. (2006). Language specific and universal markedness: An information-theoretic approach. *Linguistic Society of America Annual Meeting: Colloquium on Information Theory and Phonology*.
- [6] McCarthy, J., & Prince, A. S. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In M. González (Ed.), *Proceedings of the north east linguistic society* 24.
- [7] Hume, E. (2003). Language specific markedness: The case of place of articulation. *Studies in Phonetics, Phonology & Morphology*.
- [8] Hume, E. (2008). Markedness and the language user. *Phonological studies*, 11.
- [9] Comrie, B. (1986). Markedness, grammar, people, and the world. *Markedness* (pp. 85–106). Springer.
- [10] Blevins, J. (2006b). Syllable: Typology. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (second). Elsevier.
- [11] Bybee, J. (2010). Markedness: Iconicity, economy, and frequency. In J. J. Song (Ed.), *The oxford handbook of linguistic typology*. Oxford University Press.
- [12] Hume, E. (2004). Markedness: A predictability-based approach. *Annual Meeting of the Berkeley Linguistics Society 30: General Session and Parasession on Conceptual Structure and Cognition in Grammatical Theory*, 182–198. <https://doi.org/10.3765/bls.v30i1.948>
- [13] Stonham, J. T. (1990). *Current issues in morphological theory* (Doctoral dissertation). Stanford University.
- [14] Zec, D. (2007). The syllable. In P. De Lacy (Ed.), *The cambridge handbook of phonology* (pp. 161–194). Cambridge University Press.
- [15] McCarthy, J., & Prince, A. S. (1995). Faithfulness and reduplicative identity. *Papers in Optimality Theory*, 10. <https://doi.org/10.7282/T31R6NJ>
- [16] McCarthy, J. (1979). *Formal problems in semitic phonology and morphology* (Doctoral dissertation). MIT. Cambridge, MA.
- [17] Marantz, A. (1982). Re reduplication. *Linguistic inquiry*, 13(3), 435–482.
- [18] Saba Kirchner, J. (2010). *Minimal reduplication* (Doctoral dissertation). University of California, Santa Cruz.
- [19] Urbanczyk, S. (1996). Morphological templates in reduplication. *North East Linguistics Society*, 26(1).
- [20] Urbanczyk, S. (2006). Reduplicative form and the root-affix asymmetry. *Natural Language & Linguistic Theory*, 24(1), 179–240.
- [21] McCarthy, J., & Prince, A. S. (1986). *Prosodic morphology* [Unpublished Manuscript, University of Massachusetts, Amherst & Brandeis University].
- [22] Koch, H. (1984). The category of ‘associated motion’ in kaytej. *Language in central Australia*, 1(1), 23–34.
- [23] Panther, F., & Harvey, M. (2020). Associated path in Kaytetye. *Australian Journal of Linguistics*, 40(1), 74–105. <https://doi.org/10.1080/07268602.2019.1703644>
- [24] Panther, F. (2021). *Topics in Kaytetye phonology and morpho-syntax* (Doctoral dissertation). University of Newcastle
- [25] Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In J. Coleman (Ed.), *Computational phonology. third meeting of the ACL special interest group in computational phonology* (pp. 49–56).
- [26] Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of memory and language*, 40(3), 374–408.
- [27] Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. *Phonetic interpretation: Papers in laboratory phonology vi* (pp. 58–74). Cambridge University Press.
- [28] Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36(3), 481–487.
- [29] Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in english and arabic. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. John Benjamins.
- [30] Richtsmeier, P. T. (2011). Word-types not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology*, 2, 157–183.
- [31] Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., & Needle, J. (2020). Non-Māori-speaking new Zealanders have a Māori proto-lexicon. *Scientific reports*, 10(1), 1–9.
- [32] Turpin, M., & Ross, A. (2012). *Kaytetye to English dictionary*. IAD Press.
- [33] Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: The case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3), 859–896.
- [34] Pimentel, T., Roark, B., & Cotterell, R. (2020). Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8, 1–18.
- [35] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., & Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1), 31–40.
- [36] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2021). Package ‘lme4’. R package version.
- [37] R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.