STRESS PREDICTORS IN A PAPUAN MALAY RANDOM FOREST

Constantijn Kaland[1], Nikolaus P. Himmelmann[1], Angela Kluge[2]

[1]Institute of Linguistics, University of Cologne, Germany
[2]SIL International
{ckaland, sprachwissenschaft}@uni-koeln.de, angela_kluge@sil.org

ABSTRACT

The current study investigates the phonological factors that determine the location of word stress in Papuan Malay, an under-researched language spoken in East-Indonesia. The prosody of this language is poorly understood, in particular the use of word stress. The approach taken here is novel in that random forest analysis was used to assess which factors are most predictive for the location of stress in a Papuan Malay word. The random forest analysis was carried out with a set of 17 potential word stress predictors on a corpus of 1040 phonetically transcribed words. A complementary analysis of word stress distributions was done in order to derive a set of phonological criteria. Results show that with two phonological criteria the stress location for almost all words in the corpus could be explained.

Keywords: word stress, Papuan Malay, prosody, phonology, random forest.

1. INTRODUCTION

Little is known about the prosody of Papuan Malay and recent research has only begun to investigate this language empirically. An elaborate grammar as the result of years of fieldwork is to date the most comprehensive description [11]. Many languages in Indonesia are only described in grammars and often subject to impressions of (Western) authors. Therefore, more empirical research is needed to advance our understanding of the variation in Indonesian languages and beyond. As for Papuan Malay, its prosody is not yet fully understood. For example, prosodic prominence is reported as irrelevant concept at the phrase level [19], although there are clear acoustic differences supporting the difference between stressed and unstressed syllables ([9],[10]). A novel statistical approach is reported here to reveal the extent to which phonological factors play a role in word stress placement.

Word stress in Papuan Malay occurs mostly on the penultimate syllable, except when this syllable contains /ɛ/ [11]. For example, in /tɛ.'kan/ (‘to press’) stress moves to /kan/ because /ɛ/ could reduce to schwa, which cannot be stressed. Thus, in /'tu.kan/ (‘craftsman’) stress occurs on the default penultimate syllable. This distribution makes Papuan Malay similar to other Trade Malay varieties [16] where stress has been reported to be mostly penultimate; e.g. Manado [25], Kupang [23], Larantuka [12] and the North Moluccan varieties Tidore [30] and Ternate [14]. Similarly, varieties of Indonesian have been analysed as having penultimate stress. Work on Indonesian has shown the importance of distinguishing language varieties. Where Toba Batak listeners rated manipulated stress cues as less acceptable when these did not occur on the penultimate syllable, Javanese listeners had no preference whatsoever [4].

As for the Trade Malay varieties, Ambonese appeared to not make use of word stress [15], counter to earlier claims [29]. That is, in an acoustic analysis and a re-analysis of the vowel inventory, allegedly minimal stress pairs turned out to be rather segmentally different. Recent work on Papuan Malay, however, found consistent acoustic support for the word stress claims in duration, formant displacement and spectral tilt ([9],[10]). Substantial differences between the Trade Malay varieties are likely, given that they are spoken in areas far apart from each other. Thus, since the contact with Malay in periods of trading, the different varieties underwent their own developments. Papuan Malay in particular can be considered a relatively young Trade Malay variety, which originated from Ambonese Malay possibly around the fifteenth century. Papuan Malay mainly started spreading to larger areas in (West-)Papua during the colonial period in the second half of the 18th century [11].

Although there is acoustic evidence for regular penultimate stress in Papuan Malay ([9],[10]), more research is needed on the phonological nature of word stress. In particular, the exact role of /ɛ/ in stress distributions is unclear. As noted in [11], /ɛ/ “does not condition ultimate stress”. Among the words which have /ɛ/ in the penultimate syllable more than a third has penultimate stress. Nevertheless, in almost all words with ultimate stress (~10% of the Papuan Malay words in [11]), /ɛ/ appears in the penultimate syllable. Therefore, the current study explores whether there are more factors that affect the mobility of stress in Papuan Malay. By analysing phonological properties of syllables, the role of word stress in the prosody of

Papuan Malay can be better understood. This investigation is carried out using a random forest analysis and a distribution analysis.

Random forest analysis is a classification method based on the construction of a large number of decision trees [3]. In order to assess which variable splits (classifies) the data best, trees are constructed on the basis of random data- and variable-subsets. Random forests are particularly useful to determine the predictive value of a large set of variables and a small number of observations. Compared to other statistical methods, random forests are better able to account for overfitting and collinearity between predictors. Random forest analyses have only recently been introduced into the field of linguistics [28], and phonetics and phonology (e.g. [5], [1], [6], [2]). The method is promising as notions such as prominence, stress or phonological weight tend to correlate with a large number of acoustic and/or linguistic variables. Random forests could help to reveal underlying mechanisms of linguistic structure, by providing powerful generalizations based on a relatively small set of data from the field. The predictive value of a certain variable in a random forest is expressed by means of variable importance. The absolute variable importance values are irrelevant, as they are randomly generated (hence random forest). Therefore, the interpretation of variable importance generally relies on the relative differences between the respective values [22].

A random forest analysis is carried out in the current study to investigate which factors determine the mobility of word stress in Papuan Malay. The following sections report how this analysis was carried out (section 2), which phonological criteria could be derived (section 3) and how the results can be interpreted (section 4).

2. METHODS

In addition to the random forest analysis a distribution analysis of word stress location was done. Both were performed on a corpus of Papuan Malay words. The distribution analysis allows to assess how well the most predictive factors divide the corpus into penultimate and ultimate stress.

2.1. Corpus

The corpus in [11] provided phonetic transcriptions of spoken Papuan Malay words, including indications of word stress, word class and English gloss. No frequency data was available for the words in the corpus. For the purposes of this study, only words classified as Papuan Malay roots were selected (Appendix A.1, [11]), excluding the large number of loanwords in this language. In this way, potential influences from stress patterns originating from other languages were avoided. Furthermore, the corpus consisted of two-syllable words only to obtain a homogeneous set (words with one syllable (N = 46) or more than two syllables (N = 73) were relatively infrequent). Thus, the representativeness of the corpus was compromised to a minimal extent. An overview of the number of words per word class is given in Table 1. Note that words which translate to adjectives in English are expressed by stative verbs in Papuan Malay. For example, /bɛ.'sar/ (‘big’ - litt. ‘be big’) is labelled as verb in the corpus.

Table 1: Distribution of word classes in the corpus

Word class	Count	Word class	Count
V(erb) other	7	Adverb	32
V bi(valent)	341	Noun	355
V mono stative	205	Function (all)	49
V mono dynamic	51	Total:	1040

2.2. Predictors

Of interest to the current analysis are phonological factors that make a syllable likely to be stressed. From the literature /ɛ/ is known to be realized as schwa in Trade Malay varieties [16]. As for its phonology, schwa has lower sonority compared to most other vowels. Indeed, analyzing syllables in terms of their sonority levels can explain stress placement crosslinguistically (e.g. [17]). Also syllable structure can affect the sonority of a syllable. Although vowel nuclei are often most determining, in some languages the onset or coda of the syllable affect its sonority (e.g. [8]).

Therefore, in the current study a set of predictors was chosen (in italic) that potentially affect syllable sonority. These included the syllable *structure* in terms of consonantal and vowel segments, from which the *openness* of the syllable and the actual segments in the *onset*, *nucleus* or *coda* were derived. Papuan Malay has five vowels (/a/, /ɛ/, /ɔ/, /i/ and /u/) and 17 consonants (stops: /p/, /b/, /t/, /d/, /k/, /g/; affricates: /tʃ/, /dʒ/; nasals: /m/, /n/, /ŋ/; fricatives: /s/, /h/; rhotic: /r/; approximants: /l/, /j/, /w/). The predictor *manner* of articulation was derived from the actual segments, with plosives at the low end and open vowels on the high end of the sonority scale. Furthermore, *word class* was included which in some languages correlates with word stress placement (e.g. in English: “permit” (noun) and “to permit” (verb) form a minimal stress pair).

2.3. Statistical analysis

The analysis was done in R [18] using the package “ranger” [31], which offers a computationally less intensive way to perform random forests compared to packages such as “party” [26] or “randomForest” [13]. The response variable in the random forest was stress location (2 levels: penultimate, ultimate). The predictors were syllable structure (6 levels: CCV, CCVC, CV, CVC, V, VC), onset (18 levels: /b/, /tʃ/, /d/, /g/, /h/, /dʒ/, /k/, /l/, /m/, /n/, /ŋ/, /p/, /r/, /s/, /t/, /w/, /j/, no onset), nucleus (5 levels: /a/, /ɛ/, /i/, /ɔ/, /u/), coda (12 levels: /j/, /k/, /l/, /m/, /n/, /ŋ/, /p/, /r/, /s/, /t/, /w/, no coda), openness (2 levels: open, closed), manner of articulation in onset/coda (6 levels: plosive, fricative, nasal, rhotic, approximant, no onset/coda), manner of articulation in nucleus (3 levels: open, mid, close) and word class (7 levels: see Table 1). In addition, a control-predictor gloss (the English translation of each word) was added. Gloss is not expected to be of any predictive value and therefore should have a low variable importance. Except for word class and gloss all predictors were included for both the first and second syllable in the word (total: 18 predictors).

The number of trees in the analysis was increased in steps of 1000, starting from 1000 trees. The variable importance of the factors reached a stable ranking around 5000 trees. To obtain a robust result, the final number of trees was set to 10000 [22]. The number of randomly preselected predictors was set to the square root of the total number of predictors in the analysis (√18), following the method in [27].

The distribution analysis (Table 2) consisted of counts; 1 for each word with penultimate stress and counting 0 for each word with ultimate stress. The ratio of penultimate/ultimate stresses was then calculated by taking the average of all counts. The two analyses combined appeared particularly helpful to interpret the variable importance values, as their absolute values are not indicative (section 1).

3. RESULTS

Two factors stand out as predictors for the location of stress in Papuan Malay (Figure 1): the nucleus in the first syllable (nucleus_1) and the manner of articulation of the nucleus in the first syllable (nucleus_MAN_1). Other predictors showed considerably lower variable importance values, although the manner of articulation of the nucleus in the second syllable (nucleus_MAN_2) as well as the nucleus of the second syllable (nucleus_2) appeared more predictive than the lowest ranked ones. Given the hypothesized irrelevance of control predictor gloss (ranked 12/18), predictors with similar or lower ranking have little to no predictive value. Indeed, from the fifth ranked predictor (w_class) onwards the variable importance values hardly vary (and yield 0) compared to higher ranked ones.

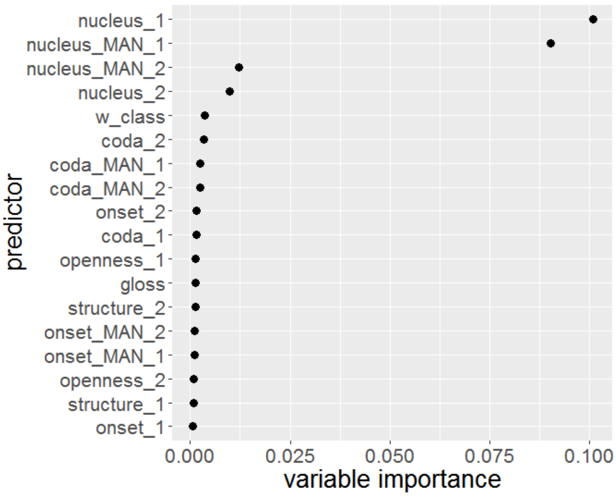

Figure 1: Variable importance plot with the predictors (1 for first syllable, 2 for second syllable) ranked from high (top) to low (bottom).

Table 2: Penultimate/ultimate stress ratio for the four most predictive factors in the random forest analysis (n = nucleus, MAN = manner of articulation, 1 = first syllable, 2 = second syllable).

n	MAN	n_1	n_MAN_1	n_MAN_2	n_2
/a/	open	1.00	1.00	.86	.86
/ɛ/	mid	.37	.57	.95	.94
/ɔ/		1.00			.95
/i/	close	.99	.99	.92	.93
/u/		.99			.90

The predictor nucleus in the first syllable showed the lowest ratio of penultimate stresses for /ɛ/ (Table 2). All other vowel nuclei in the first syllable were mostly stressed. Manner of articulation of the nucleus in the first syllable showed the lowest ratio for mid vowels. This was a reflection of the effect of /ɛ/, as stress was always penultimate when /ɔ/ was the nucleus of the first syllable. Note however, that /ɔ/ occurs in only 11% of the words in the corpus [11]. Thus, the variable importance of manner of articulation was mainly a reflection of the effect of /ɛ/ rather than /ɔ/. As for the ratios of manner of articulation of the nucleus in the second syllable, the lowest ratio of penultimate stress cases was obtained for open vowels. The highest ratio obtained for mid vowels indicates that stress was mostly penultimate when the nucleus of the second syllable was a mid-vowel. Note that /a/ is the only open vowel in Papuan Malay, explaining why nucleus_MAN_2 and nucleus_2 had similar effects (Figure 1).

To phonologically explain the ultimate stress cases, three criteria were formulated (Table 3). First, ultimate stress is mainly found when /ɛ/ occurred in the first syllable, confirming [11]. The three exceptions to this criterion are /ki.'tɔŋ/ (1 pl.), /ku.'mur/ (‘rinse mouth’) and /kus.'kus/ (‘cuscus’), see also [11] (p. 96). Note that /ki.'tɔŋ/ is short for /ki.'tɔ.raŋ/ ([1], p. 326), which has penultimate stress (from /'ki.ta/ and /'ɔ.raŋ/, litt. ‘us humans’). Re-evaluation of /kus.'kus/ showed that it could be analysed as Malay loanword [20], indicating that its inclusion in the corpus might not have been justified.

Table 3: Word counts after applying criterion that decreased the penultimate stress ratio / increased the ultimate stress ratio (Table 2). Exceptions = ultimate stress cases not following the criterion.

Criterion	Penult	Ult	Exceptions
Total	932	108	-
n_1 = /ɛ/	61	105	/ki.'tɔŋ/ /ku.'mur/ /kus.'kus/
n_MAN_2 ≠ mid	25	100	/tʃɛ.'rɛj/ /sɛ.'rɛj/ /dʒɛ.'lɛk/ /dʒɛm.'pɔl/ /sɛ.'dɔt/
n_2 = /a/	16	65	…

Second, 61 words had /ɛ/ in the first syllable and penultimate stress. From these words, 36 had a mid-vowel (/ɛ/ or /ɔ/) in the second syllable. Five exceptions to this criterion had ultimate stress, with /ɛ/ in the first syllable and a mid-vowel in the second syllable; /tʃɛ.'rɛj/ (‘to divorce’), /sɛ.'rɛj/ (‘lemongrass’), /dʒɛ.'lɛk/ (‘be bad’), /dʒɛm.'pɔl/ (‘thumb’), /sɛ.'dɔt/ (‘to suck’). Note that [ɛj] in /tʃɛ.'rɛj/ and /sɛ.'rɛj/ is analysed as realisation of underlying /aj/ due to the liquid in the onset of the second syllable ([11], p.84). With /ɛ/ in the first syllable, underlying /a/ could make the second the preferred syllable for stress. The status as native root of /dʒɛm.'pɔl/ and /sɛ.'dɔt/ is doubtful, given their report as Javanese/Sundanese loanwords ([7],[24]).

Third, the presence of an open vowel (/a/) in the second syllable increases the likelihood of ultimate stress. However, from the words with /ɛ/ in the first syllable and /a/ in the second syllable, 65 had ultimate stress. Given that there were 108 ultimate stress cases in total (Table 3), the open vowel in the second syllable did not predict stress placement as strongly as the first two criteria. In other words, the open vowel in the second syllable was of minor importance and could only explain a small additional number of stress cases after the main criteria were applied. This result is reflected in the large variable importance difference between the first two predictors and the lower ranked ones (Figure 1).

4. DISCUSSION

The results are best summarized by assuming that the default position of word stress in Papuan Malay is the penultimate syllable. When the penult contains /ɛ/, stress shifts to the ultimate syllable only when the ultimate does not contain a mid-vowel. This result indicates that /ɛ/, and in the ultimate syllable also /ɔ/, generally reject stress, although 25 words had stress on a penultimate syllable that contained /ɛ/. Furthermore, /a/ attracted stress to a limited extent, although it did not predict a stress shift.

The results are in line with the literature on Trade Malay with respect to the role of /ɛ/ (schwa) in stress placement ([11],[16]). Furthermore, the role of /a/ as stress attractor is compatible with phonological accounts that distinguish open and close vowels as more and less sonorous respectively [21]. Note, however, that the infrequently stressed mid-vowels in Papuan Malay cannot be explained on the basis of openness as main correlate of vowel sonority. The results rather support a minimal and universally adopted version of the sonority hierarchy [17].

This study has shown that random forests provide an insightful analysis of which phonological factors play a role in stress placement. It is worth stressing that without the complementary distribution analysis (Table 2), the role of the predictors was difficult to interpret. Moreover, the direction of the effect of the most predictive factors in the random forest analysis could be understood when interpreting the stress ratios. The predictive power of the random forest analysis is particularly clear from the relatively small number of exceptions with ultimate stress (*N* = 8) after applying the first two criteria in Table 3. In fact, the analysis revealed that three of these words were loanwords, which should not have been included in the corpus. For another three words alternative explanations could be found, indicating that their stress pattern was not necessarily counter to the phonological criteria (section 3). As for the exceptions with penultimate stress after applying the first criteria (*N* = 25), we cannot provide alternative explanations or additional criteria that explain why stress did not move to the ultimate syllable in these cases. Nevertheless, 25 of the 932 penultimate stress cases and 8 of the 108 ultimate stress cases constitute less than four percent of all words in the corpus. Although additional phonological criteria could theoretically be derived from the remaining highest ranked predictors in the random forest analysis, these have the risk of generating more exceptions than explained cases (Table 3).

5. ACKNOWLEDGEMENTS

Research for this paper was funded by the German Research Foundation (DFG), SFB1252 “Prominence in Language”, project A03 “Prosodic prominence in cross-linguistic perspective”, University of Cologne.

REFERENCES

[1] Arnold, D., Wagner, P., & Baayen, R. H. (2013). Using generalized additive models and random forests to model prosodic prominence in German. In F. Bimbot et al. (Eds.), *Proceedings of Interspeech 2013*, pp. 272–276.

[2] Baumann, S. & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics* 70, 20-38.

[3] Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32.

[4] Goedemans, R.W.N. & Van Zanten, E. (2007). Stress and accent in Indonesian. In V. J. van Heuven & E. van Zanten (eds.), *Prosody in Indonesian languages*, 35–62. Utrecht: LOT.

[5] Grafmiller, J. & Shih, S. (2011). New approaches to end weight. *Proceedings of Variation and Typology: New trends in Syntactic Research*, 25–27. Helsinki, Finland.

[6] Grice, M., Savino, M., Caffo, A., & Roettger, T. B. (2015). The tune drives the text - Schwa in consonant-final loanwords in Italian. *Proceedings of the 18th International Congress of Phonetic Sciences*, 10-14. Glasgow, UK.

[7] Haspelmath, M. & Tadmor, U. (2009). *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wold.clld.org, accessed on 2018-11-26.

[8] Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: The University of Chicago Press.

[9] Kaland, C.C.L. (2018). Spectral tilt as a correlate of Papuan Malay word stress. *Proceedings of Speech Prosody 2018*, 339-343. Poznan, Poland.

[10] Kaland, C.C.L. (2019). Acoustic correlates of word stress in Papuan Malay. *Journal of Phonetics, 74*, 55-74.

[11] Kluge, A. (2017). *A grammar of Papuan Malay*. Studies in Diversity Linguistics 11. Berlin: Language Science Press.

[12] Kumanireng, Th. Y. (1993). *Struktur Kata dan Struktur Frasa Bahasa Melayu Larantuka* [Word structure and Phrase Structure in Larantuka Malay]. Ph.D. dissertation, University of Indonesia.

[13] Liaw, A. & Wiener, M. (2002). Classification and Regression by RandomForest. *R News* 2, 18-22.

[14] Litamahuputty, B. (2012). *Ternate Malay: Grammar and texts*. LOT Dissertation Series 307. Utrecht: LOT.

[15] Maskikit-Essed, R. & Gussenhoven, C. (2016). No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology* 33, 353-389.

[16] Paauw, S. H. (2008). *The Malay contact varieties of Eastern Indonesia: A typological comparison*. PhD dissertation. Buffalo: State University of New York.

[17] Parker, S. (2002). *Quantifying the sonority hierarchy*. Ph.D. dissertation, University of Massachusetts, Amherst.

[18] R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Computer program, v. 3.4.0, https://www.r-project.org, retrieved 21-04-2017.

[19] Riesberg, S., Kalbertodt, J., Baumann, S., & Himmelmann, N. P. (2018). On the perception of prosodic prominences and boundaries in Papuan Malay. In S. Riesberg, A. Shiohara, & A. Utsumi (eds.), *A cross-linguistic perspective on information structure in Austronesian languages*. Berlin: Language Science Press.

[20] Scott, C.P.G. (1897). The Malayan Words in English. *Journal of the American Oriental Society*, 18, 49-124.

[21] Selkirk, E. (1984). On the Major Class Features and Syllable Theory. In M. Aronoff and R. Oehrle (Eds.), *Language Sound Structure*, 107-136, Cambridge: MIT Press.

[22] Shih, S. (2013). *Random Forests, for Model (and Predictor) Selection*. Unpublished course material. UCLA 251: Variation in Phonology. Downloaded from: http://www.bcf.usc.edu/~shihs/shih_random forests.pdf

[23] Steinhauer, H. (1983). Notes on the Malay of Kupang. In J. T. Collins (ed.), *Studies in Malay dialects: Part II* (NUSA – Linguistic Studies of Indonesian and other Languages in Indonesia 17), 42–64. Jakarta: Badan Penyelenggara Seri NUSA, Universitas Katolik Atma Jaya.

[24] Stevens, A. M., & Schmidgall-Tellings, A. E. (2010). *A comprehensive Indonesian-English dictionary*. Athens, OH: Ohio University Press.

[25] Stoel, R. B. (2007). The intonation of Manado Malay. In V.J. van Heuven & E. van Zanten (Eds.), *Prosody in Indonesian Languages*, 117-150. Utrecht: LOT.

[26] Strobl, C., Hothorn, T. & Zeileis, A. (2009). Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. *The R Journal*, 1(2), 14-17.

[27] Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rational, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods* 14(4), 323-348.

[28] Tagliamonte, S., & Baayen, R.H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2), 135-178.

[29] Van Minde, D. (1997). *Malayu Ambong: phonology, morphology, syntax*. PhD dissertation, Leiden University.

[30] Van Staden, M. (2000). *Tidore: A Linguistic Description of a Language of the North Moluccas*. Ph.D. dissertation, Leiden University.

[31] Wright, M. N. & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77, 1-17.