Laboratory Phonology
Journal of the Association for Laboratory Phonology

Kaland, C. (2024). When recall gets stressful: comparing Papuan Malay and German listeners' lexical storage of word stress. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 15(1), pp. 1–27. DOI: https://doi.org/10.16995/labphon.11695

Open Library of Humanities

When recall gets stressful: comparing Papuan Malay and German listeners' lexical storage of word stress

Constantijn Kaland, Institute of Linguistics / Phonetics, University of Cologne, Germany, ckaland@uni-koeln.de

Recent studies indicated that Papuan Malay, spoken in the Eastern provinces of Indonesia, has regular penultimate word stress. Only when schwa occurs in the penultimate syllable, stress is ultimate, making the pattern highly predictable. Acoustic, lexical and perception studies showed that these stress patterns offer disambiguating cues that Papuan Malay listeners can use in word recognition. It is however poorly understood to what extent these predictable patterns are stored lexically in this language, and in *fixed* stress languages in general. The current study investigates this question by means of stress recall tasks comparing Papuan Malay with German, the latter being analyzed as a *free* (lexical) stress language. To this end, a critical review of the previous literature is given regarding the methodological comparability of stress recall tasks across languages. The tasks used in the current study replicate the stimuli and procedures used in three previous studies. Results show that Papuan Malay listeners have worse stress recall performance than German listeners, suggesting that Papuan Malay stress is not stored lexically. The outcomes are discussed with respect to the diagnosis of word stress in Papuan Malay and its typological context.

1. Introduction

The diagnosis of word stress in Indonesian languages has been a challenging research topic due to the wealth of language diversity in the area and the poor availability of empirical and quantitative studies (e.g., Odé, 1994; Goedemans & van Zanten, 2014; Himmelmann, 2023). Recent experimental studies have crucially extended this line of research and either countered or supported impressionistic claims on the existence of word stress. A particularly interesting case in this context is test comment an Eastern Indonesian Trade Malay variety that has received the most attention in quantitative studies on word stress so far. It was claimed impressionistically that this language has regular penultimate word stress, and ultimate word stress when the penultimate syllable has a schwa (Kluge, 2017). Acoustic, perceptual and lexical studies confirmed this claim, showing that spontaneous Papuan Malay speech provides clear acoustic cues to word stress that listeners can use in word recognition (Kaland, 2019, 2020, 2021; Kaland et al., 2021).

The conclusions for Papuan Malay so far are intriguing from a psycholinguistic perspective. That is, studies have shown that listeners generally do not need to store highly regular and predictable stress patterns in their mental lexicon as attested by poor performance in recall tasks across languages (e.g., Peperkamp et al., 2010). In a gating task, however, Papuan Malay listeners were shown to be able to guess well above chance level from which word a gated syllable was taken (Kaland, 2021). That syllable crucially contained stress information as the only cue to know from which word it was taken. These results seem to indicate that listeners used lexically stored stress information to successfully identify the words, despite the high predictability of stress in their language.

The case of Papuan Malay is important for our understanding of word stress, a notion that has been often defined and approached from the perspective of well-studied languages, such as English, Dutch or German. English stress in particular was reported to be poorly representative for Western-Germanic languages. This problem has been addressed in the literature and led to an explicit call for the inclusion of more *fixed-stress* languages in the study of word stress perception (e.g., Cutler, 2005). In order to provide the crucial cross-linguistic context for the results, the current study compares Papuan Malay to German, the latter being a language with weight-sensitive *free* stress (e.g., Domahs et al., 2014). Although the current study is not the first on word stress in Papuan Malay, it offers a crucial addition to the previous work: It investigates to what extent Papuan Malay listeners are able to recall stress sequences in a task that has been applied cross-linguistically in previous research. In this way, the results in this line of research are comparable among each other and allow for embedding in a typological context. Specifically, the results offer a novel, nuanced perspective on the diagnosis of word stress in underdescribed languages, which was traditionally centered around the existence of minimal stress pairs in

Indonesian languages, allowing only a narrow (lexical) definition of word stress (e.g., Odé, 1994; Maskikit-Essed & Gussenhoven, 2016; Kaland et al., 2021 for discussions and examples).

The remainder of this section discusses the state of the art of stress research on Papuan Malay (Section 1.1), thereafter a literature overview is given of highly similar recall tasks that were carried out on different languages (Section 1.2). The extent to which these tasks are actually comparable across studies and languages is discussed in Section 1.3. Finally, Section 1.4 outlines the research question and hypotheses of the current study.

1.1. Papuan Malay stress in production and perception

An acoustic study investigated several temporal, spectral and amplitudinal cues to word stress in spontaneously produced Papuan Malay monologues from 19 speakers (Kaland, 2019). The analysed words were all disyllabic, which matches the most common word length in this language, and were taken from nonfinal positions in the intonation unit (IU). It has been shown that IU-final words exhibit rather exceptional prosodic features, i.e., the largest f0 movements, final lengthening, etc. (Kaland & Baumann, 2020). Results showed that duration per phoneme (syllable duration divided by number of phonemes) was the strongest acoustic correlate of word stress (stressed = longer), followed by vowel displacement (stressed = more peripheral), and spectral tilt (stressed = shallower intensity roll-off towards higher frequencies). All these cues also showed larger acoustic differences between stressed and unstressed syllables for ultimate stress compared to penultimate stress.

The role of the acoustic cues in word recognition was tested in a series of experiments in which listeners guessed which word matched with the hummed and acoustically manipulated syllable sequence they were hearing (Kaland, 2020). The acoustic manipulations rendered the stress pattern in the disyllabic sequences either as penultimate or ultimate using the acoustic values for each of the cues as found in production (Kaland, 2019). The participants' task was to match the manipulated sequences with the correct stress pattern of the word. Results showed higher correctness scores were obtained for patterns presented in isolation than for those in a phrase, and higher ones for ultimate stress than for penultimate stress. The latter result reflected the stronger acoustic marking of ultimate stress compared to penultimate stress in the production analysis in Kaland (2019). However, scores were overall just above chance level, indicating a minimal benefit from the cues when tested separately (53%-60% correct).

Two lexical analyses (Kaland et al., 2021) investigated Papuan Malay word stress on the basis of a representative word list of native roots (a "lexicon" of approximately 1000 words from Kluge, 2017). First, it was investigated whether stress information could help to disambiguate between words. This was done by counting word embeddings, i.e., words that are embedded in longer ones. For example in English, *bee* is an embedding in words such as *belay* and *beanie*.

When stress information is taken into account, *bee* (stressed) is only an embedding in *beanie*, matching the initial stressed syllable, and no longer in *belay*, not matching the initial unstressed syllable (e.g., Cutler et al., 2004; Cutler & Pasveer, 2006). The amount of reduction in word embeddings as a result of taking stress information into account is an indication of the extent to which stress information can facilitate word recognition. Thus, listeners could discard potential word candidates when taking stress information into account, with more reduction in the number of embeddings indicating a larger degree of facilitation. It was found that in Papuan Malay, the number of embeddings could be reduced to the same extent as in English and to a lesser extent than in Spanish, Dutch, or German. In a second lexical analysis, the phonological factors determining the position of stress in the word were investigated. This was done by comparing more than 20 variables relating to the segmental and syllabic makeup of the words. It was confirmed that /ɛ/ in the penultimate syllable, which reduces to schwa at the surface, is the main reason for stress to shift to the ultimate position. Additionally /ɔ/ was found to reject stress in the ultimate syllable as well (i.e., no shift from penultimate to ultimate). Taken together, these two phonological rules could be subsumed under the rejection of stress by mid vowels (/ɛ/ and /ɔ/ being the only mid vowels in Papuan Malay; Kluge, 2017), together explaining more than 96% of the stress patterns in the wordlist.

In order to investigate the extent to which Papuan Malay listeners could actually make functional use of the word stress cues in word recognition, a gating task was carried out (Kaland, 2021). Gating tasks have been used in word recognition experiments by presenting word parts of increasing length to listeners (e.g., Cotton & Grosjean, 1984). For Papuan Malay for example, the syllable /bɛ/ was presented to listeners in a task where they had to indicate whether that syllable was taken from ['bɛ.bɛk], in which it was stressed, or from [bɛ.'ban], in which it was unstressed. As control gates, listeners were presented either with a single phoneme [b] (chance level responses expected) or with the entire word (100% correct expected). The task presented the target word written on the screen and listeners indicated which out of two auditory gates matched the target word. Listeners thus heard a gate from a stressed syllable *and* from an unstressed one before giving their response. Results showed that listeners could correctly identify the target word in more than 80% of the cases based only on the syllable, indicating that they had successfully used the acoustic cues to stress in this task (i.e., well above chance level).

A crucial remaining question from the latter study is whether listeners were successful because they made use of lexically stored stress information, or whether they could use the acoustic contrast between the two presented gates to match what they heard with the target on the screen. The latter option does not necessarily demand lexically stored stress information, as listeners could retrieve the crucial cues from their auditory working memory, having heard those cues in the experiment. In other words, the gating task did not challenge listeners' memory of stress information. Notably, participants also did not score 100% correct, as in the control gate

presenting the entire word. This indicates that stress information helped listeners, but not to the extent that is expected when that information is unambiguously coupled to the word. It is therefore not entirely clear to what extent Papuan Malay stress patterns are stored lexically. As further outlined in the next section, the idea is that highly predictable patterns do not demand lexical storage. The question is therefore whether Papuan Malay listeners are able to store stress information in a task in which their memory is directly tested.

1.2. Testing lexical stress storage in sequence recall tasks

Discrimination tasks were carried out in which listeners were presented with three auditory stimuli and had the task to decide whether the third stimulus was more similar to the first or the second (henceforth *ABX discrimination tasks*). These showed that French listeners made many more errors distinguishing nonwords that only differ in the position of stress compared to Spanish listeners (Dupoux et al., 1997). The difference was attributed to whether word stress has a contrastive function in these languages. In French, the position is fixed on the final syllable, whereas in Spanish, it is highly variable, giving rise to minimal stress pairs that can only be successfully disambiguated when listeners are able to use their acoustic cues. Note that there are different analyses of the final (word) prominence in French, i.e., it has “fixed stress” (Di Cristo, 1998, p. 196), “no primary word stress at all” (Van der Hulst, 2012, p. 1515), it is a language for which “the need for w-final accents is not hard to establish” (Gussenhoven, 2004, p. 258), or “the lexical representation of a word does not include metrical or tonal properties” (Delais-Roussarie et al., 2015, p. 65). French listeners were reported to exhibit *stress deafness*, i.e., not being literally deaf to the acoustic cues, but unable to use them in word recognition (Dupoux et al., 1997). In the remainder of this paper I therefore refrain from using this term.

The Spanish-French difference in performance in the ABX task was hypothesized to originate from differences in the ability to memorize varying stress patterns. The French listeners were expected to not have such experience from their every day use of their language, whereas Spanish listeners would need to do so all the time. There was, however, considerable overlap in the performance of participants, with some French participants being as successful as Spanish ones and some Spanish participants showing as much difficulty as French ones (Dupoux et al., 1997). A different paradigm was developed in order to test the hypothesized differences more directly in stress memory (Dupoux et al., 2001). This paradigm concerned a recall task of nonce word sequences that differed in phonemes (e.g., ['ku.pi] or ['ku.ti]) or in stress (e.g., ['mi.pa] or [mi.'pa]). Sequences consisted of two to six nonce words, and listeners showed overall more recall difficulty as sequences got longer. For each subsequent experiment in Dupoux et al. (2001) more variation in the stimulus material was added, i.e., synthesized pitch variation in addition to the other acoustic cues, variation in the vowel makeup of the nonce words, and the addition of another speaker producing the nonce words. With each additional type of variation, French

listeners showed worse recall performance for the stimuli that had stress differences, whereas the Spanish listeners' performance was stable throughout the series of experiments. In a final experiment, all acoustic variation was taken away by testing French and Spanish participants using a single token from the stimulus material. The results showed similar performance for both language groups, in that French listeners performed clearly better compared to the previous experiments that had more sources of acoustic variation.

The study explained the results by three perceptual strategies to store incoming speech in the auditory working memory (Dupoux et al., 2001). The first strategy concerns explicit categorization by matching the auditory input with stored representations in the lexicon, which gets more difficult with longer sequences. The second is a mismatch detection mechanism between two recent subsequent units, i.e., whether they were the same or not. The second strategy worked well for the French listeners when the nonce words only differed for the crucial experimental cues (i.e., segmental/stress) and had no other naturalistic acoustic variation (i.e., pitch, vowels, speaker's voice). The third strategy concerns the ability to store phonological representations of stress patterns, which is only available for listeners of a stress language (i.e., Spanish).

Phoneme/stress recall tasks (henceforth *PSRTs*) involving multiple sources of acoustic variation have been reported in several studies since Dupoux et al. (2001). In what follows, the main findings of these studies are summarized and a schematic overview of them is given in **Table 1**. The accuracy in each of the PSRTs is expressed using an *index* proposed in Peperkamp & Dupoux (2002), which subtracts the error rate (percentage) for phoneme contrasts from the error rate for stress contrasts. A critical discussion on the cross-study comparability is given in Section 1.3, providing a summary of the main findings of all discussed studies in **Table 1**.

1.2.1. Peperkamp & Dupoux (2002)

A series of PSRTs was carried out on several languages without contrastive stress to test whether the information needed to extract stress rules affects listeners' ability to recall the patterns. A distinction was made between languages for which phonetic information is sufficient to extract the stress rules (French, Finnish), ones for which phonetic, phonological and function word information is needed (Hungarian) and ones for which content words need to be entirely segmentable (i.e., lexical boundaries need to be known; Polish). Spanish was included as a control language with contrastive stress. Results confirmed that with more information needed to extract the stress rule, listeners are better able to recall stress contrasts, as they were used to match the stress information in the speech signal with their stored representation of the word. Crucially, the outcome of this study shows remarkable differences in how stress is derived from (and stored in) the speech signal within the class of languages that do not have lexically contrastive stress. It thus appears that this class can be successfully subdivided according to the hypothesized stress extraction criteria.

1.2.2. Dupoux et al. (2008)

Given that French listeners had shown the worse recall performance in PSRTs in previous literature, it was tested to what extent French late learners of Spanish would be affected by their inability to perceive stress contrasts. With L1 French and L1 Spanish as control groups, it was shown that French late learners of Spanish patterned with the L1 French group, thus not showing any learning effect in the (in)ability to perceive stress contrasts. This was furthermore interpreted as evidence for the importance of language acquisition in learning stress contrasts. That is, this ability is very difficult to acquire if it has not been learned before the age of 10, as was the case for the French late learners in the study.

1.2.3. Peperkamp et al. (2010)

In another typological comparison of languages using PSRTs, Standard French was distinguished from Southeastern French. In the latter variety, a small number of stress contrasts is available due to the presence of schwa word-finally. The other languages involved were Hungarian, Finnish, Polish and Spanish. Specific attention was given to several aspects of stress: its domain (French: [phonological] phrase; others: word), the need to store stress information with the word in the lexicon (no: French, Polish; yes: Finnish, Hungarian, Spanish), variability of the stress position (invar.: Std. French, Finnish, Hungarian; var.: SE French, Polish, Spanish) and number of lexical exceptions to the stress rules (none: French, Finnish, Hungarian; some: Polish; many: Spanish). The results (**Table 1**) were categorised into three degrees of PSRT performance, distinguished mainly by stress predictability and number of lexical exceptions. Thus, when stress is highly predictable and shows virtually no exceptions, recall is the worst (French, Finnish, Hungarian). When stress is highly predictable, but there are some lexical exceptions, recall is possible to a moderate extent (Polish). When stress is not predictable and has many lexical exceptions, recall performance is the best (Spanish). Thus, stress domain and variability of the stress location seemed to be of secondary importance for the PSRT performance.

1.2.4. Haake et al. (2013)

German children with a typical or impaired (specific language impairment: SLI) language development were tested in a PSRT. Stress contrasts were hypothesized to be particularly problematic for the SLI listeners. To assess the origin of these hypothesized difficulties, an additional auditory processing assessment was carried out in both groups. This assessment concerned the discrimination of duration and intensity contrasts such that listeners needed to identify the longer or louder non-speech sound from a pair. On both acoustic contrasts, SLI children showed significantly lower accuracy than the typically developing ones, and one subgroup of SLI listeners showed particular difficulty processing duration contrasts (SLI1) compared to the other SLI listeners (SLI2). As for the PSRT using sequences of two or four nonce words, two types of

stress contrasts were tested: either between antepenultimate and penultimate stress (['pi.ku.ma] vs. [pi.'ku.ma]) or between penultimate and ultimate stress ([pi.'ku.ma] vs. [pi.ku.'ma]). Overall lower recall performance was shown by either SLI group compared to the typically developing listeners (**Table 1**). Typically developing children furthermore showed a higher recall accuracy for antepenultimate-penultimate contrasts than for penultimate-ultimate contrasts.

1.2.5. Heisterueber et al. (2014)

German adult speakers were tested using sequences of two nonce words, either with a phoneme contrast or with a stress contrast. The phoneme contrast concerned place of articulation ([ku.pa.mi] vs. [ku.pa.ni]) or both place and manner of articulation ([ku.mi.ta] vs. [ku.mi.fa]). The stress contrasts concerned either penultimate-ultimate contrasts (as in Haake et al., 2013) or antepenultimate-ultimate contrasts (['ka.ti.mu] vs. [ka.ti.'mu]). Participants were pretested for their basic auditory processing (pitch, duration, intensity) and basic memorizing performance (letter sequence recall). In the PSRT, recall accuracy as well as brain activity (fMRI) were recorded. The participant groups were divided into two subgroups; ones that performed either above (*good*) or below (*poor*) the median accuracy for the stress contrasts. The results showed that both groups had similar recall performance on the phoneme contrasts. However, only the poor performers showed worse recall for stress contrasts. The good performers showed better recall for stress contrasts than for phonemic ones (**Table 1**). The brain imaging results showed significant differences in regions associated with phonological processing (left middle temporal gyrus) between the two participant groups, further corroborating the observation of considerable individual differences in recall performance between participants.

1.2.6. Correia et al. (2015)

European Portuguese was tested in a PSRT because it has variable stress, vowel reduction and duration as its main cues, and crucially: limited co-occurrence of stress and pitch accents. PSRTs with and without vowel quality as a cue were carried out and showed that the absence of vowel quality worsened listeners' recall performance. The results also showed that in nuclear accented positions, recall was worse than for stress contrasts in postnuclear positions. The latter effect was explained as the result of accentual lengthening interfering with stress lengthening, i.e., listeners perceived stress differences better when duration was unambiguously a cue to stress and not shared with accent. Results showed the importance of prosodic context (acoustic cues, phrase-level accents) for recall performance.

1.2.7. Rahmani et al. (2015)

In a typological study of stress contrast perception in Persian, the role of morphology in stress placement was investigated. Morphological information is needed to extract stress

placement rules, but it is crucially missing in early stages of acquisition and therefore no lexical information needs to be stored for Persian stress, also termed *postlexically contrastive*. Japanese and Dutch were included because tone or stress information, respectively, needs to be stored lexically in order to distinguish words (upper baseline), whereas French and (standard) Indonesian were included because they have highly predictable non-lexical stress or no stress at all, respectively (lower baseline). Results showed that Dutch and Japanese listeners' recall performance was similar and as such clearly better than listeners of Persian, French and Indonesian. Open access to the experimental setup and stimulus materials of Rahmani et al. (2015) was provided and used in subsequent studies (below).

1.2.8. Bruggeman (2020)

Without evidence for stress from speech production in Tashlhiyt Berber and Moroccan Arabic, it was investigated how well listeners in these languages perform in a PSRT, replicating the methodology of Rahmani et al. (2015). Results patterned with those of Persian, French and Indonesian, lending further support for the absence of lexical stress in Tashlhiyt Berber and Moroccan Arabic. It was also found that speakers in the stimulus materials differed in the extent to which they used duration and f0 as stress cues. That is, the Dutch female speaker used a non-standard f0 rise and minimal duration differences to cue stress (cf. the Dutch male speaker). There was, however, no evidence that these differences had biased the outcomes of the studies.

1.2.9. Lialiou et al. (2023)

A PSRT replicating the one in Rahmani et al. (2015) was carried out with bilingual listeners of Maltese and Maltese English, who were either Maltese- or Maltese English-dominant. This was done to test whether Maltese has weight-sensitive stress, a claim lacking consistent and in-depth acoustic investigations. The distinction between the respective dominant languages was made in order to test whether Maltese English-dominant listeners would perform better than Maltese-dominant ones, given that English is generally analysed as a weight-sensitive stress language. Analyses concerned modelling of probabilities of obtaining a correct answer from each Maltese listener group as well as those from Rahmani et al. The results for both listener groups patterned with the worst performing ones in Rahmani et al.: Persian, French and Indonesian. Thus, the outcomes indicated that Maltese listeners do not store stress information lexically. Note that the index measure proposed in Peperkamp & Dupoux (2002) showed a somewhat different outcome when comparing the two studies. The comparability of the PSRTs is further discussed in Section 1.3.

1.2.10. Peperkamp & Brazeal (2023)

The robustness of the inability to perceive stress contrasts in French listeners was tested again using explicit training on the stimuli with a stress contrast, receiving feedback on their performance. There were six training sessions in total, which were spread over multiple days. Another group of participants did not receive training (controls). The PSRT was carried out before and after the training sessions (*pretest* and *posttest*, respectively). Results indicated a small consistent improvement for the posttest in both participant groups, which was explained as a result of task familiarization rather than a genuine ability to discriminate stress patterns. The study thus confirmed the strong resistance in French listeners to recall stress information (e.g., Dupoux et al., 2008).

1.3. Comparability of PSRTs

Note that the reported measures of recall accuracy differed among the discussed studies (**Table 1**). The index proposed in Peperkamp & Dupoux (2002) is based on error rates, which can be straightforwardly obtained from correctness rates (i.e., as percentage: error rate = 100 − correct rate). The index is the result of subtracting phoneme error rates from stress error rates, as a measure of listeners' performance (index = stress error rate − phoneme error rate). Thus, the index takes the errors made for phoneme contrasts as a baseline to express the degree of difficulty listeners had with stress contrasts. This measure thus accounts for baseline differences in the perception of phoneme contrasts across languages. It is in itself intriguing that in the typological studies, using highly comparable PSRT paradigms, baseline (phonemic) error rates differed among some languages (Peperkamp & Dupoux, 2002; Peperkamp et al., 2010; Rahmani et al., 2015). The index thus abstracts over number of items and conditions, and as such seems to be suitable to compare recall accuracy across studies. Note that the index, when showing negative values, also accurately captures that recall performance was better for stress contrasts than for phoneme contrasts (Dupoux et al., 2001; Peperkamp & Dupoux, 2002; Dupoux et al., 2008; Heisterueber et al., 2014). However, the index does not reflect overall differences among listeners, i.e., due to language impairment (Haake et al., 2013). That is, children with poor recall performance for both segmental and stress contrasts show a similar index value compared to the typically developing ones (**Table 1**).

It is also important to note that the studies on Tashlhiyt Berber, Moroccan Arabic, and Maltese (Bruggeman, 2020; Lialiou et al., 2023) concluded that listeners in those languages pattern with the lowest performing ones in Rahmani et al. (2015). However, their index values are in fact similar to the best performing listener groups (Dutch and Japanese) in Rahmani et al. This is a puzzling observation given that identical stimuli and procedures were used across all three studies. Note that none of the three studies reported index values directly, i.e., they applied mixed effects modelling on the correctness rates with, among others, *contrast* (phoneme/stress)

as a factor. It is beyond the scope of this study to apply this type of modelling on the older studies in **Table 1**. For this reason, index values (computed post hoc if not explicitly reported) are chosen to express the results using the same metric for all studies, as they allow for (limited) comparison of the PSRT studies.

The index values suggest that Tashlhiyt Berber, Moroccan Arabic, and Maltese listeners perform better than the French listeners in earlier studies (e.g., Dupoux et al., 2001; Peperkamp & Dupoux, 2002; Dupoux et al., 2008; Peperkamp et al., 2010). Given some controversial reports in the literature on the status of word stress in Tashlhiyt Berber, Moroccan Arabic, and Maltese (e.g., Gordon & Nafi, 2012; Boudlal, 2001; Azzopardi, 1981), the index values in **Table 1** might be open to different interpretations, i.e., falling somewhat in between languages that maximally rely on lexically stored stress information (e.g., Spanish) and languages that do not at all (e.g., French). At the same time, French listeners in Peperkamp & Brazeal (2023) show lower index values than in any of the previous studies involving French participants. Given the high comparability of the methods applied to French participants, it is reasonable to assume that there are multiple sources of variation affecting the index values across studies. Within-study comparisons of the index values are likely to be more reliable than across-study comparisons.

Thus, the studies discussed in the previous sections are comparable to a limited extent, although they all made use of a paradigm that tested both phonemic and stress contrasts using recall tasks. There are potentially multiple aspects of the studies that limit their comparability. For example, stimulus materials were generally adapted to the language in order to guarantee that they did not match existing words or to test specific additional prosodic aspects such as type of stress pattern (Haake et al., 2013; Heisterueber et al., 2014) and phrase position (Correia et al., 2015). Furthermore, studies differed in their experimental setup in that the number and type of training rounds prior to the PSRT varied (cf. Dupoux et al., 2001, Heisterueber et al., 2014, Rahmani et al., 2015). Sequence length is likely another source of variation affecting the index values in **Table 1**. It is, however, unclear how exactly sequence length affects the average index values. One could hypothesize that with longer sequences (e.g., 5 or 6 nonce words), language differences come out stronger in the index values as listeners have higher error rates for longer sequences (i.e., large range of index values) compared to shorter sequences (e.g., ≤ 4 nonce words), but this pattern does not hold across the studies (see **Table 1**). By averaging over sequence lengths, the index value does not capture the steepness of the error rate increase when sequences get longer. This rate might also vary across languages, as can be observed in the results reported in Rahmani et al. (2015). Another source of variation in the index values might come from phrase prosodic expectations of the listeners. These vary largely between the studied languages, even more so when taking into account how they might have interacted with word prosodic ones. The PSRT involves an item familiarisation phase and, given the length of the items, they are likely stored as lexical items comparable to words. However, presented in

a sequence and with varying prosodic patterns, listeners' recall performance could have been influenced by the prosodic expectations of larger units (i.e., phrases). In fact, phrase prosodic context (nuclear, postnuclear) was shown to affect European Portuguese listeners' recall scores in a PSRT (Correia et al., 2015). Furthermore, some studies reported a high level of individual variation across listeners' performance (Peperkamp et al., 2010; Heisterueber et al., 2014; Lialiou et al., 2023; also discussed in Dupoux et al., 1997), which might be a reflection of general recall abilities, such as overall memory or auditory processing abilities. Those might naturally vary between participants and could also contribute to the baseline differences in phoneme recall mentioned above.

Study (part)	**Seq. length**	**Language**	**Index**
Dupoux et al. (2001), Exp. 4	2, 3, 4, 5, 6	French	38.8
		Spanish	−7.5
Peperkamp & Dupoux (2002), Exp. 1	2, 3, 4, 5, 6	French	38.1
—, Exp. 1		Finnish	24.0
—, Exp. 2		Hungarian	23.7
—, Exp. 2		Polish	11.6
—, Exp. 1		Spanish	−4.4
Dupoux et al. (2008), Exp. 1	2, 4	French	51.7
		French L1 Spanish L2	50.2
		Spanish	−4.5
Peperkamp et al. (2010)	2, 5	Std. French	50.4
		SE French	40.9
		Hungarian	40.0
		finnish	37.5
		Polish	22.9
		Spanish	1.3

(Contd.)

Study (part)	Seq. length	Language	Index
Haake et al. (2013)	2, 4	German (child typical)	16.1
		German (child SLI1)	35.8
		German (child SLI2)	17.3
Heisterueber et al. (2014)	2	German (poor recall)	14.3
		German (good recall)	−7.4
Correia et al. (2015), Exp. 2, nuclear	2, 5	EU Portuguese	28.0
Rahmani et al. (2015)	3, 4, 5	French	37.3
		Persian	36.1
		Indonesian	35.5
		Dutch	24.4
		Japanese	19.0
Bruggeman (2020), Ch. 8	3, 4, 5	Moroccan Arabic	30.9
		Tashlhiyt Berber	24.6
Lialiou et al. (2023)	3, 4, 5	Maltese English (M dom.)	24.0
		Maltese English (ME dom.)	20.9
Peperkamp & Brazeal (2023), trainees	2, 3, 4, 5, 6	French (pre-test)	28.1
		French (post-test)	26.5

Table 1: Overview of the literature reporting sequence recall tasks with nonce words differing in phonemes or stress. *Index* refers to the error rate (percentage) for phoneme differences subtracted from the error rate for stress differences (higher index = more stress recall difficulty), averaged over all tested sequence lengths. The rows are sorted (1) ascending chronologically per study and (2) descending by index within each study.

1.4. Research question and hypotheses

In spite of the limitations on the strict comparability, it becomes clear from the literature that PSRTs provide a thoroughly developed and cross-linguistically tested experimental approach to

investigate the extent to which listeners can memorize word stress patterns. Given that lexical stress storage remains to be investigated for Papuan Malay (Section 1.1), the aim of the current study is to investigate stress pattern recall in listeners of this language. For comparability with the most recent investigations covering typologically different languages, the current study makes uses of the stimulus materials provided in Rahmani et al. (2015) and a new version of the experimental setup developed for Bruggeman (2020) and Lialiou et al. (2023). The details of this approach are further outlined in Section 2. For comparability with well-studied languages, the PSRT in the current study is also carried out with German listeners. German is reported as more representative of Germanic stress than English, mainly due to the acoustic realization and the cues listeners attend to (e.g., Cutler, 2005; Cooper et al., 2002). Duration was reported as the main acoustic correlate of word stress in German (Dogil, 1999). The research question answered in the current study is therefore:

RQ: To what extent do Papuan Malay and German listeners recall stress patterns from memory?

Answering this research question sheds new light on the nature of Papuan Malay stress patterns. These were shown to be readily available in the speech signal and useful to listeners for word disambiguation (Kaland, 2019, 2021; Kaland et al., 2021). However, the distribution of word stress patterns is highly regular and predictable, which demands little to no lexical storage of stress information. It is therefore unclear how well Papuan Malay listeners are able to recall stress patterns, and the hypotheses for this language are therefore twofold. If listeners are highly proficient in recalling stress patterns, it is likely due to their native ability to store stress information. However, if listeners show difficulties recalling stress patterns, this is taken as an indication that their native language does not require them to do so. A rough estimation of the possible range within which the hypothesized Papuan Malay index may fall can be obtained by considering the other languages tested in previous studies. Papuan Malay listeners' performance in a PSRT could be similar to Polish listeners' (e.g., Peperkamp & Dupoux, 2002; Peperkamp et al., 2010), as both languages have highly regular word stress patterns and a clearly defined set of exceptional patterns. Note that exceptional patterns differ in nature between Papuan Malay and Polish. That is, in Papuan Malay (exceptional) ultimate stress depends entirely on the segmental makeup of the default stress location and is found in native roots (e.g., Kluge, 2017; Kaland et al., 2021). In Polish, however, (exceptional) ante-penultimate stress occurs mainly in borrowings (e.g., Comrie, 1976). Thus, Papuan Malay listeners most likely rely on an ability to distinguish the two main stress patterns, which have a more native status (i.e., historically longer part of their perception) than in Polish. This hypothesis is furthermore supported by Polish listeners' tendency to accept penultimate stress in borrowings (e.g., Domahs et al., 2012). It could therefore be that Papuan Malay listeners are slightly better at stress pattern recall than Polish ones. Whether

this subtle difference is detectable in a PSRT remains to be seen. Note that the Polish index values varied between 11.6 (Peperkamp & Dupoux, 2002) and 22.9 (Peperkamp et al., 2010), possibly due to task differences such as stimulus material and sequence lengths. It is furthermore expected that Papuan Malay listeners outperform the Indonesian ones in Rahmani et al. (2015), given that the Indonesian standard variety tested in that study is reported to be stressless (e.g., Van Zanten & Van Heuven, 1998). Thus, insofar as the previous studies are comparable, the hypothesized range of the index value for Papuan Malay listeners lies approximately between 11.6/22.9 (Polish) and 35.5 (Indonesian).

By comparison, German listeners are expected to have a better recall of stress patterns than the Papuan Malay ones, as German stress patterns are less predictable and require lexical storage to a larger extent (Domahs et al., 2014). It is expected that German listeners' performance is similar to the Dutch ones in Rahmani et al. (2015) (see **Table 1**).

2. Method

The PSRT described here used the OpenSesame (Mathôt et al., 2012) version created in Bruggeman (2020) and used in Lialiou et al. (2023) to implement an identical experiment in PsyToolkit (Stoet, 2010, 2017). This was done to allow for remote experimenting, which was needed in this study in order to be able to recruit the Papuan Malay participants. If not explicitly mentioned below, all other aspects of the setup, stimuli and procedure are identical to the PSRT described in Rahmani et al. (2015). Given the importance of reproducibility in general and the specific need to improve comparability of PSRTs across languages (Section 1.3), an online and offline version of the experiment and stimuli used in this study are available for future testing with other languages (Kaland, 2023).

2.1. Participants

In total 26 native speakers of Papuan Malay completed the experiment: 20 female, 6 male; mean age: 24; age range: 18 – 35. Some participants spoke other languages apart from Papuan Malay: Indonesian (2), Yali (2), Biak (1) and English (1). All of them learned Papuan Malay from birth and reported to speak it in daily life (with family, at school, at work). Participants who did not pass the training round (Section 2.4) or did not provide responses to all test stimuli were not included in the above number and not included for data analysis (8).

In total 26 native speakers of German completed the experiment: 14 female, 12 male; mean age: 31; age range: 20–64. Some of them reported to speak a dialect in addition to their standard variety of German: Bergisch (1), Badisch (1), Fränkisch (1) and Sächsisch (1).

All participants carried out the experiment in a quiet room using headphones. None of them reported having hearing problems.

2.2. Stimuli

Stimuli consisted of 48 audio files with spoken versions of nonce words, balanced for *contrast* (phoneme, stress), *speaker gender* (male, female), and *speaker language* (Dutch, Persian). Each contrast was generated by two nonce words (A and B). That is, the phonemic contrast occurred between ['mu.ku]^{A} and ['mu.nu]^{B}, and the stress contrast occurred between ['nu.mi]^{A}/[nu.'mi]^{B}. Thus, the phonemic contrast was based on a consonantal difference in the second unstressed syllable: [ku] or [nu]. The stress contrast was based on a combination of cues: duration, intensity, f0, and formant frequencies (with higher values obtained for the stressed syllable than for the unstressed syllable). Note that the choice for Dutch and Persian speakers in the stimulus materials was primarily made in Rahmani et al. (2015), in particular to test the recall of Persian stress patterns using Dutch as a control language. The paradigm was suitable to test other languages as well and, crucially, the stimuli are nonce words in Papuan Malay (not occurring in any of the word lists or conversation transcripts in Kluge, 2017), and in German. Detailed acoustic characteristics of the stimuli are described in Rahmani et al. (2015, Table 2) and f0 contours are analysed in detail in Bruggeman (2020, p. 161). There were three tokens of each item, resulting in a total of 4 (items) $\times 2$ (languages) $\times 2$ (genders) $\times 3$ (tokens) $=$ 48 words.

2.3. Design

The PSRT tested the recall of sequences consisting of three, four or five nonce words (five sequences per length). The sequences were designed so that there was either a phoneme or a stress contrast present, and were generated so that repeated items (AA or BB) were different versions of that item, as produced by the speakers. The sequences used for each contrast in the PSRT are listed in **Table 2** and are identical to the ones in Rahmani et al. (2015).

Length	**Sequences**				
3	AAB	ABA	ABB	BAA	BAB
4	ABAA	ABBA	BAAB	BABB	BBAB
5	ABAAB	ABABB	ABBAB	BABAA	BABBA

Table 2: Nonce word sequences per length (N words) as used in the PSRT test block with A and B referring to each item member of the contrast (either phoneme or stress).

2.4. Procedure

Before the start of the experiment, participants were asked which language(s) they spoke to assess their level of native language proficiency and multilingualism. Thereafter, they received instructions about the experimental task. The experiment was run using PsyToolkit (Stoet,

2010, 2017). The procedure described in Rahmani et al. (2015, p. 6) was implemented in the PsyToolkit online environment (HTML-based), such that the experiment could be run remotely via a web browser (Kaland, 2023). Instructions were translated from English into Papuan Malay. The contrasts were tested in two block orders: half of the participants started with the phoneme contrast, the other half with the stress contrast.

Each part consisted of a training session, a warm-up round and the actual test block. In the training session, participants learned to associate member A of the contrast with key "1" and member B with key "2". This was done by instructing them to press any of the keys and listen to all items associated with that key (12 items for each key). Then, participants pressed any of the keys and listened to a single item chosen randomly from the set associated to that key. They could press either key as often as they wanted until they indicated that they had learned the item-key associations. After that, participants heard a single item, which was randomly chosen from the set, and pressed the key that they just learned was associated to that item. Feedback was given on whether participants correctly or incorrectly identified the item in order to facilitate their learning of the button-response association. After they had identified eight items correctly, the training session ended.

In the warm-up round participants listened to sequences of two nonce words (AA, AB, BA, BB, for each language set). Their task was to identify the sequence by pressing the correct key combination. They needed to complete all eight sequences without errors before they could continue. Each sequence was presented (repeatedly) until participants had correctly identified it.

In the test block, participants identified sequences of three, four or five nonce words (**Table 2**), which were presented in random order, different for each participant. Within each sequence, items were chosen randomly from the same speaker gender and language and therefore only varied in contrast and version (no items were repeated within sequences). Participants received no feedback on their performance during the test block. A test block consisted of 30 trials (5 sequences $\times$ 3 lengths $\times$ 2 languages).

2.5. Statistical analysis

A generalized linear mixed model (GLMM) analysis was run on the correctness scores (0 for incorrect, 1 for correct), with *contrast* (two levels: phoneme, stress), *sequence length* (successive differences contrast coded, three levels: 3, 4, 5), *speaker language* (two levels: Dutch, Persian), *speaker gender* (two levels: male, female), *block order* (two levels: phoneme first, stress first) *experiment language* (two levels: Papuan Malay, German) as fixed factors, and with participant and item as random intercepts. The following formula was used: glmer(score $\sim$ contrast + seq. length + spk.lg + spk.gender + block.order + exp.lg + (1|item) + (1|pp), data, contrasts = list(seq.length = contr.sdif), family = binomial).

In addition, the index proposed by Peperkamp & Dupoux (2002) was calculated for each sequence length by subtracting the error rate for the phoneme contrasts from the error rate for the stress contrasts.

3. Results

The results are the accuracy scores and index values in **Table 3**, and the summary of the GLMM analysis for each of the factors in **Table 4**. Recall accuracy scores showed that listeners performed significantly better for phoneme contrasts than for stress contrasts. Performance significantly decreased with each increase in sequence length. As for the language of the listeners in the experiments, results showed overall significantly worse recall for Papuan Malay than for German. The latter outcome also shows in the mean index values (**Table 3**).

The language of the speaker in the stimuli did not affect the recall accuracy. As for gender of the speaker, recall was better for male ones than for female ones. Whether participants started with stress contrasts or with phoneme contrasts (block order) did not have an effect on recall accuracy.

The index values calculated per sequence length (**Table 3**) indicate a steady increase in stress recall difficulty when sequences got longer for Papuan Malay listeners. For German listeners, however, the index value was lower for four-word sequences than for three-word sequences, and higher again for five-word sequences.

Language	Sequence length	Stress	Phoneme	Index	Mean index
Papuan Malay	3	48.08	79.23	31.15	33.97
	4	36.92	71.54	34.62	
	5	18.08	54.23	36.15	
German	3	77.31	85.00	7.69	10.51
	4	74.23	79.23	5.00	
	5	53.08	71.92	18.85	

Table 3: Recall rate (percentage) per language, sequence length and contrast type (stress/phoneme). Index values are mean stress error rates subtracted from mean phoneme error rates (higher index = more recall difficulty). Mean index is the mean of the indexes for the three sequence lengths.

Factor	*β*	*SE*	*z*	*p*
(Intercept)	0.57	0.29	1.97	= 0.05
Contrast: phoneme	1.31	0.09	14.33	<0.001
Seq.length: 4–3	−0.45	0.19	−2.30	<0.05
Seq.length: 5–4	−0.92	0.19	−4.78	<0.001
Spk.language: Persian	−0.14	0.09	−1.55	n.s.
Spk.gender: male	0.17	0.09	1.88	=0.06
Block.order: stress first	0.29	0.31	0.91	n.s.
Exp.language: Papuan Malay	−1.37	0.31	−4.48	<0.001

Table 4: Results of the GLMM. (n.s. = not significant).

4. Discussion and conclusion

This study showed that Papuan Malay listeners have more difficulty recalling stress patterns than German ones. This can be seen in the lower overall recall rates and in the higher relative error rates as expressed by the index value (**Table 3**). The German index value obtained in this study is similar to the ones reported for German in earlier work (Haake et al., 2013; Heisterueber et al., 2014). Given the otherwise limited comparability of index values across studies (Section 1.3), the comparisons drawn in this discussion concern the studies using identical stimuli and setup (i.e., Rahmani et al., 2015; Bruggeman, 2020; Lialiou et al., 2023).

The recall rates for phoneme contrasts were overall worse for Papuan Malay listeners than for German listeners, a pattern also observed between listeners of Dutch and Japanese on the one hand and listeners of Indonesian, Persian and French on the other (Rahmani et al., 2015). Similarly, recall for phoneme contrasts was worse for languages in which stress recall was also worse, e.g., Tashlhiyt Berber < Moroccan Arabic (Bruggeman, 2020) and Maltese-dominant < Maltese English-dominant (Lialiou et al., 2023). That is, recall of phoneme contrasts, although generally easier than recall of stress contrasts, is facilitated when listeners have experience with stress contrasts. The Papuan Malay index value is slightly lower than the Indonesian one obtained from Rahmani et al. The hypothesis that Papuan Malay listeners would outperform the Indonesian ones is therefore not clearly met. The hypothesized similarity with the Polish performance in Peperkamp & Dupoux (2002) and Peperkamp et al. (2010) (Section 1.4) also

does not hold, as Papuan Malay listeners showed clearly worse recall performance. The German listeners in the current study's PSRT performed significantly better than the Papuan Malay ones. The German result can be ascribed to the use of word stress in this language, as hypothesized in Section 1.4.

It is furthermore interesting to observe that the German listeners showed a minimal decline in recall performance for stress contrasts between the three-word and four-word sequences (**Table 3**). This effect is unexpected in that the overall recall accuracy in this study, and in studies that tested other languages, more clearly decreased with longer sequences. It is unlikely that this is a technical artifact of the method used in this study, as a similar pattern would have been observed for Papuan Malay. It is not clear why this result is observed for German and why only for stress contrasts and not for phonemic ones. Note that the Dutch results in Rahmani et al. (2015) also show a minimal difference in stress recall accuracy scores between three- (69.33%) and four-word (68.67%) sequences. Crucially, no such similarity was found for any of the other languages or for the phonemic contrasts in that study. In Lialiou et al. (2023) a smaller decline in stress recall accuracy between three-word and four-word sequences was found for Maltese English-dominant speakers (63.50% to 56.50%) than for Maltese-dominant speakers (54.00% to 44.40%). These observations seem to suggest that the type of stress in Western-Germanic languages played a role in the relatively high recall of stress contrasts in four-word sequences.

A post hoc inspection was done on the stress recall accuracy per participant and per tested sequence in the German results of the current study. This reveals a high amount of individual variability among participants. Some recalled four-word sequences with the best accuracy compared to the other lengths (N = 7), others recalled stress contrasts in three- and four-word sequences comparably well (N = 7) and the rest showed an overall decline in accuracy when sequences got longer (i.e., the pattern observed for other non-Germanic languages, N = 12). When inspecting the accuracy per sequence (**Table 5**), generally better performance was observed when words with stress on the final syllable 1) occur more often in the sequence, 2) occur two times directly after each other (note the mismatch detection mechanism argued for in Dupoux et al. (2001), see Section 1.2) and 3) occur as the last word(s) in the sequence (i.e., are more recent). **Table 5** suggests that these three aspects need to be interpreted as global trends, as they do not all apply to the same extent for all sequence lengths. Nevertheless, together they indicate that recall performance is boosted by the occurrence of final (iambic) stresses.

Although German word stress assignment is free in the phonological sense (it could occur almost anywhere in the word), the overall majority of German (and Dutch and English) words have word-initial stress (e.g., Baayen et al., 1995; Cutler, 2005). Work on other languages has shown that listeners tend to be more sensitive to less frequent stress patterns (Sulpizio

Length	Sequence	Recall
3	AAB	88.46
	ABA	73.08
	ABB	84.62
	BAA	71.15
	BAB	69.23
4	ABAA	73.08
	ABBA	80.77
	BAAB	65.38
	BABB	73.08
	BBAB	78.85
5	ABAAB	44.23
	ABABB	57.69
	ABBAB	53.85
	BABAA	44.23
	BABBA	65.38

Table 5: Recall (percentage) for the stress contrasts by German listeners.

& McQueen, 2012; Domahs et al., 2012; Domahs et al., 2013; Kaland, 2020). In this line of reasoning, the sensitivity for deviant stress patterns, i.e., their perceptual salience, facilitates the storage in auditory working memory. However, such a facilitation effect does not directly explain why recall is better for sequences of four words than for sequences of three words. This difference could be explained when calculating the amount of finally stressed words per sequence as a percentage. The highest percentages are reached in sequences of four words. That is, B (intended key press: "2" for final stress) occurs maximally 66.67% in three-word sequences (in two out of five times), maximally 75% in four-word sequences (in two out of five times), and maximally 60% in five-word sequences (in three out of five times), see **Table 2**. Thus, the relatively higher percentages of finally stressed words in four-word sequences could have led to overall better recall compared to three- and five-word sequences.

Two further points remain unexplained. First, why would the irregular pattern boost facilitation in certain (Germanic) languages more than in others? Second, is the perceptual mechanism responsible for this outcome only related to the kind of word stress in the language at hand, or also to auditory working memory and to general storage strategies? The answers are speculative without further investigation, but they seem to confirm that speech perception mechanisms are shaped by one's native language (Cutler, 2012). An additional word of caution is also in order here. When the four-word stress bias is indeed due to the nature of word stress in Western-Germanic languages, the index value proposed in Peperkamp & Dupoux (2002) becomes (even) less comparable across studies and languages (see also the discussion in Section 1.3). That is, subtracting the phoneme contrast error rate acts as a baseline for expressing the relative stress recall difficulty. When, however, stress and phoneme recall are affected differently, depending on the language, and on the length and composition of the sequence, the index value will not reflect this straightforwardly. Thus, the index value could indicate better recall for longer sequences (as for German in **Table 3**), where this actually reflects different decline rates in recall accuracy between stress and phoneme contrasts.

Coming back to the question of word stress in Papuan Malay, the current outcomes require a discussion of how they fit in the context of previous studies indicating that this language should be analysed as a stress language (Section 1.1). The studies discussed in the introduction show that Papuan Malay word stress is produced, perceived and functional. The current results rather indicate that Papuan Malay listeners do not have experience in memorizing stress patterns. Note, however, that this does not necessarily mean that their language *does not have* regular word stress. Previous studies showed that the predictability of word stress patterns determines the recall abilities of listeners (Peperkamp et al., 2010). Or simply put, with highly predictable stress patterns there is no need to store the stress information with the word in the lexicon. Listeners of a predictable stress language are therefore not required to do so and have shown worse recall performance compared to listeners of less predictable stress languages. The type of stress in Papuan Malay is therefore likely to be "assigned" at a later stage in word production, i.e., after lexical retrieval. This explanation would be consistent with the results in the previous studies on Papuan Malay (Kaland, 2019, 2021; Kaland & Baumann, 2020). It is also consistent with the lack of minimal stress pairs in Papuan Malay (e.g., Kluge, 2017). This type of "postlexical" stress can, however, still be functional at the word level, i.e., to facilitate word disambiguation (Kaland et al., 2021) and word recognition (Kaland, 2021). In retrospect, word recognition success in the gating task was most likely the result of having heard the stressed and unstressed syllables directly before giving a response (Kaland, 2021), thus only probing listeners' ability to detect stress differences from working memory, not from their lexical storage, as discussed in Section 1.1.

It is also expected that the type of word stress in Papuan Malay would facilitate word *segmentation* due to its regularity (currently under investigation). In context, the present results therefore indicate that the extent to which listeners are able to memorize stress information is just one aspect of the notion of word stress, and as such provides a partial contribution to the diagnosis of a language as a stress language. Multiple aspects need to be taken into account, ranging from acoustic production and perception, phonological rules, and lexical function, to word recognition and lexical storage. All these aspects closely relate to how predictable word stress patterns are for listeners. This study furthermore confirmed earlier work in that the predictability of word stress patterns correlates negatively with listeners' lexical storage abilities (Peperkamp et al., 2010), with listeners used to unpredictable patterns being able to use their lexical storage better (German) than listeners used to predictable patterns (Papuan Malay).

Additional files

The additional files can be found as follows:

- **df.csv.** Data. DOI: https://doi.org/10.16995/labphon.11695.s805
- **df.R.** Script. DOI: https://doi.org/10.16995/labphon.11695.s806
- **Experimental setup.** OSF: https://osf.io/xwmka/

Acknowledgements

Research for this paper was funded by the German Research Foundation (DFG) – Project-ID 281511265 – SFB 1252. The author thanks Anna Bruggeman and Maria Lialiou for help setting up the experimental procedure, the Papuan Malay Bible Translation Team (Tim Penerjema Alkitab Melayu Papua), Sophia Meier, Alessandra On and Jonathan Reich for help participant recruitment and experiment facilitation, and two anonymous reviewers for their comments. The experiments reported in this paper have been conducted following protocols and informed consent practices in compliance with the Helsinki Declaration, with approval of the Papuan Malay Bible Translation Team and the Faculty of Arts and Humanities of the University of Cologne. Informed consent was obtained from each individual participant prior to participation. Note that Kaland & Lialiou (2024) appeared after acceptance of the current article and was therefore not included in the overview in **Table 1**.

Competing interests

The author has no competing interests to declare.

References

Azzopardi, M. (1981). *Phonetics of Maltese: Some areas relevant to the deaf*. https://era.ed.ac.uk/handle/1842/19041

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). CELEX2. DOI: https://doi.org/10.35111/GS6S-GM48

Boudlal, A. (2001). *Constraint Interaction in the Phonology and Morphology of Casablanca Moroccan Arabic* [PhD Thesis]. Université Mohammed V. https://rucore.libraries.rutgers.edu/rutgers-lib/38436/

Bruggeman, A. (2020). *Lexical and postlexical prominence in Tashlhiyt Berber and Moroccan Arabic* [PhD Thesis]. Universität zu Köln. https://kups.ub.uni-koeln.de/11189/

Comrie, B. (1976). Irregular stress in Polish and Macedonian. *International Review of Slavic Linguistics*, *1*, 227–240.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of Lexical Stress on Lexical Access in English: Evidence from Native and Non-native Listeners. *Language and Speech*, *45*(3), 207–228. DOI: https://doi.org/10.1177/00238309020450030101

Correia, S., Butler, J., Vigário, M., & Frota, S. (2015). A Stress "Deafness" Effect in European Portuguese. *Language and Speech*, *58*(1), 48–67. DOI: https://doi.org/10.1177/0023830914565193

Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics*, *35*(1), 41–48. DOI: https://doi.org/10.3758/BF03205923

Cutler, A. (2005, January). Lexical Stress. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 264–289). Blackwell Publishing Ltd. DOI: https://doi.org/10.1002/9780470757024.ch11

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. The MIT Press. DOI: https://doi.org/10.7551/mitpress/9012.001.0001

Cutler, A., Norris, D., & Sebastián-Gallés, N. (2004). Phonemic repertoire and similarity within the vocabulary. In S. Kin & M. J. Bae (Eds.), *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech 2004-ICSLP)* (pp. 65–68). Sunjijn Printing Co. DOI: https://doi.org/10.21437/Interspeech.2004-61

Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. In R. Hoffmann & H. Mixdorff (Eds.), *Speech Prosody 2006* (Vol. 40). TUD press. DOI: https://doi.org/10.21437/SpeechProsody.2006-2

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., & Yoo, H.-Y. (2015, June). Intonational phonology of French: Developing a ToBI system for French. In S. Frota & P. Prieto (Eds.), *Intonation in Romance* (1st ed., pp. 63–100). Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199685332.003.0003

Di Cristo, A. (1998). Intonation in French. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems: A Survey of twenty languages* (pp. 195–218). Cambridge University Press.

Dogil, G. (1999). The phonetic manifestation of word stress in Lithuanian, Polish, German, and Spanish. In H. v. d. Hulst (Ed.), *Word Prosodic Systems in the Languages of Europe* (pp. 273–311, Vol. 4). De Gruyter. DOI: https://doi.org/10.1515/9783110197082.1.273

Domahs, U., Genc, S., Knaus, J., Wiese, R., & Kabak, B. (2013). Processing (un-)predictable word stress: ERP evidence from Turkish. *Language and Cognitive Processes*, *28*(3), 335–354. DOI: https://doi.org/10.1080/01690965.2011.634590

Domahs, U., Knaus, J., Orzechowska, P., & Wiese, R. (2012). Stress "deafness" in a Language with Fixed Word Stress: An ERP Study on Polish. *Frontiers in Psychology*, *3*. DOI: https://doi.org/10.3389/fpsyg.2012.00439

Domahs, U., Plag, I., & Carroll, R. (2014). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. *The Journal of Comparative Germanic Linguistics*, *17*(1), 59–96. DOI: https://doi.org/10.1007/s10828-014-9063-9

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A Destressing "Deafness" in French? *Journal of Memory and Language*, *36*(3), 406–421. DOI: https://doi.org/10.1006/jmla.1996.2500

Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress "deafness". *The Journal of the Acoustical Society of America*, *110*(3), 1606–1618. DOI: https://doi.org/10.1121/1.1380437

Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress 'deafness': The case of French learners of Spanish. *Cognition*, *106*(2), 682–706. DOI: https://doi.org/10.1016/j.cognition.2007.04.001

Goedemans, R., & van Zanten, E. (2014). No Stress Typology. In J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller & E. Van Zanten (Eds.), *Above and Beyond the Segments* (pp. 83–95). John Benjamins. DOI: https://doi.org/10.1075/z.189.07goe

Gordon, M., & Nafi, L. (2012). Acoustic correlates of stress and pitch accent in Tashlhiyt Berber. *Journal of Phonetics*, *40*(5), 706–724. DOI: https://doi.org/10.1016/j.wocn.2012.04.003

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511616983

Haake, C., Kob, M., Willmes, K., & Domahs, F. (2013). Word stress processing in specific language impairment: Auditory or representational deficits? *Clinical Linguistics & Phonetics*, *27*(8), 594–615. DOI: https://doi.org/10.3109/02699206.2013.798034

Heisterueber, M., Klein, E., Willmes, K., Heim, S., & Domahs, F. (2014). Processing word prosody – behavioral and neuroimaging evidence for heterogeneous performance in a language with variable stress. *Frontiers in Psychology*, *5*. DOI: https://doi.org/10.3389/fpsyg.2014.00365

Himmelmann, N. P. (2023). On the comparability of prosodic categories: Why 'stress' is difficult. *Linguistic Typology*, *27*(2), 341–361. DOI: https://doi.org/10.1515/lingty-2022-0041

Kaland, C. (2019). Acoustic correlates of word stress in Papuan Malay. *Journal of Phonetics*, *74*, 55–74. DOI: https://doi.org/10.1016/j.wocn.2019.02.003

Kaland, C. (2020). Offline and online processing of acoustic cues to word stress in Papuan Malay. *The Journal of the Acoustical Society of America*, *147*(2), 731–747. DOI: https://doi.org/10.1121/10.0000578

Kaland, C. (2021). The perception of word stress cues in Papuan Malay: A typological perspective and experimental investigation. *Laboratory Phonology*, *12*(1), 1–33. DOI: https://doi.org/10.16995/labphon.6447

Kaland, C. (2023). Stress 'deafness' task PsyToolkit. https://osf.io/xwmka/

Kaland, C., & Baumann, S. (2020). Demarcating and highlighting in Papuan Malay phrase prosody. *The Journal of the Acoustical Society of America*, *147*(4), 2974–2988. DOI: https://doi.org/10.1121/10.0001008

Kaland, C., Kluge, A., & Van Heuven, V. J. (2021). Lexical analyses of the function and phonology of Papuan Malay word stress. *Phonetica*, *78*(2), 141–168. DOI: https://doi.org/10.1515/phon-2021-2003

Kaland, C., & Lialiou, M. (2024). Quantity-sensitivity affects recall performance of word stress. *Interspeech 2024*, 4238–4242. DOI: https://doi.org/10.21437/Interspeech.2024-182

Kluge, A. (2017). *A grammar of Papuan Malay*. Language Science Press. DOI: https://doi.org/10.17169/langsci.b78.35

Lialiou, M., Bruggeman, A., Vella, S., Grech, S., Schumacher, P., & Grice, M. (2023). Word-level prominence and "stress deafness" in Maltese-English bilinguals. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 132–136). Guarant International.

Maskikit-Essed, R., & Gussenhoven, C. (2016). No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology*, *33*(2), 353–389. DOI: https://doi.org/10.1017/S0952675716000154

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. DOI: https://doi.org/10.3758/s13428-011-0168-7

Odé, C. (1994). On the perception of prominence in Indonesian. In C. Odé, V. J. Van Heuven & E. Van Zanten (Eds.), *Experimental studies of Indonesian prosody* (pp. 27–107). Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië, Rijksuniversiteit Leiden.

Peperkamp, S., & Brazeal, J. (2023). Lasting stress 'deafness' after auditory training: French listeners revisited. *Proceedings of the 20th International Congress of Phonetic Sciences*, 1–5. https://hal.science/hal-04197465

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology*, *7* (pp. 203–240). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110197105.1.203

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, *38*(3), 422–430. DOI: https://doi.org/10.1016/j.wocn.2010.04.001

Rahmani, H., Rietveld, T., & Gussenhoven, C. (2015). Stress "Deafness" Reveals Absence of Lexical Marking of Stress or Tone in the Adult Grammar (I. Berent, Ed.). *PLOS ONE*, *10*(12), e0143968. DOI: https://doi.org/10.1371/journal.pone.0143968

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. DOI: https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. DOI: https://doi.org/10.1177/0098628316677643

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*(1), 177–193. DOI: https://doi.org/10.1016/j.jml.2011.08.001

Van der Hulst, H. (2012). Deconstructing stress. *Lingua*, *122*(13), 1494–1521. DOI: https://doi.org/10.1016/j.lingua.2012.08.011

Van Zanten, E., & Van Heuven, V. J. (1998). Word stress in Indonesian; Its communicative relevance. *Bijdragen tot de Taal-, Land- en Volkenkunde/Journal of the Humanities and Social Sciences of Southeast Asia and Oceania*, *154*. DOI: https://doi.org/10.1163/22134379-90003908