Laboratory Phonology
Journal of the Association for Laboratory Phonology

Riesberg, S., et al. 2020 Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 8, pp. 1–35. DOI: https://doi.org/10.5334/labphon.192

JOURNAL ARTICLE

# Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay

Sonja Riesberg[1,2], Janina Kalbertodt[1], Stefan Baumann[1] and Nikolaus P. Himmelmann[1]

[1] Universität zu Köln, Cologne, DE

[2] ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, AU

Corresponding author: Sonja Riesberg (sonja.riesberg@uni-koeln.de)

This paper shows how the Rapid Prosody Transcription method (RPT, cp. Cole & Shattuck-Hufnagel, 2016) can be utilized when investigating the prosodic systems of a little-described language. We report the results of a set of perception experiments on the prosody of Papuan Malay, which support the claim made in earlier (production) studies that Malayic varieties appear to lack stress (i.e., lexical stress as well as post-lexical pitch accents). We show that inter-rater agreement of speakers of Papuan Malay is much lower for prosodic prominences than for boundaries when rating their native language. However, they show higher agreement when asked to rate prominences in German. Most importantly, they seem to make use of the same acoustic cues as a German control group. We therefore conclude that while Papuan Malay indeed seems to lack post-lexical pitch accents, speakers of Papuan Malay appear to be able to perceive the accentual prominences characteristic of German.

**Keywords:** Rapid Prosody Transcription; Papuan Malay; German; prosodic prominence; stress

## 1. Introduction

The prosodic systems of most of the roughly 7,000 languages still spoken today are virtually unknown. Until very recently, prosodic analyses have not been included as standard in the kind of basic grammatical descriptions produced by fieldworking linguists. This has improved to some extent in the last decade or so, and there are now a number of practical guides as to how to include prosody in language documentation and description (Himmelmann, 2006; Himmelmann & Ladd, 2008; Jun & Fletcher, 2014). However, the task of documenting the world's prosodic diversity remains considerable, and tools that may facilitate and accelerate this work are still urgently needed.

This contribution explores the usefulness in this regard of the *Rapid Prosody Transcription* (RPT) method, whereby naïve (i.e., linguistically untrained) listeners rate audio recordings of spoken language for prosodic prominences and boundaries. This approach was developed by Jennifer Cole and colleagues (e.g., Cole et al., 2010a; Cole et al., 2010b; see Cole & Shattuck-Hufnagel, 2016, p. 7–13 for a concise introduction and further references), following up on earlier work in the tradition of the perception-oriented approach to intonation developed in Eindhoven as summarized in 't Hart et al. (1990); see also Himmelmann et al. (2018), p. 210–213 for further details and references. The present article follows up on a suggestion by Cole and Shattuck-Hufnagel (2016, p. 11) that "RPT can even be used to explore prosody, from the perspective of the listener, in languages for which the prosodic phonology has not yet been worked out, and such data may be then used as the basis for developing more articulated grammatical models."

The language used as a test case is Papuan Malay, a Trade Malay variety spoken in eastern Indonesia. A first test run using RPT with Papuan Malay speakers, reported in Riesberg et al. (2018), showed that Papuan Malay speakers do not agree on instances of prosodic prominences, while at the same time they are able to agree on prosodic boundaries. The experiments reported in the present paper continue this work and attempt to further clarify which prosodic events are salient for speakers of Papuan Malay. In doing so, we also extend the RPT methodology by exploring its crosslinguistic applicability. That is, Papuan Malay speakers not only provide judgments on naturally produced Papuan Malay, but also on naturally produced German data. Moreover, as a control and for comparative purposes, the same experiments are also run with German participants who, accordingly, provide judgments on both German and Papuan Malay recordings.[1]

There are two reasons for trying out this crosslinguistic setup. First, it is well known that judgments on prosodic events are strongly influenced by non-prosodic factors, including syntax, semantics, and pragmatics (e.g., Cole et al., 2010a). Having listeners with no command of the target language provide judgments removes these potentially confounding factors and forces them to exclusively pay attention to prosody. Second, having speakers with different native languages provide judgments on the same dataset allows for a direct comparison of the kinds of prosodic events that provoke a reaction.

The core feature of the RPT method is to have naïve listeners mark prosodic prominences and boundaries in short segments of recordings, thus providing a first glimpse as to which (prosodic as well as non-prosodic) factors might be of relevance to them in this regard. It should be obvious that such an approach cannot establish 'facts' of the type 'language X makes use of pitch accents' or 'speakers of language X hear durational differences as marking lexical stress.' Instead, the ideal outcome of RPT experiments is to be able to develop a set of hypotheses that can subsequently be tested with other methods. This first step towards generating hypotheses is thus the goal of the present contribution, rather than a fully worked-out analysis of the prosodic system of Papuan Malay.

The paper is structured as follows. Section 2 provides some background concerning what is known about prosody in Malayic varieties so far, highlighting those points where they appear to differ from better-studied European languages. Section 3 presents essential information on data collection and methods of analysis. Section 4 is dedicated to reporting the core results, with relatively little discussion, which follows in Section 5. Note that there are two appendices which provide further details on data and methods as well as the statistical models underlying the result summary in Section 4. Given that we are comparing two speaker groups in two tasks in two languages each, the empirical design of the study is fairly complex, comprising eight inter-rater agreement results tables showing analyses for up to 16 potentially contributing factors. We have therefore tried to focus our presentation on the most significant results, relegating fuller documentation to Appendix B.

## 2. Prosody in Papuan Malay

Papuan Malay (henceforth PM) is a trade variety of Malay spoken in the two easternmost provinces of Indonesia—Papua Barat and Papua—by approximately 1.2 million speakers (see Kluge, 2017). It is spoken mostly in the coastal areas, and to a lesser extent in the

[1] Riesberg et al. (2018) already contains a first brief comparison with results of a study in which German speakers used the RTP method for rating German (Baumann & Winter, 2018). Baumann and Winter's study, however, used read sentences rather than excerpts from natural spoken language and thus does not provide the best basis for comparison. The current paper, therefore, replicates the comparison between Papuan Malay and German with a proper comparison set; that is, by comparing the same data type. Note also that the Kappa values by Papuan Malay raters presented in this paper differ slightly from the values presented in Riesberg et al. (2018). This is due to the fact that the current analyses are based on a reduced data set (40 excerpts, rather than 56 excerpts, as in Riesberg at al., 2018).

mountainous inland. The Indonesian half of New Guinea, with its more than 270 indigenous languages, is linguistically highly diverse, and most speakers are at least bilingual. PM serves as the lingua franca in this area, and most native speakers speak PM in addition to one or more local languages.

With regard to prosody, Malayic varieties are widely believed to fall into two groups, western and eastern. Western varieties, which include Standard Indonesian as well as Standard Malaysian Malay, are not further discussed here.[2] PM belongs to the Eastern varieties, which also include Ambon and Manado Malay. These varieties are widely claimed to make use of word-level prominence distinctions and to have a default linking pattern for an IP-final edge tone combination in which the prefinal edge tone is linked to the penultimate syllable, and the final boundary tone is linked to the final syllable. In a minority pattern, both tones are linked to the final syllable. Consequently, these languages are usually analyzed as having penultimate stress, with a small number of lexical bases, often loan words, having final stress (e.g., van Minde, 1997 for Ambon Malay; Stoel, 2007 for Manado Malay).

This is also the analysis proposed for PM by Kluge (2017). In order to support it, Kluge 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 (Kluge, 2017, p. 63). Kluge concludes that 964 (90%) of all words have penultimate stress (including both open and closed penultimate syllables), and only 108 (10%) have stress on the final syllable. Of the 108 words that displayed ultimate stress, 105 (97%) contained the front open-mid vowel /ɛ/ (cognate with schwa in the Malayic varieties in which schwa is part of the phoneme inventory) in the penultimate syllable. Yet, Kluge notes that /ɛ/ does not condition ultimate stress; as for 61 of those words with penultimate stress the stressed syllable also contained an /ɛ/ (cp. Kaland et al., 2019 for further discussion and details). Importantly, Kluge's analysis is exclusively based on her own auditory judgment, which is shared by other speakers of Germanic languages, including the authors of the present article. The analysis receives some, though not particularly strong, support from a thorough investigation of acoustic parameters by Kaland (2019), who finds that the syllables considered as stressed by Kluge differ from unstressed ones with regard to duration, formant displacement, and spectral tilt. Note that for the related variety Ambon Malay, Maskikit-Essed and Gussenhoven (2016) claim not to be able to find any evidence for word-level prominences.

As for phrase-level prosody, no detailed investigations are yet available for PM (but see Himmelmann, 2018 for some preliminary observations). Maskikit-Essed and Gussenhoven's (2016) work on Ambon Malay is one of only two studies that address the issue of intonational structure in eastern Malayic varieties in more detail. The other study is Stoel's (2007) analysis of Manado Malay. Most importantly in the current context, the two analyses agree in that there are no post-lexical pitch accents in these varieties, the only intonational targets on the IP level being an edge tone combination occurring at the end of the IP. This combination is typically a rise-fall, but a continuous rise or a fall-rise are also possibilities. Maskikit-Essed and Gussenhoven (2016) analyze this tonal movement as a floating boundary tone (i.e., HL% for the rise-fall in an autosegmental-metrical analysis[3]). An alternative analysis, used here primarily for expository purposes,

[2] See van Heuven and van Zanten (2007) and Riesberg et al. (2018) for brief summaries of, and references to, the work on Indonesian. Himmelmann and Kaufman (in press) and Kaufman and Himmelmann (accepted) provide a short overview on what is known about prosody in Austronesian languages more generally.

[3] The present exposition is also couched in the terminology and formalisms used in the autosegmental-metrical framework for prosodic analysis (Gussenhoven, 2004; Ladd, 2008), and more specifically the Tone and Break Indices (ToBI) framework (Beckman et al., 2005). 'H' stands for a high tone, 'L' for a low tone.

would be a combination of a phrase tone (H-) and a boundary tone (L%).[4] This means, inter alia, that whenever a phrase is expanded, pitch targets move to the right and segments that are prosodically prominent when occurring at the edge are no longer prominent. Example (1) illustrates this.

(1) Papuan Malay (elicited)

a. *baju*
shirt
'shirt'

b. *baju* *mera*
shirt red
'red shirt'

As seen in **Figure 1**, all signs of F0-related prominence on baju 'shirt' disappear when it does not appear in IP-final position (as in [1b]), major pitch changes being limited to the now final word mera 'red' (in the phrase 'red shirt'). Note that the prosody in (1b) is independent of information-structural context. That is, the prosody would be the same, for example, in contrastive contexts, regardless of whether the noun (*the red* SHIRT*, not the red* SKIRT) or the adjective (*the* RED *shirt, not the* BLUE *one*) is in a contrast set. Stoel (2007, p. 121) provides a spontaneous example from Manado Malay. Similarly, Maskikit-Essed and Gussenhoven (2016) 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. They come to the conclusion that information focus in Ambon Malay is not expressed by means of prosody.

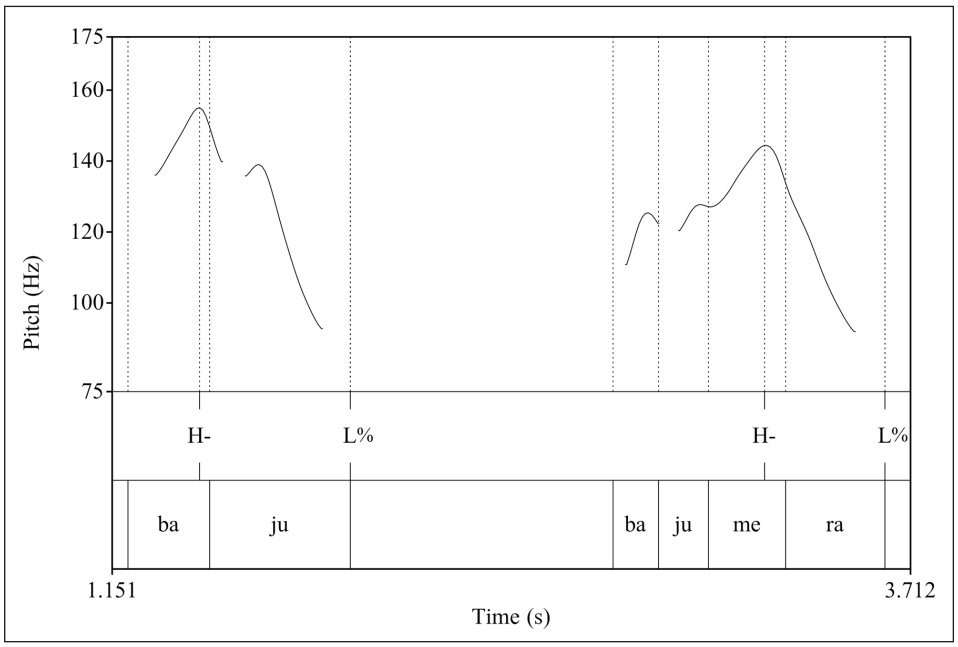


**Figure 1:** F0 contours and annotation tiers for syllables and ToBI-style intonation analysis of the two Papuan Malay phrases in (1).

[4] Stoel (2007) essentially proposes this analysis, though using slightly different terminology. See Himmelmann et al. (2018) for further details and comments.

To briefly summarize the state of the art regarding the prosody of PM and closely related Trade Malay varieties: On the phrase level, major prosodic events, specifically major pitch changes, appear to happen mostly at the right edge of IPs. These changes typically consist of at least two tonal targets and occur within a two-syllable window from the end of the IP. While the exact function is unclear, it is very likely that these edge tone combinations play a role in marking the boundaries of IPs. On the word level, there are indications for word-based prominence distinctions, but it remains an open question what kind of prominence distinctions, if any, are to be assumed on this level. All evidence investigated so far is acoustic or involves auditory judgments by Western researchers. It is not clear at all what the speakers themselves hear.

The present study therefore takes a perceptual approach to questions of Malayic prosody. It explores the evidence generated by the RPT methodology with regard to the PM prosodic system and whether this evidence allows us to develop some more specific hypotheses about this system that can be further tested with different methodologies. Specifically, we ask:

1. Does RPT provide evidence as to which kind of prosodic events are salient for PM speakers?
2. What kind of hypotheses, if any, regarding the structure of PM prosody does the RPT evidence suggest? And how do these compare to what we already know about it?

As already mentioned in the introduction, we investigate these questions by applying the RPT methodology in two conditions. In one condition, the familiar-language condition, native speakers provide annotations on their native languages, i.e., languages they are familiar with and in which they understand the segments they have to annotate prosodically. In the other condition, the unfamiliar-language condition, they annotate materials from a language they are not familiar with and in which they do not understand the recordings they are asked to annotate. The hypothesis motivating the unfamiliar-language scenario is that in this scenario the annotators, lacking an understanding of the syntax and semantics of the tokens to be annotated, are forced to focus on prosodic events only. Thus, inasmuch as this scenario generates interpretable data, it should provide evidence for the kinds of acoustic events the annotators are able to identify in unfamiliar speech. Of course, as also already pointed out above, RPT judgments are holistic and do not provide direct evidence for phonological categories, but they may provide interesting pointers for forming hypotheses about phonological categories, as we will see in Section 5.

## 3. Methods and Participants

Rapid Prosody Transcription (RPT) is a method for collecting data on the perception of prosody in a rather quick and easy manner. Ordinary listeners who are naïve with respect to prosodic analysis listen to excerpts of audio recordings. They are given minimal instructions 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 asked to either 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). Alternatively, the experiment can be run in a digital version (e.g., by using LMEDS; see Mahrt, 2016). In this setup, participants click on the word they perceive as prominent, and the respective word then changes colour from black to red. For the boundary experiment, clicking on the word after which participants wish to mark a boundary places a vertical line after the word (see Appendix A for further details and illustrations).

The advantage of this method is its simplicity and directness, which provides us with prosodic judgments that are based on the listeners' holistic perception of form and function. Importantly, however, and as noted by Cole and colleagues, the elicited prominence and boundary judgments are not made on prosodic grounds alone when pertaining to a familiar (native) language. Instead, they are also influenced by morpho-syntactic, semantic, and pragmatic factors. Accordingly, some of our variables (for example, the distinction between content and function words) pertain to these other levels of description. As mentioned above, we expect these factors to be irrelevant in the unfamiliar-language condition, e.g., when PM speakers rate German data. Inasmuch as they are still found to be significant—as indeed they are (cp. Section 4.2)—they must be epiphenomena, reflecting acoustic properties of grammatical or semanto-pragmatic categories.

### *3.1. Familiar- and unfamiliar-language conditions*

We carried out RPT experiments with native speakers of German (GE) and PM. These speakers were asked to provide judgments on two datasets: one German and one PM dataset. Speakers from both populations provided judgments on both datasets. As judgments for prosodic prominences and boundaries were provided in separate experiments, the paper reports on $2 \times 2 \times 2 = 8$ datasets in total:

Familiar-language Condition

1. PM_PM_p: PM sample rated by PM speakers for prominences
2. PM_PM_b: PM sample rated by PM speakers for boundaries
3. GE_GE_p: GE sample rated by GE speakers for prominences
4. GE_GE_b: GE sample rated by GE speakers for boundaries

Unfamiliar-language Condition

5. GE_PM_p: GE sample rated by PM speakers for prominences
6. GE_PM_b: GE sample rated by PM speakers for boundaries
7. PM_GE_p: PM sample rated by GE speakers for prominences
8. PM_GE_b: PM sample rated by GE speakers for boundaries

As indicated in this list, we will be using short forms to refer to the different datasets: The first abbreviation refers to the language of the dataset, the second abbreviation denotes the population that provided the judgments, and the third codes whether we are dealing with the boundary or the prominence task. GE_PM_p thus reads 'the German dataset is rated for prominences by PM speakers.'

Before carrying out the experiments, we piloted the unfamiliar-language condition with both PM and GE speakers. *Prima facie*, it is not obvious that speakers can actually work on materials from an unfamiliar language as it is minimally necessary to be able to identify words in fluent speech. That is, the RPT method presupposes that the participants can actually read the transcript of the segment they are listening to and identify individual words in it. Fortunately, the orthographic conventions for PM and German are similar enough that speakers could basically apply the orthographic rules they learned for their native language. With the exception of the digraph <ng> for the velar nasal, PM orthography is almost phonemic, with each symbol corresponding to exactly one sound, and in fact corresponding to the sound most commonly also represented by the same symbol in German. So the task for the German participants was not particularly difficult and none of them reported problems in following the transcript when listening to the recorded speech. For the PM participants, the German orthography was a little more challenging as the phoneme to grapheme

correspondence is not always transparent. Yet, all PM participants in the pilot study stated that they had no problems identifying the words in the transcript. Still, it is clear that lack of familiarity not only with the spoken language but also with its orthographic representation contributed to overall worse results in the unfamiliar-language condition.

### *3.2. Participants*

The raters in our perception studies were first-year linguistics students at the University of Cologne (Germany), and students of English Language and Literature or Anthropology at the Universitas Papua (UNIPA) in Manokwari (West Papua, Indonesia). None of them had any experience in prosodic analysis or reported any hearing or reading problems at the time the experiments were conducted. Taking part in the experiments was an optional component of the class work. The varying participant numbers seen in **Table 1** reflect different class sizes.

Altogether, 146 native speakers of German and 93 native speakers of PM participated in the experiments. **Table 1** summarizes how participants were distributed across the different tasks. Here we can see that both populations, German and PM, included more female than male raters, which basically reflects the gender ratio among students in the humanities.

Within the group of 146 German speakers, there were 10 bilinguals, but in all cases German was their dominant language of communication at the time.

Of the 93 native speakers of PM, 77 were bilingual in PM and Standard Indonesian, and the remaining 16 spoke another local language in addition. Bilingualism in PM and Standard Indonesian is the default in this group. Due to the fact that national media (television, radio broadcasts, and newspapers) and the majority of formal education is in Standard Indonesian, every Indonesian university student has a command of Standard Indonesian to varying degrees. All participants stated that PM was (one of) their first language(s) and that PM was their first language of communication at home and at university, as well as in everyday interaction with their friends. The command of English among these students is very basic, usually being limited to a few hundred lexical items and the most rudimentary syntactic constructions. Importantly, English prosody is not part of the curriculum and students are not familiar with English stress and prosodically marked information focus.

The GE_PM_p and the GE_PM_b experiments were originally conducted with 27 and 25 participants, respectively. The results of three participants (one in GE_PM_p and two in GE_PM_b), however, were excluded from the analyses, because the metadata they gave was unclear. On further inquiry, it turned out that the three subjects had actually started learning PM only 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 PM at the time of the experiment.

**Table 1:** Participant numbers in the experiments.

| Familiar-language condition | Male/Female | Total | Unfamiliar-language condition | Male/Female | Total |
|---|---|---|---|---|---|
| PM_PM_p | 10/12 | 22 | GE_PM_p | 8/18 | 26 |
| PM_PM_b | 7/15 | 22 | GE_PM_b | 10/13 | 23 |
| GE_GE_p | 7/19 | 26 | PM_GE_p | 9/41 | 50 |
| GE_GE_b | 6/14 | 20 | PM_GE_b | 6/44 | 50 |

### 3.3. Stimuli and procedure

The participants annotated 40 excerpts of audio recordings (the same in the prominence and the boundary experiments). The excerpts were presented as audio files on a computer and the participants listened to them wearing headphones. For the German dataset, only 39 excerpts were used in the final analyses. This is due to a mistake in one of the transcripts which was only discovered after the experiment had been conducted.

All excerpts, PM as well as GE, were taken from a corpus of natural monologic speech consisting of re-tellings of Chafe's *Pear Movie* (Chafe, 1980). Two of the authors, SR and NPH, were involved in recording and transcribing these re-tellings. Himmelmann et al. (2018, p. 214–220) provide further details on the corpus.

The excerpts varied in length, ranging from one to 15 seconds. For PM, the total number of words was 539, which were divided into excerpts ranging from six to 48 words. More than half were between 10 and 15 words long. In the case of GE, a total of 637 words were divided into excerpts of between six and 33 words, the majority being between 12 and 19 words long. The number of words in the PM and GE datasets was not the same primarily because the two languages differ with regard to the frequency of function words, which is considerably higher in German than in PM. This is because the latter does not make use of definite and indefinite articles, for example (see also Himmelmann et al., 2018, p. 224f). In both data sets, there were two short excerpts which involved only a single intonational phrase (IP). The majority of excerpts consisted of two to three IPs, the remainder involving four or five IPs. The boundary judgments are based on the consensus between multiple expert raters making use of acoustic evidence, as further detailed in Himmelmann et al. (2018, p. 217f).

The PM dataset included speech from 19 different native speakers of Papuan Malay (10 female, nine male); the GM set was spoken by 19 native speakers of German, 12 of whom were female and seven male. Further details on the procedure, including the instructions for participants in their original wording, are provided in Appendix A. Note that the instructions for the two groups of speakers could not be literally translated since various concepts, including the core concept of *prominence*, do not have exact translation equivalents in German and Indonesian. Furthermore, the German participants completed the tasks online on their own computers. They thus received their instructions solely in written form and were not able to ask questions. Due to lack of a stable internet connection and a sufficient number of laptops, the Papuan Malay participants had to complete experiments in a lab setting, listening to the audio files on computers provided by the Centre of Endangered Languages Documentation (CELD) and recording their prominence and boundary judgments on printed transcripts. Importantly, Papuan Malay experimenters were present in this setup who briefly explained the task in Papuan Malay and answered the participants' questions where they occurred. This was the case especially in the prominence experiment because the concept of *prominence* or *stress* is not familiar to speakers of Papuan Malay. In answering such questions, the CELD staff members provided one or two spontaneous illustrations in Papuan Malay, which involved a rising pitch, or once or twice a rising gesture. Note that these illustrations of 'highlighting' or 'sticking out' were invented by native speakers and not by the German researchers.

### 3.4. Test variables

We tested the influence of a number of prosodic and morpho-syntactic cues that have been found to have an effect on native listeners' perception of prominence or boundaries in other (generally West Germanic) languages. For each test word in both datasets, we investigated the following—partly continuous-valued and partly discrete—prosodic factors: *Word duration* (in ms), *mean duration of syllables* (PM only) and *duration of*

*lexically stressed syllable* (GE only; both in ms), *duration of the last syllable within a word* (in ms), *minimum, maximum*, and *mean pitch* (in Hz), absolute *pitch range* (in semitones, st), *number of syllables* (both abstract phonological and actually realized), and *presence of a pause*. 'Pause' was defined as segmental silence with a duration of more than 100 ms. Intensity measures, i.e., measures of loudness, have not been included in the analysis since the excerpts were not recorded via headsets. Thus, the distance between mouth and microphone varied within utterances, so that variation with respect to intensity could not be regarded as being intended by the speakers. Consequently, if intensity is not a reliable linguistic factor in production, listeners who base their prominence judgments on this factor make use of an artefact and not a proper prosodic factor.

An increase in duration, pitch height, and pitch range has 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 have been shown to trigger the perception of a phrase break (e.g., Turk & Shattuck-Hufnagel, 2007). In addition, we analyzed *intonational categories* in both languages. For German, these comprise the accent types as postulated in the GToBI framework (Grice et al., 2005). As there are no comparable pitch accents in PM, we included a rough, pseudo-categorical measure of pitch change between words in order to see whether these changes are at all relevant for boundary or prominence perception in PM. For this measure, we took the final F0-value within a word and the initial F0-value in the subsequent word and calculated the tonal range in semitones (st). In a second step, we classified the range values as either constituting a rise, a fall, or a level intonation. A tonal range of more than $+1$ st was considered a rise; a tonal range of more than $-1$ st was considered a fall. Everything in between was accordingly considered to be level intonation. This kind of measure is employed because it has been shown before that (at least in West Germanic languages) a tonal reset between words, i.e., a 'rise' in the sense of our pseudo-categorical measure, is regularly perceived as the beginning of a new phrase.

Furthermore, we analyzed the morpho-syntactic cues *part-of-speech* (POS), *part-of-speech class* (i.e., content words versus function words), whether the word is the *last verbal argument* in an intonational phrase, and *syntactic break* (three levels: no break, 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, p. 31), while the last verbal argument in a sentence is of importance when it comes to focus projection; that is, in the default intonation of a broad focus sentence, the last verbal argument receives the nuclear accent (Uhmann, 1988, p. 66), which is the prosodic prominence that is decisive for the pragmatic meaning of the whole utterance.

Finally, we correlated the boundary ratings of our participants (i.e., the *b*-scores; see below) with the prominence ratings (the *p*-score).[5] We did this because the available research on PM suggests that prosodic boundaries in PM are marked by an edge tone combination which includes a major rise (and hence sounds prominent, at least to German

[5] There are different ways of identifying the prosodic boundaries in the excerpts in order to provide a basis for the correlation. We tested different options in this regard, including expert ratings by the four authors, and found that very similar results were obtained by the different methods. We chose the *b*-scores provided by the participants as bringing in expert ratings by the German authors could potentially be interpreted as biasing boundary definitions towards German perception. Note that the *b*-scores were provided throughout by a different group of raters than the *p*-scores, hence the *p*-scores are not simple replications of the *b*-scores.

listeners; cp. Section 2 above). Thus, there is the possibility that PM raters associate boundaries with prominences. In German, as just mentioned, the last and most important prominence, namely the nuclear accent, also in many cases occurs *close to* the boundary. Hence, we might find a correlation between boundaries and prominences here as well. As the prominence judgments in both groups by far outnumbered the boundary judgments, we did not investigate the converse correlations (i.e., predicting *b*-scores by *p*-scores).

### *3.5. Data analysis*

All experiments consisted of binary classification tasks. In the prominence experiment, participants had a binary choice for each word in the transcript to rate it as either prominent or non-prominent. In the boundary experiment, 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. If raters did set a boundary after the last word—an option that was technically possible in both the print and the online setup—these choices were ignored in the analysis. The PM set consisted of 539 words in total; thus each participant produced 539 data points in the prominence experiment and 499 data points in the boundary experiment. The German set contained 637 words, resulting in 637 data points for the prominence experiment and 597 data points for the boundary experiment.

For the statistical analysis of these data, a mixed effects logistic regression was performed on the raw data, i.e., on the binary values we received from participants, 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 created only single effect models (e.g., only maximum pitch or part-of-speech, but not both variables) with random intercepts for speaker, sentence, and rater. Subsequently, odds ratios were calculated to enable a comparison of the factors by means of the effect magnitude, which allowed us 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' κ provides a single coefficient as a measure of agreement across all raters. Cohen's κ 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 marked a word as prominent or set a boundary between two words with respect to the total number of participants. **Figures 3** (PM raters) and **4** (GE raters) show *p*- and *b*-scores for the PM example sentence (2), illustrated by a *praat* (Boersma & Weenink, 2017) image in **Figure 2**. The higher the value, the more participants perceived a word as prominent (grey line with triangles) or perceived a boundary after the respective word (black line with squares). Recall that no *b*-score was calculated for the last word of an excerpt.

(2) Papuan Malay

| *yang* | *tiga* | *orang* | *ini* | *pegang* | *topi* | *satu* |
|---|---|---|---|---|---|---|
| REL | three | person | DEM | carry | hat | one |

"The three people are carrying a hat."

In comparison, the *p*- and *b*-scores of the German example sentence (3), illustrated in **Figure 5**, is given in **Figures 6** (GE raters) and **7** (PM raters).

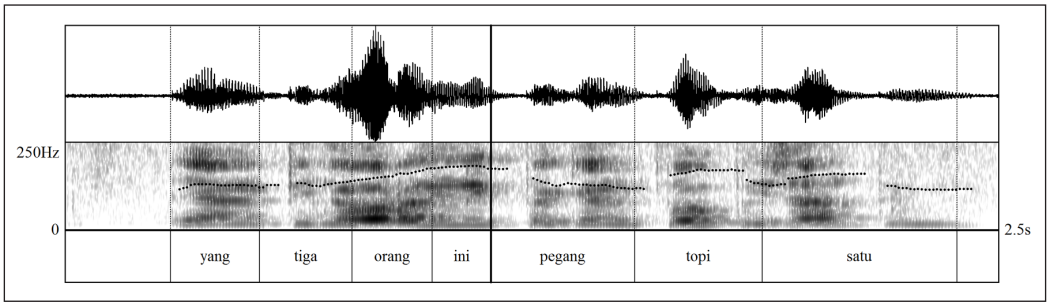


**Figure 2:** Wave form, spectrogram, and F0 for the PM excerpt in (2).

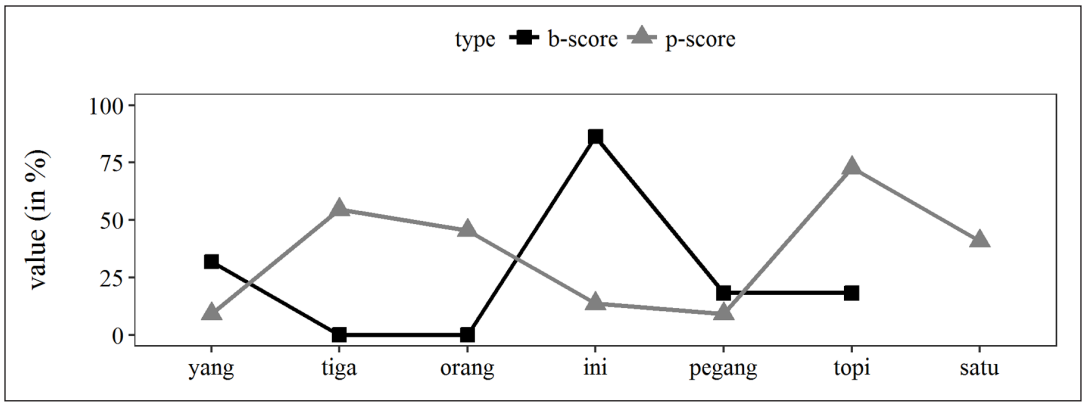


**Figure 3:** *P*- and *b*-scores for the PM excerpt in (2), PM raters judging PM example.

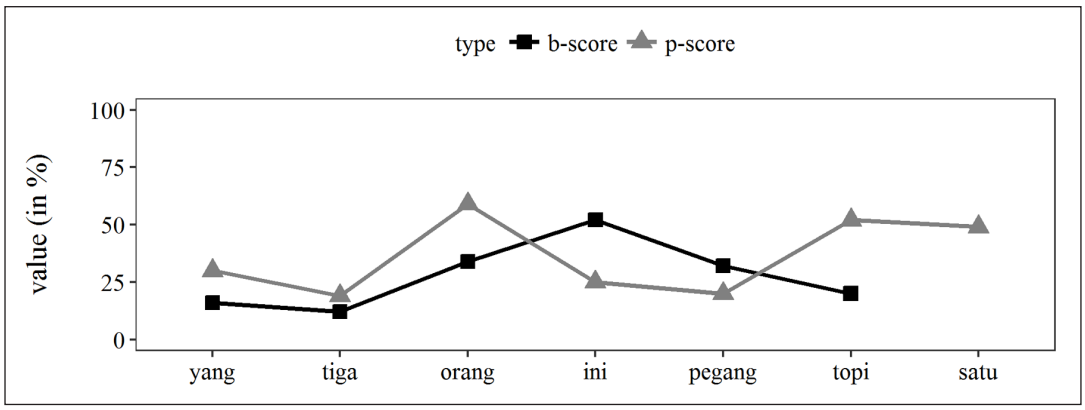


**Figure 4:** *P*- and *b*-scores for the PM excerpt in (2), German raters judging PM example.

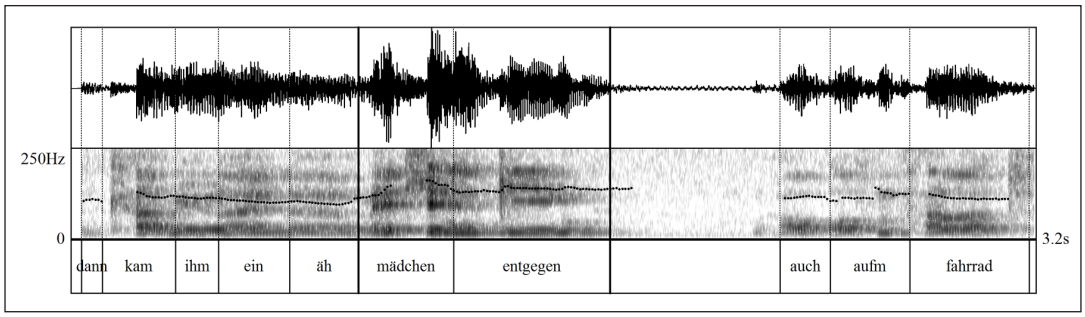


**Figure 5:** Wave form, spectrogram, and F0 for the German excerpt in (3).

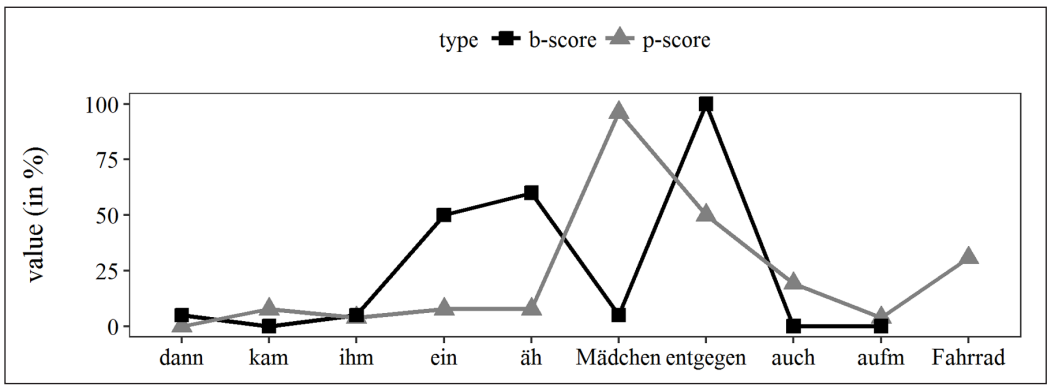


**Figure 6:** *P*- and *b*-scores for the German excerpt in (3), German raters judging German example.

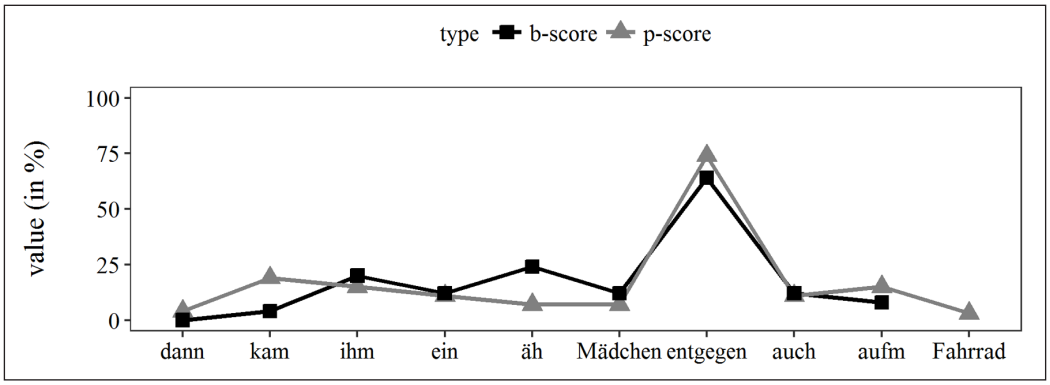


**Figure 7:** *P*- and *b*-scores for German excerpt in (3), PM raters judging German example.

(3) German

| *dann* | *kam* | *ihm* | *ein* | *äh* | *Mädchen* | *entgegen* | *auch* | *aufm* | *Fahrrad* |
|---|---|---|---|---|---|---|---|---|---|
| then | came | 3s.M.DAT | a | INTJ | girl | towards | also | on_a | bike |

"Then there was a girl coming towards him, also on a bike."

## 4. Results

We report results separately according to the two conditions, starting with the familiar-language condition, i.e., listeners rating their native language. In each section, we first present inter-rater agreement results and then look at factors that appear to have contributed to the judgments made by the subjects.

### *4.1. Familiar-Language Condition*

#### 4.1.1. Inter-rater agreement

The main finding concerning overall agreement within each group (cp. Fleiss' κ scores in **Table 2**) is that PM annotators perform significantly worse when providing prominence judgments compared to the German annotators, as well as compared to their own performance in providing boundary judgments.

As illustrated by the German data, agreement in prominence judgments tends to be generally lower than agreement in boundary judgments, a result also found in all RPT studies published to date (see Cole & Shattuck-Hufnagel, 2016, p. 7 for references). Clearly, however, the difference in the case of the PM raters is of a different magnitude and suggests that they find it difficult to agree on the location of prominent words in the short excerpts of natural speech used in this task. Note that the PM boundary results

rule out an explanation pointing to a basic lack of understanding of the experimental procedure.

Without doubt, PM speakers are less familiar with experimental setups and, even more fundamentally, with processing written speech than their German counterparts. This may be a major factor contributing to the generally lower agreement scores for PM raters than for the German participants that we will see throughout this paper. The lower κ score for PM raters judging boundaries compared to the German raters exemplifies this difference.

To test whether the low score of the PM participants in the prominence experiment was simply due to very low agreement between some individual participants, we also calculated Cohen's κ scores for each rater pair. In **Table 3**, the pair-wise inter-rater agreement is summarized for all four experiments, using the agreement categories postulated by Landis and Koch (1977), who characterize κ 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.

The figures in **Table 3** make it clear that the low Fleiss' κ score for the Papuans' prominence judgments in PM indicates a general failure to agree on which words in an excerpt are prominent. Most pairs are in the slight category and almost no pair is found in the higher three agreement categories. Notice also the high coefficient of variation, indicating that prominence judgments are highly variable across the group.

**Table 2:** Fleiss' κ scores for prominences and boundaries in the familiar-language condition.

| | **Prominences** | **Boundaries** |
|---|---|---|
| PM raters on PM | 0.106 | 0.408 |
| | z = 37.3 | z = 139 |
| German raters on German | 0.475 | 0.579 |
| | z= 216 | z =195 |

**Table 3:** Cohen's κ scores for prominences and boundaries in the familiar-language condition.

| **Pair-wise agreement** | **PMs on PM** | | | | **Germans on German** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Prominences 22 raters** | | **Boundaries 22 raters** | | **Prominences 26 raters** | | **Boundaries 20 raters** | |
| | **Pairs** | **%** | **Pairs** | **%** | **Pairs** | **%** | **Pairs** | **%** |
| None | 26 | 11.26% | 6 | 2.60% | 0 | 0.00% | 0 | 0.00% |
| Slight | 153 | 66.23% | 36 | 15.58% | 24 | 7.38% | 0 | 0.00% |
| Fair | 50 | 21.65% | 48 | 20.78% | 50 | 15.38% | 7 | 3.68% |
| Moderate | 2 | 0.87% | 105 | 45.45% | 198 | 60.92% | 100 | 52.63% |
| Substantial | 0 | 0.00% | 36 | 15.58% | 48 | 14.77% | 82 | 43.16% |
| (almost) perfect | 0 | 0.00% | 0 | 0.00% | 5 | 1.54% | 1 | 0.53% |
| | 231 | 100.00% | 231 | 100.00% | 325 | 100.00% | 190 | 100.00% |
| Mean Cohen's κ | 0.1196 | | 0.4131 | | 0.4771 | | 0.5829 | |
| *SD* | 0.1008 | | 0.1993 | | 0.1496 | | 0.0889 | |
| Coefficient of variation | 0.8428 | | 0.4825 | | 0.3136 | | 0.1525 | |

In the other three experiments, the majority of all pairs are found in the moderate category, with 15% or more also showing substantial agreement. In the case of boundary judgments, note that there is a considerable difference between PM and German participants regarding the coefficient of variation, reflecting the fact that the behaviour of PM raters in this task is much more heterogeneous than in the German group. Almost 40% of the PM pairs here are found in the lower three agreement categories, in which there are less than 4% of the German pairs. Similar observations hold when comparing the German prominence agreement scores with the German boundary judgments.

### 4.1.2. Factors influencing *p*- and *b*-scores

#### *4.1.2.1. Distribution of p- and b-scores*

Before turning to the factors that impact the perception of prominences and boundaries, we will shortly evaluate the distributions of *p*- and *b*-scores for each group in the familiar-language condition. An ideal outcome of an RPT experiment would be a bimodal distribution of *p*- and *b*-scores, with a mode (most frequent value within the data) of 0% and a second peak around 100%. The reason for the mode to be 0% instead of 100% is that there are typically more words in a sentence *not* marked for prominence or followed by a boundary than words that *are* marked. Importantly, then, a *p*- (or *b*-)score of 0% does not indicate that there was no agreement among participants, but rather that they agreed on the *absence* of a prosodic prominence (or boundary, respectively) for that particular word. The values in between the extremes should occur considerably less frequently, resulting in a shallow plateau connecting the extremes.

However, as shown in **Figure 8**, the distribution of *p*-scores in the PM data (upper left panel) does not look bimodal, but rather like a log-normal distribution. Contrary to expectations, the mode is 13.6% accordance (82 words, i.e., 15.2% of the data) instead of 0%, which makes up only 3% of the data (16 words). This pattern is the major reason for the relatively high coefficient of variation mentioned above. In addition, a consensus of 100% regarding the presence of a prominence is never achieved in the native prominence ratings of PM: The highest agreement rate is 81.8%, and this only holds for 0.5% of the data. This distribution shows a high degree of variability of judgments in the PM

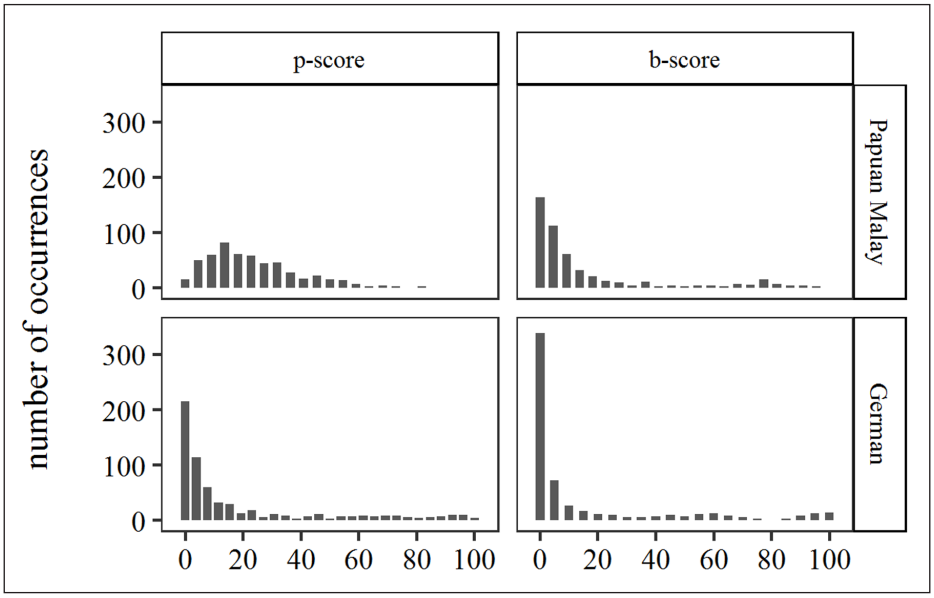


**Figure 8:** Distribution of *p*- and *b*-scores in the familiar-language condition. The first column shows the *p*-score, the second column the *b*-score for PM (upper row) and German (lower row) judged by native raters.

prominence data, indicating once more that PM speakers have difficulties in agreeing on the location of prominent words.

The distribution of *b*-scores in the PM data (cp. **Figure 8**, upper right), in contrast, looks more like the expected pattern described above: The mode is 0% with 164 words (32.9% of the data) and the highest *b*-score achieved is 95.5%. However, this highest *b*-score is again achieved only three times in the whole dataset (499 words) and thus makes up only 0.6% of the data. Although this distribution still shows a considerable amount of variability, it is clear that native speakers of PM perform much better when judging boundaries instead of prominences.

In comparison, the native German judgments on prominence and boundaries show much less variability: For both *p*- and *b*-scores the mode is 0%, with 215 words (33.7% of the data) in the prominence experiment[6] and 339 words (56.7% of the data) in the boundary experiment. Complete agreement regarding the presence of a boundary is observable for both experiments, but this value is only achieved five times (0.8% of the data) in the prominence experiment and 15 times (2.5% of the data) in the boundary experiment.

#### *4.1.2.2. Linguistic variables affecting p- and b-scores (familiar-language condition)*

Altogether, 16 possible parameters were tested for their effects on prominence and 15 for the boundary ratings (the *b*-score itself being investigated as one factor impacting on the *p*-score). **Figures 9–12** provide an overview of the odds ratios of all variables tested for both languages (see Appendix B for the model likelihood ratios for all factors investigated). In the following, we will not discuss each factor in detail, but rather confine ourselves to the most robust trends observable in the data.

Comparing the two languages, we find a major difference in that the odds ratio of most variables is much higher for German than for PM, which also mirrors the lower degree of variability and the higher agreement in the German inter-rater values presented in the

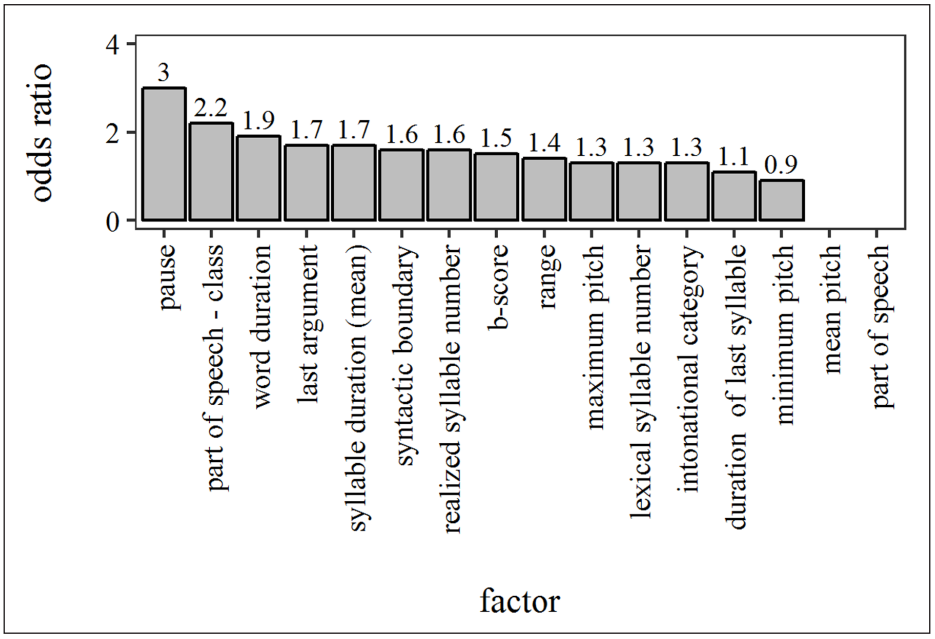


**Figure 9:** Variables affecting prominence ratings by PM listeners judging PM, ordered according to odds ratios (missing odds ratios indicate that no significant effect was found).

[6] The distribution of the *p*-scores is very similar to the German RTP study by Baumann and Winter (2018). Still, the *p*-score of 0 found in this study is somewhat higher, namely 45.2%.

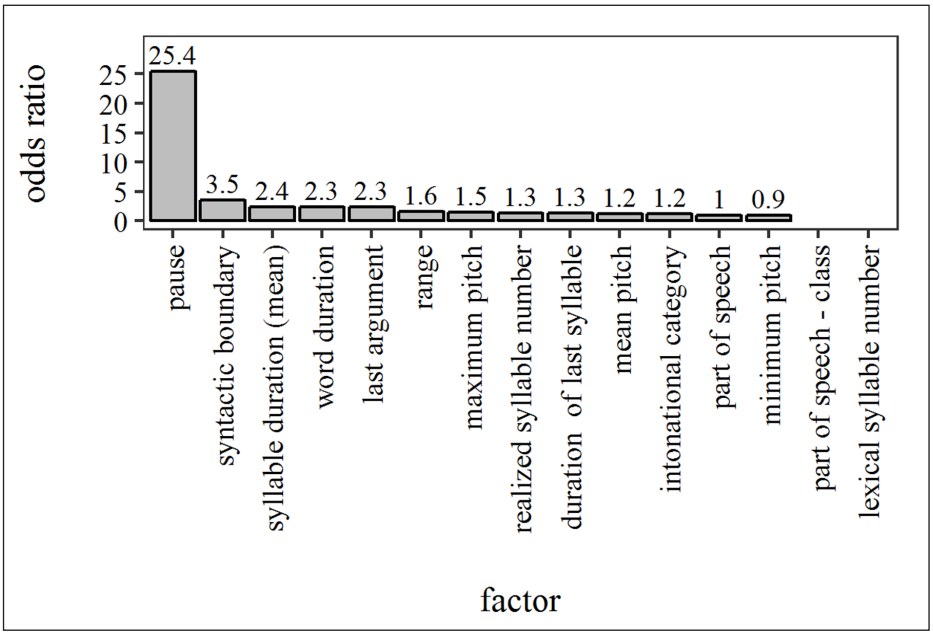


**Figure 10:** Variables affecting boundary ratings by PM listeners judging PM, ordered according to odds ratios (missing odds ratios indicate that no significant effect was found).

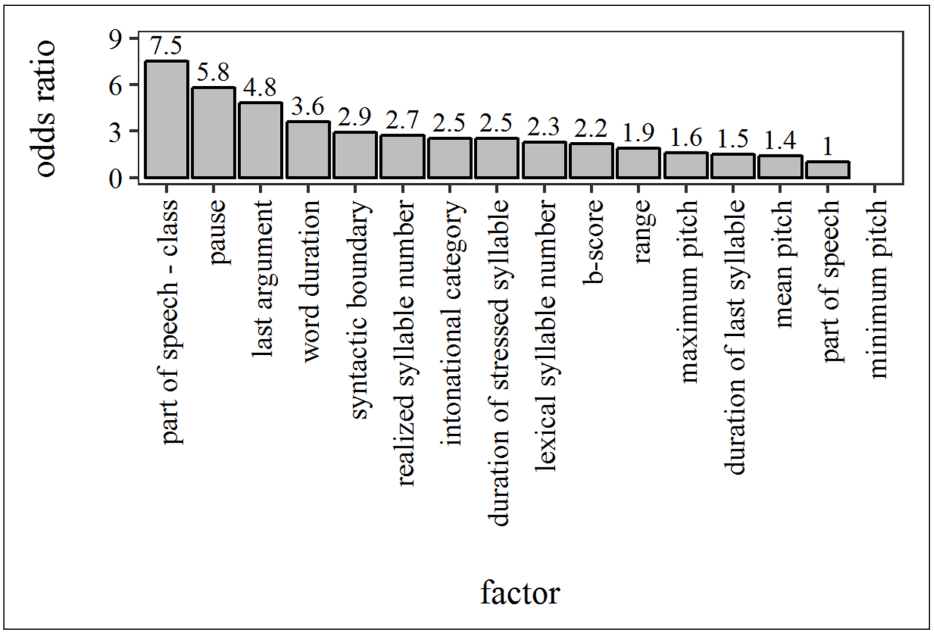


**Figure 11:** Variables affecting prominence ratings by German listeners judging German, ordered according to odds ratios (missing odds ratios indicate that no significant effect was found).

preceding sections. This is particularly obvious for the odds ratio of PAUSE in the German boundary judgments (see the *b*-scores for the German example in **Figure 6** above), but it also holds for all other factors.

In addition to PAUSE, durational measures and syntactic boundaries play a role in boundary judgments in both languages. This is not surprising as it is well known that syntactic phrasing often correlates with prosodic phrasing, to some extent mutually influencing each other. Longer durations probably reflect final lengthening, which in all likelihood is a universal correlate of intonational phrase boundaries. Otherwise, the order of the lower-ranked factors differs quite considerably between the two languages

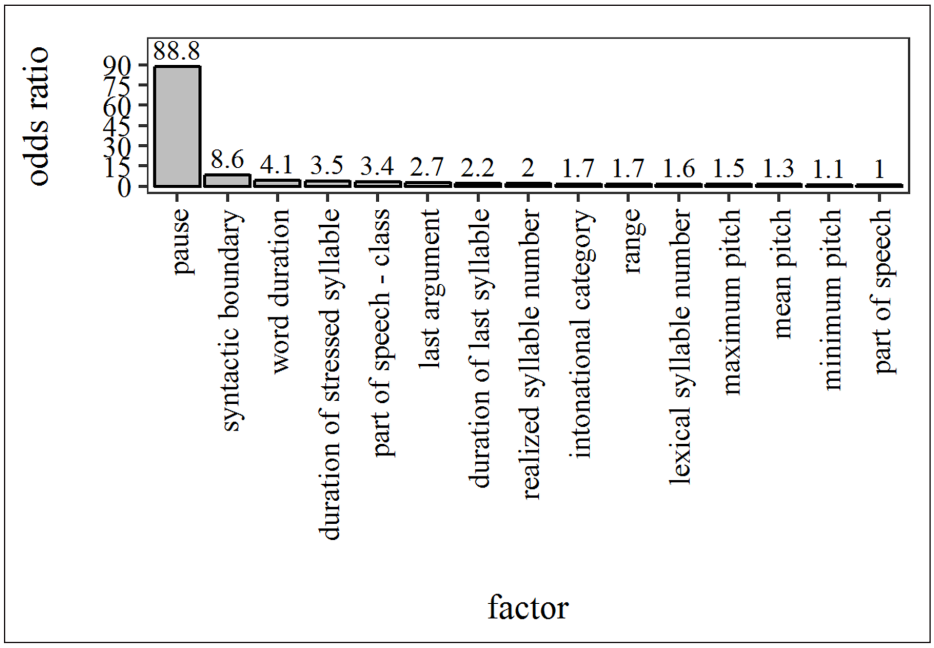


**Figure 12:** Variables affecting boundary ratings by German listeners judging German, ordered according to odds ratios.

with regard to boundary judgments. However, the odds ratios are all fairly low and the differences between the factors very small, making it unlikely that these differences are truly significant.

The fact that PAUSE plays a role in both boundaries and prominences is a commonality across the two groups, though with the difference that it is ranked second for prominence judgments in German but first for prominence judgments in PM. At first glance, it might appear rather odd that pause plays a role in prominence perception. However, there are two phenomena that are tightly connected to the presence of a pause, and these are domain-initial strengthening (e.g., Fougeron & Keating, 1997) and domain-final lengthening (e.g., Turk & Shattuck-Hufnagel, 2007). A pause is usually produced in between two intonational phrases (IPs) and hence signals both the end of the first IP and the beginning of the second IP. Domain-initial strengthening affects the first segment(s) within a prosodic unit, whereas final lengthening affects the last segment(s) of such a unit. In both cases, the affected segments are produced with increased duration. This lengthening might lead to the impression of prominence, similar to the lengthening that happens when a word is accented. Therefore, the presence of a pause might trigger the perception of a prominence at the very beginning and the very end of a phrase, although only indirectly.

Another commonality is that POS CLASS is relevant for prominence judgments in both languages, reflecting the fact that function words are often not judged to be prominent. As for other factors influencing prominence judgments, the highest-ranked factors are almost the same (i.e., LAST ARGUMENT, WORD DURATION, SYNTACTIC BOUNDARY; see high *p*-score for *Mädchen* ['girl'] in the German example in **Figure 6** above: The content word [noun] *Mädchen* is the last argument in the phrase). However, for PM the odds ratios for all factors other than the top-ranked PAUSE (and perhaps POS CLASS) are all both very close to each other and fairly low, and hence do not allow for more far-reaching conclusions. The odds ratios for the top-ranked factors for German, on the other hand, differ somewhat more clearly. Another major difference between the two languages pertains to the fact that INTONATIONAL CATEGORY appears in the middle range in German, but towards the lower end in PM.

In fact, German accent types are the strongest predictors for prominence judgments in German, confirming the results of a recent RPT study on German (see Baumann & Winter, 2018). Using the same set of main accent types proposed in the German ToBI system (and labelled by the same two annotators in both studies), **Figure 13** shows that the mean *p*-score increases steadily in accordance with the accent type, starting with no accent (mean *p*-score: 6.9%), via monotonal and falling accents and ending with rising accents (mean *p*-score of L+H*[7]: 68.9%). The only difference to the results in Baumann and Winter is the reversed order of high accents (which were judged as more prominent in the previous study) and low accents (which were judged as less prominent).

When evaluating each accent type separately, we can observe that the odds for being perceived as prominent increase as the accent gets steeper (with the exception of L*). In addition, rising accents are clearly preferred over falling accents. **Table 4** lists the odds ratios for the accent types displayed in **Figure 13**.

These results clearly show that the most reliable cue for German natives for prominence perception is a bitonal accent, as even the falling accent H+L* increases the odds for 'prominent' more strongly (9.4 to 1) than the strongest non-tonal parameter, i.e., POS CLASS (7.5 to 1), cp. **Figure 11** (note that in **Figure 11** all accent types, including *no accent*, are comprised in the factor INTONATIONAL CATEGORY, hence the low overall odds ratio for this factor).

Given that bitonal, and especially rising, accents have such a strong impact, the question arises why there is no (strong) effect of pitch range observable in this data. A possible

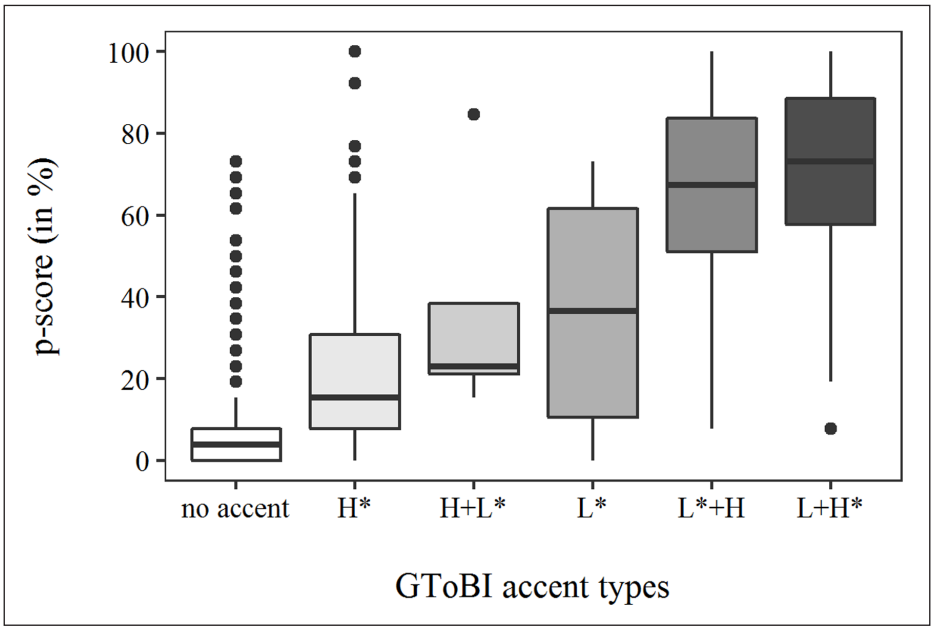


**Figure 13:** The effect of accent type on *p*-scores for German listeners judging German.

**Table 4:** Odds ratios per accent type in GE_GE prominence perception (the model likelihood ratio is $\chi^2(4) = 4649.8$ with $p < 2.2e^{-16}$).

| Accent Type | H* | L* | H+L* | L*+H | L+H* |
|---|---|---|---|---|---|
| Odds Ratio | 4.7:1 | 7.3:1 | 9.4:1 | 43.9:1 | 44.8:1 |

[7] As standard in ToBI-style intonation analysis, a star indicates association with an accented syllable.

explanation for this apparent mismatch is that, first, the pitch range of each word was calculated by taking the absolute pitch minimum and the absolute pitch maximum within a word (correcting for micro-prosody)—but these values did not necessarily mark the edges of a tonal movement (especially for words that were not accented). Second, pitch accents in German are not marked by tonal movement alone, but usually coincide with a longer duration and enhanced loudness of the accented syllable. This combination of factors marking pitch accents makes it difficult to correlate the results for pitch range with the results for accent type presented here.

While INTONATIONAL CATEGORY (accent type) also plays a role in the German boundary judgments, its influence is much weaker compared to the prominence judgments. The only exception is falling accents, which show much less variation and correlate with a high *b*-score (mean of 70%). Here, the fall appears to lead to the impression of finality of an utterance. The significant contribution of falling accents to the boundary perception data is also shown by the odds ratios for each accent type, which increase strongest for the falling H+L* accent; namely by roughly 53 to 1 (cp. **Table 5**). This means that falling accents are the second most important factor for German boundary judgments (after PAUSE, cp. **Figure 12**). Rising accents are less important than syntactic boundaries, but more important than word duration.

Like other Malayic varieties, PM does not make use of metrically anchored pitch accents (cp. Section 2). In order to see whether a difference in INTONATIONAL CATEGORY nevertheless plays a role in the perception of prominences and boundaries, the amount of tonal movement between adjacent words was measured in semitones (st) and then divided into the classes 'level' (less than +/− 1 st difference between the offset of word x and the onset of word x+1), 'fall' (more than −1 st difference), and 'rise' (more than +1 st difference). **Figures 14** and **15** show, however, that the effects of tonal movements

**Table 5:** Odds ratios per accent type in GE_GE boundary perception (the model likelihood ratio is $\chi^2(4) = 714.8$ with $p < 2.2e^{-16}$).

| Accent Type | H* | L* | L+H* | L*+H | H+L* |
|---|---|---|---|---|---|
| Odds Ratio | 1.5:1 | 2.9:1 | 4.6:1 | 6.9:1 | 52.9:1 |

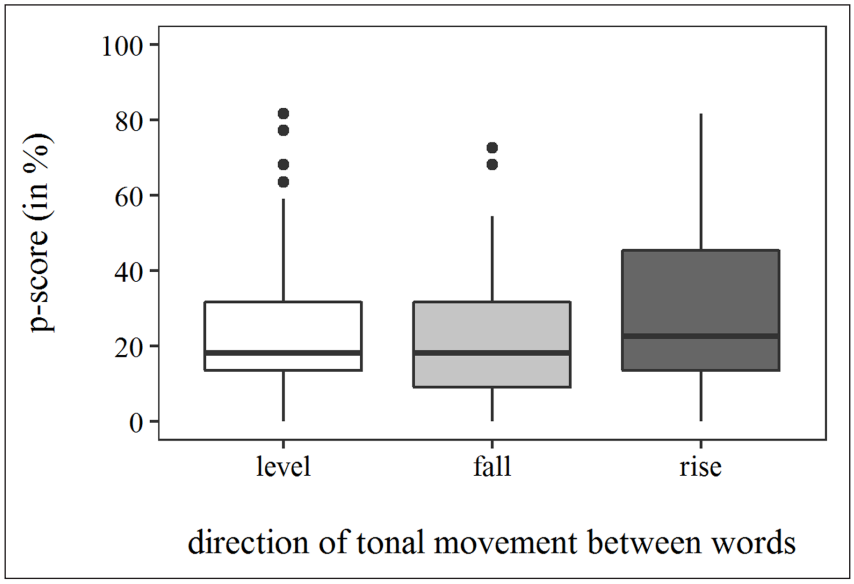


**Figure 14:** The effect of tonal movement on *p*-scores, PM listeners judging PM.

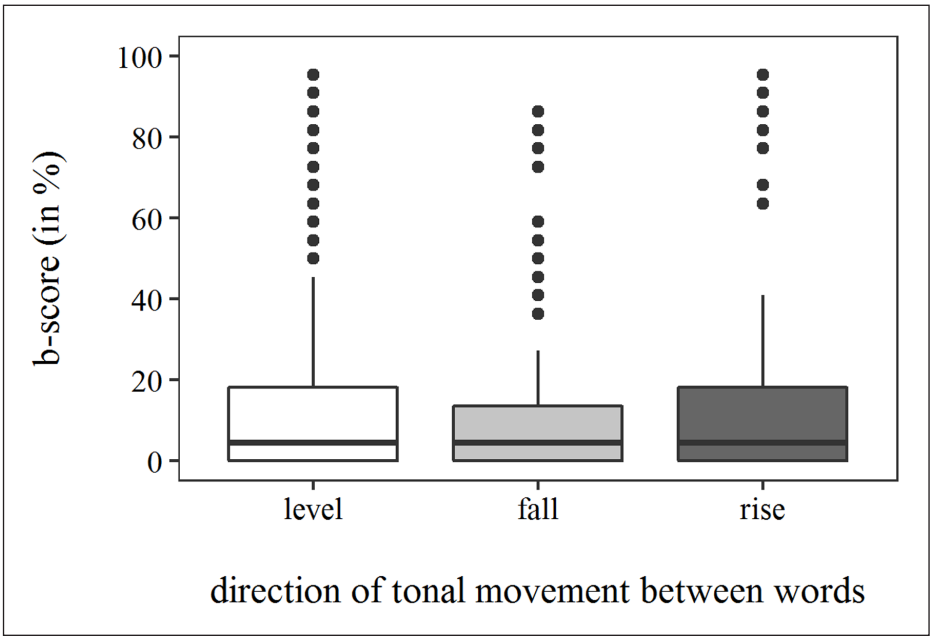


**Figure 15:** The effect of tonal movement on *b*-scores, PM listeners judging PM.

across words are only marginal in both tasks (odds ratio prominence: 1.3 to 1; odds ratio boundary: 1.2 to 1, cp. also **Figures 9** and **10**).

### *4.2. Unfamiliar-Language Condition*

#### 4.2.1. Inter-rater agreement

When participants judge tokens from unfamiliar languages, the overall agreement scores are generally somewhat lower than in the case of judgments provided for a familiar language, as shown in **Table 6**. This is hardly surprising as there are at least two factors that render the annotation tasks more difficult in the case of unfamiliar languages. On the one hand, participants will encounter occasional problems in identifying words in the transcript, especially in cases of fast or otherwise reduced speech. On the other hand, prosodically relevant events such as pitch or durational changes tend to come in a language-specific form and may thus not be easily heard as such. This is probably less of a problem in the case of pauses, but is certainly relevant in the case of subtler rhythmic or melodic changes.

Note that unfamiliarity with the language of the recordings is equivocal with regard to complicating the annotation task, as already hinted at above and further discussed and exemplified in Himmelmann et al. (2018). It complicates the task, because it makes it more difficult to identify words and the kind of prosodic events one expects to hear. On the other hand, however, it makes it easier, as it reduces the factors influencing the prosody judgments and thus eliminates the possibility of conflicts between prosodic and non-prosodic factors.

In light of the results for the familiar-language condition in **Table 2**, there are two surprising outcomes of the unfamiliar-language condition. First, against the general trend described above, PM speakers judging prominences in the German data do considerably better than when judging prominences in their native language (Fleiss' κ score 0.283 versus 0.106). This would suggest that PM speakers are able to pick up on at least one cue for prosodic prominence. But why, then, is there so little agreement regarding prominence in their native language? This could be due to the fact that prominence is not prosodically marked in PM. Alternatively, or in addition, PM speakers may also be strongly influenced

by non-prosodic factors in their own language, which may give rise to behaviour that is overall more heterogeneous. The data presented in the following sections will provide some evidence in support of the first hypothesis.

The second somewhat surprising outcome of the unfamiliar-language condition is that, across both participant groups, the scores for boundary and prominence judgments are much closer than in the familiar-language condition, where the scores for prominences are considerably lower than those for boundaries (cp. **Table 2**). However, it may very well be that the similarity between the boundary and prominence scores in the unfamiliar-language condition does not necessarily reflect a substantial similarity in rating behaviour, but is in fact coincidental. This may be the case for PM, as suggested by the pair-wise comparisons shown in **Table 7**, which are distributed differently across the Koch and Landis agreement categories and also differ with regard to the coefficient of variation. Thus, for the PM participants, the great majority of the pairs is in the fair category for prominences (71%), with rather small sets in the slight and moderate categories (16% and 13%, respectively). For boundaries, on the other hand, only slightly more than half of the pairs are in the fair category (55%) and considerably more pairs are found in the slight and moderate categories (about 21% each). There are even three pairs in the substantial category and one in the none category.

In the case of the German raters' scores, there are similar, but much less clear tendencies. For boundaries, over 60% of the pairs achieve moderate and substantial agreement,

**Table 6:** Fleiss' κ scores for prominences and boundaries in unfamiliar-language condition.

| | **Prominences** | **Boundaries** |
|---|---|---|
| PM raters on German | 0.283 | 0.289 |
| | z = 129 | z = 112 |
| German raters on PM | 0.399 | 0.415 |
| | z= 324 | z =324 |

**Table 7:** Cohen's κ scores for prominences and boundaries – unfamiliar-language condition.

| **Pair-wise agreement** | **PMs on German** | | | | **Germans on PM** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Prominences 26 raters** | | **Boundaries 23 raters** | | **Prominences 50 raters** | | **Boundaries 50 raters** | |
| | **Pairs** | **%** | **Pairs** | **%** | **Pairs** | **%** | **Pairs** | **%** |
| none | 0 | 0.00% | 1 | 0.40% | 0 | 0.00% | 0 | 0.00% |
| slight | 53 | 16.31% | 55 | 21.74% | 15 | 1.22% | 21 | 1.71% |
| fair | 231 | 71.08% | 140 | 55.34% | 602 | 49.14% | 458 | 37.39% |
| moderate | 41 | 12.62% | 54 | 21.34% | 595 | 48.57% | 652 | 53.22% |
| substantial | 0 | 0.00% | 3 | 1.19% | 13 | 1.06% | 94 | 7.67% |
| (almost) perfect | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | 325 | 100.00% | 253 | 100.00% | 1225 | 100.00% | 1225 | 100.00% |
| Mean Cohen's κ | 0.2918 | | 0.4021 | | 0.2952 | | 0.4310 | |
| *SD* | 0.0885 | | 0.0890 | | 0.1312 | | 0.1155 | |
| Coefficient of variation | 0.3034 | | 0.2215 | | 0.4446 | | 0.932681 | |

while for prominences, agreement is split almost equally across the fair and moderate categories. Here, the coefficient of variation also varies drastically between the two tasks, with unusually high variation in the case of boundary judgments, a point to which we will return in the next section. Still, as also discussed in the next section, the *b*-scores of the German raters are the best predictors for their *p*-scores (but recall from Section 3 that *b*- and *p*-scores were provided by different individuals). Inasmuch as boundary perception influences the German participants' prominence perception, the lack of major difference in agreement results for prominence and boundary judgments may thus not be coincidental, but rather reflects a factual interrelation.

As we will also see in the next section, while some prosodic factors influence both prominence and boundary judgments, both the ranking and the effect size differ quite clearly between the two, supporting the assessment that the overall similarities in agreement results for boundaries and prominences are at least partially coincidental, especially in the case of PM.

#### 4.2.2. Factors influencing *p*- and *b*-scores

##### *4.2.2.1. Distribution of p- and b-scores*

As in the case of the familiar-language condition, we also find a major deviation from the expected bimodal distribution in the unfamiliar-language condition. Here, however, it is the German participants judging PM boundaries who achieve a modal value of 4% instead of 0%, as seen in **Figure 16**. That is, the German participants had difficulties in agreeing on where *not* to put a boundary. More than half of all words (265 words, 53%) were considered to be located at a boundary by 10% or less of the German listeners, giving rise to the very high coefficient of variation seen in **Table 7**.

The other three datasets—PM speakers judging German prominences and boundaries, and German speakers judging PM prominences—are more in line with the basic expectations in that the modal values are 0%. However, all four datasets show a high degree of variability in that most *p*- and *b*-scores lie between the values of 2% and 20%. Especially for the German subjects, this pattern sharply contrasts with the results in the native condition (cp. **Figure 8**), where the variability within this range was considerably lower.

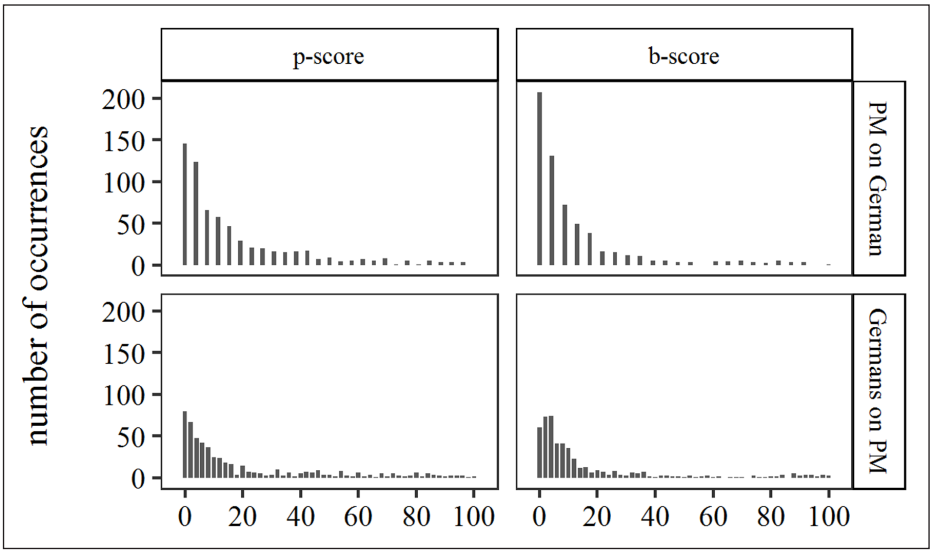


**Figure 16:** Distribution of *p*- and *b*-scores in the unfamiliar-language condition. The first column shows the *p*-score, the second column the *b*-score for PM speakers judging German (upper row) and Germans judging PM (lower row).

Given this high variability, it is to be expected that all the factors considered in the following sections will be relatively weak, as indeed is shown by the figures.

#### *4.2.2.2. Linguistic variables affecting p- and b-scores (unfamiliar-language condition)*

**Figures 17–20** show the odds ratios of all variables tested for both languages: PM speakers judging German prominences and boundaries (top row) and German speakers judging PM prominences and boundaries (bottom row). Appendix B provides the model likelihood ratios for all investigated factors. Again, we will keep the discussion rather short, focussing only on the most conspicuous patterns.

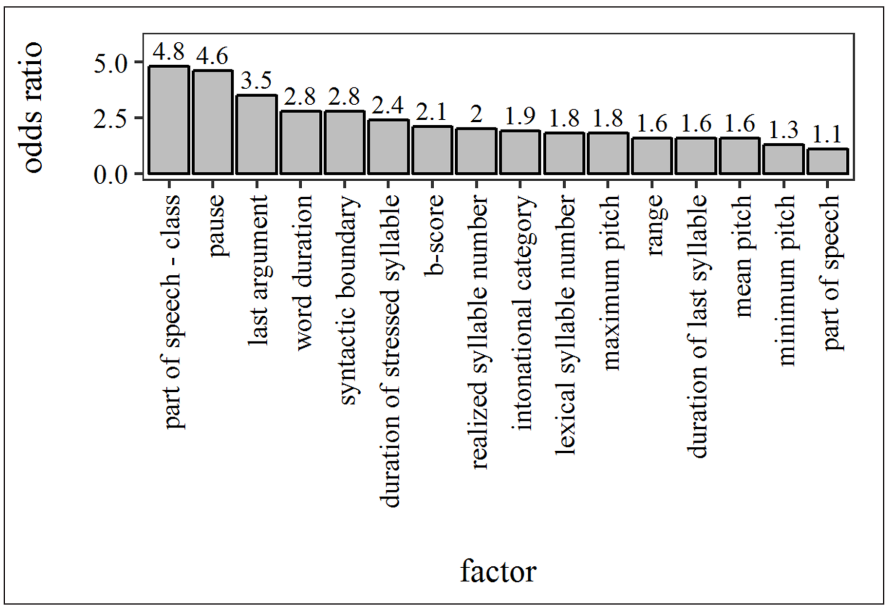


**Figure 17:** Variables affecting prominence ratings by PM listeners judging German, ordered according to effect size (indicated by odds ratios).

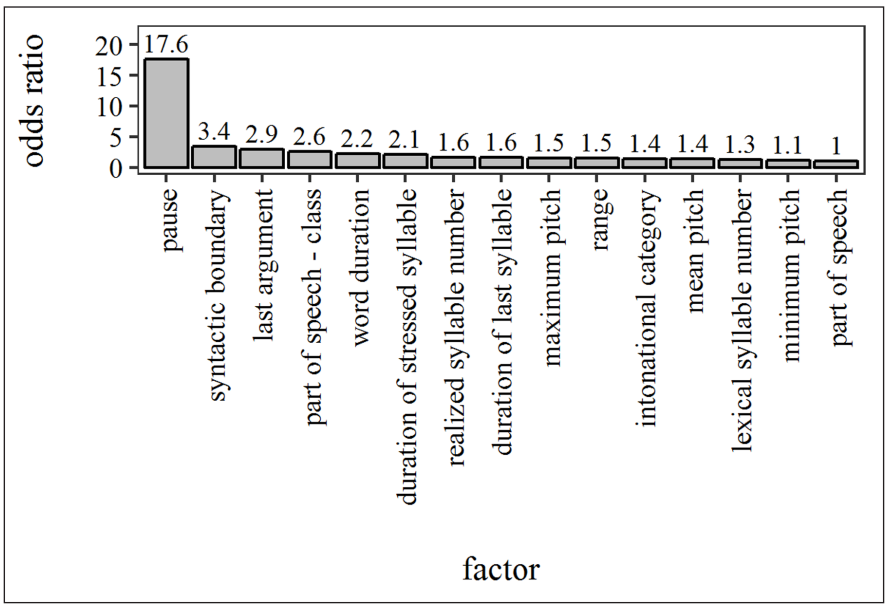


**Figure 18:** Variables affecting boundary ratings by PM listeners judging German, ordered according to effect size (indicated by odds ratios).

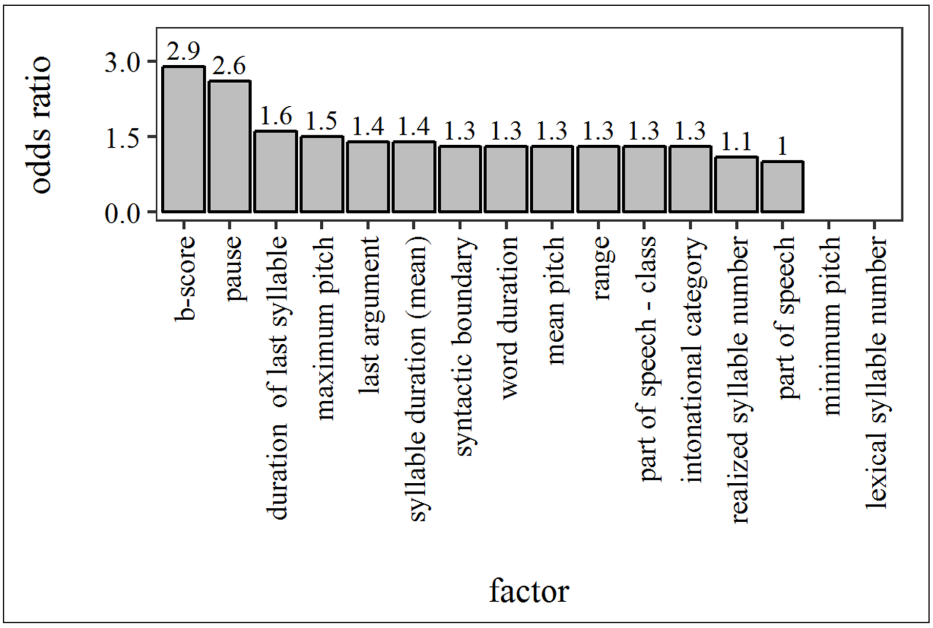


**Figure 19:** Variables affecting prominence ratings by German listeners judging PM, ordered according to effect size (indicated by odds ratios; missing odds ratios indicate that no significant effect was found).

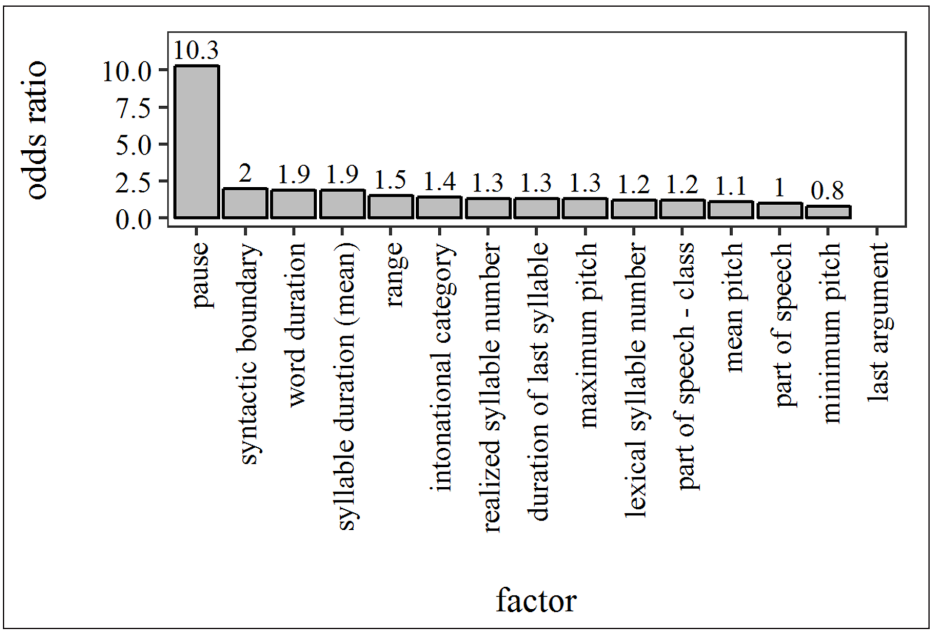


**Figure 20:** Variables affecting boundary ratings by German listeners judging PM, ordered according to effect size (indicated by odds ratios; missing odds ratios indicate that no significant effect was found).

As in the familiar-language condition, PAUSE is the most important prosodic factor for prominence judgments and boundary judgments for both speaker groups, being ranked first or second in all four datasets (see **Figure 7** above, illustrating PM raters' judgments of a German example). This is to be expected in the case of boundaries, but surprising in the case of prominences. Here, German listeners judging PM prominences actually appear to be influenced primarily by boundary phenomena. The variable B-SCORE refers to the fact that PM words marked as preceding a boundary in the boundary experiment are about three times more likely to be marked for prominence by German listeners. Overall, this

means that inasmuch as speakers of German hear prominences in PM speech at all, they do this preferably at prosodic boundaries.

The fact that PAUSE yielded much lower odds ratios in the unfamiliar- than in the familiar-language condition can be explained by problems in identifying pauses: Since participants listened to the excerpts only twice and had to simultaneously read the transcript and mark the position of boundaries, it is likely that they did not perceive all the pauses that a native speaker—who knows the words and the ways they can be realized (e.g., segment deletion)—would perceive. As a consequence, different non-native raters would miss different pauses and this, in turn, would lead to an overall lower correlation of PAUSE and B-SCORE and to lower odds ratios.

The fact that SYNTACTIC BOUNDARIES are ranked second for boundary judgments performed by each listener group suggests that in both PM and German syntactic and prosodic boundaries tend to coincide. More importantly, perhaps, the fact that the morpho-syntactic factor SYNTACTIC BOUNDARY is stronger than individual prosodic parameters suggests that prosodic boundary marking consists of a cluster of factors, of which PAUSE is of course the most salient.

While the relative (lack of) importance of the factors influencing boundary judgments is similar across the two listener populations, the ranking of the factors influencing prominence judgments differs considerably, except that in both groups PAUSE is ranked second. As already noted above, the judgments by the German listeners appear to be most strongly influenced by boundary phenomena, such as DURATION OF LAST SYLLABLE, LAST ARGUMENT, and SYNTACTIC BOUNDARY, in addition to PAUSE and B-SCORE. Boundary-related phenomena also play a role for the PM listeners (e.g., LAST ARGUMENT, SYNTACTIC BOUNDARY), but here POS CLASS is ranked highest. As is well known, function words in German are rarely stressed and tend to form prosodic words with a preceding or following content word. Prosodic words tend to be stressed on the content word part. Given that PM listeners of course do not know the difference between content and function words, the top rank of POS CLASS thus suggests that that content words are in fact produced as more prominent by the German speakers.

This conjecture is supported by a closer look at INTONATIONAL prominence distinctions in German. The distribution depicted in **Figure 21** clearly shows an increase in mean

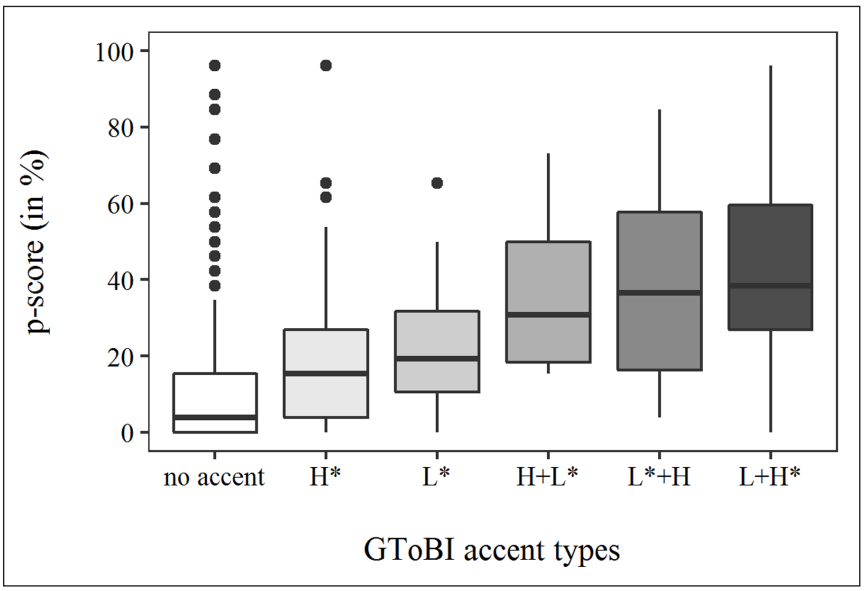


**Figure 21:** The effect of accent type on *p*-scores for PM speakers judging German.

*p*-scores, depending on the type of accent. This strongly suggests that PM listeners are sensitive to the difference between accent types in German.

Importantly, the behaviour of the PM listeners mirrors the German (native) listeners' behaviour (cp. **Figure 13** above), though on a much lower effect level. When there is no tonal accent, the mean *p*-score is rather low at 10.5%, increasing via monotonal accents (H* 19.5% p-score, L* 22.6% *p*-score) through the bitonal falling accent to bitonal (high) rising accents, with a mean *p*-score of 43.2% for L + H*.

**Table 8** shows that the odds ratios for each accent type indicate that pitch accents increase the odds for 'prominent' and that rising accent types are clearly the strongest cue for prominence judgments made by the PM listeners (odds ratio of 7.7 to 1 as opposed to an odds ratio of 4.8 to 1 in the case of POS CLASS, cp. **Figure 17**).

The sensitivity to intonational cues is also found in the results for the boundary task for PM listeners. While for prominences PM listeners reacted most strongly to rising accents, when the task was to judge prosodic boundaries they reacted most strongly to falling accents. The distribution of odds ratios in **Table 9** mirrors the one for German listeners in **Table 5** above.

On the other hand, intonational cues do not seem to play a major role for German listeners judging PM data. Recall from Section 3 above that there are no accentual distinctions in PM and that – for the present paper – we are only looking at a coarse equivalent of the tonal movements in the German data by comparing level, falling, and rising transitions between words (INTONATIONAL CATEGORY for PM). As in the case of PM listeners judging examples from their native language, German listeners of the PM data also showed hardly any effects, with very high variability both for prominence and boundary judgments. That is, the odds for 'prominent' increase by only 1.2 to 1 in the case of a falling contour and by 1.7 to 1 in the case of a rising contour, compared to a level intonation. The odds for 'boundary' decrease with falling contours by 0.9 to 1, but increase by 1.7 to 1 for rises.

In interpreting the results for accent type, it is important to keep in mind that pitch accents in German are not only marked by tonal movement, but also by enhanced loudness and longer duration of the accented syllable. PM listeners thus appear to react to the complex mixture of prosodic parameters correlating with pitch accents rather than to single parameters such as PITCH RANGE and WORD/SYLLABLE DURATION. This is also indirectly shown by the fact that in the unfamiliar-language condition many of the top-ranked factors are morpho-syntactic in nature rather than prosodic, as already noted above (e.g., POS CLASS, LAST ARGUMENT). As the listeners are unaware of the relevant morpho-syntactic features in the unfamiliar-language condition, we interpret this finding as pointing towards a holistic perception of prosodic prominence-lending features. In this view, PM listeners appear to be able to detect prosodic prominence in German as a holistic phenomenon, in sharp contrast to their inability to do the same in their native

**Table 8:** Odds ratios per German accent type in PM prominence perception (the model likelihood ratio is $\chi^2(4) = 1383.7$, with $p < 2.2e^{-16}$).

| Accent Type | H* | L* | H+L* | L*+H | L+H* |
|---|---|---|---|---|---|
| Odds Ratio | 2:1 | 2.4:1 | 5.7:1 | 6.1:1 | 7.7:1 |

**Table 9:** Odds ratios per German accent type in PM boundary perception (the model likelihood ratio is $\chi^2(4) = 474.7$, with $p < 2.2e^{-16}$).

| Accent Type | H* | L* | L+H* | L*+H | H+L* |
|---|---|---|---|---|---|
| Odds Ratio | 1.5:1 | 2.6:1 | 3.8:1 | 4.9:1 | 9.9:1 |

language. It is not very likely that the fact that the PM participants are students of English is of any relevance in this regard, because the level of English competence is rather low, as mentioned in Section 3 above.

Against this background, the results for German listeners judging PM prominences clearly suggest that there is no prominence marking in PM similar to that in German—otherwise the German listeners would have been able to pick up on it. More specifically, the difference between content words and function words does not appear to have clear prosodic correlates in PM, at least, none which are related to prominence distinctions.

Before discussing the implications of these results, we briefly turn to a comparison of the judgments by native and non-native speakers on the same materials. That is, we compare the judgments by PM listeners on German excerpts with the judgments by German listeners on the same data, and compare the judgments by PM listeners with those by German listeners on the PM data.

### *4.3. Comparing native and non-native judgments*

In the preceding section we saw that the PM and the German participants appear to be influenced, at least in part, by the same factors when providing judgments on recordings in the unfamiliar language. This suggests that the prosodic factors found to be relevant in this regard are cross-linguistically identifiable. Being sensitive to the same factors, however, does not necessarily mean that participants from both groups located prominences and boundaries in exactly the same places. In order to get a rough idea whether there are commonalities in the prominence and boundary judgments across the two groups, we also computed inter-rater agreement scores across both groups, comparing each PM participant with each German participant working on the same task (i.e., providing prominence judgments on PM recordings or providing boundary judgments on German recordings, etc.). **Table 10** provides the results of this pair-wise agreement comparison.

In interpreting the data in **Table 10**, it is important to keep in mind that intergroup agreement can be maximally as good as the within-group agreement. Hence, it should not come as a surprise that there is hardly any agreement with regard to prominence judgments

**Table 10:** Cohen's κ scores pair-wise agreement between native and non-native raters.

| Pair-wise agreement | Prominences | | | | Boundaries | | | |
|---|---|---|---|---|---|---|---|---|
| | PM data 22 PMx50 GER | | German data 26 GERx26 PM | | PM data 22 PMx50 GER | | German data 20 GERx23 PM | |
| | Pairs | % | Pairs | % | Pairs | % | Pairs | % |
| none | 179 | 16.27% | 1 | 0.15% | 7 | 0.64% | 3 | 0.65% |
| slight | 906 | 82.36% | 96 | 14.20% | 945 | 85.91% | 38 | 8.26% |
| fair | 15 | 1.36% | 482 | 71.30% | 148 | 13.45% | 243 | 52.83% |
| moderate | 0 | 0.00% | 97 | 14.35% | 0 | 0.00% | 165 | 35.87% |
| substantial | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 11 | 2.39% |
| (almost) perfect | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |
| | 1100 | 100.00% | 676 | 100.00% | 1100 | 100.00% | 460 | 100.00% |
| Mean Cohen's κ | 0.0631 | | 0.3008 | | 0.1377 | | 0.3543 | |
| *SD* | 0.0604 | | 0.0951 | | 0.0570 | | 0.1326 | |
| Coefficient of variation | 0.9569 | | 0.3161 | | 0.4140 | | 0.3743 | |

on the PM data, since the PM participants hardly agreed on these among themselves (cp. **Table 2**). Germans, in contrast, primarily perceive prominences at prosodic boundaries in the PM data (cp. **Figure 19**).

As for the German data, the intergroup agreement scores are mostly in the *fair* category for both prominence and boundary judgments, but a sizable number of pairs is also found in the *moderate* category, in particular for boundary judgments. Overall, the agreement on boundaries is somewhat stronger, as also shown by the mean Cohen's κ scores (0.3543 for boundaries versus 0.3008 for prominences). While one should be careful not to over-interpret these results, the figures seem to indicate that the perception and interpretation of prosodic events (as boundaries or prominences) is not totally different across the two groups. Or, put differently, some aspects of German prosody can apparently be picked up by non-native speakers who are completely unfamiliar with the language.

Given that PM participants agreed reasonably well on boundaries in their native language (Fleiss' κ score of 0.408, cp. **Table 2**) and German participants agreed on PM boundaries even to a slightly higher degree (Fleiss' κ score of 0.415; cp. **Table 6**), it is somewhat surprising that intergroup agreement is comparatively low, as shown by the mean Cohen's κ score of 0.1377 in **Table 10**. Note also that PAUSE and SYNTACTIC BOUNDARY were found to be the two strongest factors influencing both the native (**Figure 10**) and the non-native (**Figure 20**) behaviour. In all likelihood, the low intergroup agreement here is due to the inability of the German listeners to agree on the absence of a prosodic boundary. As noted above, the mode for *b*-scores for this group was not 0%, as expected, but 4% (**Figure 16**). Only 60 out of 499 words (12%) were unanimously judged as not being followed by a boundary. And more than half of all words (265 words, 53%) were considered to be located at a boundary by 10% or less of the German listeners. On the other hand, when we look at those places where both PM and German listeners agree to a level of 70% or more on the presence of a boundary, there is substantial intergroup agreement.

## 5. Discussion

The results reviewed in the preceding section have implications for a number of issues, both with regard to specific issues in the prosodic analysis of PM and German and with regard to more general methodological and conceptual issues in prosody research. Our focus here is on the implications regarding PM prosody, addressing our core question of whether the RPT method provides useful input for the prosodic analysis of little-known languages. Other issues will only be mentioned in passing.

We begin with prominences, where the results speak to the perception of prominences by PM listeners and the possibility that the PM prosodic system lacks systematic prominence marking. The fact that PM speakers have difficulties in agreeing on prominences in their native language (**Table 2**) gives rise to the hypothesis that PM lacks post-lexical pitch accents of the kind characteristic of the German prosodic system—a system that has the option of prosodically highlighting specific words and syllables across an utterance.

The unfamiliar-language condition in the present study provides further support for this hypothesis. The fact that PM listeners show better agreement with regard to prominences in German (**Table 7**) makes it very unlikely that the same kinds of prosodic events occur in the PM data and are not picked up by them. The evidence for the claim that PM and German listeners react to the same kinds of prosodic events when judging prominences in the German data is as follows. For one, there is fair intergroup agreement in this task, as shown in **Table 10**. More importantly, however, PM listeners appear to be guided by exactly the same factors as their German counterparts when judging prominences in German, though with a lower effect magnitude. The most important factors here are INTONATIONAL CATEGORY (i.e., accent type, especially rising accents), POS CLASS, PAUSE,

and LAST VERBAL ARGUMENT, which are in the same ranking for both groups (cp. **Figure 11** in combination with **Table 4** and **Figure 17** in combination with **Table 8**). In short, RPT data strongly suggest the hypothesis that PM lacks the kind of post-lexical pitch accents known from Germanic languages, in line with other varieties of Malay for which this has also been suggested in the literature (cp. Section 2). At the same time, PM listeners appear to be able to perceive the accentual prominences characteristic of German.

The latter observation suggests the further hypothesis that the inability to agree on prominences in one's native language does not necessarily imply that speakers of such a language are stress 'deaf,' to use the term coined by Peperkamp and Dupoux (2002). As just mentioned and illustrated in greater detail in Section 4.2, PM listeners appear to be able to perceive the same kinds of prominences in German data as the native German listeners do. In doing this, they do not make use only of presumably universal prosodic cues such as PAUSE. Instead, they appear to rely also on much more language-specific cues such as accent types. In this regard, however, it is important to be clear on the fact that German accent types do not involve only the definitional pitch changes, but usually make a word stand out through a combination of various prosodic parameters, including increased duration and intensity. That is, we do not believe that our data suggest the hypothesis that PM listeners actually perceive German accentuation in the same way as German listeners do. Instead, German accentuation appears to involve a highly salient combination of prominence-lending features which, as a holistic phenomenon, can also be perceived by non-native listeners.

This is not the place to further explore the implications of the idea that some language-specific prominence phenomena, such as the German accent types, are possibly perceptible by non-native listeners. Nevertheless, it should be obvious that this idea suggests a whole line of further investigation into the interplay between language-specific and potentially universal factors in the structure and use of prosodic prominences. See Himmelmann et al. (2018) for an exploration of this general issue with regard to the perception of prosodic boundaries, including a proposal for a distinction between (potentially universal) phonetic and (language-specific) phonological intonational phrase boundaries.

Note that the preceding argument does not necessarily imply that there are no systematic prominence distinctions whatsoever in the PM prosodic system. There might be prominence distinctions of a kind that is very different from the kind of prominence characterizing German accentuation. The RPT evidence in this regard is inconclusive. The fact that there is so little agreement among the PM listeners themselves speaks against a hypothesis along these lines. However, at least in theory, it might be the case that non-prosodic factors play a key role when PM listeners judge PM data for prominences and that these factors overlay existing (weak) prosodic cues for prominence. This idea, however, would require the additional assumption that the presumed non-prosodic factors are very disparate (otherwise, the Fleiss' κ score in **Table 2** would be higher). That is, some listeners are influenced by position in the clause, others follow information structural concerns, and so on.

The unfamiliar-language condition here again provides an interesting pointer. Given the lack of agreement among the PM speakers, it may come as a minor surprise that German listeners judging PM data, who are not 'distracted' by non-prosodic factors, achieve agreement values that are not in a totally different range from the Fleiss' κ score achieved in the familiar-language condition (Fleiss' κ of 0.475 for judging German prominences versus 0.399 for judging PM prominences). As noted in Section 4.2, German listeners seem to be primarily guided by prosodic boundaries when judging PM prominences (cp. **Figure 19**). This allows for two, not mutually exclusive interpretations. First, the PM prosodic system includes boundary-related prominences, perhaps similar to what has

been called an *edge* or *phrase tone* (or *phrase accent*) in other prosodic systems. The cues for these boundary-related prominences include the ones that are typically associated with prominences in German, especially tonal movement. Second, lacking the familiar cues for prominences, German speakers tend to hear words at prosodic boundaries as prominent, presumably because the pragmatically most relevant prominences in German (i.e., nuclear accents) often occur close to a phrase boundary.

The first option raises the question of why PM listeners do not also judge words at boundaries as prominent. Answers to this question require further acoustic and perceptual studies which go beyond the scope of this paper. For our current purposes it is sufficient to note that the RPT data provide important pointers for such further investigations. These include, for example, the following questions: Is there acoustic evidence for prominences at PM boundaries? (The available data suggest that there is such evidence, including major pitch changes and longer durations.) Do PM listeners actually perceive the potentially prominence-lending cues as such? If so, why do they not consider words occurring at boundaries as prominent (and what are the non-prosodic confounds, if any)?

The second option also suggests a strand of further research, not specifically concerning PM, but rather German and prominence perception more generally. In what way are prosodic boundaries, and specifically pauses, prominence-lending? Note that PAUSE is also among the most important factors influencing German listeners' judgment of prominences in German data (**Figure 11**). While the odds ratio (5.8:1) is substantially lower than the odds ratio for the main factor (i.e., rising accents), PAUSE is still among the more important factors in this regard. To date, PAUSE—as a factor occurring *after* the word in question—has usually not been investigated alongside the typical prominence-lending factors (but see Dahan & Bernard, 1996, who looked at the use of both pre- and post-target pauses in the production and perception of emphasis in read French). Our RPT data, however, suggest that (post-target) PAUSE plays a role for naïve listeners, at least when judging short excerpts of spontaneous narrative speech.

Turning to boundary judgments, we see that these basically conform to what has been found in RPT work on other languages (e.g., Mo & Cole, 2010; Smith, 2011, Smith & Edmunds, 2013; Jyothi et al., 2014; Pintér et al., 2014). For both groups, there is moderate to substantial agreement in the familiar-language condition, and fair to moderate agreement in the unfamiliar-language condition. PAUSE is by far the most significant predictor for boundaries in all conditions. For the German data, falling accents also strongly correlate with boundary judgments, and this applies to both German and PM listeners (cp. **Tables 5** and **9**). In the case of PM listeners, this is somewhat surprising as pitch measures have not been found to be significant predictors when they judge boundaries in their native language (the same is true for the German listeners when judging PM recordings; cp. **Figures 10** and **20**). This, however, may be an artefact of the way we defined our pitch measures (cp. Sections 3 and 4.1.2). Importantly, and unlike the German data, which were annotated for accent types, we did not include more global annotations for pitch movements in the PM data, because no reasonably comprehensive analysis of PM intonation is available to date. Therefore, it is unclear at this point whether a systematic annotation of intonational phrase-final boundary tones would not reveal them to be a further major factor in influencing boundary judgments.

As in the case of prominences, the results of the unfamiliar-language condition suggest the hypothesis that listeners can pick up relevant prosodic cues without understanding the utterances for which they are providing judgments, and without being familiar with the prosodic system (see Himmelmann et al., 2018 for a similar finding). This hypothesis is supported by the fair to moderate agreement scores in the unfamiliar-language condition

(**Table 7**), as well as the fact that almost the same factors play a role in the judgments by native and non-native listeners. Importantly, the intergroup agreement scores (**Table 10**) show that native and non-native listeners are in fact identifying the same boundaries. This is clearly so in the case of PM. In the case of German, this is in principle also true. However, here the relevant score in **Table 10** is relatively low (Fleiss' κ of 0.1377) because German listeners differed substantially as to where *not* to put a boundary, while at the same time agreeing largely on where to put a boundary (see Section 4.3 for details).

Finally, a note on methodology is in order. The preceding discussion, to our minds, makes it clear that the RPT methodology in fact provides important clues when approaching a prosodic system that has not been investigated before. The unfamiliar-language condition has been found to be particularly revealing, as it provides a means to probe the phonetic events to which listeners of the investigated variety are sensitive. Furthermore, the observation that morpho-syntactic factors are found to be significant in this condition—even though speakers clearly cannot know where syntactic boundaries are or identify function words in an unfamiliar language—suggests that a possibly complex set of prosodic parameters is associated with these morpho-syntactic factors. In this way, the pointers provided by the unfamiliar-language condition are clearly useful in designing further research into the prosody of the language at hand.

With regard to the unfamiliar-language condition, it should be noted that we do not want to suggest that this can be usefully applied in all circumstances. There are quite a few factors involved, such as the ability of participants to read transcripts in an unfamiliar language and to identify individual words in these transcripts when listening to the corresponding recording. However, we believe that wherever it is feasible to carry out RPT experiments in the unfamiliar-language condition, it promises useful results.

As already pointed out in the introduction, the RPT method does not 'prove' the presence or absence of a particular prosodic category. Its major merit, when applied to little-known languages, is its ability to generate hypotheses and point to fruitful avenues for further research. Furthermore, when carried out more systematically across speech communities with different prosodic systems, we believe that it will generate important evidence for the perceptibility of phonetic cues for prosodic categories. Some phonetic cues will turn out to be widely perceptible regardless of familiarity with a given prosodic system, whereas others will be found to be highly language-specific.

## 6. Conclusion

In this article, we argue that the Rapid Prosody Transcription method (Cole & Shattuck-Hufnagel, 2016) is a helpful tool for gaining a first insight into the prosodic system of a little-known language, using Papuan Malay as an example. The extent to which naïve listeners can agree on (pre-theoretically defined) prosodic prominences and boundaries and the factors influencing their judgments provide important pointers to the kinds of prosodic events that may play a role in the language at hand. We have further argued that the first insights into prosodically relevant distinctions provided by RPT methodology are considerably enriched and refined by including an unfamiliar-language condition. That is, speakers not only provide judgments on prominences and boundaries in their native language, but also on a language they do not know (but for which they are able to link example excerpts they hear to the corresponding written transcripts). We illustrate this claim here with German data and the inclusion of a complementary dataset with results for German listeners providing judgments for the same dataset as the PM listeners.

The unfamiliar-language condition is particularly useful for a first attempt to distinguish between non-prosodic and prosodic factors influencing the judgments on prominences

and boundaries. Furthermore, it points to phonetic events the listeners are sensitive to—both individual factors such as pitch changes or durational differences, and more complex events such as German pitch accents.

With regard to our test case PM, the following observations and suggestions illustrate the usefulness of RPT methodology for probing the prosody of little-known languages:

- PM listeners agree on prosodic boundaries to an extent that is similar to that found in RPT studies on other languages, but they show very little agreement on prosodic prominences when judging excerpts from their native language.
- PM listeners agree on prominences when judging excerpts from the unfamiliar language German and they do so using the same cues that German listeners use when providing judgments on their native German, in particular pitch accent types. This finding suggests that a) there are no post-lexical pitch accents in PM similar to those in German (otherwise we would expect PM listeners to be able to agree on these to the same extent that they do with regard to the German data); b) PM listeners are not 'stress deaf' in the sense that they are unable to detect prosodic prominences of all kinds; c) some aspects of German pitch accents, as holistic phenomena involving durational and intensity cues in addition to pitch changes, can be perceived by speakers of languages who are not familiar with this type of prosodic prominence from their own language.
- German listeners agree on prosodic prominences in PM to an extent that is in the same range as the agreement they show for judging prominences in their native German. The prominences they perceive are typically located at prosodic boundaries and in fact appear to be cued primarily by these boundaries. This suggests the need to further investigate a) the interrelation between boundary and prominence perception, and b) the question of why PM listeners do not perceive prosodic prominences at the same locations.
- PM and German listeners are able to agree on prosodic boundaries in languages not familiar to them, using the same kinds of cues that native speakers of these languages use when judging prosodic boundaries.

## Abbreviations

b-score = boundary score; DAT = dative; DEM = demonstrative; GE = German; INTJ = interjection; OR = odds ratio; M = masculine; PM = Papuan Malay; p-score = prominence score; POS = part of speech; REL = relative pronoun; RPT = Rapid Prosody Transcription; s = singular

## Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** Instructions for participants. DOI: https://doi.org/10.5334/labphon.192.s1
- **Appendix B.** Result Details. DOI: https://doi.org/10.5334/labphon.192.s2

## Acknowledgements

The empirical work for this paper 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). The writing of this paper was supported by the German Research Foundation (DGF) within the Collaborative Research Centre 1252 Prominence in Language (Projects A01 Intonation and Attention Orienting, A03 Prosodic

Prominence in Cross-linguistic Perspective, and B05 Prominence Related Structures in Austronesian Symmetrical Voice Languages) at the University of Cologne.

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 in Papua. Special thanks to Claudia Wegener for computing the kappa statistics, to Isabel Compes, Gabriele Schwiertz, and Lea Krause for help with the experiments in Germany, and to Katherine Walker for improving English grammar and style. We are thankful to the two anonymous referees and to associate editor Aoju Chen for helpful comments. Thanks also to LabPhon admin Kip Wilson for efficiently handling the submission process.

## Competing Interests

The authors have no competing interests to declare.

## Author contributions

All authors designed research and collaborated on the final manuscript; NPH and SR carried out experiments; JK analyzed data; SR drafted Sections 1–3, JK section 4, and NPH Sections 5 and 6; SR and JK produced the appendices with input from all other authors.

## References

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

**Baumann, S.,** & **Winter, B.** (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics, 70*, 20–38. DOI: https://doi.org/10.1016/j.wocn.2018.05.004

**Beckman, M. E., Hirschberg, J.,** & **Shattuck-Hufnagel, S.** (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology I: the phonology of intonation and phrasing* (pp. 9–54). Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199249633.003.0002

**Boersma, P.,** & **Weenink, D.** (2017). Praat: Doing phonetics by computer. *Computer programme, v. 6.0.28*. Retrieved from http://www.praat.org/

**Büring, D.** (2012). Predicate integration – phrase structure or argument structure? In I. Kucerova & A. Neeleman (Eds.), *Contrasts and positions in information structure* (pp. 27–47). Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511740084.003

**Chafe, W.** 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

**Cole, J., Mo, Y.,** & **Baek, S.** (2010a). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes, 25*, 1141–1177. DOI: https://doi.org/10.1080/01690960903525507

**Cole, J., Mo, Y.,** & **Hasegawa-Johnson, M.** (2010b). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology, 1*, 425–452. DOI: https://doi.org/10.1515/labphon.2010.022

**Cole, J.,** & **Shattuck-Hufnagel, S.** (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology, 7*(1): 8, 1–29. DOI: https://doi.org/10.5334/labphon.29

**Dahan, D.,** & **Bernard, J.-M.** (1996). Interspeaker variability in emphatic accent production in French. *Language and Speech*, *39*(4), 341–374. DOI: https://doi.org/10.1177/002383099603900402

**Fougeron, C.,** & **Keating, P.** (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728–3740. DOI: https://doi.org/10.1121/1.418332

**Grice, M., Baumann, S.,** & **Benzmüller, R.** (2005). German intonation in autosegmental-metrical phonology. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 55–83). Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199249633.003.0003

**Gussenhoven, C.** (2004). *The phonology of tone and intonation*. Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511616983

**Hart, J. 't, Collier, R.,** & **Cohen, A.** (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511627743

**Himmelmann, N. P.** (2006). Prosody in language documentation. In J. Gippert, N. P. Himmelmann & U. Mosel (Eds.), *Essentials of language documentation* (pp. 163–181). Berlin: Mouton de Gruyter.

**Himmelmann, N. P.,** & **Kaufman, D.** (in press). Prosodic systems: Austronesia. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Prosody*. Oxford, England: Oxford University Press.

**Himmelmann, N. P.,** & **Ladd, D. R.** (2008). Prosodic description: An introduction for fieldworkers. *Language Documentation & Conservation, 2*, 244–274.

**Himmelmann, N. P., Sandler, M., Strunk, J.,** & **Unterladstetter, V.** (2018). On the universality of intonational phrases in spontaneous speech – a cross-linguistic interrater study. *Phonology, 35*(2), 207–245. DOI: https://doi.org/10.1017/S0952675718000039

**Jun, S.-A.,** & **Fletcher, J.** (2014). Methodology of studying intonation: From data collection to data analysis. In S.-A. Jun (Ed.), *Prosodic typology II: The phonology of intonation and phrasing* (pp. 493–519). Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199567300.003.0016

**Jyothi, P., Cole, J., Hasegawa-Johnson, M.,** & **Puri, V.** (2014). An investigation of prosody in Hindi narrative speech. *Proceedings of Speech Prosody, 7*, 623–627. DOI: https://doi.org/10.21437/SpeechProsody.2014-112

**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., Himmelmann, N. P.,** & **Kluge, A.** (2019). Stress predictors in a Papuan Malay random forest. *Proceedings of the International Congress of Phonetic Sciences* (ICPhS XIX), Melbourne, 2871–2875.

**Kaufman, D.,** & **Himmelmann, N. P.** (accepted). Suprasegmental phonology. In A. Adelaar & A. Schapper (Eds.), *The Oxford guide to the Western Austronesian languages*. Oxford, England: Oxford University Press.

**Kluge, A.** (2017). *A grammar of Papuan Malay*. Berlin, Germany: Language Science Press.

**Ladd, R. D.** (2008). *Intonational phonology* (2nd ed.). Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511808814

**Landis, J. R.,** & **Koch, G. G.** (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. DOI: https://doi.org/10.2307/2529310

**Mahrt, T.** (2016). *LMEDS: Language markup and experimental design software*. Retrieved from https://github.com/timmahrt/LMEDS

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

**Mo, Y.,** & **Cole, J.** (2010). Perception of prosodic boundaries in spontaneous speech with and without silent pauses. *Journal of the Acoustical Society of America, 127*, 1956. DOI: https://doi.org/10.1121/1.3384972

**Peperkamp, S.,** & **Dupoux, E.** (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 203–240). Berlin, Germany: Mouton de Gruyter.

**Pintér, G., Mizuguchi, S.,** & **Tateishi, K.** (2014). Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. *Proceedings of Interspeech,* 2014, 544–547.

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

**Riesberg, S., Kalbertodt, J., Baumann, S.,** & **Himmelmann, N. P.** (2018). On the perception of prosodic prominences and boundaries in Papuan Malay. In S. Riesberg, A. Shiohara & A. Utsumi (Eds.), *A cross-linguistic perspective on information structure in Austronesian languages* (pp. 389–414). Berlin, Germany: Language Science Press.

**Rietveld, T. C. M.,** & **Gussenhoven, C.** (1985). On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics, 15*, 273–285. DOI: https://doi.org/10.1016/S0095-4470(19)30571-6

**Smith, C. L.** (2011). Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models. *Proceedings of the 3rd International Conference on the Discourse-Prosody Interface 2009*, Paris, France, 335–349.

**Smith, C. L.,** & **Edmunds, P.** (2013). Native English listeners' perceptions of prosody in L1 and L2 reading. *Proceedings of Interspeech* 2013, Lyon, France, # IS130507, 235–238.

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

**Turk, A. E.,** & **Shattuck-Hufnagel, S.** (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*, 445–472. DOI: https://doi.org/10.1016/j.wocn.2006.12.001

**Uhmann, S.** (1988). Akzenttöne, Grenztöne und Fokussilben. Zum Aufbau eines phonologischen Intonationssystems für das Deutsche. In H. Altmann (Ed.), *Intonationsforschungen* (pp. 65–88). Tübingen, Germany: Niemeyer.

**van Heuven, V. J.,** & **van Zanten, E.** (2007). Concluding remarks. In V. J. van Heuven & E. van Zanten (Eds.) *Prosody in Indonesian Languages* (pp. 191–202). Utrecht, the Netherlands: LOT.

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

**How to cite this article:** Riesberg, S., Kalbertodt, J., Baumann, S., & Himmelmann, N. P. 2020 Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 11(1): 8, pp. 1–35. DOI: https://doi.org/10.5334/labphon.192

**Submitted:** 22 December 2018 **Accepted:** 07 May 2020 **Published:** 01 July 2020