On the perception of prosodic prominences and boundaries in Papuan Malay

Sonja Riesberg, Janina Kalbertodt, Stefan Baumann, Nikolaus P. Himmelmann (Universität zu Köln)

This paper reports the results of two perception experiments on the prosody of Papuan Malay. We investigated how native Papuan Malay listeners perceive prosodic prominences on the one hand, and boundaries on the other, following the Rapid Prosody Transcription method as sketched in Cole & Shattuck-Hufnagel (2016). Inter-rater agreement between the participants was shown to be much lower for prosodic prominences than for boundaries. Importantly, however, the acoustic cues for prominences and boundaries largely overlap. Hence, one could claim that inasmuch as prominence is perceived at all in Papuan Malay, it is perceived at boundaries, making it doubtful whether prosodic prominence can be usefully distinguished from boundary marking in this language. Our results thus essentially confirm the results found for Standard Indonesian by Goedemans & van Zanten (2007) and various claims regarding the production of other local varieties of Malay; namely, that Malayic varieties appear to lack stress (i.e. lexical stress as well as post-lexical pitch accents).

1. Introduction

Papuan Malay (henceforth PM) is a local variety of Indonesian/Malay, spoken in the two easternmost provinces of Indonesia – Papua Barat and Papua – by approximately 1,200,000 speakers (see Kluge 2014). It is spoken mostly in the coastal areas, and to a lesser extent in the mountainous inland. Indonesian Papua, with its more than 270 indigenous languages, is linguistically highly diverse, and most speakers are at least bilingual. Papuan Malay serves as the lingua franca in this area, and most native speakers speak PM in addition to one or more local languages.

This paper reports on two perception experiments that investigate the contribution of prosody with respect to how native speakers of PM perceive prosodic prominences and boundaries in natural speech. It thus stands alongside a growing number of recent papers that discuss the prosodic systems of different varieties of Indonesian, such as, for example, the study by Goedemans & van Zanten (2007) on Standard Indonesian and, most recently, the paper by Maskikit-Essed & Gussenhoven (2016) on Ambonese Malay.

For a long time, the standard assumption has been that (Standard) Indonesian displays lexical stress on the penultimate syllable, unless this syllable contains a schwa, in which case stress falls on the final syllable (cf. Alieva et al. 1991; Cohn 1989). Secondary stress has been claimed to fall on the first syllable and every odd syllable thereafter, but never on the one adjacent to the syllable that carries the main stress (Cohn & McCarthy 1994). Other authors have claimed that schwa in (some varieties of) Indonesian can be stressed just as well as any other vowel (Halim 1974; Laksman 1994).

However, in a growing number of studies, the claim that Indonesian displays lexical stress on the penultimate syllable has been challenged. While some authors found that there is a preference for stress to occur on the penultimate syllable but free variation – especially in longer words – is possible (cf. van Zanten 1994; van Zanten & van Heuven 2004), other authors came to the conclusion that there is no lexical stress at all (Zubkova 1966; Odé 1997). Especially in more recent publications, it has been pointed out that the aforementioned disagreement as to whether or not Indonesian displays lexical stress is probably due to the fact that 'Indonesian' as a reasonably homogeneous language does not exist. Around 700 indigenous languages are spoken in the Republic of Indonesia (cf. Ethnologue), with the great majority of people being at least bilingual, speaking a local language in addition to (some variety of) Indonesian. Often, Indonesian is learned as a second language, usually from the age of six or seven, when children enter primary school and are exposed to Indonesian as the language of education. Furthermore, in addition to standard Indonesian and the indigenous languages, local varieties of Malay are spoken in many regions of Indonesia (e.g. Ambonese

Malay, Jambi Malay, Kupang Malay, Manado Malay, Papuan Malay, etc.). Often, these local varieties of Malay take the place of standard Indonesian and are the major means of everyday communication. It is thus very likely that studies on lexical stress in 'Indonesian' are based on data from speakers with different substrate dialects and languages, which means that the contradictory results of such studies are probably due to the different prosodic properties of these substrates. More recent studies therefore make an effort to control for the linguistic background of the participants in their experiments. Yet even these more recent studies provide results that are not straightforward to interpret, an assessment which is also valid of our study, as further detailed in § 5. This is in part due to the fact that more recent studies – even if they control for substrate influence – continue to have issues regarding the naturalness of the word tokens under investigation (often loan words four syllables in length or even longer) and adequate sampling. Many studies rely on non-natural lab speech, often produced by a single speaker, and evaluated by only a few more.

Goedemans & van Zanten (2007), for example, conducted a carefully designed perception experiment with two groups of participants with different linguistic profiles: one group consisted of speakers of Indonesian with Javanese as their substrate language, the other group consisted of speakers of Indonesian who were additionally native speakers of Toba Batak. These two languages were chosen because Toba Batak is said to exhibit clearly defined stress, while Javanese is said to have only weak stress, the location of which lacks consensus in the literature (Goedemans & van Zanten 2007: 40). As stimuli, the authors recorded material from one Toba Batak Indonesian speaker and one Javanese Indonesian speaker. This material was manipulated such that presumably prominence-lending phonetic cues, i.e. pitch excursions, duration and intensity, would occur on different syllables. It was then judged for acceptability by listeners of the two different groups. The Javanese listeners did not show any preference for stress on either the penultimate or the ultimate syllable for both the Javanese Indonesian and the Toba Batak Indonesian stimuli. The Toba Batak listeners, on the other hand, clearly preferred penultimate stress in the Toba Batak speech data, but showed no clear preferences for the Javanese data. Goedemans & van Zanten interpret these results as evidence against lexical stress in Javanese Indonesian. Though their experiment was explicitly *not* designed to investigate prominence above the word level, they do observe that phrasal prominence always occurs close to the boundary. They come to conclude that "the distinction between accent lending and boundary marking intonation movements is very difficult to make in Indonesian" (Goedemans & van Zanten 2007: 57).

One of the few studies that address the issue of phrasal prominence in more detail is the work by Maskikit-Essed & Gussenhoven (2016) on Ambonese Malay (see also Stoel 2007 on Manado Malay, Himmelmann 2010 on Waima'a, and Clynes & Deterding 2011 on Brunei Malay). Maskikit-Essed & Gussenhoven conducted a production experiment with four native speakers of Ambonese Malay. They recorded 80 mini-dialogs consisting of read question-answer pairs, which contained eight target nouns in different positions (phrase- and IP-final as well as phrase- and IP-medial) and were controlled for different focus conditions. In these eight target words, no evidence for (post-)lexical stress in the putative stressed syllables (ultimate or penultimate, depending on the word) was found. Furthermore, the phrase-final pitch movement, which is a typical feature of declarative mood in many languages in the area (Himmelmann 2010: 67), is not tied to either the final or the prefinal syllable. Rather, it is sensitive to the available space and tends to be timed earlier when the word is longer. Finally, Maskikit-Essed & Gussenhoven tested two focus conditions, one in which the phrase-final target word was in focus, and one in which it occurred in post-focal position, i.e. a focal element preceded the phrase-final target word. In the latter condition, the authors could not find any signs of reduction of the post-focal target words, either in duration or in pitch height. Furthermore, the pitch contours were similar, not only on the target words, but also over the whole sentences (Maskikit-Essed & Gussenhoven 2016: 372). Taking these results together, Maskikit-Essed & Gussenhoven come to the conclusion that information focus in Ambonese Malay is not expressed by means of prosody.

For Papuan Malay, Kluge (2014) recorded 1,072 words in two different carrier sentences, one in which the target word occurs clause-finally, and one in which it appears in clause-medial position.[1] Kluge concludes that 964 (90%) of all words have penultimate stress

[1] The two carrier sentences Kluge used are: *Sa blum taw ko pu kata itu, kata xxx.* 'I don't yet know that word of yours, the word xxx' and *Ko pu kata xxx itu, sa blum taw.* 'Your word xxx, I don't know yet.' (Kluge 2014: 57).

(including both open and closed penultimate syllables), and only 108 (10%) show stress on the final syllable. Of those 108 words that displayed ultimate stress, 105 (97%) contained the front open-mid vowel /ɛ/ (the equivalent of Indonesian schwa) in the penultimate syllable. Yet, Kluge notes that /ɛ/ does not condition ultimate stress, as for 65 of those words with penultimate stress, the stressed syllable also contained an /ɛ/. In addition, three words with ultimate stress contained /i/ and /u/ vowels in the penultimate syllable (Kluge 2014: 89).

Based on this analysis, Papuan Malay would appear to be very similar to Ambonese Malay as presented in the grammar by van Minde (1997), where it is claimed that Ambonese Malay has regular penultimate stress, with a small number of lexical items showing ultimate stress. Note that in both grammars, the analysis is based primarily on the auditory impression of the Western researcher who hears one or the other syllable as more prominent. It is unclear what Ambon and Papuan Malay native speakers actually hear. The present study is a first exploration of this question. Recall from above that Maskikit-Essed & Gussenhoven (2016) did not find clear acoustic evidence for (lexical) stress or (post-lexical) pitch accents in Ambonese Malay. Hence, it may very well be the case that Western ears tend to hear these languages according to the categories they know from their own prosodic systems, and not necessarily based on the 'objectively' available acoustic input. That is, if Maskikit-Essed & Gussenhoven's (2016) findings hold up to further scrutiny, the phrase-final pitch movement in Ambon Malay that is heard by Western researchers as being clearly located on either the penultimate or the ultimate syllable is actually most often (i.e. in terms of the measurable acoustic cues) located somewhere in between the final two syllables and thus, strictly speaking, is not properly anchored to either one, but rather to the phrase-final boundary.

In targeting perception rather than production, the current study takes up the line of research pioneered by Leiden phoneticians in the 1990s, though with a somewhat different methodology (see the book edited by van Heuven & van Zanten 2007 for a summary). With regard to these studies, Papuan Malay would appear to be most similar to Toba Batak, for which a system with predominantly penultimate and occasional ultimate stress has also been reported, though possibly with a higher functional load, as a fair number of minimal stress pairs are claimed to exist (Roosman 2007: 92f provides a succinct summary of the literature). Unfortunately, Roosman (2007) does not investigate prominence perception by Toba Batak speakers of their native tongue. Moreover, the work by Goedemans & van Zanten (2007) discussed above only looks at the perception of different varieties of Indonesian by native Toba Batak speakers. Hence, the results here will not be directly comparable with the results reported by the Leiden group. It will nevertheless begin to sketch out one of the constellations not yet investigated in detail, i.e. the native perception of a prosodic system which – to Western ears – appears to have a fairly clear lexical stress system with predominantly penultimate stress.

In concluding these introductory remarks, it bears emphasizing that although much of the literature – and consequently also parts of this introduction – makes reference to phonological categories, including in particular '(lexical) stress', such categories only make sense as part of a comprehensive analysis of the prosodic system of a given language. Since such an analysis does not yet exist for Papuan Malay, the main purpose of the current chapter is to provide perceptual data for a more comprehensive investigation of the Papuan Malay system, which in addition will require a rigorous and detailed acoustic analysis, a task currently being undertaken by one of the authors (Himmelmann).

The present chapter is structured as follows: §2 describes the experimental setup and methods, before §3 and §4 report on the results of the two experiments (on prominences and boundaries, respectively). §5 summarizes the findings and draws some preliminary conclusions on the interrelation between the perception of prosodic cues and their interpretation by native listeners of Papuan Malay.

2. Methods

Given the growing amount of evidence in the literature to support the assumption that the prosodic systems of different varieties of Malay differ significantly from the better-known European systems, we wanted to address the question of how native speakers of one of these varieties – Papuan Malay – interpret prosodic cues if required to judge the presence or absence of prominences and boundaries. We therefore conducted two perception experiments using the *Rapid Prosody Transcription* (RPT) method, as introduced in different papers by Jennifer Cole and colleagues (cf. Mo et al. 2008; Cole et al. 2010a; Cole et al. 2010b, Cole & Shattuck-Hufnagel 2016:7–13). In the RPT setup, ordinary listeners that are naïve with respect to prosodic analysis listen to excerpts of audio recordings. They are given

minimal instructions (see below) and are allowed to play the audio recordings only twice. On a printed transcript of the recorded excerpts, in which punctuation and capitalization have been removed, the participants are either asked to underline those words which they perceive as prominent (prominence experiment), or to draw a vertical line after the word which they perceive to be the last word of a prosodic unit (boundary experiment).

The advantage of this method is its simplicity and directness, providing us with coarse-grained linguistic data: prosodic judgments by untrained listeners, which are based on the listeners' holistic perception of form and function. As noted by Cole and colleagues, the prominence and boundary judgments elicited in this task are clearly not based exclusively on prosodic factors, but also include morpho-syntactic, semantic and pragmatic factors. Our main concerns here are prosodic factors, but some of our variables (for example, the distinction between content and function words) also target these other levels.

Subjects

The raters of our perception study were 44 native speakers of Papuan Malay (22 for the prominence experiment, 22 for the boundary experiment). Of the 22 participants in the prominence experiment, 15 were female. 15 were bilingual in Papuan Malay and standard Indonesian, and 7 participants were additionally proficient in another local language. Of the 22 participants in the boundary experiment, 12 were female. 17 subjects were speakers of Papuan Malay and standard Indonesian, and 5 spoke another local language in addition. All 44 participants were students at the Universitas Papua (UNIPA) in Manokwari (West Papua), aged between 18 and 28 years. All participants stated that Papuan Malay was (one of) their first language(s)[2] and that Papuan Malay was their first language of communication at home and at university, as well as when talking to friends. None of them had any experience in prosodic analysis or reported any hearing or reading problems.

Stimuli and procedure

The participants annotated 56 excerpts of audio recordings (the same for both the prominence and the boundary experiment). These excerpts were taken from a corpus of natural speech, consisting of speakers re-telling Chafe's *Pear Movie* (Chafe 1980) and playing the *Tangram Task*.[3] Excerpts thus consisted of both monologues (the pear movie recordings) and dialogues (the tangram recordings). They were of varying lengths, ranging from 1 to 15 seconds, and included 28

[2] Four further participants that took part in the prominence experiment were excluded from the results because they had learned Papuan Malay only at a later age when they entered primary school. They were therefore not considered native speakers, even if they had lived in Manokwari for several years and their dominant language was Papuan Malay at the time of the experiment.

[3] The *Tangram Task* is an elicitation method that involves two speakers negotiating whether the picture described by speaker one is the same as the picture given to speaker two.

different native speakers of Papuan Malay (17 female, 11 male).

Instructions for the participants of the experiments were, as stated above, minimal. They included a short written description of what we mean by *prominence* and *boundaries*, respectively. For the prominences, it was explicitly stated that underlining more than one word per excerpt was allowed. No audio examples were given, but both instructions contained a written example that illustrated how to mark either prominences or boundaries, and how choices could be corrected, if necessary (see Appendix A for the original instructions in Indonesian, and Appendix B for English translations).

The data in (1) show an example of one of the excerpts, including glosses and translation (1a), and how it was presented to the participants of the experiment (1b). (1c) shows the prominence choices made by one of the participants (RW, female, 23 years), (1d) indicates the boundary positions perceived by another participant (JGL, female, 25 years).

(1) Papuan Malay

a. *yang tiga orang ini pegang topi satu*
REL three person DEM carry hat one
'The three people are carrying a hat.'

b. *yang tiga orang ini pegang topi satu*

c. *yang tiga orang ini pegang topi satu*

d. *yang /tiga orang ini /pegang topi satu*

Test variables

We tested the influence on the native listeners' judgments of a number of prosodic and morpho-syntactic cues which have been found to have an effect on prominence or boundary perception in other (generally West Germanic) languages. For each test word in both experiments, we investigated the following prosodic factors: *word duration* (in ms), *mean duration of syllables* (in ms), *duration of the last syllable within a word* (in ms), *minimum, maximum* and *mean pitch* (in Hz), absolute *pitch range* (in semitones), *number of syllables* (both abstract phonological and actually realized) as well as *presence of a pause*. An increase in duration, pitch height and pitch range have been shown in many studies to correlate with higher perceived prominence in Germanic languages (e.g. Cole et al. 2010a; Rietveld & Gussenhoven 1985), while presence of a pause and domain-final lengthening has been shown to trigger the perception of a phrase break (e.g. Turk & Shattuck-Hufnagel 2007).

Furthermore, we analyzed the morpho-syntactic cues *part-of-speech* (POS), *part-of-speech class* (i.e. content words vs. function words), whether the word is the *last verbal argument* in the excerpt, and *syntactic break* (three levels: no, weak or strong break). The label *weak break* was assigned to sentence-medial words that were followed by a subordinate clause (e.g. relative clause), while the label *strong break* was assigned to sentence-final words. Again, all these structural factors were chosen from a European point of view, since West Germanic languages are known to be sensitive to these parameters. In English and German, function words are usually less prominent than content words (Büring 2012: 31), while the last verbal argument in a sentence is of importance when it comes to focus projection, i.e. in the default intonation of a broad focus sentence, the last verbal argument receives the nuclear accent (Uhmann 1988: 66).

In addition to these linguistic factors, we correlated the experiment's outcome with an expert rating of prosodic boundaries, which represents the consensus judgments of the authors, all of them German natives. Boundaries in this version are based on the consensus of at least three of the four authors. In a pre-test, this expert rating was statistically analyzed with the same factors examined for the native raters, showing strong influences of pause, overall word duration, mean syllable duration and syntactic structure. The effect of syntactic structure is somewhat surprising, as two of the authors do not know the language and thus have no understanding of the syntax.

Data analysis

Both experiments consisted of binary classification tasks. In the prominence experiment (Experiment I), participants had a binary choice for each word in the transcript to rate it as either prominent or non-prominent. In the boundary experiment (Experiment II), there was a choice for each consecutive pair of words to either place a boundary between them or not. That is, for an excerpt containing n words, there were $(n - 1)$ consecutive word pairs and thus $(n - 1)$ potential boundaries the rater had to decide upon, since no judgment was needed after the last word of an excerpt. Given that our set of 56 excerpts consisted of 730 words altogether, each participant thus

produced 730 data points in the prominence experiment and 674 data points in the boundary experiment.

For the statistical analysis of these data, a mixed effects logistic regression was performed using the *lme4-package* (Bates et al. 2015) in *R* (R Core Team 2015), which suits both continuous and categorical input variables. As this study is exploratory in nature, we only created single effect models (e.g. only *maximum pitch* or *part-of-speech,* but not both variables) with random effects for speaker, sentence and rater. Subsequently, odds ratios were calculated to enable a comparison of the factors by means of effect size in order to determine which cue had the strongest influence on the raters' judgments.

We further calculated both the Fleiss' kappa coefficient (plus its z-normalized score) and Cohen's kappa. Fleiss' kappa provides a single coefficient as a measure of agreement across all raters. Cohen's kappa calculates agreement between an individual pair of raters for each word/consecutive pair of words, comparing the labels (i.e. prominent – non-prominent, and boundary – no-boundary, respectively).

In addition, we calculated the prominence-score (p-score) and the boundary-score (b-score), which serve as relative measures representing the ratio of subjects that underlined a word, i.e. that perceived a word as prominent, or drew a vertical line, i.e. perceived a prosodic break, with respect to the total number of participants. An example showing p- and b-scores is given in Figure 1.

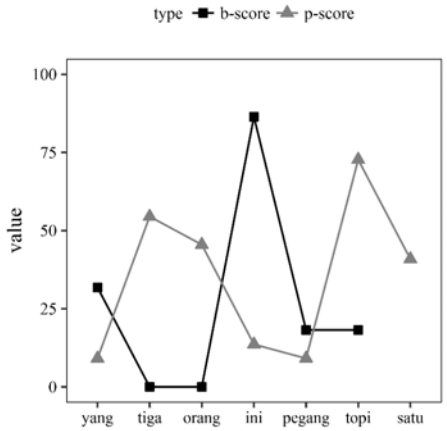

Figure 1: P- and b-scores for one PM excerpt (cf. (1) above). The higher the value, the more participants perceived a word as prominent (gray line with triangles) or perceived a boundary after the respective word (black line with squares). Recall that no b-score has been calculated for the last word of an excerpt.

3. Results of the Prominence Experiment

3.1. Inter-rater and multi-rater agreement

As mentioned above, we measured the overall inter-rater agreement for both experiments by calculating Fleiss' and Cohen's kappa coefficients. These two measures allow us to compare the performance of the two rater groups between the two experiments. They also make it possible to compare our results with similar studies that used RPT to investigate native speakers' perception of prominences in American English and in German.

The Fleiss' kappa score we calculated for the prominence experiment amounts to 0.103 (z = 42.1), a value that turns out to be surprisingly low in cross-linguistic comparison. In Table 1, we compare the PM inter-rater scores from the prominence experiment with those of two comparable studies on American English (Mo et al. 2008; Cole et al.

2010a) and German (Baumann & Winter 2015). The study by Cole and colleagues used spontaneous conversational speech from the *Buckeye Corpus,* which consists of interviews with adult speakers of American English from Columbus (Pitt et al. 2007). Baumann & Winter's study, on the other hand, used read sentences that displayed different focus structures and information status categories. Both made use of the RPT method as described above. The comparison clearly shows that the PM listeners perform significantly worse in the prominence task than English or German listeners.

	German	**English**	**PM**
Fleiss' Kappa	0.53	0.42	0.103
z	244	20.4	42.1

Table 1: Fleiss' kappa for prominences in German, American English, and PM rating studies.

The slightly higher agreement of German raters compared to English raters is probably due to the different data types used in the respective experiments, i.e. read speech versus spontaneous conversational data. Considering that the naturalness of the stimuli might have an effect on the raters' level of agreement in their perception of prominences, the PM scores are probably best compared with the English scores. Still, the difference between English raters, with a Fleiss' kappa score of 0.42, and Papuan Malay, with a kappa score of only 0.103, is also striking.

To test whether the low score of the PM raters in the prominence experiment was just due to very low agreement between some individual participants, we calculated Cohen's kappa scores for every single rater pair. In Table 2, the pair-wise inter-rater agreement is summarized, using the agreement categories postulated by Landis & Koch (1977), who characterize kappa values between 0 – 0.20 as slight agreement, 0.21 – 0.40 as fair, 0.41 – 0.60 as moderate, 0.61 – 0.80 as substantial, and 0.81 – 1 as (almost) perfect agreement.

	Prominences	
inter-rater agreement	**pairs**	**percentage**
none	25	10.82 %
slight	164	71.00 %
fair	40	17.32 %
moderate	2	0.87 %
substantial	0	0.00 %
(almost) perfect	0	0.00 %
	231	100 %

Table 2: *Inter-rater agreement categories (based on Cohen's kappa scores) for PM subjects in the prominence experiment.*

As we can see, more than 80 % of the pairs showed either 'slight' or no agreement, and for only about 17 % of pairs was the agreement 'fair'. The picture gained by the Fleiss' kappa study is thus confirmed. As we will see in §4.1, both the multi-rater agreement and the pair-wise inter-rater agreement in the prominence experiment is much lower than in the boundary experiment.

3.2. Factors determining perceived prominence

As already indicated by the low kappa values above, we observed a high degree of variability in the listeners' judgments, leading to predominantly low p-scores. In fact, the modal value in our data was a p-score of 13.6 %, as shown in Figure 2. There was not a single item (out of 726[4]) that *all* raters considered prominent, the highest p-score being 81.8 % (18 out of 22 participants agreeing on assigning prominence to a given word), which was achieved only three times. Furthermore, there were only twenty words which all participants

[4] Four items had to be discarded because no pitch features could be calculated.

judged as *not* prominent (out of 726 words altogether).

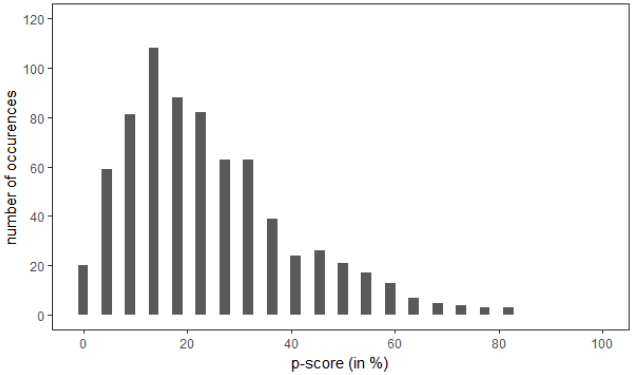

Figure 2: Distribution of p-scores in the PM data.

When examining which of the 14 test variables influenced the perception of prominence, only *part-of-speech* was *not* found to have a significant effect on prominence judgments ($X^2(1) = 0.6444$, $p = 0.4221$). Note, however, that the actual effect sizes of the various factors were found to be rather small, as indicated by the odds ratios. An odds ratio of 1 usually indicates that there is no change in the odds of receiving a certain outcome when manipulating the test variable. An odds ratio bigger than 1 indicates an increase in the odds of getting a certain outcome (cf. Field et al. 2012: 320, 923), in our case a prominence response. We have excluded variables with extremely small odds ratios from further consideration, in order to concentrate on those effects that are most likely to have noticeable effects on prominence judgments. Our threshold was set to an odds ratio of 1.5 to 1. This procedure led to the exclusion of all measures relating to pitch (*maximum*, *minimum*, *mean pitch* and *pitch range*) in addition to *part-of-speech*, *number of syllables (phonological)* and *duration of the last syllable*.

The strongest effect was found for *pause* ($X^2(1) = 156.26$, $p < 0.0001$), increasing the odds of observing a prominence response in the presence of a pause as opposed to the absence of a pause by 2.7 to 1 (logit estimate: 1.01, SE = 0.08). Figure 3 shows the relation between prominence judgments on a word and a subsequent pause.

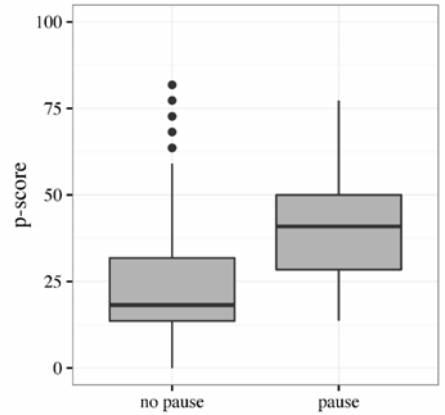

Figure 3: P-scores as a function of presence or absence of a subsequent pause.

The second most influential factor for the perception of prominence by native speakers of Papuan Malay was *part-of-speech class* ($X^2(1) = 329.3$, $p < 0.0001$), i.e. content vs. function word, as displayed in Figure 4. Being presented with a content word as opposed to a function word increases the odds of observing a positive response for prominence by 2.1 to 1 (logit estimate: 0.73, SE = 0.04).

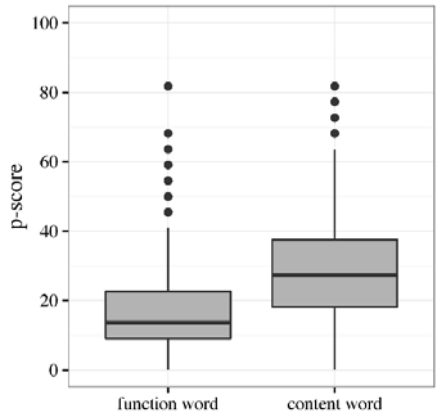

Figure 4: P-scores as a function of part-of-speech class.

As a third factor, overall *word duration* had an impact on the prominence ratings ($X^2(1) = 857.16$, $p < 0.0001$). In this continuous parameter, a change by one standard deviation increases the odds of a prominence response by 1.9 to 1 (logit estimate: 0.62, SE = 0.02). Figure 5 shows this effect as a tendency of longer words to reach a higher p-score, i.e. the longer the word, the more participants marked it as prominent.

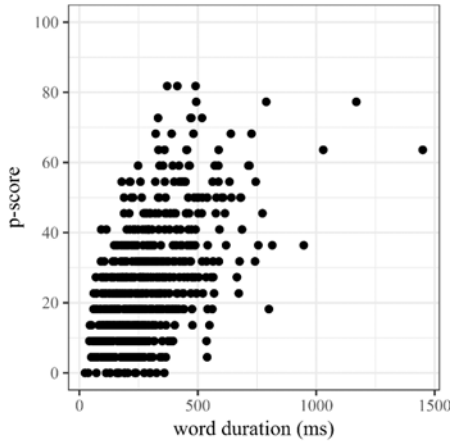

Figure 5: P-scores as a function of word duration.

The remaining four test variables were found to be more marginally relevant and clearly overlap with one of the three preceding variables. Thus, *mean syllable duration* and *number of syllables (actually realized)* – both with odds ratios of 1.6 to 1 – are obviously related to word duration. Similarly, *syntactic break* (odds ratio 1.5 to 1) and *last verbal argument* (odds ratio 1.6 to 1) often overlap with pauses.

4. Results of the Boundary Experiment

4.1. Inter-rater and multi-rater agreement

The first result to note with regard to inter-rater agreement is that our participants performed much better in the boundary experiment (Experiment II) than they did in the prominence experiment (Experiment I). That is, inter-rater agreement was much higher in the former than in the latter. Table 3 repeats the Fleiss' kappa scores for the prominence experiment (cf. §3.1) and contrasts them with the scores for the boundary experiment.

Prominences	Boundaries
0.103	0.407
z = 42.1	z = 160

Table 3: Fleiss' kappa scores for prominences and boundaries in PM.

Comparing the boundary scores again with Mo et al.'s (2008) RPT results for American English, we see that – in contrast to the prominence scores – English listeners and Papuan Malay listeners are not too far apart in their perception of boundaries: 0.544 for American English vs. 0.407 for Papuan Malay.

As with the prominence experiment, we additionally looked at the pair-wise inter-rater agreement. Table 4 summarizes the Cohen's kappa values by using the agreement categories of Landis & Koch (1977). Compared with the results of the prominence experiment (repeated in the second and third columns of Table 4), we see a clear difference between the two experiments: while in the prominence experiment more than 80 % of all rater pairs showed either no or only slight agreement, only about 18 % of the rater pairs showed such low agreement in the boundary experiment. Instead, the majority of pairs who participated in the boundary experiment (more than 60 %) showed moderate or even substantial agreement.

	Prominences		**Boundaries**	
inter-rater agreement	**pairs**	**percentage**	**pairs**	**percentage**
none	25	10.82 %	4	1.73 %
slight	164	71.00 %	37	16.02 %
fair	40	17.32 %	51	22.08 %
moderate	2	0.87 %	106	45.89 %
substantial	0	0.00 %	33	14.29 %
(almost) perfect	0	0.00 %	0	0.00 %
	231	100 %	231	100 %

Table 4: Inter-rater agreement categories (based on Cohen's kappa scores) for PM subjects in both experiments.

4.2. Factors determining perceived boundaries

As we have seen in the previous section, the overall agreement of raters is better in the boundary experiment than in the prominence experiment. This is reflected in Figure 6, where we observe a modal value of 0, which is to be expected as there are usually many more word pairs with no boundaries between them than ones where the two words are separated by a boundary. We can also observe a longer tail to the right, indicating that higher scores are reached than in the prominence experiment. That is, the participants agreed more on the position of boundaries than on the position of prominences (top scores: 95.5 % as opposed to 81.8 %). However, even though the agreement among raters is higher for boundaries than for prominences, complete agreement (on the *presence* of a boundary) is never achieved.

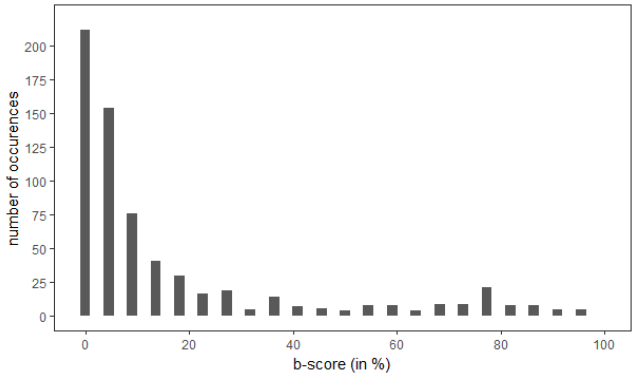

Figure 6: Distribution of b-scores in the PM data.

When correlating the multiple possible factors introduced in §2 with the outcome, the only variable that does not reach significance is *part-of-speech class* ($X^2(1) = 0.7962$, p = 0.3722). In the same way as with

the prominence results, however, we will concentrate only on the strongest effects, indicated by odds ratios bigger than 1.5 to 1. The variables not considered further are the morpho-syntactic parameters *part-of-speech* and *last argument*, and the duration/syllable number measures *duration of last syllable*, *number of syllables (phonological)* and *number of syllables (actually realized)*. This also includes two of the pitch measures, i.e. *minimum* and *mean pitch*, but note that the other two pitch measures (*maximum pitch* and *pitch range*) are also only marginally effective (odds ratio 1.6 to 1 for *pitch range* and odds ratio 1.7 to 1 for *maximum pitch*).

The most significant factor affecting the perception of a boundary in this experiment was the presence of a *pause* ($X^2(1) = 1519$, $p < 0.0001$). As illustrated in Figure 7, the presence of a pause in contrast to a non-interrupted signal increased the odds of a positive response for boundary by 22.9 to 1 (logit estimate: 3.13, SE = 0.09).

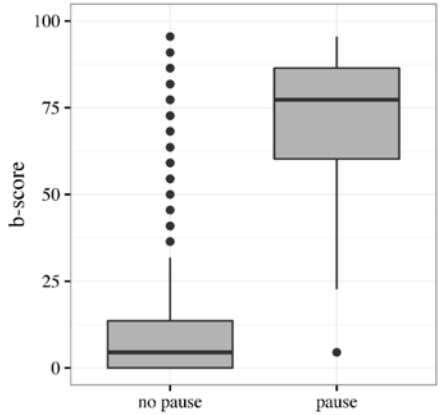

Figure 7: B-score as a function of pause.

Although much weaker, another major effect on the perception of boundaries was found in the *syntactic structure* of the utterances ($X^2(1) = 1514.2$, $p < 0.0001$). As Figure 8 indicates, the type of syntactic break influences the perception of a boundary. Thus, a change of one unit increases the odds of observing a boundary response by 3.4 to 1 (logit estimate: 1.23, SE = 0.03). The effect size can be explained by the amount of variability shown in the plot and the overly coarse values for this parameter. Thus, there are quite a few instances where participants agreed on the presence of a boundary even though there was no major (clausal) syntactic break. Such boundaries typically involve a clause-internal syntactic break such as the right edge of a topic or subject NP. Recall that the syntactic break parameter only distinguishes subordinate clause and sentence boundaries from no boundary (= all syntactic boundaries within a clause). Furthermore, participants did not always agree on perceiving a prosodic boundary at sentence boundaries (= strong syntactic breaks), which in part is due to the fact that sentence boundaries are often not easy to determine in spontaneous discourse. The high variability, especially in cases of a strong syntactic break, leads to a relatively small effect of *syntax*, although the mean values of the two categories *weak* and *strong break* are far apart from each other.

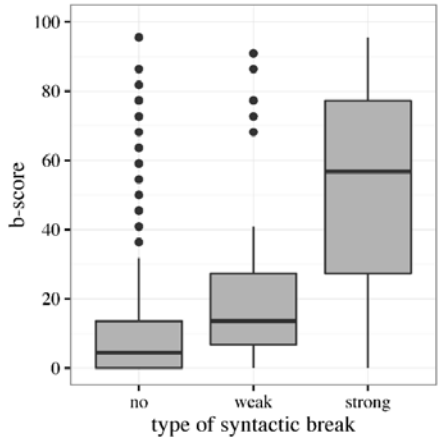

Figure 8: B-score as a function of syntactic structure.

Next to *pause* and *syntax*, the *mean duration of syllables* was found to be the third most important factor for the perception of boundaries, but the effect here is relatively weak ($X^2(1) = 1415.1$, $p < 0.0001$; see Figure 9). Thus, a change of one standard deviation increases the odds of getting a boundary response by 2.6 to 1 (logit estimate: 0.95, SE = 0.029).

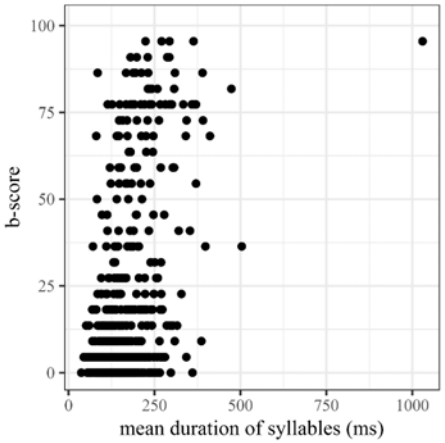

Figure 9: B-score as a function of mean syllable duration.

We found almost the same effect size for the parameter *word duration* ($X^2(1) = 1423.7$, $p < 0.0001$), where a change of one standard deviation increases the odds of observing a boundary response by 2.5 to 1 (logit estimate: 0.90, SE = 0.028).

5. Discussion

If we compare the results gained in the two RPT-experiments, we find a high degree of variability for prominence judgments, but less variability for boundary judgments. The lack of agreement with regard to prominence judgments is reflected in overall low Kappa values and low p-scores (best p-score achieved: 81.8 %; modal value: 13.6 %). The considerably stronger agreement in the perception of prosodic boundaries is shown by much higher Kappa values and more consistent b-scores at both ends of the scale (best b-score achieved: 95.5 %; modal value: 0 %).

When correlating the native judgments of Experiment I and II with the various parameters that might affect the perception of prominences and boundaries, respectively, we observe an interesting pattern (see Table 5): the two *prosodic* factors most important in influencing prominence and boundary ratings are basically the same; namely, *pause* and *word duration/mean syllable duration*.[5] Apart from the considerable difference in effect size with regard to the parameter *pause*, the major difference between the two experiments pertains to the *non-prosodic* factor found to be most influential for the relevant judgment. *Part-of-speech class* was found to be a relevant cue for prominence but not for boundaries. This is not surprising, as content words are generally claimed to be more prominent than function words, due to their higher semantic weight or structural strength (see Büring 2012) and to their (commonly) lower word frequency (see Cole et al. 2010a). In contrast, *syntactic structure* becomes more important when it is the participants' task to judge the position of boundaries. This, again, is in line with findings for other languages (see, for example, Cole et al. 2010b).

Table 5 lists the three most important factors determining prominence and boundary judgments in descending order. Importantly, and somewhat surprisingly from a European perspective, the relevant prosodic factors are not only (almost) the same across both experiments, but the ranking is also the same, i.e. *pause* in first and *word duration/mean syllable duration* in third position.

[5] For boundaries, *word duration* is the fourth most influential parameter with an odds ratio of 2.5, and *mean syllable duration* is the third-most effective cue, with an odds ratio of 2.6. As this difference is extremely small, we regard these two factors as equally effective with regard to prosodic boundaries. In the case of prominences, the difference between *word duration* (odds ratio of 1.9) and *mean syllable duration* (odds ratio 1.6) is somewhat more pronounced, but still not very large.

Experiment I: Prominences		Experiment II: Boundaries	
	OR		**OR**
pause	2.7	pause	22.9
part-of-speech class	2.1	syntactic structure	3.4
word duration	1.9	mean syllable duration /word duration	2.6/2.5

Table 5: Major effects for both experiments in terms of their effect size (odds ratio =OR).

When comparing the odds ratios, it is obvious that the effect sizes for the prominence-lending parameters are smaller than their counterparts for boundary perception. The rather small effect sizes are linked to the very high degree of variability in the prominence ratings (low Kappa values). That is, the effects these parameters may have on prominence judgments clearly do not lead to substantial agreement with regard to these judgments. In fact, the high degree of variability raises the question of whether the notion of prominence makes any sense to PM speakers, a point we will return to below.

For boundaries, by contrast, the effects seem to be more robust. Furthermore, the variability observed here is within the range of variability observed for other languages (see section 4.1 above). Major phonetic cues for prosodic boundaries are pauses and longer word and syllable durations, which are widely attested cross-linguistically (see, for example, Turk & Shattuck-Hufnagel 2007). In fact, with regard to boundary perception, native hearing and Western auditory analysis appear to be quite similar, as revealed by the comparison of the expert rating by the four authors (cf. §2) and native listeners' judgments in Figure 10. In most cases in which experts did not perceive a prosodic boundary, the native raters also tended towards the perception of no boundary, which is indicated in the plot by larger dots for lower b-scores. However, in some cases where non-native experts did not perceive a boundary, there was a considerable agreement among native raters that they perceived a boundary. The opposite pattern can be observed for instances in which the experts did perceive a boundary: there are fewer instances of low b-scores but (slightly) more instances of higher b-scores. Statistically, this pattern is mirrored by a strong correlation for the perception of boundaries between the two groups ($X^2(1) = 2949.9$, $p < 0.0001$). When the experts observed a boundary in contrast to no boundary, the odds of a boundary response by the native listeners increased by 26.8 to 1 (logit estimate: 3.3, SE = 0.07), which is higher than the strongest factor influencing native speakers' boundary judgments (i.e. *pause* with an odds ratio of 22.9).

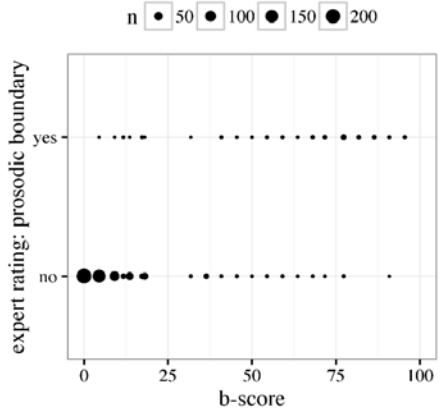

Figure 10: Correlation between non-native (German) experts' and native (PM) listeners' boundary perception (indicated by binary scores and b-scores, respectively). The size of the dots indicates the number of compatible observations (the more observations, the larger the dot).

Given the very weak inter-rater agreement results for prominence judgments and the fact that the same prosodic cues appear to play a role in judging prominences and boundaries, we tentatively conclude that the perception of prominence is to some extent conflated with the – more clearly conceptualized – perception of (prosodic) boundaries in PM. This conclusion is in line with similar observations quoted from the literature in section 1, which also raise doubts as to the feasibility of separating prosodic prominences from prosodic boundaries in other Malayic varieties.

It should be noted, however, that there is no perfect match between prominences and boundaries in that natively perceived boundaries are not reliable predictors for prominences. Testing the effect of b-scores on p-scores, we found an odds ratio of only 1.1 to 1, although the likelihood-ratio test revealed significance. The small effect size is mainly due to the fact that substantially more prominences were marked than boundaries. Recall from §3.2 that only 20 of the 726

words (i.e. 2.75 %) occurring in the test items were unanimously judged to *lack* prominence. In contrast, of the 674 non-final words in the test utterances, 212 were unanimously judged *not* to precede a boundary (i.e. 31.45 %).

Perhaps the most surprising result of our preliminary exploration is the fact that pitch-related parameters do not appear to play a role for PM speakers in judging prominences and boundaries. Recall from §3.2 and §4.2 that only *maximum* and *mean pitch* were found to be marginally effective in the case of boundary judgments, but well below the more effective parameters *pause*, *syntactic break*, and *mean syllable/word duration*. This finding is particularly relevant because the claims of Western researchers regarding lexical stress differences in Malayic varieties appear to be primarily based on differences in pitch alignment, with high pitch targets being heard as located on either the penultimate or the ultimate syllable of a word. The production study by Maskikit-Essed & Gussenhoven (2016) for Ambonese Malay already questioned whether there is in fact a clear alignment of pitch targets with respect to syllable boundaries. Our study suggests that, although modulations of pitch are clearly present (acoustically as well as perceptually to the Western ear), these do not appear to play a role either in the perception of boundaries or in the marking of prosodic prominences in PM – and possibly other Malayic varieties.

In fact, prosodic prominence may not be a relevant category in PM and other Malayic varieties in general, which would thus represent further instances of what has been termed *stress deafness* (see e.g. Peperkamp & Dupoux 2002, Dupoux et al. 2010 for French). However, our results are not directly comparable with this line of work as the methods used quite clearly differ. It is also far from clear whether stress deafness is a homogeneous phenomenon. Hence, it may turn out that the French and PM cases only partially overlap, if at all.

We need more data to answer the question of whether PM listeners are really insensitive to prominence-lending pitch modulations. This includes the further question of whether they do not respond to pitch modulations at all, i.e. also when rating languages that are known to primarily use pitch in the marking of prosodic prominence (a study presenting German stimuli to Papuan Malay listeners is currently under way). If we were to find higher prominence scores when PM listeners rate German data, the present results would only support the conclusion that pitch modulations are not systematically employed in prominence marking in PM, thus confirming similar findings in the literature reported in §1. If prominence scores by PM listeners prove to be similar across different languages, this would suggest a more general account in terms of stress deafness for PM listeners.

We would also like to add a cautionary note regarding the notion of '(post-)lexical stress' as it has been used in this chapter and in much of the previous descriptive and experimental literature. Inasmuch as 'stress' is understood to be a phenomenon that pertains to the phonologically organized highlighting of a syllable relative to adjacent ones by way of modulating phonetic parameters such as pitch and duration, the current study supports the conclusions of earlier studies that lexical stress is not part of the prosodic system of Malayic varieties. As pointed out in the introduction, the cases of Papuan and Ambon Malay are particularly interesting in this regard, because pitch modulations here appear to be – both acoustically and perceptually to the Western ear – very regular and clearly anchored to different syllables (penultimate or final), unlike in the Indonesian spoken by Javanese native speakers, where pitch modulations are much more variable.

In this context, it should be noted that it is very well possible that in PM pitch targets are clearly aligned with syllables, in contrast to Maskikit-Essed & Gussenhoven's (2016) claims for Ambonese Malay. If this were to be the case, we would need a stress-like notion to be able to account for differing alignments of pitch targets with penultimate and final syllables which, however, would differ from the standard understanding of 'lexical stress', as this distinction does not appear to be perceived as a prominence distinction by native speakers.

While it thus seems very likely that prosodic prominence is organized differently in these languages, a number of phenomena may still need to be accounted for in stress-related metrical categories. To give just one more example, Kluge (2014) makes the occasional reference to stress distinctions in discussing segmental alternations in PM. An example is the observation that /s/ is only palatalized in unstressed syllables (Kluge 2014: 73). If one denies lexical stress distinctions in the standard sense given above, one needs to identify another factor that adequately constrains the palatalization rule. Furthermore, lexical stress in the sense of phonologically organized prominence distinctions is of course not the only possible prosodic organization at word level. Foot structure, for example, may be evident in terms of phenomena not directly reflected in phonetic differences. Thus, it

should be clearly understood that denying the existence of lexical stress in these languages does not mean that there is no word-prosodic organization at all.

Acknowledgements

Work on this chapter by Riesberg and Himmelmann was generously supported by the Volkswagen Foundation within the scope of the project “Documentation Summits in the Central Mountains of Papua” (Az 85892). We are grateful to the Centre of Endangered Languages Documentation (CELD) in Manokwari, particularly to Yusuf Sawaki, Jean Lekeneny and Anna Rumaikeuw, for providing support and the facilities for conducting the experiments. Special thanks to Jan Strunk and Christoph A. Bracks for computing the kappa statistics, and to Katherine Walker for improving style and grammar.

Abbreviations

b-score = boundary-score; DEM = demonstrative; PM = Papuan Malay; p-score = prominence score; REL = relative pronoun; RPT = Rapid Prosody Transcription

Appendix A

Instructions for Experiment I (Prominences)

Pertama-tama kami mengucapkan terima kasih karena Anda bersedia berpartisipasi dalam eksperimen tentang bagaimana Anda memahami bahasa. Jawaban yang Anda berikan tidak ada yang salah atau benar karena semuanya bergantung pada rasa bahasa.

Dalam berbicara seseorang akan mengucapkan beberapa atau banyak kata dalam sebuah kalimat dengan nada yang lebih menonjol dibandingkan dengan kata-kata lain yang terdapat dalam kalimat tersebut. Kata-kata dengan nada yang menonjol ini biasanya dapat dirasakan oleh pendengarnya. Tugas Anda adalah menandai (menggarisbawahi) kata-kata yang nadanya Anda dengar lebih menonjol dibandingkan dengan kata-kata lain dalam rekaman kalimat yang akan Anda putar.

Berikut ini Anda akan diputarkan 56 kalimat. Setiap kalimat juga akan disajikan dalam bentuk tertulis. Untuk mulai silakan klik **Contoh 1**, dst.

Tugas Anda adalah menggarisbawahi **semua** kata yang nadanya Anda anggap lebih menonjol (mis. lebih tinggi) dibandingkan dengan kata-kata lain pada setiap rekaman kalimat yang Anda dengarkan. Silakan garis bawahi kata tersebut dengan cara seperti ini:

Dia melihat sapi

Dalam hal ini, Anda dimungkinkan untuk memilih lebih dari satu kata pada setiap rekaman kalimat!

Dia melihat sapi dan kuda makan rumput

Anda dapat memutar setiap rekaman kalimat sebanyak dua kali. Akan tetapi, tidak memungkinkan untuk menghentikan rekaman pada saat contoh kalimat sedang diputar.

Jika Anda harus mengoreksi pilihan Anda, silakan coret kata yang telah Anda garis bawahi dengan cara seperti ini:

~~*sapi*~~

Selamat mengikuti eksperimen ini!

Instructions for Experiment II (Boundaries)

Ketika seseorang berbicara, dia akan membagi ucapan mereka menjadi potongan-potongan. Potongan-potongan tersebut membentuk kelompok kata-kata yang memudahkan pendengar untuk memahami ucapan pembicara. Potongan-potongan tersebut penting terutama saat pembicara memproduksi ucapan yang panjang.

Contoh potongan yang mungkin Anda ketahui adalah potongan nomor ketika Anda memberi tahu nomor telepon Anda kepada orang lain. Biasanya, Anda tidak setiap kali memberi satu nomor (0, 8, 1, 3 …), tetapi Anda akan memotong nomor hp tersebut menjadi kelompok-kelompok yang terdiri atas dua, tiga, atau empat angka (081, 358, 772 …).

Untuk rekaman yang akan Anda dengar, Anda diminta untuk

menandai potongan dengan cara menyisipkan garis tegak lurus atau vertikal (pada cetakan) untuk bagian yang Anda dengar sebagai satu potongan. Batas antara dua potongan tidak harus sama dengan lokasi tempat Anda akan menulis tanda koma, titik, atau tanda baca lainnya. Jadi, Anda harus benar-benar hati-hati mendengar ujaran dan tandai batas yang Anda dengar sebagai akhir sebuah potongan.

Sebuah potongan mungkin saja berupa satu kata, atau mungkin terdiri atas beberapa kata, dan ukuran (jumlah kata) dalam setiap potongan dari para pembicara bisa saja berbeda-beda dalam satu ujaran. Beberapa ujaran mungkin Anda dengar konsisten, yaitu terdiri atas satu potongan saja. Jika demikian, Anda tidak perlu menandai batas potongan.

Contoh:

081|358|772…

0813|5877|2…

Bapak saya | sudah datang

Bapak | saya sudah datang

Appendix B

Instructions for Experiment I (Prominences)

First of all, we want to say thank you for participating in our experiment on how people perceive language. There is no right or wrong answer - we are just interested in your innate sense of language.

When talking, people will stress or emphasize some words within a sentence more than others. These stressed words can usually be perceived by the hearer. Your task in this experiment is to point out (underline) all words that you perceive to be more emphasized compared to the rest of the utterance in the recordings that we will play to you.

You will hear 56 sentences. You will also receive each sentence as a written transcript. To start, please click **Example 1**, and so on.

Your task is to underline **all** words that you perceive to stick out (e.g. because they are higher/louder) compared to the other words in each recording that you will hear. Please underline your choice in the following way:

He sees a cow

It is possible to choose more than one word for each recording!

He sees a cow and a horse eating grass

You can play each recording twice. It will not be possible to stop the recording while it is playing.

If you want to make a correction to your choice, please cross out the underlined word.

~~*cow*~~

Enjoy the experiment!

Instructions for Experiment II (Boundaries)

When people speak, they chunk their utterances into units. These chunks of words help the hearer to understand the utterance. They are especially important if the speaker produces longer, coherent speech.

An example you might be familiar with is the chunking of digits when giving somebody your telephone number. Instead of spelling one digit after another (0, 8, 1, 3 …), it is common to divide the number into units consisting of two, three, or four digits each (081, 358, 772 …).

For the recordings you will hear, you are asked to mark those chunks by inserting a vertical line (on the printout) to divide what you perceive to be a unit. The boundary between two chunks does not necessarily have to coincide with where one would write a comma, a full stop or any other punctuation, so please listen carefully and draw the line where you hear the end of one unit.

One unit might consist of one word only, or it can contain several words - the size of a unit might vary from utterance to utterance. Some recordings might consist of one unit only. If this is the case, you don't have to draw a boundary.

Examples:

081|358|772…

0813|5877|2…

let's eat grandpa

let's eat | grandpa

References

Alieva, Natalia F. & Arakin, Vladimir D. & Ogloblin, Alexander K. & Sirk, Yu H. 1991. *Bahasa Indonesia: Descripsi dan teori*. Yogyakarta: Kanisius.

Bates, D. & Maechler, M. & Bolker, B. & Walker, S. 2015. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8, <URL: http://CRAN.R-project.org/package=lme4>.

Baumann, Stefan & Winter, Bodo. 2015. Comparing prosodic and non-prosodic factors in naïve listeners' prominence judgments. (Paper presented the conference Phonetics and Phonology in Europe (PaPE), Cambridge 29 June 2015.)

Büring, Daniel. 2012. Predicate Integration - Phrase Structure or Argument Structure? In Kucerova, Ivona & Neeleman, Ad (eds.), *Contrasts and Positions in Information Structure*. Cambridge: Cambridge University Press. 27-47.

Chafe, Wallace. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, New Jersey: Ablex

Clynes, Adrian & Deterding, David. 2011. Standard Malay (Brunei). *Journal of the International Phonetic Association* 41(2). 259–268.

Cohn, Abigail C. 1989. Stress in Indonesian and bracketing paradoxes. *Natural Language and Linguistic Theory* 7, 167–216.

Cohn, Abigail C. & McCarthy, John J. 1994. Alignment and parallelism in Indonesian phonology. *Working Papers of the Cornell Phonetics Laboratory* 12, 53–137

Cole, Jennifer & Mo, Yoonsook & Hasegawa-Johnson, Mark. 2010a. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1. 425–452.

Cole, Jennifer & Mo, Yoonsook & Baek, Soondo. 2010b. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and*

Cognitive Processes 25. 1141–1177.

Dupoux, E., Peperkamp, S. & Sebastián-Gallés, N. (2010). Limits on bilingualism revisited: stress 'deafness' in simultaneous French-Spanish bilinguals. Cognition, 114, 266-275.

Ethnologue – Languages of the World. 2016. SIL International Publications. https://www.ethnologue.com/

Field, Andy & Miles, Jeremy & Field, Zoë. 2012. *Discovering statistics using R*. LosAngeles et al.: Sage.

Goedemans, Rob & van Zanten, Ellen. 2007. Stress and accent in Indonesian. In van Heuven, Vincent J. & van Zanten, Ellen (eds.), *Prosody in Indonesian languages*, 35-63. Utrecht: LOT

Halim, Amran. 1974. *Intonation in relation to syntax in Bahasa Indonesia*. Jakarta: Djambatan.

Himmelmann, Nikolaus P. 2010. Notes on Waima'a intonation. In Ewing, Michael & Klamer, Marian (eds.), *East Nusantara: Typological and areal analyses*, 47-69. Canberra: Pacific Linguistics.

Kluge, Angela. 2014. *A grammar of Papuan Malay.* Utrecht: LOT.

Ladd, Robert D. 2008. *Intonational phonology.* 2nd edn. Cambridge: Cambridge University Press.

Laksman, Myrna. 1994. Location of stress in Indonesian words and sentences. In Odé, Cecilia & van Heuven, Vincent J. (eds.) *Experimental studies of Indonesian prosody* (Semaian 9. Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië), 108-139. Leiden: Leiden University.

Landis, J. Richard & Koch, Gary G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Lewis, M. Paul & Simons, Gary F. & Fennig, Charles D. (eds.). 2016. *Ethnologue: Languages of the World*, Nineteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Maskikit-Essed, Raechel & Gussenhoven, Carlos. 2016. No stress, no pitch accent, no prosodic focus: The case of Ambonese Malay. *Phonology* 33. 353-389.

Mo, Yoonsook & Cole, Jennifer & Eun-Kyung Lee. 2008. Naïve listeners' prominence and boundary perception. ISCA Archive, http://www.isca-speech.org/archive.

Odé, Cecilia. 1997. On the perception of prominence in Indonesian: An experiment. In Odé, Cecilia & Stokhof, Wim(eds.), *Proceedings of the Seventh International Conference on Austronesian Linguistics*, 151-166. Amsterdam: Rodopi.

Peperkamp, Sharon & Dupoux, Emanuel. 2002. A typological study of stress 'deafness'. In: Gussenhoven, Carolos & Warner, Nancy (eds.), Laboratory Phonology 7, 203-240. Berlin: Mouton de Gruyter.

Pitt, Mark A.& Dilley, Laura & Johnson, Keith & Kiesling, Scott & Raymond, William & Hume, Elizabeth, et al. 2007. *Buckeye corpus of conversational speech* (second release). Columbus, OH: Department of Psychology, Ohio State University.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna, Austria. Version 3.2.2.

Rietveld, Toni C. M. & Gussenhoven, Carlos. 1985. On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics* 15. 273-285.

Roosman, Lilie. 2007. Melodic structure in Toba Batak and Betawi Malay word prosody. In van Heuven, Vincent J. & van Zanten, Ellen (eds.), *Prosody in Indonesian languages*, 89-116. Utrecht: LOT

Stoel, Ruben B. 2007. The intonation of Manado Malay. In Vincent J. van Heuven & Ellen van Zanten (eds) *Prosody in Indonesian Languages*. Utrecht: LOT. 117–150.

Turk, Alice E. & Shattuck-Hufnagel, Stefanie. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35. 445–472.

Uhmann, Susanne. 1988. Akzenttöne, Grenztöne und

Fokussilben. Zum Aufbau eines phonologischen Intonationssystems für das Deutsche. In Altmann, Hans (ed.), *Intonationsforschungen*, 65-88. Tübingen: Niemeyer.

van Heuven, Vincent J. & van Zanten, Ellen. 2007. Concluding remarks. In van Heuven, Vincent J. & van Zanten, Ellen (eds.), *Prosody in Indonesian languages*, 191-202. Utrecht: LOT.

van Minde, Don. 1997. *Malayu Ambong: phonology, morphology, syntax*. Leiden: University of Leiden. (Doctoral dissertation.)

van Zanten, Ellen. 1994. The effect of sentence position and accent on the duration of Indonesian words: A pilot study. In Odé, Cecilia & van Heuven, Vincent J. (eds.), *Experimental studies of Indonesian prosody* (Semaian 9. Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië), 140-180. Leiden: Leiden University.

van Zanten, Ellen & van Heuven, Vincent J. 2004. Word stress in Indonesian: Fixed or free? *NUSA Linguistic Studies of Indonesian and other Languages in Indonesia* 53, 1-20.

Zubkova, Ludmila G. 1966. *Vokalizm Indonezijskogo jazyka*. St Petersburg: Leningrad University. (Doctoral dissertation.)