# On the robustness of intonational phrases in spontaneous speech – a cross-linguistic interrater study

Nikolaus P. Himmelmann, Meytal Sandler, Jan Strunk & Volker Unterladstetter[1]

Universität zu Köln

## Abstract

This study is concerned with the identifiability of intonational phrase (IP) boundaries across familiar and unfamiliar languages. Four annotators independently segmented a substantial corpus of more than three hours of spontaneous narrative speech into IPs. The corpus included recordings from their native German, but also from three languages from eastern Indonesia they had never heard before. The results show significant interrater agreement across the corpus as a whole as well as for each subcorpus. Hence, IP boundaries can be reliably identified across familiar and unfamiliar languages, which in turn suggests that IP-phrasing has its roots in a non-language specific, universal property of spoken language.

A closer look at the boundaries where annotators did not agree shows a) that annotators differ in their segmentation strategies, some focusing on melodic cues, others paying greater attention to rhythmic cues; and b) that the more a boundary is based on melodic cues only, the more difficult it becomes to identify it.

## 1. Introduction

Spoken language is produced in chunks delimited by prosodic cues such as a coherent intonation contour and pauses. These chunks are recognized in all models of prosodic analysis, albeit by differing names and definitional criteria. Widely known are *tone group* (Halliday 1967) and *intonation unit* (Chafe 1980 passim) next to *intonation(al) phrase*, the term used here and in most work applying an autosegmental-metrical approach to prosody (Shattuck-Hufnagel & Turk 1996: 206; Ladd 2008). They also play a role in models of speech production (Levelt 1989) and are basic units of analysis in the type of discourse and conversation analysis inspired by Chafe (1987, 1994).

Intonational phrases (henceforth IPs) are widely held not to pose particular problems of identification. Thus, for example, Shattuck-Hufnagel & Turk (1996: 211) note that "[p]erceptually, the boundaries of an Intonational Phrase are quite clear, …". And Chafe (1994: 62) writes:

> In spite of problematic cases, intonation units emerge from the stream of speech with a high degree of satisfying consistency, not just in English, but in all languages I have been able to observe and in fact in all styles of speaking, …

However, to date this assumption has not been put to empirical scrutiny in ways standard in many disciplines and research areas concerned with segmentation or classification tasks, i.e. by evaluating

[1] Author contribution:
NPH: overall design of study and paper, main author of sections 1, 2, 6, 7 and final revision of all other sections; contributor to consensus version. MS: contributor to interrater study (including recordings) and consensus version; main administrator of interrater study. JS: all statistical analyses, draft of section 4, contributor to interrater study design and consensus version. VU: draft of sections 3 and 5, contributor to interrater study (including recordings) and consensus version; main implementer of disagreement study.
Research for this paper was funded by grant nr. 01UG1240A from the German Federal Ministry for Education and Research (*Bundesministerium für Bildung und Forschung* (BMBF)) to Nikolaus P. Himmelmann.

interrater agreement. As briefly reviewed in section 2, there have been a few interrater studies involving IPs, but these are typically limited in the following two regards: 1) they involve quite short (< 30 seconds) examples specifically recorded for the task or excerpted from longer recordings; and 2) they usually combine a number of tasks, i.e. labeling prosodic boundaries and labeling prosodic prominences (e.g. pitch accents).

The current study, in contrast, is exclusively concerned with IP boundaries and involves the segmentation of a corpus of more than three hours of spontaneous narrative speech, consisting of 60 recordings which are between 01:24 and 10:06 minutes long (cp. Table 1 below). Most importantly, however, it is also concerned with an issue which, to the best of our knowledge, has not been addressed in the literature to date, but is hinted at in the quote from Chafe above. This is the issue of whether IPs are identifiable across languages, specifically whether non-native listeners are able consistently to identify IP boundaries in languages they are not familiar with, i.e. without being able to understand the utterances to be segmented and without familiarizing themselves with the specifics of the prosodic system of the language in question.

This research question presupposes the hypothesis that some features of IP boundaries are universally recognizable. This hypothesis has some initial plausibility because at least two types of criteria for IP boundaries can be easily applied across natural languages, and in fact are widely used in this way. First, there are melodic cues to IP boundaries, specifically the overall coherence of the intonation contour. Second, there are rhythmic cues to IP boundaries, including (non-hesitation) pauses, domain-final lengthening and domain-initial anacrusis. Note that the first criterion also holds for tone languages where boundary tones may be lacking or difficult to perceive.

The two types of boundary cues are essentially independent of each other. A coherent intonation contour may be interrupted by a (short) pause but still be heard as a coherent contour. Similarly, a coherent intonation contour may contain a lengthened segment which, however, is not heard as unit-final. Consequently, using these two types of criteria for segmentation, on the one hand, involves the assumption that the units delimited by each cue type overlap to a significant degree, so that it is useful to make use of both types. On the other hand, using two types of criteria has the potential for conflicting results and hence the question of how to resolve them, an issue we will return to repeatedly (but see in particular section 6).

The current study thus differs from other interrater studies primarily with regard to its cross-linguistic perspective. The material to be segmented is roughly comparable between languages, as it consists of retellings of the Pear Film (Chafe 1980) in four different languages: German, the native language of the annotators; Papuan Malay, the language of everyday communication in the major centers of West Papua (Indonesia); Wooi, an Austronesian language spoken on Yapen Island in West Papua; and Yali, a Papuan highland language from West Papua. Two of the authors have first-hand working experience with the West Papuan languages. All the other annotators taking part in the experiment were unfamiliar with these languages.

The core questions to be answered by this study then are as follows:

Q1. Do the segmentation results for the corpus as a whole and for each of the languages involved show above-chance interrater agreement according to standard kappa metrics for annotation tasks? That is, can IPs be reliably identified in spontaneous narrative speech across a number of differing languages using just the two types of cues mentioned above (melodic and rhythmic)?

Q2. Is there significant variation in interrater agreement values for familiar and unfamiliar languages? What are possible reasons for (the lack of) such variation?

Q3. Are melodic and rhythmic boundary cues handled differently by different annotators? How do such differences impact on interrater agreement? Is it possible and useful to work only with one type of cue?

As for the second question, there are two ways in which the familiarity with a language may become relevant in the segmentation task and influence interrater agreement. First, it could be the case that the prosodic cues used as segmentation criteria come in language-specific forms and are thus more readily recognised in familiar than in unfamiliar languages. Prima facie, such language-specific forms are less likely for pauses, but they have some initial plausibility for melodic cues as well as unit-final lengthening and anacrusis (the latter, inter alia, presuppose that segment boundaries are easily identified). If there are in fact such language-specific forms, this would predict significantly worse interrater agreement results for unfamiliar languages, *unless* these effects are offset by some other factor (e.g. the usefulness of pauses as boundary cues).

Second, as is well-known from the literature (e.g. Cole *et al.* 2010b), prosodic boundary perception is not only influenced by prosodic factors, but also by non-prosodic ones, in particular syntactic structure and semantic and pragmatic coherence. Thus, there is a strong tendency of IP boundaries to overlap with clause boundaries and a concomitant tendency to hear IP boundaries at clause boundaries. The unfamiliar language-condition completely removes the potential influence of non-prosodic factors, theoretically with two possible outcomes. On the one hand, interrater agreement could be significantly less strong in the case of unfamiliar languages because of the missing non-prosodic information. However, as non-prosodic information brings in a totally different layer of factors, it also increases the potential for conflict between different types of segmentation cues (cp. Ladd 2008: 288–290). As a consequence, interrater agreement in familiar languages could, on the other hand, be worse than in unfamiliar ones, as in the latter case annotators are forced to focus exclusively on prosodic cues.

The third question partially overlaps with the second one, but highlights the potential for conflicts and different segmentation strategies resulting from the fact that a number of different prosodic cues are used as criteria for prosodic boundaries. Here it is to be expected that interrater agreement is highest when multiple cues point to the same location for a potential boundary and that it is lower when they diverge. A prototypical example of the latter are sequences of IPs without intervening pauses, where diagnosing a boundary primarily depends on melodic cues.

The paper is structured as follows. Section 2 briefly reviews previous interrater studies concerned with the identification of IP boundaries and highlights the points where the current study diverges from these. It also provides further details on the boundary cues focused on here and their complex interrelationship. Of particular importance in this regard is the fact that, on the one hand, melodic cues take precedence over rhythmic cues in cases of conflict. On the other hand, rhythmic cues, in particular pauses, are more easily and robustly identifiable. Section 3 provides details on the task design and the data.

The empirical core of this study is presented in sections 4–6, all of which are concerned with statistically significant patterns in the segmentation data. The detailed interrater agreement results provided in section 4 show that there is indeed robust interrater agreement for the corpus as a whole as well as for individual languages. There is, however, also some variation both with regard to the languages investigated and the individual student annotators participating in this study. This variation points to the fact that partially different strategies are used by the annotators in the segmentation task, and that different boundary cues present varying kinds of difficulties. In order to determine more precisely the factors giving rise to disagreements, section 5 supplies a systematic analysis of all boundaries where disagreements occur. These can be shown to fall into a smallish number of subtypes, most importantly misinterpreted major pitch changes and IP boundaries without pauses. As (the lack of) pauses are found to be a major factor in determining interrater agreement, section 6, finally, takes a closer look at the distribution of pauses in the corpus and compares a purely pause-based segmentation with one using both melodic and rhythmic cues.

Section 7 discusses the theoretical import of our results for current concepts of IPs and their functions. The results provide clear support for the claim that some cues for IP boundaries are universal, and that speakers thus in principle are in a position to identify IP-like chunks even in languages they are not familiar with. Importantly, however, this does not mean that the *structure* of IPs as phonological units is necessarily also universal. Rather, our results support a view of prosodic categories (and perhaps phonological categories more generally) which are partially universal inasmuch as they are grounded in the mechanics of speaking, but partially also language-specific inasmuch as they reflect the contingencies of historical developments in the language-specific grammaticalization of prosodic features.

Section 8 summarizes.

## 2. Prosodic interrater agreement studies and their targets

Interrater agreement studies concerned with prosodic phenomena can be roughly classified into two types. One type targets an annotation scheme of prosodic categories. It requires a theoretical understanding of these categories, and practical training for handling them. A recent example is the study by Breen *et al.* (2012), who compare two such prosodic annotation schemes, the less widely known Rhythm and Pitch (RaP) system, and the well-established and widely used To(nes and) B(reak) I(ndices) system (Silverman *et al.* 1992, Pitrelli *et al.* 1994; cp. the two volumes ed. by Jun (2005, 2014) for recent surveys and cross-linguistic applications). They also present a very useful survey of previous interrater studies of this type and of the methodological issues involved in them, which do not need to be repeated here (see also Cole *et al.* 2010b: 1143–1145).

Of major importance for the current study is the fact that this type of study targets language-specific phonological categories, i.e. tonal targets and prosodic boundaries of different types. The annotation schemes tested may differ in the consistency and directness of the auditory and acoustic evidence for the targeted category, but the decisions are clearly about (abstract) phonological categories and not about phonetic events. Part of the training component defining this type of study is the provision of examples illustrating typical auditory and acoustic correlates of the intended categories. Furthermore, labelers are usually provided with acoustic data (minimally wave-form and F0 contour) in addition to audio files when working on their task.

The other type of study targets the perception of prosodic prominences and boundaries by naïve listeners (i.e. listeners who have no expertise in prosodic theory and annotation), and then investigates which properties correlate with the points in the transcript marked by them as prominences or boundaries. The focus is usually on phonetic cues (e.g. pitch changes) but may also include grammatical (word class, syntax) and semantic or pragmatic information. A prototypical study along these lines is Mo *et al.* (2008)[2] with detailed analytical follow-ups in Cole *et al.* (2010a) on phonetic factors, and Cole *et al.* (2010b) on syntactic (and more generally, non-prosodic) factors. In this study, more than 70 undergraduate students of linguistics marked prosodic prominences and boundaries in 18 short (20 seconds long) excerpts of spontaneous American English, based solely on their auditory impressions. The instruction regarding prominences and boundaries they received is summarized as follows:

> A prominent word is defined as a word that is “highlighted for the listener, and stands out from other non-prominent words”, while a chunk is defined as a grouping of words “that helps the listener interpret the utterance”, and that chunking is “especially important when the speaker produces long stretches of continuous speech”. (Mo *et al.* 2008: 736)

In Mo *et al.* (2008), the annotators marked their prominences and boundaries on printouts of the transcripts of the excerpts, which included word boundaries, speech errors and dysfluencies, but no punctuation or capitalization. The findings of this study most relevant for the current investigation are: a) there is significant (i.e. well above chance) interrater agreement with regard to boundaries with a mean Cohen’s kappa agreement coefficient of 0.582 across all pairs of transcribers (the values with regard to prominence marking are much lower); b) there is significant variation with regard to both the speakers, where Fleiss’ kappa coefficients (measuring agreement between all listeners at the same time) range from 0.35 - 0.95, and the listeners, with some agreement pairs only reaching a Cohen’s kappa agreement coefficient as low as 0.24, while others agree to a very large extent, as reflected in a Cohen’s kappa coefficient of 0.85.

[2] The method has its origin in the perception-oriented approach to intonation developed in Eindhoven as summarized in 't Hart, Collier & Cohen (1990). Work on boundary perception in this framework is illustrated by de Pijper & Sanderman (1994); see Sanderman (1996) for more detailed discussion. Streefkerk (2002) contains an overview and demonstration of work on prominence perception in this tradition.

In some ways, Buhmann *et al.* (2002) is a very similar study, based on Dutch corpus data. However, their procedure is different in a number of important regards. First, while working with non-expert annotators, they include a relatively intensive training period in which, after having received the instructions and an illustration of the task, the annotators first worked through a learning corpus of 15 minutes, receiving feedback on their performance on various levels. Second, the test corpus was substantially larger than the corpus used in most other studies, consisting of more than 8000 words (45 minutes) of read, scripted and unscripted (spontaneous) speech. Third, an on-line working environment was used in carrying out the task, which included the audio-visual display of waveforms as well as time-aligned text files. Finally, the test corpus was already pre-segmented into pause-bounded phrases of a length of roughly ten seconds, using automatically detected pauses larger than 0.5 seconds as indicators for strong prosodic phrase boundaries. Given the intensive training and the pre-segmentation, it does not come as a surprise that Buhmann et al. obtain fairly high coefficients of interrater agreement. In the case of boundaries, the Cohen's kappa coefficients for interrater pairs range from 0.695 to as high as 0.884 (Buhmann *et al.* 2002: 782).

As for prosodic boundaries, Buhmann *et al.* (2002: 779) simply speak of "breaks", thus clearly targeting a non-technical category which presumably is part of the non-expert understanding of spoken language. They distinguish strong and weak breaks, which are defined as follows:

1. **Strong breaks** (symbol '||') are defined as severe interruptions of the normal flow of speech. They are typically realized as a clear pause or even an inhalation.

   Ex: *he was there || and so was his girl-friend*

2. **Weak breaks** (symbol '|') are defined as weak but still clearly audible interruptions of the speech flow. Although no real pause is observed, it is clear that the words (or parts of a word) straddling the break are not connected the way one would expect them to be in fluent speech. In case of doubt between a strong and a weak break, the human transcriber is instructed to choose for a weak break.

   Ex: *I can tell you | this was un|be|lievable* (Buhmann *et al.* 2002: 780f)

Note that while the instructions in Mo *et al.* (2008) focus on a presumed function of chunking (cp. "that helps the listener interpret the utterance" in the quote above), here the focus is clearly on auditory impressions, with an explicit emphasis on pauses and not explicitly appealing to coherent melody contours.

The current study clearly belongs to the second type in that it targets the perception of prosodic boundaries by non-expert listeners. It differs from the preceding studies in some aspects of procedure, as further detailed in the following section. But there are in particular two major points of difference which warrant further attention here. The most important difference is that our study brings in a new variable, i.e. different languages, and compares the performance of annotators across familiar and unfamiliar languages. As already mentioned in the introduction, this task design presupposes that some aspects of the chunking of speech or the "interruptions of the normal flow of speech", as Buhmann *et al.* put it in the quote above, occur across languages and can be auditorily identified across languages.

As for their cross-linguistic occurrence, it will suffice to point out that there is probably no discussion of the intonation of a particular language which does not make reference to the coherence of the melody setting off one IP from the adjacent ones. Furthermore, Fletcher (2010), a very thorough review of the literature on speech timing and rhythm, provides a wealth of references for tempo changes (2010: 540-547) and pauses (2010: 573-575) as cross-linguistically widely attested cues for boundaries.

As for their cross-linguistic identifiability, however, matters are different, as this has not been explored systematically to this date and is the very topic of this investigation. Hence, it is of particular importance what cues we made use of and how we explained them to the annotators. This is the second point where the present study in part diverges from Mo *et al.* (2008) and Buhmann *et al*. (2002). Our written instructions regarding IP boundary cues, handed out to the annotators and explained once verbally, read as follows:[3]

> Your task is to segment an audio recording containing the narration of a short film into **intonational phrases**, i.e. into sequences that are perceivable as a **distinct unit** by means of a coherent melody/a coherent pitch contour.
>
> <u>To keep in mind</u>
> Boundaries between two intonational phrases are typically characterised by two features:
> 1. an interruption of the **rhythmic** delivery by a (sometimes only very short) pause, **lengthening** of the last segment at the end of a unit and/or **increased articulation rate** at the beginning of a new unit (*anacrusis*);
> 2. a disruption of the pitch contour/melody line: a **pitch jump** (up or down) between the end of a unit and the beginning of the subsequent one; intonational phrases often exhibit a constant decline in fundamental frequency which at the boundary of a unit is reset to the default pitch level of the speaker in a given context (*reset*). This is typically followed by another decline in fundamental frequency (*declination*).
>
> Pauses, however, may sometimes also occur within an intonational phrase, e.g., if the speaker is searching for a word or corrects him/herself = hesitation pauses. Hesitation pauses are often filled (*umh*, etc.) but not necessarily so. What is important is that the pitch levels before and after a

[3] The instructions in the original German:
Ihre Aufgabe ist es, eine Audio-Aufnahme mit der Nacherzählung eines kurzen Films in **Intonationseinheiten** einzuteilen, d.h. in Abschnitte, die durch eine kohärente Melodie/einen kohärenten Tonhöhenverlauf als **eine Einheit** erkennbar sind.
<u>Wissenwertes</u>
Grenzen zwischen zwei Intonationseinheiten zeichnen sich dabei in der Regel durch zwei Dinge aus:
1. eine **rhythmische** Unterbrechung durch eine (ggf. auch nur sehr kurze) Pause, die **Dehnung** des letzten Segments am Ende einer Einheit und/oder die **beschleunigte Produktion** am Anfang einer neuen Einheit (*Anakrusis*);
2. durch eine Unterbrechung im Tonhöhenverlauf/in der Melodie: einen Tonhöhensprung (nach oben oder unten) zwischen dem Ende der einen und dem Beginn der folgenden Einheit; oft zeichnet sich eine Intonationseinheit durch einen kontinuierlichen Abfall der Grundfrequenz aus, der an eine Einheitsgrenze auf die Normaltonlage des Sprechers zurückgesetzt wird (*reset*). Daraufhin folgt typischerweise ein erneuter Abfall der Grundfrequenz (*declination*).

Pausen können allerdings manchmal auch innerhalb einer Intonationseinheit auftreten, z.B. wenn der Sprecher / die Sprecherin nach dem folgenden Wort sucht oder sich korrigiert = Verzögerungspausen. Verzögerungspausen sind oft, aber nicht notwendig gefüllt (*ähm* etc). Wichtig ist, dass der Tonhöhenverlauf vor und nach der Pause nahtlos aneinander anschließt, es mithin nicht zu einem Neueinsatz der Melodie kommt, sondern die vor der Pause begonnene Kontur fortgesetzt wird.

hesitation pause fit together continuously. That is, rather than a new onset of the melody line, the original pitch contour is continued after the pause.

Along with these explanations, the annotators were presented with five audio examples of boundary cues between two intonational phrases in order to exemplify typical configurations at IP boundaries. The examples were taken from a short personal narrative in German that is not part of the corpus used in the segmentation task. They were played several times, illustrating the following typical boundary configurations:

1. two IPs set off by a clear melodic break (clearly audible new onset by downward jump in pitch after strongly rising boundary tone) accompanied by a pause of 240ms and greatly reduced intensity of the second IP.
2. two IPs set off primarily by a clear melodic break only (new onset by downward jump in pitch after strongly rising boundary tone) accompanied by a very short (70ms), but noticeable period of silence.
3. two IPs in direct sequence without any intervening silence, but with final lengthening at the end of the first IP and a clear melodic break (falling boundary tone followed by upward jump in pitch).
4. one IP with an internal hesitation pause of 690ms after which the pitch is resumed at approximately the same level as before the hesitation.
5. two IPs involving minor unit-internal hesitations and no intervening pause, but a clear melodic break (major upward jump in pitch) and anacrusis at the beginning of the second IP.

Like the Buhmann *et al.* study, annotators were thus very clearly instructed to follow prosodic cues for boundaries only, but unlike Buhmann *et al.*, a clear distinction was made between melodic and rhythmic cues. Furthermore, and in line with widespread assumptions in the literature, the primacy of melodic coherence is emphasized. But note that while our instructions go into a moderate degree of technical detail, we did not make direct reference to analytical constituents of melodic contours such as boundary tones, despite the fact that all languages in our corpus make use of them.[4] There are two reasons for this decision. First, we do not want to presume that boundary tones are universally attested in natural languages. Second, the concept of a boundary tone only makes sense as part of an overall theoretical model, knowledge of which we could not, and did not want to, presuppose on the side of the annotators participating in this study.

Importantly, as just mentioned, melodic and rhythmic cues are not of equal theoretical status, and both are ambivalent as boundary cues. As for melodic coherence, it would appear to be uncontroversial that it is the primary cue for IPs, both theoretically and auditorily. This is clear from the fact that melodic coherence is the only obligatory boundary cue, while all other boundary cues are strictly speaking optional (though final lengthening and pauses may occur quite frequently and regularly). Furthermore, rhythmic cues in part depend on, and can be overridden by, melodic coherence. Lengthening is heard as unit-final only if such an interpretation is coherent with the melody (otherwise, it may be heard as

[4] Reference to boundary tones in the description of the examples used for instruction purposes has been added only to make it easier for the expert reader to identify the type of example we have used. In the actual instructions, the focus was on the auditory impression.

particular emphasis on the syllable where it occurs). Similarly, pauses are heard as boundary pauses only when the melodic contour appears to have reached its projected endpoint. Thus, in principle, one could entertain the idea that IP boundaries should exclusively be defined and identified in terms of melodic coherence.

However, among all boundary cues, melodic coherence is the most difficult to explicate and define precisely, and to be perceived consistently when paying conscious attention to it (as in a segmentation task such as the present one). Models of intonation such as the autosegmental-metrical one may be quite successful in modeling actually occurring contours, but they do not explicate – and provide little practical support for explicating to non-expert listeners – the coherence of contours. Using such models in a segmentation task basically means that the participants have to be trained in recognizing (the components of) the contours provided for by the model, as has been done in the first type of interrater study mentioned at the beginning of this section.

In practical terms, the jumps in pitch occurring between off- and onsets of IPs often are not larger than the micro-perturbations caused by obstruents. And while we just noted the fact that the identification of rhythmic boundary cues in part depends on their alignment with the melodic contour, the reverse also holds: The identification of a coherent contour in part depends on its interplay with rhythmic cues. The clearest example for this interdependence is the fact that there are limits to the length of a pause across which a melody can be heard as coherent. While the exact length may vary depending on language, culture, and speaker, coherent contours rarely span pauses longer than one second. Furthermore, a possible melodic endpoint tends to be heard as an actual melodic endpoint more clearly and easily when accompanied by segmental lengthening and followed by a pause.

It thus seems to be the case that while theoretically there is a primacy of melodic cues, in practical-operational terms a relation of mutual reinforcement exists: the more cues, melodic and rhythmic, come together, the clearer, and possibly also stronger, the boundary. With "practical-operational" we here refer primarily to the segmentation task at hand. However, we do not think that it is very speculative to assume that this also holds for speaker-hearers engaged in the actual production and comprehension of fluent speech.

The ambivalence of pauses as boundary indicators arises from the fact that they may occur both in between and within IPs. There is thus a need to distinguish between IP *external* and *internal* pauses. External pauses are pauses that occur between two subsequent IPs. According to a widespread view (e.g. Goldman-Eisler 1968, Levelt 1989, Chafe 1994), they usually arise because speakers need some time to plan the next IP (and are hence also called planning pauses) but they may in some cases also be used more deliberately as an IP boundary signal. External pauses also often give the speaker the opportunity to breathe. Internal pauses, in contrast, are pauses that occur, usually unintentionally, during the course of production of an IP. They most often result from production difficulties, such as problems with lexical access (finding a particular word), self-corrections, etc., and are hence also often called hesitation pauses.

In practical-operational terms, pauses are probably the easiest IP boundary cue to identify. Because of this, external pauses, when correctly interpreted as such, are an important practical cue for IP boundaries. The occurrence of lots of internal pauses, in contrast, may render identification of IP boundaries more difficult as they can be misinterpreted as IP boundary cues, especially when the hearer does not understand the content of a given segment.

To summarize, our study focusses on prosodic boundary cues and, in the case of languages unfamiliar to the annotators, actually forces them to exclusively pay attention to them. While theoretically, melodic coherence is the primary cue for IP boundaries, in practical (and probably also in online-processing) terms, both melodic and rhythmic cues are, and have to be, used in identifying IP boundaries. They mutually reinforce each other when occurring temporally aligned, but they may give rise to disagreements when not synchronized. Pauses have a special status because their identification tends to be relatively easy and to allow for great consistency, but they are not unequivocal boundary cues because of the occurrence of IP-internal pauses.

## 3. Data and procedure

The corpus used for the segmentation task in this study consists of sixty retellings of the Pear Film, a six-minute film made in 1975 for the cross-linguistic study of the cognitive, cultural and linguistic aspects of narrative production (Chafe 1980).[5] The sound track does not contain any linguistic utterances (or music for that matter) but simply consists of the sounds associated with the depicted actions (such as picking pears or crashing into a rock with a bike).

The recording procedure we used in the construction of our corpus was for one person to watch the pear film on a laptop screen and then tell it to another person who had not seen the film before. The interlocutor was instructed to behave 'naturally' in accordance with the context of retelling a movie, i.e. to ask clarification questions and to provide feedback whenever and wherever appropriate. While all interlocutors engaged in appropriate (verbal and non-verbal) backchanneling, only very few actually asked clarification questions, never exceeding three questions in one telling. All verbal utterances of the interlocutor are included in the recordings and transcripts used for this study, but they are not included in the segmentation task. Only the narrators' speech is segmented into IPs.

The sixty pear stories are told in different languages, primarily German and three languages from Eastern Indonesia, the major field site of the first author. Table 1 provides details for the corpus, which is partitioned into three groups for processing and presentation purposes, each comprising twenty stories. This corpus was originally compiled for the AUVIS project (hence the name of the corpus).[6] The main goal of this project was to explore possibilities of automatically annotating and searching audio and

[5] See also http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm.

[6] AUVIS = *Audiovisual data-mining using event segmentation in multimodal language data as an example*, cp. https://tla.mpi.nl/projects_info/auvis/ for more information. As a case study for realistic search scenarios, the project involved an exploration of the alignment between gestural, prosodic and grammatical units. In gesture research, all annotation is standardly done by multiple annotators, which was one reason to work with multiple annotators for the prosodic annotation as well.

video streams of unannotated or only partially annotated recordings from unrelated languages, with a particular focus on underdocumented and underresourced languages. For practical and explorative purposes, the corpus also includes smallish samples from a few additional varieties, i.e. Kölsch (the German dialect spoken in Cologne), English, and Waima'a, an Austronesian language from East Timor. Segmentation results for these varieties do not differ from the results obtained for the four main languages and are therefore included in our overall statistics. They are excluded from those parts of the study specifically concerned with cross-linguistic comparison, because they are too small for valid statistical inferences.

Table 1: The AUVIS Pear Film Corpus (interrater version[7])

| No. of recordings | | Total length | Mean length | Total number of words |
|---|---|---|---|---|
| **Group I: Germanic** | | | | |
| German (DEU) | 18 | 53m 28s | 02m 58s | 8,836 |
| Kölsch (KSH) | 1 | 02m 31s | 02m 31s | 286 |
| English (ENG) | 1 | 10m 06s | 10m 06s | 1,418 |
| **Subtotal** | **20** | 01h 06m 05s | 05m 12s | 10,540 |
| **Group II: Papuan Malay** | | | | |
| Papuan Malay (PMY) | **20** | 01h 04m 00s | 03m 12s | 10,373 |
| **Group III: Eastern Indonesian** | | | | |
| Wooi (WBW) | 12 | 34m 53s | 02m 54s | 3,557 |
| Waima'a (WMH) | 2 | 08m 15s | 04m 08s | 1,406 |
| Yali (YAC) | 6 | 17m 42s | 02m 57s | 2,007 |
| **Subtotal** | **20** | 01h 00m 50s | 03m 20s | 6,970 |
| **Total** | **60** | **03h 10m 55s** | **03m 55s** | **27,883** |

The first group in Table 1 consists of eighteen pear film recordings in (Standard colloquial) German, one recording in the vernacular dialect of Cologne (Kölsch) and one recording in (American) English. Six of these recordings were done with analog audio and video recorders in the nineteen nineties and therefore are of somewhat lower quality, especially with regard to the video (which did not play a role in the current study). The remaining ones were recorded in 2012 with up-to-date audio/video equipment[8] for the specific purposes of the AUVIS project. At the time of recording, the speakers involved were mostly students in their early twenties at the Universität zu Köln. Five recordings involve more mature speakers (30-50 years old).

The second group comprises recordings in Papuan Malay, the *lingua franca* of West Papua, the western half of the island of New Guinea governed by Indonesia (see Kluge 2014 for a recent description). The pear film recordings in Papuan Malay were recorded at the Center for Endangered

[7] This version of the *AUVIS Pear Film Corpus* differs from the version used in gesture-related studies with regard to one German retelling, which was replaced by another one at a later point, when it became apparent that the narrator of the former retelling was aware of the fact that the study was concerned with gestures.

[8] All recent recordings in Cologne and Indonesia were done with a Sony digital video recorder (e.g. HDR-CX730E or similar type) mounted on a tripod and an external microphone (in most instances, a stereo on-camera condenser microphone).

Languages Documentation (CELD) in Manokwari, the capital of the province of *Papua Barat* (West Papua). The narrators as well as their interlocutors were all of approximately equal age (early to mid-twenties) and enrolled as English students at the local university.

The third group consists of three lesser-known languages of Eastern Indonesia, for which language documentation corpora have been compiled in documentation projects based in Cologne. Two of these languages, Wooi (Kirihio *et al.* 2009–2015) and Waima'a (Belo *et al.* 2002–2006), are Austronesian languages (Central-Eastern Malayo-Polynesian branch) spoken in coastal settings in West Papua and East Timor, respectively. Both speech communities are small (less than 3,000 speakers each), multilingual and currently shifting to regional standards (Papuan Malay and Tetum, respectively). The pear film recordings in Wooi and Waima'a have all been recorded in the field sites and are generally of a lower quality than the recordings done at the CELD (more background noises of different kinds). The age of the Wooi speakers is more mixed than in the other language groups, covering a range from speakers in their early twenties to mature speakers of 50 years and older. The third language, Yali (Riesberg *et al.* 2012-2016), is a Papuan language (Trans-New-Guinea phylum) spoken in the highlands of West Papua. The number of speakers is somewhat higher (around 10,000) and only younger generations are multilingual in varieties of Malay (both Standard Indonesian and Papuan Malay, to differing degrees). The recordings were made at the CELD in Manokwari with young native speakers in their early twenties who were enrolled as students at the local university or (in one case) as a secondary school student.

Prior to the current study, all sixty pear story recordings had been transcribed by native speakers of the respective languages using the ELAN program.[9] For current purposes, all information pertaining to the temporal alignment of the transcription to the audio stream was eliminated and a plain text version was created. The task of the annotators was independently to segment the recordings into IPs on the basis of the audio stream and the plain text script. That is, for each recording, the annotators received the WAVE file (but no video file), a plain text file containing the transcript without any hints with regard to prosodic phrasing (no punctuation, line breaks, paragraphs, capitals, etc.), and a (largely empty) ELAN file. Note that in distinction to most other perception studies to date (cp. section 2 above), the transcript also did not contain any indications with regard to dysfluencies, but it did contain a representation of unclear segments which could not be transcribed. These were indicated by small x's, with roughly one x per unidentifiable syllable.

The ELAN file given to the annotators contained two annotation tiers, one for the narrator, and another one for the interlocutor. In order to facilitate orientation within the recording, we left the utterances of

[9] All transcriptions were checked by one of the authors or another member of the documentation team in Cologne working on the language in question. We thank Sonja Riesberg for her help with the Yali data. See http://dobes.mpi.nl/projects/waimaa/ (DoBeS Waima'a project), http://dobes.mpi.nl/projects/wooi/ (DoBeS Wooi project), and http://dobes.mpi.nl/projects/celd/ (DoBeS Central Papuan Summits Languages project including a documentation of Yali) for full acknowledgements and further information on working procedures in the documentation projects.
ELAN is a multimedia annotation tool for multi-modal research, see further http://tla.mpi.nl/tools/tla-tools/elan/.

the interlocutors (which were very few anyway) in place and also included them on separate lines in the plain text transcription file. The tier for the narrator was left blank. After identifying a stretch of the audio stream which they assumed to form an IP, the annotators' task was to copy the respective portion of the transcript from the plain text file and paste it into the appropriate selection on the narrator tier in ELAN.

The workflow was explained verbally to the annotators. The instructions regarding the boundary cues they were to pay attention to were given to them in written form and also explained verbally, as further detailed in section 2 above.

Annotators worked on the task on their own, without any time constraints (some taking less than a week per package, others close to a month). They received the recordings in packages per group, starting with Group I (Germanic), then Group II (Papuan Malay), and finally Group III (Eastern Indonesian languages). The order of the recordings in a group was alphabetical in accordance with the abbreviated names of the narrators, except for Group II, which was arranged in such a way that male and female narrators followed each other roughly in alternating order. In the Germanic part of the corpus, alphabetic ordering already resulted in a well-mixed sequence of female and male narrators. Most narrators in the Eastern Indonesian part of the corpus are men, except for Waima'a (two females). The sequence here was Wooi first, then Waima'a, and finally Yali.

Four linguistics students were recruited for this task and paid a fixed rate for each delivery package. They were students in different linguistics programs at the University of Cologne with varying degrees of familiarity with prosodic analyses, as further detailed in Table 2.

Table 2: A brief characterization of the student annotators and the authors' consensus version

| | |
|---|---|
| **R1** | Bachelor student (female) in Linguistics, basic introduction to prosody as part of introductory courses of BA program |
| **R2** | Master student (male) in Linguistics, basic introduction to prosody as part of introductory courses of BA program |
| **R3** | Master student (female) in Linguistics, basic introduction to prosody as part of introductory courses of BA program |
| **R4** | Master student (female) in Linguistics, specialising in phonetics, writing MA thesis on prosodic topic at the time of involvement in the project |
| **CONS**/Authors | each session independently segmented by 2 of the authors, divergences discussed and resolved, final check by first author |

In addition and as also indicated in Table 2, the authors produced a consensus version which, importantly, is based on specific hypotheses regarding the phonological structure of IPs in each of the languages investigated. This version was produced in several steps. First, each recording was independently segmented by two of the three last-named authors. Second, the three last-named authors compared their segmentations and produced a first consensus version by resolving disagreements through relistening and discussion. As a third and final step, this version was checked by the first author with a special focus on problematic cases and on overall consistency in instances where the exact placement of the boundary is arguably arbitrary (see section 5 for further explanation). In contrast to the four student annotators R1 to R4, the authors made regular use of instrumental evidence, usually in the form of F0 and intensity plots produced by PRAAT (Boersma 2001, Boersma & Weenink 2015) in order to decide particularly intricate cases. Given that the consensus version is based on phonological hypotheses regarding the structure of IPs in each language and was done by annotators with expert training in prosody (to varying degrees) and, in the case of NPH and VU, with first-hand knowledge of the languages and their prosodic systems, we decided that the consensus version (CONS) could be treated as the ground truth in the later course of the study, against which the performance of the other annotators can be evaluated.

To conclude this overview of data and methods, we briefly comment on our statistical procedures. Since the task of the annotators was to segment into IPs a given transcription of a recording, which we provided to them in a practical orthography including word boundaries, we can treat the IP segmentation task as a simple binary classification task: Between each consecutive pair of words in the transcription, the annotators can either posit an IP boundary or not. For a transcription containing *n* words, there are (*n* - 1) consecutive word pairs and thus (*n* - 1) potential IP boundaries about which the annotators have to decide.[10] We solely focus here on this binary classification and disregard the exact location in which the annotators put an IP start or end boundary on the ELAN time line.[11]

When evaluating interrater agreement, we cannot simply compare the raw agreement between annotators to a baseline assuming equal probabilities of 0.5 for positing or not positing an IP boundary between two consecutive words. Instead, we have to take into account the fact that there are many more non-boundaries between words than boundaries, that is, a boundary is much less likely than a non-boundary (the average length of IPs in our consensus segmentation (CONS) is 4.26 words, SD = 2.79 words). We therefore use the standard kappa measures of interrater agreement that incorporate information about the relative frequency of the different categories (in our cases, *boundary* vs. *non-boundary*). In order to assess agreement between all annotators, we will use Fleiss' kappa (Fleiss 1971), which can be used to evaluate interrater-reliability between multiple annotators at once. In addition, we will also compare the student annotators' segmentations individually to our consensus segmentation (CONS), which we consider as ground truth, using the more traditional Cohen's kappa (Cohen 1960) for comparisons between two annotators as well as well-known measures from information retrieval, namely, the error rate, precision, recall and f-score (the harmonic mean of precision and recall) (cp. van Rijsbergen 1979).

[10] In practice, annotators occasionally forgot to copy and paste a word from the transcription into the ELAN time line or accidently copied one word twice. For our evaluation, we had to correct these copy-and-paste errors by occasionally adding or deleting a word. This was usually unproblematic because the intended IP boundaries were still clear due to the temporal alignment of the IP segments created by the annotator in ELAN with the audio signal. Moreover, the number of these copy-and-paste errors is relatively low: The sloppiest annotator (R3) made 200 copy-and-paste errors in all, amounting to about 3 errors per recording.

[11] The gestural annotation of the AUVIS corpus, in contrast, which was not used in this study, was evaluated using an interrater agreement metric that incorporates information about exact start and end times and temporal overlap (Holle & Rein 2013).

Where appropriate, we will evaluate differences in interrater agreement between languages as well as the segmentation accuracy of individual annotators on different subsets of the corpus by calculating means and variances of these measures on the basis of the 60 individual recordings in our corpus and by comparing them using non-parametric statistical tests. In most cases, we will use the so-called Wilcoxon-Mann-Whitney rank sum test (Wilcoxon 1945; Mann & Whitney 1947) for unpaired samples. In section 6, however, we chose the so-called Wilcoxon signed-rank test (Wilcoxon 1945) for paired samples to compare one student annotator's segmentation to two different reference segmentations. We assume the conventional significance level of $p \leq 0.05$ throughout.

## 4. Interrater agreement results on the corpus as whole and on individual languages

In this section, we first look at overall agreement between the annotators on the entire corpus in order to assess the validity and reliability of the IP as a universally identifiable unit. Second, we compare the segmentations of individual annotators to our consensus (CONS) segmentation, taking the latter as ground truth, in order to look for possible differences in the segmentation behavior of individual annotators. Third, we compare interrater agreement on individual languages in order to answer the question of whether annotators agree equally on the segmentation of IPs across different languages.

The whole corpus comprises 27,883 words. Since the start of the first IP and the end of the last IP in a narration always coincide with the first and last words and are thus given by definition, we excluded them from the evaluation and therefore have to consider 27,823 potential IP boundaries in all (one less than the number of words for each of the sixty recordings). Table 3 provides an overview of the IP segmentations created by the five annotators (i.e. four students and the authors' consensus version) and shows that the corpus was divided into roughly 6,800 IPs on average, resulting in a mean IP length of about four words.

Table 3: Overview of IP segmentation by annotator

| Annotator | IPs | Mean length of IPs (in words) | Std. dev. of length of IPs (in words) |
|---|---|---|---|
| **R1** | 8,441 | 3.29 | 2.05 |
| **R2** | 7,898 | 3.51 | 2.20 |
| **R3** | 5,159 | 5.35 | 3.84 |
| **R4** | 5,864 | 4.72 | 2.95 |
| **CONS (ground truth)** | 6,499 | 4.26 | 2.79 |
| **(grand) mean** | 6,772 | 4.09 | 2.82 |

Table 4 shows that the annotators, including CONS, exhibit a high agreement with regard to the exact location of IP boundaries. The table provides the number and percentage of cases in which *x* of the five annotators, R1 - R4 and CONS, posit an IP boundary, ranging from 0 for places where no annotator has posited a boundary, to 2 for cases where two annotators have placed a boundary and three annotators have not, to 5 for places where all annotators have assumed an IP boundary. There are 21,574 totally unanimous cases in which all five annotators agree, corresponding to a raw agreement of 77.54%.

Table 4: Overall agreement on the IP segmentation of the whole AUVIS corpus

| | **Number of cases in which *x* annotators posit a boundary** | | | | | | |
|---|---|---|---|---|---|---|---|
| ***x*** | **0** | **1** | **2** | **3** | **4** | **5** | **total** |
| **count** | 17,593 | 2,641 | 1,193 | 1,031 | 1,384 | 3,981 | 27,823 |
| **%** | 63.23% | 9.49% | 4.29% | 3.71% | 4.97% | 14.31% | 100.00% |
| | **all unanimous cases**<br>21,574 | | **raw agreement**<br>77.54% | | **Fleiss' kappa**<br>0.71<br>(z = 375, p < 0.001) | | |

However, as mentioned at the end of section 3, this type of task involved a strong bias for non-boundary decisions. Hence, the relevant baseline with which to compare the raw agreement score of 77.54% is not simply a random assignment of boundaries with a probability of 0.5. Rather, we have to calculate Fleiss' kappa for multiple annotators (Fleiss 1971) in order to assess whether the interrater agreement is robustly above chance level and how much so. On the entire corpus, all five annotators achieved a Fleiss' kappa score of 0.71, which is statistically significantly greater than zero according to the *z*-score and can be regarded as substantial agreement according to an oft-cited table by Landis & Koch (1977). When we only consider the four student annotators R1 - R4 and exclude CONS from the analysis, we also obtain a raw agreement score of 78.21% and a statistically robust and substantial interrater agreement (κ = 0.68, n = 27,823, *z* = 277, p < 0.001). These results on the whole corpus show that recordings of spontaneous speech in different languages can be segmented into IPs reliably even by non-expert annotators without special training.

If we now take our own consensus segmentation (CONS) as the ground truth and compare individual student annotators' segmentations to it, we obtain the results presented in Table 5. The upper part of Table 5 provides true positives (both the student annotator and CONS assume a boundary), false positives (the student annotator assumes a boundary but CONS does not), true negatives (neither the student annotator nor CONS posits a boundary), and false negatives (the student annotator does not posit a boundary but CONS does) as raw numbers, from which the evaluation measures in the lower half of the table are calculated. It is apparent that individual student annotators' segmentations agree quite well with the authors' consensus segmentation with error rates ranging from 10.25% for annotator R1 to 6.06% for annotator R4, f-scores ranging from 79.25% for annotator R3 to 86.24% for annotator R4, and most importantly values of Cohen's kappa statistics (overall) ranging from 0.74 for annotator R3 to 0.82 for annotator R4, all of which are highly statistically significantly above chance (zero).[12] All of the four student annotators are thus able to provide a reliable segmentation into IPs that agrees to a large extent with the expert segmentation created by the authors.

[12] R1 (κ = 0.74, n = 27,823, *z* = 125, p < 0.001), R2 (κ = 0.75, n = 27,823, *z* = 126, p < 0.001), R3 (κ = 0.74, n = 27,823, *z* = 125, p < 0.001), and R4 (κ = 0.82, n = 27,823, *z* = 138, p < 0.001).

Table 5: Comparison of annotators to ground truth on the whole AUVIS corpus

| | **Annotator** | | | |
|---|---|---|---|---|
| **Measure** | **R1** | **R2** | **R3** | **R4** |
| **true positives** | 5,984 | 5,797 | 4,572 | 5,279 |
| **false positives** | 2,397 | 2,041 | 527 | 525 |
| **true negatives** | 18,987 | 19,343 | 20,857 | 20,859 |
| **false negatives** | 455 | 642 | 1,867 | 1,160 |
| **error rate** | 10.25% | 9.64% | 8.60% | 6.06% |
| **precision** | 71.40% | 73.96% | 89.66% | 90.95% |
| **recall** | 92.93% | 90.03% | 71.00% | 81.98% |
| **f-score** | 80.76% | 81.21% | 79.25% | 86.24% |
| **Cohen's kappa (overall)** | 0.7393 | 0.7481 | 0.7392 | 0.8237 |
| **Mean kappa per narration** | 0.7422 | 0.7437 | 0.7381 | 0.8241 |
| **Std. dev. of kappa per narration** | 0.0903 | 0.0711 | 0.0920 | 0.0630 |
| **Coeff. of variation of kappa** | 0.1217 | 0.0956 | 0.1246 | 0.0765 |

Student annotators nonetheless differ amongst each other in their tendency to either assume more or fewer IP boundaries than CONS: R1 and R2 posit relatively many IP boundaries (cf. Table 3) and therefore segment the recordings into relatively short IPs, which results in high recall values above 90% but lower precision values slightly above 70% for these two annotators (cf. Table 5). Annotators R3 and R4, in contrast, assume relatively fewer IP boundaries and therefore longer IPs (cf. Table 3), which is reflected in Table 5 in high precision values of about 90% as well as lower recall values of approx. 71% and 82%, respectively.

The last three rows in Table 5 additionally provide the mean and standard deviation as well as the coefficient of variation of Cohen's kappa for each annotator calculated per narration (that is, the sixty recordings in our corpus; cf. Table 1). These figures are useful to determine whether one or more of the four annotators exhibits a higher interrater agreement with CONS than the others. From the overall Cohen's kappa values in Table 5, it appears that R4 agrees better with CONS than the other three annotators. Non-parametric Wilcoxon-Mann-Whitney tests comparing pairs of annotators based on the sixty Cohen's kappa values for individual recordings confirm this impression.[13] No comparison between the three annotators R1, R2, and R3, in contrast, yields any evidence for statistically significant differences between their mean Cohen's kappa values.[14]

Moreover, R4 also has the lowest coefficient of variation (defined as the standard deviation divided by the mean; cf. the last row of table 5), namely, 0.08 vs. 0.12 for R1, 0.10 for R2, and 0.12 for R3, which means that R4 is also the most consistent of the four annotators with regard to agreement with CONS across all 60 narrations. This is in all likelihood related to the fact that R4 is the only one of the four student annotators who has had an in-depth training in phonetics and prosodic analysis, albeit not specifically for the present study. That is, unsurprisingly perhaps, training and experience in prosodic

[13] R4's mean Cohen's kappa per narration of 0.8241 is statistically significantly higher than R1's of 0.7422 ($W = 864$, $z = -4.91$, $p < 0.001$), significantly higher than R2's kappa of 0.7437 ($W = 651$, $z = -6.03$, $p < 0.001$), and also significantly higher than R3's kappa of 0.7381 ($W = 798$, $z = -5.26$, $p < 0.001$).

[14] R1 vs. R2: $W = 1,894$, $z = 0.49$, $p = 0.624$; R1 vs. R3: $W = 1,760$, $z = -0.21$, $p = 0.836$; R2 vs. R3: $W = 1,970$, $z = 0.89$, $p = 0.374$.

analysis improve the quality and consistency of IP segmentation results and also interrater agreement, at least, when an expert segmentation serves as ground truth.

Still, the overall results indicate that there is a well-above chance agreement between annotators of different levels of expertise in determining IP boundaries in an extensive corpus of spontaneous narrative speech, including recordings in the native language of the annotators as well as in languages unfamiliar to them. This suggests the hypothesis that some boundary cues for IPs, including the ones highlighted in the instructions for the segmentation task (section 2), can be applied reliably and consistently over a range of familiar and unfamiliar languages. To further scrutinize this hypothesis, we now turn our focus to individual languages contained in our corpus and possible differences with regard to the interrater-reliability of IP segmentation on these subcorpora.

Table 6 provides raw agreement figures as well as Fleiss' kappa values for the four larger language subcorpora in our corpus. Interrater agreement turns out to be remarkably similar across these four languages, the value for each language also being similar to the overall Fleiss' kappa score of 0.71 for the whole corpus (Table 4). The highest Fleiss' kappa value is attained on the Yali subcorpus (raw agreement = 77.31%, κ = 0.75, n = 2,001, *z* = 106, p < 0.001), which also happens to be the smallest subcorpus among the four subcorpora. This is followed next by Wooi (raw agreement = 76.81%, κ = 0.74, n = 3,545, *z* = 139, p < 0.001), and then German (raw agreement = 81.71%, κ = 0.72, n = 8,818, *z* = 214, p < 0.001). Interrater agreement on Papuan Malay is somewhat lower (raw agreement = 73.82%, κ = 0.68, n = 10,353, *z* = 219, p < 0.001). The test statistics thus confirm that there is indeed substantial agreement between the five annotators' segmentations of each of these four subcorpora.[15]

[15] The results on the three minor subcorpora in the AUVIS corpus, Cologne German, English, and Waima'a are fully in line with the results for the larger subcorpora: Cologne German (raw agreement = 88.07%, κ = 0.82, n = 285, *z* = 44, p < 0.001), English (raw agreement = 80.45%, κ = 0.72, n = 1,417, *z* = 85, p < 0.001), and Waima'a (raw agreement = 75.85%, κ = 0.67, n = 1,404, *z* = 80, p < 0.001).

Table 6: Interrater agreement for individual languages in the AUVIS corpus

| German | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Number of cases in which *x* annotators posit a boundary** | | | | | | |
| ***x*** | **0** | **1** | **2** | **3** | **4** | **5** | **total** |
| **count** | 6,274 | 602 | 268 | 314 | 429 | 931 | 8,818 |
| **%** | 71.15% | 6.83% | 3.04% | 3.56% | 4.87% | 10.56% | 100.00% |
| | **all unanimous cases** | | **raw agreement** | | **Fleiss' kappa** | | |
| | 7,205 | | 81.71% | | 0.72 (z = 214, p < 0.001) | | |

| Papuan Malay | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Number of cases in which *x* annotators posit a boundary** | | | | | | |
| ***x*** | **0** | **1** | **2** | **3** | **4** | **5** | **total** |
| **count** | 6,033 | 1,201 | 544 | 454 | 511 | 1,610 | 10,353 |
| **%** | 58.27% | 11.60% | 5.25% | 4.39% | 4.94% | 15.55% | 100.00% |
| | **all unanimous cases** | | **raw agreement** | | **Fleiss' kappa** | | |
| | 7,643 | | 73.82% | | 0.68 (z = 219, p < 0.001) | | |

| Wooi | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Number of cases in which *x* annotators posit a boundary** | | | | | | |
| ***x*** | **0** | **1** | **2** | **3** | **4** | **5** | **total** |
| **count** | 2,009 | 411 | 175 | 97 | 139 | 714 | 3,545 |
| **%** | 56.67% | 11.59% | 4.94% | 2.74% | 3.92% | 20.14% | 100.00% |
| | **all unanimous cases** | | **raw agreement** | | **Fleiss' kappa** | | |
| | 2,723 | | 76.81% | | 0.74 (z = 139, p < 0.001) | | |

| Yali | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Number of cases in which *x* annotators posit a boundary** | | | | | | |
| ***x*** | **0** | **1** | **2** | **3** | **4** | **5** | **total** |
| **count** | 1,173 | 204 | 79 | 43 | 128 | 374 | 2,001 |
| **%** | 58.62% | 10.19% | 3.95% | 2.15% | 6.40% | 18.69% | 100.00% |
| | **all unanimous cases** | | **raw agreement** | | **Fleiss' kappa** | | |
| | 1,547 | | 77.31% | | 0.75 (z = 106, p < 0.001) | | |

While there is thus substantial agreement for all subcorpora, the results are not totally even across all four languages. The difference between Papuan Malay and the other languages just noted is statistically significant, whereas there is no evidence for any statistically significant differences amongst German, Wooi, and Yali.[16] These results suggest that the familiar vs. unfamiliar language distinction is not the most important factor in determining interrater agreement. Or, to put this somewhat differently, it does not seem to be necessary to understand spontaneous speech in order to be able to consistently segment it into IPs, provided that the boundary cues are identifiable clearly enough. One question that naturally arises from this state of affairs is what distinguishes the Papuan Malay subcorpus from the other two Papuan subcorpora in this regard. As we will see in section 6, the distribution of pauses – rather than, say, particular characteristics of Papuan Malay phonology – appears to be of major import here.

[16] The by-narration means and standard deviations for the four languages are as follows: German (mean κ = 0.71, SD = 0.04, n = 18), Papuan Malay (mean κ = 0.68, SD = 0.05, n = 20), Wooi (mean κ = 0.73, SD = 0.05, n = 12), and Yali (mean κ = 0.75, SD = 0.04, n = 6). Test results showing a lower interrater agreement on Papuan Malay compared to the three other languages are as follows: German: W = 251, *z* = -2.08, p = 0.038; Wooi: W = 60, *z* = -2.34, p = 0.019; and Yali: W = 21, *z* = -2.37, p = 0.016. This is to be compared to the lack of evidence for differences among the remaining three languages, cp. German vs. Wooi: W = 84, *z* = -1.02, p = 0.325; German vs. Yali: W = 30, *z* = -1.60, p = 0.119; Wooi vs. Yali: W = 29, *z* = -0.66, p = 0.553.

But before we explore this issue more closely, let us see whether the overall agreement values for the individual languages also hold for the individual student annotators as it may very well be the case that the statistical patterns for individual annotators diverge from the overall pattern just discussed. Table 7 gives a basic overview of the number and the average length of IPs occurring in the segmentations by the different annotators on the different languages.

Table 7: Number and mean length (in words) of IPs per annotator and language

| **German** | | | | **Papuan Malay** | | | |
|---|---|---|---|---|---|---|---|
| **Anno-tator** | **IPs** | **Mean length of IPs (in words)** | **Std. dev. of length of IPs (in words)** | **Anno-tator** | **IPs** | **Mean length of IPs (in words)** | **Std. dev. of length of IPs (in words)** |
| **R1** | 2,238 | 3.93 | 2.71 | **R1** | 3,502 | 2.95 | 1.49 |
| **R2** | 1,887 | 4.65 | 2.92 | **R2** | 3,214 | 3.21 | 1.68 |
| **R3** | 1,085 | 8.03 | 4.72 | **R3** | 2,157 | 4.78 | 3.08 |
| **R4** | 1,583 | 5.53 | 3.48 | **R4** | 2,315 | 4.45 | 2.67 |
| **CONS** | 1,748 | 5.02 | 3.27 | **CONS** | 2,657 | 3.88 | 2.36 |
| **mean** | 1,708 | 5.13 | 3.55 | **mean** | 2,769 | 3.73 | 2.33 |

| **Wooi** | | | | **Yali** | | | |
|---|---|---|---|---|---|---|---|
| **Anno-tator** | **IPs** | **Mean length of IPs (in words)** | **Std. dev. of length of IPs (in words)** | **Anno-tator** | **IPs** | **Mean length of IPs (in words)** | **Std. dev. of length of IPs (in words)** |
| **R1** | 1,213 | 2.92 | 1.50 | **R1** | 612 | 3.26 | 2.08 |
| **R2** | 1,289 | 2.74 | 1.45 | **R2** | 711 | 2.81 | 1.61 |
| **R3** | 914 | 3.86 | 2.47 | **R3** | 531 | 3.75 | 2.51 |
| **R4** | 889 | 3.96 | 2.29 | **R4** | 498 | 4.00 | 2.56 |
| **CONS** | 933 | 3.78 | 2.37 | **CONS** | 551 | 3.62 | 2.48 |
| **mean** | 1,048 | 3.37 | 2.07 | **mean** | 581 | 3.44 | 2.27 |

Overall, statistical trends in Table 7 are again surprisingly similar to those in Table 3 for the whole corpus. CONS and R4 again posit a similar number of IPs and accordingly produce quite similar mean lengths of IPs also for the four individual subcorpora shown in Table 7.[17] R1 and R2 again tend to segment the recordings into shorter units compared to the other annotators. There are thus individual differences in annotator behavior that are valid across the different languages in the corpus. This may be interpreted as a first indication that overall segmentation strategies are indeed similar across the four languages.

That this is not necessarily so, however, is shown by student annotator R3. This student segments the German narrations, which she is able to understand, into very long IPs with an average length of more than eight words.[18] Boundaries here are preferably placed at clause boundaries,[19] ignoring the fact that clauses in spontaneous speech often are chunked into a number of different IPs. Example (1)

[17] In fact, of all four student annotators, R4 is also closest in behavior to CONS for the three remaining languages Cologne German, English, and Waima'a, the minor subcorpora in the AUVIS corpus.

[18] R3 also has the longest mean length of IPs in the other two languages she understands, namely, Cologne German and English. For Austronesian Waima'a, in contrast, R3 exhibits a mean length of IPs close to the overall average.

[19] While we have not investigated this systematically across the whole German subcorpus, close inspection of a number of segments drawn from different parts of it strongly suggest that it is indeed clause and sentence boundaries that R3 is orienting towards rather than the end of declination units, which occasionally, but not systematically, overlap with sentence boundaries.

illustrates a typical case where R3 fails to identify four IP boundaries in succession before she posits a boundary in agreement with the other raters right at the end of the whole clause after *entgegen*. Here, and in the following examples, each boundary posited by a student annotator is indicated with a grey pipe, boundaries in the consensus version are marked with a black pipe (the boundaries after *Mädchen* and after *auf der* are from R1 and not from R3).[20]

(1) *dann kam ihm <ein ->* (0.2) ||||
then came him(DAT) a

*ein dickes Mädchen* | *mit langen Zöpfen* ||||
a fat girl | with long pigtails ||||

*auf einem anderen Fahrrad* ||| *<auf der ->* | *<auf einer ->* |||
on a other bicycle ||| on the | on a |||

*auf der staubigen Landstraße entgegen* |||||
on the dusty country.road toward |||||
'Then a fat girl with long pigtails came riding on another bicycle towards him on the dusty country road' (DEU_pear_Alex)

In contrast, R3 is closer in segmentation behavior to the other annotators, and specifically to R4 and CONS, with regard to the three Papuan languages she is unfamiliar with. This difference suggests that R3 used different segmentation strategies in familiar vs. unfamiliar languages. Importantly, segmentation in the familiar languages more strongly takes into account non-prosodic factors, while segmentation in the unfamiliar languages – by necessity – has to rely exclusively on prosodic cues. For the other student annotators, the data in Table 7 suggest that they are more consistent in using primarily prosodic cues in segmenting familiar and unfamiliar languages. This observation supports the view mentioned in section 1 that the inclusion of non-prosodic factors in IP segmentation increases the potential for disagreements, since at least some syntactic boundaries such as sentence boundaries do not systematically correlate with IP boundaries. While sentence boundaries typically are also IP boundaries, the reverse does not hold. This is especially clear in narrative speech, where long strings of syntactically coordinated constructions (of the type: *and then … and … and …*) may occur.

The data in Table 7, however, are merely indicative with regard to the consistency of performance of each annotator across the four main languages under investigation. One way to further investigate this issue is to compare the performance of each student annotator for each language with the consensus version, as detailed in Table 8. The main observations emerging from this table are as follows:

- The Cohen's kappa values for each annotator/language pair are all well within the substantial agreement range, the lowest values being κ = 0.68 for R1/Papuan Malay and κ = 0.69 for R3/German.

[20] Further conventions in the examples: pause length is given in ( ); < > surround false starts. As noted in section 3, pauses and false start were not marked as such in the transcripts given to the student annotators. Glosses for grammatical categories: DAT – dative, DET – determiner, FILL – filler, PART – particle, PL – plural, POSS – possessive, PRTC – participle, Q – question marker, REL – relative marker, SG – singular, and TOP – topic marker.

As just noted, R3 tends to recognize prosodic boundaries in German primarily at major syntactic boundaries.

- R1 and R4 both achieve their lowest kappa values for Papuan Malay;[21] and for R3, the Papuan Malay results are also not significantly better than R3's worst results for German.[22] Likewise, while R2 has a slightly better mean kappa value for Papuan Malay than for one other language (Wooi), the difference between these two languages is not statistically significant.[23] The overall worse results for Papuan Malay are thus not due to particularly bad performances of one or two student annotators. Instead, all student annotators appear to have had particular difficulty with this subcorpus.
- With regard to the best results, the picture is somewhat more varied with most student annotators performing almost equally well for two languages: R1 has best results for German and Yali, R2 for German and Yali, R3 for Wooi and Yali, and R4 for German and Wooi. That is, while performance for the native language is among the best for three out of four student annotators, they all do (almost) equally well with regard to at least one other, unfamiliar language in the corpus.[24] Although this does not exactly replicate the simple bipartition (Papuan Malay vs. all other languages) we found across all annotators, it confirms the conclusion we drew from the data in Table 6 that the familiar vs. unfamiliar distinction is not the most important factor in determining interrater agreement.
- R4 clearly outperforms the other annotators for German and Papuan Malay and also achieves the best results for Wooi (although R3's results for Wooi are not statistically significantly worse). For Yali, she does not score the highest mean kappa value but her results are not statistically significantly worse than those achieved by R1 and R3. [25] These findings confirm her status as a more experienced annotator.
- The range of agreement values varies less than 0.20 points between the annotators for every language, with 0.08 for Yali being the lowest range, and 0.18 for German the highest.
- The range of agreement values for each annotator across all languages is also quite limited, with 0.08 being the range for both R2 and R4, and R3's 0.15 being the being the maximum range.

[21] R1's results on Papuan Malay are, however, only marginally significantly worse than those on Wooi: $W = 73$, $z = -1.83$, $p = 0.070$.

[22] Papuan Malay vs. German for R3: $W = 166$, $z = 0.40$, $p = 0.696$.

[23] Papuan Malay vs. Wooi for R2: $W = 156$, $z = 1.40$, $p = 0.170$.

[24] R1: Yali vs. German ($W = 39$, $z = 1.00$, $p = 0.343$) but note that in this case Yali is also not significantly better than Wooi ($W = 18$, $z = 1.69$, $p = 0.103$). R2: German vs. Yali ($W = 68$, $z = 0.93$, $p = 0.378$), German vs. Papuan Malay ($W = 267$, $z = 2.54$, $p = 0.010$). R3: Wooi vs. Yali ($W = 36$, $z = 0.00$, $p = 1.0$), Wooi vs. Papuan Malay ($W = 26$, $z = 3.66$, $p < 0.001$). R4: German vs. Wooi ($W = 145$, $z = 1.57$, $p = 0.124$), German vs. Yali ($W = 97$, $z = 2.87$, $p = 0.003$).

[25] German: R4 vs. R1 ($W = 57$, $z = 3.32$, $p < 0.001$). Papuan Malay: R4 vs. R2 ($W = 103$, $z = 2.62$, $p = 0.008$). Wooi: R4 vs. R3 ($W = 72$, $z = 0.00$, $p = 1.0$). Yali: R4 vs. R3 ($W = 27$, $z = -1.44$, $p = 0.180$), R4 vs. R1 ($W = 26$, $z = -1.28$, $p = 0.240$).

Table 8: Comparison of annotators to ground truth (CONS) for individual languages

| LANGUAGE | MEASURE | R1 | R2 | R3 | R4 |
|---|---|---|---|---|---|
| German | Cohen's kappa (overall) | 0.7967 | 0.7763 | 0.6960 | 0.8728 |
| | Mean kappa per narration | 0.7926 | 0.7752 | 0.6897 | 0.8641 |
| Papuan Malay | Cohen's kappa (overall) | 0.6813 | 0.7348 | 0.7064 | 0.7762 |
| | Mean kappa per narration | 0.6814 | 0.7369 | 0.7016 | 0.7818 |
| Wooi | Cohen's kappa (overall) | 0.7423 | 0.7080 | 0.8472 | 0.8525 |
| | Mean kappa per narration | 0.7335 | 0.6984 | 0.8299 | 0.8456 |
| Yali | Cohen's kappa (overall) | 0.8063 | 0.7575 | 0.8444 | 0.8062 |
| | Mean kappa per narration | 0.8148 | 0.7601 | 0.8378 | 0.7928 |

The results presented so far show robust interrater agreement with regard to the identification of IP boundaries across the whole corpus as well as for individual subcorpora. This finding, in turn, supports the hypothesis that it is possible to make consistent use of the melodic and rhythmic boundary cues focused on in our instructions (cp. section 2) across familiar and unfamiliar languages. Still, limited variation occurs, both across annotators and languages. At this point, it is not clear what factors cause this variation, with one exception: One student annotator (R3) appears to have paid more attention to semanto-syntactic cues than to prosodic cues in segmenting her native German, causing her segmentations for German, but not for the other languages, to diverge significantly from all other segmentations. The next section will take a closer look at other possible factors causing interrater disagreements in order to see whether the observed variation can be systematically accounted for.

## 5. Disagreements

This section investigates the factors rendering the identification of IP boundaries difficult and thus contributing to disagreements between annotators in tasks such as the one reported on here. Given the hypothesis that the boundary cues work the same across different languages and can be identified with the same degree of consistency across familiar and unfamiliar languages, we may expect that the disagreement patterns are largely similar across the investigated languages. This expectation presupposes that there are in fact disagreement *patterns*, i.e. that the lack of agreement between raters can be accounted for systematically. If so, cases of disagreement should be amenable to a smallish number of factors. Alternatively, cases of disagreement could be random, possibly involving a multitude of overlapping factors with no clear patterns emerging across the corpus.

As discussed in more detail in section 2, melodic coherence probably presents most difficulties for consistent identification, and pauses, while relatively easy to identify in principle, are ambivalent as boundary cues given the necessity to distinguish between IP internal and external pauses. Based on these observations, we may expect disagreements to arise in particular in the following two constellations. First, the boundary between two IPs is marked by melodic cues only. Second, disagreements may arise due to the ambivalent status of pauses (is the pause IP-internal or IP-external?). This section, then, is concerned with the following two questions:

1) Can the disagreements occurring in our segmentation task be accounted for systematically in terms of a small number of disagreement categories, specifically ones relating to difficulties in identifying the endpoint of a melodic contour and to the ambivalence of pauses?
2) Do we find the same disagreement categories across different languages or are some types of disagreement specific to a given language?

In order to investigate these questions, we looked at all instances of disagreements between the consensus version and the versions produced by the student annotators, and devised a classification scheme for them. The basic distinction here was a purely formal one: One or more student annotators posited a boundary where the consensus version has none, or vice versa. The former we call false positives. Conversely, false negatives are instances where a student annotator did not perceive a boundary, but the consensus version has one. Furthermore, a number of boundary cue constellations were defined which provide the likely reason for a disagreement to occur (e.g. dysfluency = IP-internal pause not recognized as such; full exemplification below). In a second step, all instances of disagreements were coded for one of these categories by one of the authors (VU). Finally, all codings were cursorily checked by the first author and all unclear cases were discussed and jointly settled in the group of all authors.

In line with our expectations, the disagreements occurring in our segmentation task can be roughly allocated to two factors: IP-boundaries based primarily on melodic cues and ambivalence of pauses. Each factor may lead to false positives and false negatives, resulting in five basic types of disagreements, which we now exemplify in detail. In addition, there is a residue of disagreements (< 15%) which do not fit into these five types and which we will also briefly illustrate below.

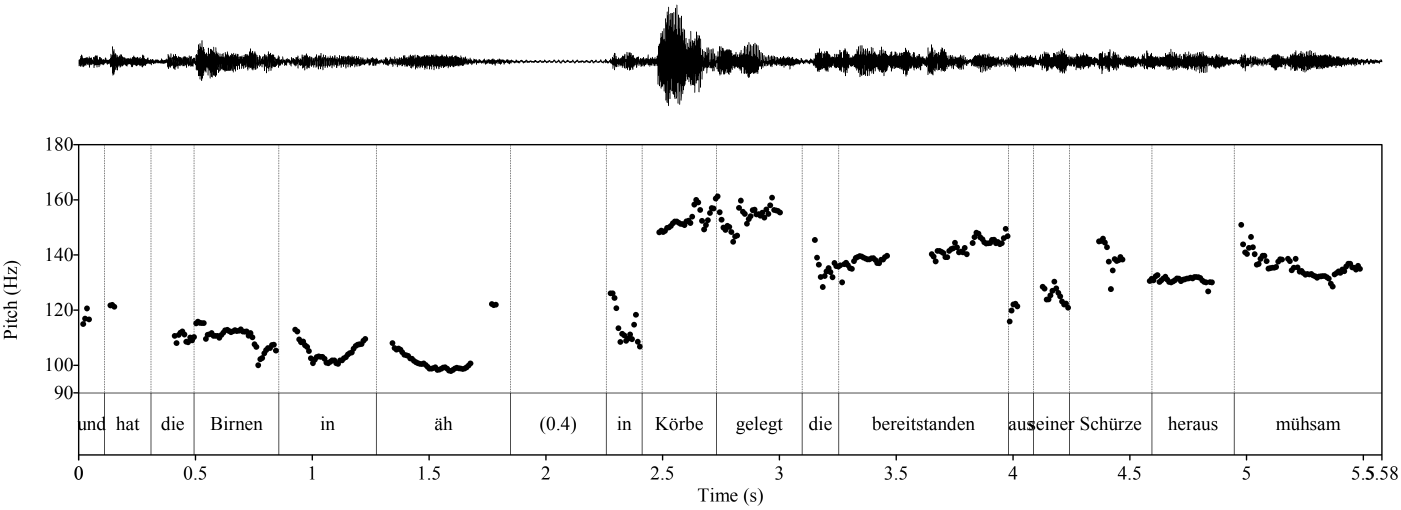
Turning to IP-boundaries indicated primarily by melodic cues, false negatives here occur in instances where the speaker produces a sequence of two or more IPs in rapid succession without intervening pauses, which we call *latching*. Example (2) from German provides an illustration (the first boundary after *äh* is dysfluency-related and not relevant at this point) where latching occurs in three IPs in a row. The main indication for IP boundaries here are pitch jumps leading to an interruption of the contour, downward after *gelegt* and *bereitstanden*, upward after *heraus* (only one student annotator here fully agrees with the consensus version).

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (2) | *und* | *hat* | *die* | *Birnen* | *<in ->* | *äh* | (0.4) | \| | *in* | *Körbe* | *gelegt* | \|\|\| |
| | and | have:3SG | the | pears | in | um | | | in | baskets | put:PRTC | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *die* | *bereitstanden* | \|\|\| | *aus* | *seiner* | *Schürze* | *heraus* | \|\| | *mühsam* | \|\|\|\|\| |
| that | stand.by:3PL | | out.of | his | apron | out | | strenuous | |

‘and (he) put the pears into uhm, into baskets, that stood there, from out of his apron, strenuously.’
(DEU_pear_Flor)

Figure 1: Waveform and $f_0$ extraction of example (2)



The converse of latching are false positives following major pitch changes. Most frequently, the pitch change in question consists in a clear rise, and a boundary would be posited by a student annotator after the word bearing the rise, where the consensus version has none. More rarely, this occurs after a fall in pitch. In both instances, it seems likely that the pitch change is heard as constituting the end of a coherent melody. Importantly, these false positives occur at places where rhythmic boundary cues are missing (in particular, there are no pauses). We therefore consider them examples of the fragility of melodic cues in a perhaps somewhat surprising guise: insecurity on the part of an annotator with regard to the presence or otherwise of a melodic boundary cue may not only lead to missing a boundary only signaled by melodic cues (the latching instances just discussed), but it may also lead to a profusion of boundaries at melodically prominent places, just to be on the safe side, as it were.

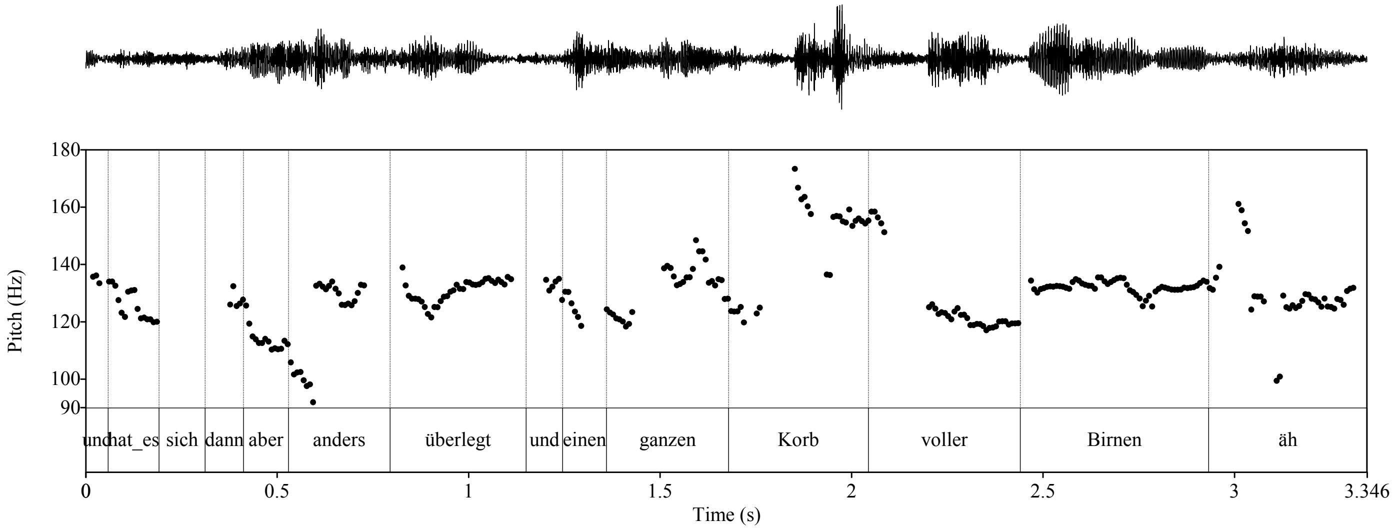
To illustrate this disagreement type in more technical terms: In a language with (post-lexical) pitch accents such as German, the pitch changes in question mark pitch accents and are probably misinterpreted by the student annotators as boundary tones (or combinations of pitch accents and boundary tones), and hence as the end of a melodic contour. Consider (3) where the narrator emphasizes the fact that after first having taken only a few pears, the protagonist decides to take a complete basket full of pears, realizing a very noticeable high target on *Korb*, as shown in Figure 2. Note that there is no rhythmic interruption after *Korb*, and no interruption of the coherent pitch contour starting with *und* and ending with *Birnen*. In this example, everyone agrees on the boundaries after *überlegt* and the hesitation particle *äh*, but one student annotator additionally posits a boundary after *Korb* and after *Birnen* (the latter illustrates a dysfluency-related disagreement further discussed below).

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (3) | *hat* | *es* | *sich* | *dann* | *aber* | *anders* | *überlegt* | ||||| |
| | have:3SG | it | himself | then | however | differently | consider:PRTC | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *und* | *einen* | *ganzen* | *Korb* | \| | *voller* | *Birnen* | \| | *äh* | ||||| |
| and | a | whole | basket | | full.of | pears | | um | |

‘but then (he) decided differently and (took) an entire basket full of pears um’ (DEU_pear_Flor)

Figure 2: Waveform and $f_0$ extraction of example (3)

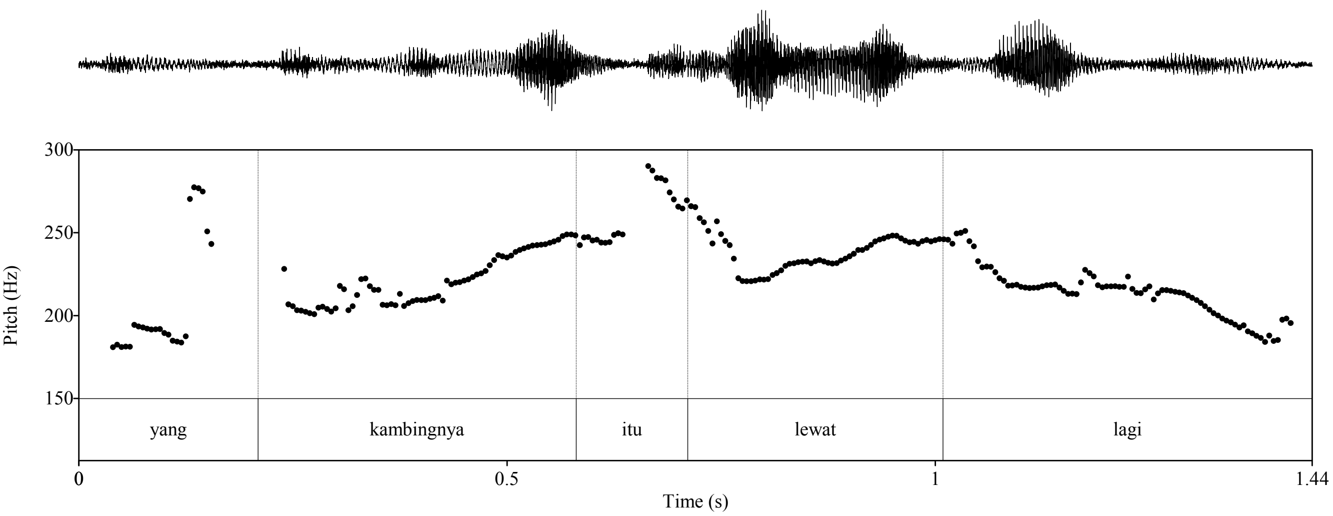


The languages from West Papua do not make use of post-lexical pitch accents, but they make use of rising pitch to mark boundaries of intermediate phrases (phrases larger than phonological words but smaller than IPs). Such intermediate phrase boundaries are generally not accompanied by major rhythmic breaks nor is the overall pitch contour interrupted.[26] A very widespread location for such boundaries to occur is at the end (on the final syllable, to be precise) of an initial topic constituent, as in (4) from Papuan Malay, where the final syllable of the demonstrative *itu* carries a high tone, followed by a fall in pitch.

(4) *yang* *kambing-nya* *itu* | *lewat* *lagi* |||||
REL goat-3SG.POSS that pass_by again
‘(but actually) that one goat, it moved on’ (PMY_pear_Miry)

[26] Cp. Himmelmann (2010) for details in Waima'a. This is areally a widespread phenomenon as shown in particular by Stoel (2005, 2006), see also the contributions in van Heuven & van Zanten (2007).

Figure 3: Waveform and $f_0$ extraction of example (4)



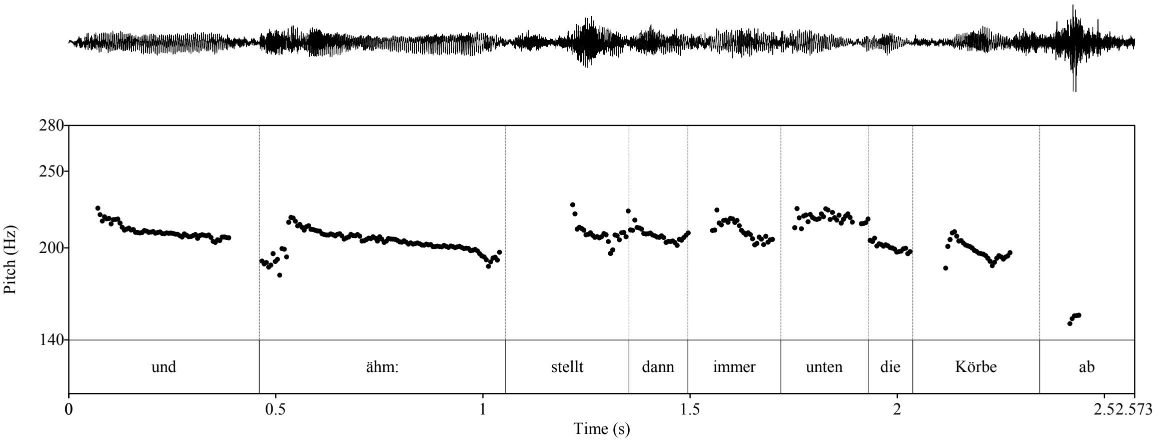
Turning to the other factor giving rise to disagreements, ambivalence of pauses, the large majority of instances in this category relates to dysfluencies. Dysfluencies are a somewhat special case, because one could argue that they are inherently ambiguous with regard to the boundary issue, as the speaker does not properly deliver an IP already in production, and either interrupts or abandons it. Consequently, dysfluencies could be handled by a convention, stipulating that all instances of dysfluency either always or never induce a boundary. While in our instructions we drew attention to the problem of IP-internal dysfluencies, we did not propose conventions for handling these instances, as these would have required major training efforts to be useful.

It is apparent in our data that student annotators used differing strategies in dealing with dysfluencies. Some tended to posit a boundary whenever a dysfluency occurred; others tended to consider the dysfluency to be part of a larger IP and rarely posited a boundary before or after it. Both strategies were not followed consistently, though. The length of the interruption clearly is an additional factor. The longer the interruption, the more likely all annotators posited a boundary. Conversely, false starts involving minimal interruptions (no silences, no *uhm*, etc.) were generally considered to form a single IP with the continuation.

As for the consensus version, we tried consistently to distinguish between hesitations (IP-internal dysfluencies) and truncations, i.e. the abandonment of a unit currently under way. This distinction is primarily based on pitch evidence, but the length of interruption also plays a role. Interruptions lasting more than one second were generally considered truncations. Otherwise, a dysfluency was considered to be IP-internal, if speech delivery was resumed after the dysfluency on (more or less) exactly the same pitch level that was reached before. The idea here is that if one were to cut out the dysfluency, the IP would display an overall coherent intonation contour, hence making it likely that the speaker actually continues with the delivery of an IP begun before the dysfluency. This is illustrated by example (5) where the $f_0$ extraction in Figure 4 clearly shows that the pitch on *stellt* continues almost exactly on the same level as it was on *und* right before the intervening hesitation.

(5) *und* | *ähm:* || *stellt* *dann* *immer* *unten* *die* *Körbe* *ab* |||||
and | umh || put:3SG then each.time down the baskets PRTC |||||
'and then (he) puts the baskets down each time (a basket is full)' (DEU_pear_Nele)

Figure 4: Waveform and $f_0$ extraction of example (5)

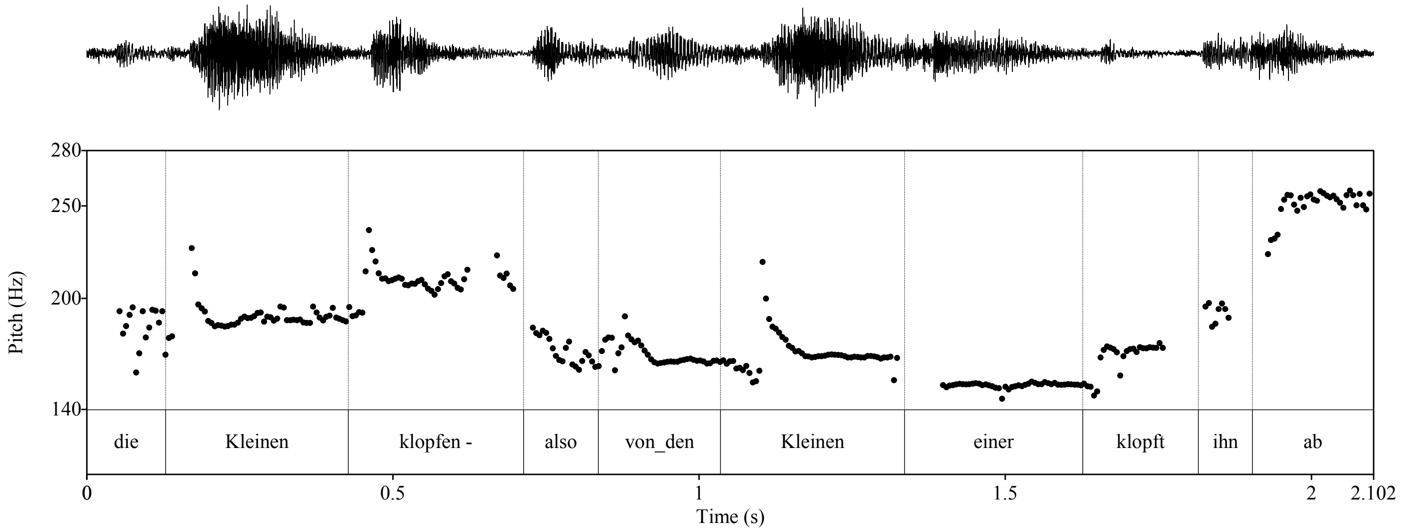


As mentioned above, student annotators applied different strategies in such instances, though never fully consistently. In the German recordings, for example, the initial hesitation pattern CONJUNCTION – FILLER – CLAUSE occurs very frequently. For this pattern, R1 tends to posit a boundary both before and after the filler, while the other student annotators often treat the filler as IP-internal or consider CONJUNCTION + FILLER to make up one IP of its own.

In instances of truncations, on the other hand, there is clear evidence for the start of a new IP, as seen for example in (6). Here the speaker aborts the utterance early on right after *klopfen* in order to make clear that only one of three children is dusting off the boy who has fallen from his bike. The truncation is clearly cued by the jump in pitch contour between *klopfen* and *also*, and the following acceleration in speech rate. In this case, the discourse particle *also* further makes explicit that the speaker has interrupted the delivery of a unit under way and initiates a repair. In this example, three of the four student annotators agree with the consensus version, probably because the overall syntactic construction is also repaired and started anew.

(6) *die* *Kleinen* *klopfen* - |||| *also* *von* *den* *Kleinen* *einer* *klopft* *ihn* *ab* |||||
the small_ones tap:3PL |||| well of the small_ones one tap:3SG him off |||||
'the small ones dust – well, one of the small ones dusts him off' (DEU_pear_Alex)

Figure 5: Waveform and $f_0$ extraction of example (6)



While there are many instances where the distinction between a hesitation and a truncation is reasonably clear, it should be noted that the distinction is to some degree arbitrary in that, for example, it would be difficult to give a principled reason for the decision to set the maximal length of IP-internal pauses at exactly one second, rather than, say, 0.9 or 1.2 seconds. Consequently, without a detailed protocol and some training, it is to be expected that dysfluencies give rise to a considerable number of disagreements in segmentation tasks, specifically false positives in the case of hesitations and false negatives in the case of truncations.

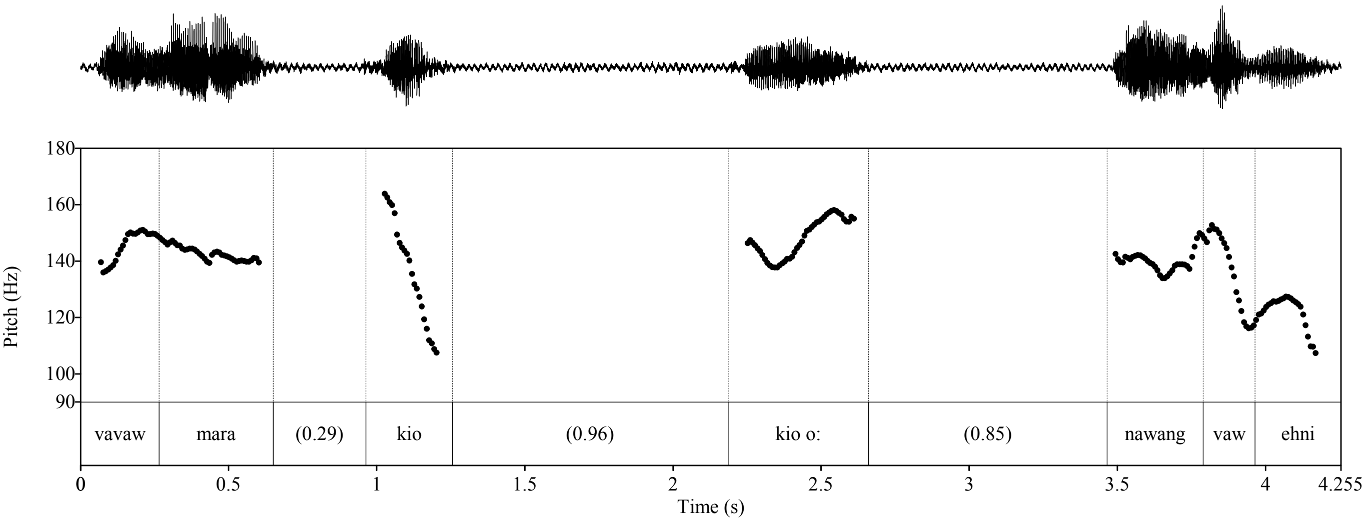
Note that there is a second type of false negatives associated with the ambivalence of pauses, i.e. cases where a student annotator appears to have interpreted as a hesitation pause what the consensus version considers a planning pause (the converse of dysfluency-related false positives). The Wooi example in (7) shows several instances of differing pause interpretation. Here we have a total of four IPs all separated from one another by pauses – a short one and two longer ones. All annotators agree that the pause between the two instances of *kio* is indeed a planning pause occurring at an IP boundary, but the other two pauses are interpreted differently. R2 has a boundary in both instances (in agreement with the consensus version), R4 in neither. R1 and R3 posit a boundary at the third pause, but consider the first, comparatively short pause a hesitation pause.

| (7) | *vavaw* | *mara* | (0.3) | \|\| | *kio* | (1.0) | \|\|\|\|\| |
|---|---|---|---|---|---|---|---|
| | *vavaw* | *mara* | | | *<i>ko* | | |
| | DET.PL | TOP | | | <3SG>take | | |

| *kio* | *o:* | (0.9) | \|\|\|\| | *nawang* | *vaw* | *ehni* | \|\|\|\|\| |
|---|---|---|---|---|---|---|---|
| *<i>ko* | *o:* | | | *nawang* | *vaw* | *ehni* | |
| <3SG>take | FILL | | | basket | DET.PL | one | |

‘so, he took, he took uh, one of the baskets’ (WBW_pear_Davi)

Figure 6: Waveform and $f_0$ extraction of example (7)



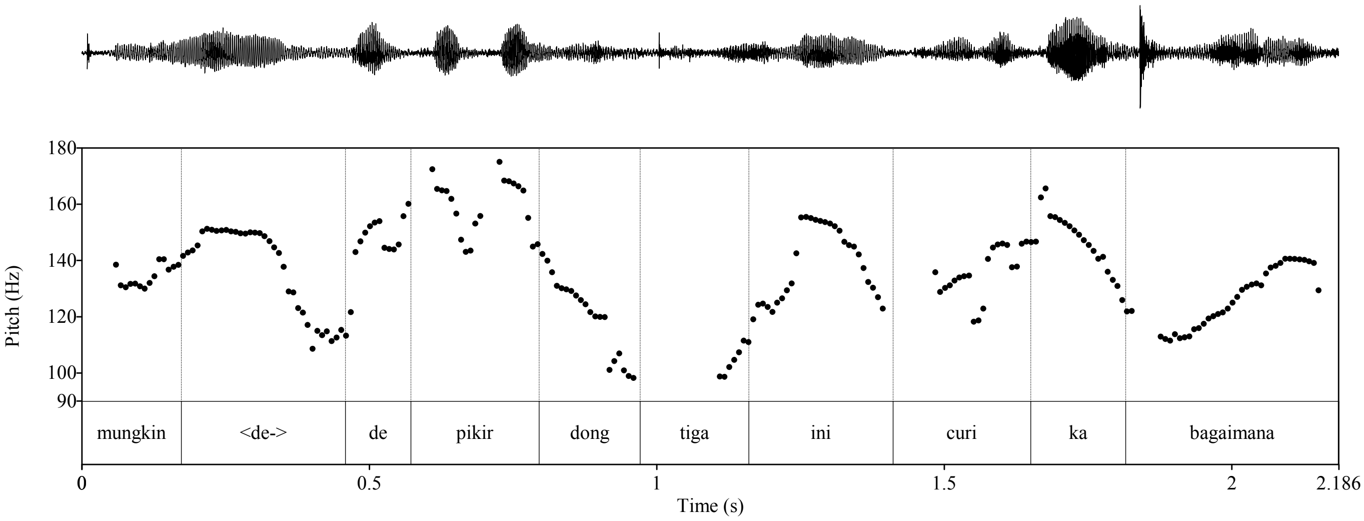
This concludes the review of the main five types of disagreements occurring in our data. There is one more type, which is completely unrelated to the nature of IPs, but which needs brief mention for the sake of completeness. We have called this type the mechanical cases, because the reason for the disagreement is essentially of a technical nature, extraneous to the issue under investigation. In these instances, the transcript provided to the student annotators and the audio recording did not match, which led to confusions as to where exactly one should place a boundary. This could be due to a simple transcription error (word missing or misspelled in the transcript) or to an unclear acoustic signal (unit-final devoicing, slurred articulation, or very high speech rate). Example (8) illustrates a transcription error from the Papuan Malay data. The transcript contains the word *yang* which is not present in the audio (and does not make sense in this context). While all annotators agreed that there is a boundary between the audible strings *ini* and *curi*, two of them decided to put it after *yang*, thereby producing a disagreement where strictly speaking there is none. Note that the large majority of mechanical cases are similar to this example in that they involve instances where all annotators posit a boundary but 'disagree' on the exact location of it.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (8) | *mungkin <de->* | ‖ | *de* | *pikir* | ⦀ | *dong* | *tiga* | *ini* | ‖❙ |
| | perhaps | ‖ | 3SG | think | ⦀ | 3PL | three | this | ‖❙ |

| | | | | | | |
|---|---|---|---|---|---|---|
| ***yang*** | ‖ | *curi* | *ka* | ❘ | *bagaimana* | ‖❙ |
| REL | ‖ | steal | Q | ❘ | how | ‖❙ |

'Maybe he thought these three had stolen (his fruits), right?' (PMY_pear_Titi)

Figure 7: Waveform and $f_0$ extraction of example (8)



As for unclear acoustic signals, it is a characteristic of many languages in West Papua that clause-final syllables, especially when they represent highly predictable clause or phrase-final particles, are regularly devoiced and thus barely audible, as in example (9) from Wooi. Although the final *pa* is not visible in either waveform or F0 extraction (cp. Figure 8), the initial closure of the bilabial stop in *pa* is still audible (though only barely so), and also visible in the accompanying video recording. Native speakers perceive and transcribe such devoiced clause-final particles without hesitation. But, of course, non-native listeners have difficulties to perceive such elements and may become confused about where to place an IP boundary, especially when the segmental string at the beginning of the next IP is not very clear. At the end of example (9), all annotators perceive a boundary. But two of the student annotators place it before the barely audible *pa*, while the others put it after it.

| (9) | *toru* | \|\|\|\|**\|** | *sebenarnya* | *toru* | \|\|\| | *mana* | *hanya* | \|\| |
|---|---|---|---|---|---|---|---|---|
|  | three |  | actually | three |  | but | only |  |

| *koru* | *ay* | *tura* | \|\| | *pa* | \|\|**\|** |
|---|---|---|---|---|---|
| two | exist:PL | stand |  | PART |  |

‘three – actually there were three (baskets) but only two were standing there’ (WBW_pear_Yuli)

Figure 8: Waveform and $f_0$ extraction of example (9)

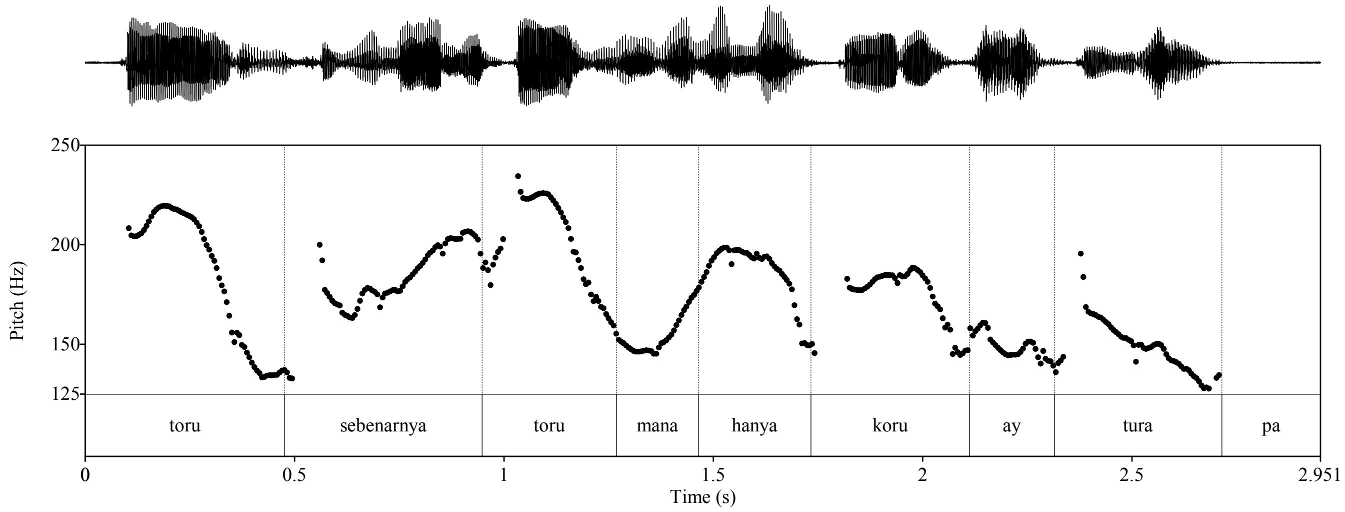


Table 9 provides an overview of the frequency of disagreements among student annotators and the consensus version in the entire corpus. The data show that roughly in one fifth (22.46%) of all boundary decisions a disagreement occurred (rightmost column of Table 9), meaning that at least one student annotator made a decision that differed from the consensus version. But the table also clearly shows that disagreements do not occur randomly. Rather, they can be classified into the relatively small number of types to be expected from the nature and interaction of the boundary cues used in the segmentation task. There is a residue of 562 instances (= 9% of all controversial boundaries) where it is unclear to us why the disagreements occur. These are simply labeled *UNCLEAR* in the final row of Table 9. Given the substantial amount of data to be segmented (> 3 hours of spontaneous narratives altogether), it is likely that a fair amount of these disagreements is simply due to oversights.

Table 9: Overview of types of disagreement in the entire corpus

| **Type of disagreement** | **Number of cases** | **% of controversial boundaries** | **% of potential boundaries** |
|---|---|---|---|
| **MELODIC CUE ONLY** | **3,472** | **55.56** | **12.48** |
| - latching (false negative) | 1,333 | 21.33 | 4.79 |
| - pitch change (false positive) | 2,139 | 34.23 | 7.69 |
| **AMBIVALENCE OF PAUSES** | **1,933** | **30.94** | **6.95** |
| - hesitation (false positive) | 1,109 | 17.75 | 3.99 |
| - truncation (false negative) | 251 | 4.02 | 0.90 |
| - planned pause (false negative) | 573 | 9.17 | 2.06 |
| **MECHANICAL CASES**[27] | **282** | **4.51** | **1.01** |
| **UNCLEAR** | **562** | **8.99** | **2.02** |
| **Total** | **6,249** | **100.00** | **22.46**<br>(of 27,823) |

[27] Note that strictly speaking, the boundaries involving mechanical cases of disagreement could, and perhaps should, have been removed from the data used in our statistical analyses. However, their number is so small

The data in Table 9 also confirm the prediction that disagreements are prone to arise in particular in those instances where melodic and rhythmic boundary cues are not well synchronized. Only the disagreement type *planned pause*, which accounts for only 9.17% of all controversial boundaries, involves boundaries where, according to the consensus version, both melodic and rhythmic cues clearly indicate a boundary (cp. example (7) above). The instances categorized as *melodic cue only* do not involve pauses, and hesitations and truncations are dysfluencies where speakers do not manage properly to deliver an IP and where melodic and rhythmic cues, while present, are consequently not properly aligned.

Having answered the first question posed in the introduction to this section in the affirmative, we now turn to our second question: Do we find the same disagreement categories across different languages or are some types of disagreement specific to a given language? Table 10 lists the total occurrences of our main disagreement types for our four main languages/subcorpora and the percentage of disagreements per type for each language.[28]

Table 10: Distribution of disagreement types over languages / subcorpora

| | **Languages / Subcorpora** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Type of disagreement** | **German** | | **Papuan Malay** | | **Wooi** | | **Yali** | |
| **MELODIC CUE ONLY** | **711** | **45.34%** | **1,761** | **67.39%** | **535** | **71.24%** | **217** | **50.35%** |
| - latching (false negative) | 441 | 28.12% | 652 | 24.95% | 73 | 9.72% | 71 | 16.47% |
| - pitch change (false positive) | 270 | 17.22% | 1,109 | 42.44% | 462 | 61.52% | 146 | 33.88% |
| **AMBIVALENCE OF PAUSES** | **749** | **47.77%** | **619** | **23.69%** | **125** | **16.64%** | **139** | **32.25%** |
| - hesitation (false positive) | 445 | 28.38% | 371 | 14.20% | 59 | 7.86% | 66 | 15.31% |
| - truncation (false negative) | 82 | 5.23% | 103 | 3.94% | 14 | 1.86% | 16 | 3.71% |
| - planned pause (false negative) | 222 | 14.16% | 145 | 5.55% | 52 | 6.92% | 57 | 13.23% |
| **UNCLEAR** | **108** | **6.89%** | **233** | **8.92%** | **91** | **12.12%** | **75** | **17.40%** |
| **Total** | **1,568** | **100.00%** | **2,613** | **100.00%** | **751** | **100.00%** | **431** | **100.00 %** |
| **MECHANICAL CASES (excluded)** | 45 | (2.79%) | 97 | (3.58%) | 71 | (8.64%) | 23 | (5.07%) |

Table 10 shows that all disagreement types are attested in all languages. Hence they are not language-specific. There are some potentially significant differences with regard to the frequencies with which each type occurs in a given subcorpus. Yet these differences are difficult to evaluate, as they in part depend on the overall frequency of the constellation of boundary cues that may give rise to a

(1.01 % of all potential boundaries) that their removal would not significantly have affected our overall results, other than slightly increasing interrater agreement rates. In the remainder of this section, however, we exclude the mechanical cases from further considerations for the reasons given in the next footnote.

[28] Mechanical cases are not included in this and the following table because the significant differences here are not relevant for the point under discussion, but would impact on the overall percentages for the other categories. Specifically, the relatively high number of mechanical cases in the Wooi subcorpus is very likely due to the strong tendency in this language to devoice clause-final particles, as illustrated with example (9) above. On the Wooi subcorpus, comparison between annotators yielded a significantly higher proportion of mechanical cases of disagreement than on the German, Papuan Malay and Yali subcorpora taken together ($\chi^2 = 46.67$, df = 1, $p < 0.001$).

disagreement. Thus, for example, if a subcorpus only contains a few examples of latching, then chances for latching-related disagreements to arise is much lower than for a subcorpus where latching abounds. The same holds for IP-internal pauses (hesitations) and truncations. As we will see in the following section, there are indeed significant differences in the distribution of pauses across the four corpora.

In one instance, however, it would appear to be warranted to draw some conclusions based on the observed frequency differences. This is the case of the significantly lower number of disagreements of the pitch-change type in the German subcorpus (17.22%) compared to all the Papuan languages (lowest percentage for Yali of 33.88%).[29] Strictly speaking, in order to ascertain whether there is indeed a significant difference here, we would have to be able to provide an operational definition of "major pitch change" across all corpora, determine the absolute number of such changes, and then see whether disagreements arise more often in the Papuan subcorpora relative to the total number of constellations with the potential for disagreement. Nevertheless, given the substantial differences in percentages and the fact that, impressionistically speaking, pitch in the German subcorpus is as varied as in the Papuan subcorpora, it may not be too speculative to claim that this difference is probably due to the familiar vs. unfamiliar language distinction. That is, pitch changes in German were less likely to be misinterpreted as boundary signals because of the student annotators' familiarity with the language. This conclusion is further corroborated by the fact that the results on the smaller subcorpora of Cologne German and English, both languages that the annotators were able to understand well, clearly pattern with those on the German subcorpus: 17.65% of disagreements were of the pitch-change type on the Cologne German subcorpus and 18.59% on the English subcorpus, percentages that are remarkably similar to the 17.22% on the German subcorpus in Table 10. In contrast, the figure of 31.89% pitch-related disagreements on the other smaller subcorpus Waima'a is higher and similar to the results obtained for the West Papuan languages.

Importantly, this difference between familiar and unfamiliar languages does not extent to all controversial boundaries which are indicated by melodic cues only. The other subtype in this category, i.e. latching, does not show a clear influence of the familiar vs. unfamiliar divide. Rather, boundaries indicated by melodic boundary cues only in general presented a difficulty for the student annotators, as shown by the data in Table 11 detailing the distribution of the disagreement types across the student annotators. For all student annotators, disagreements relating to the disagreement category *melodic cues only* make up for more than 50% of all disagreements (discounting mechanical cases) and either latching (for R3 and R4) or pitch change (for R1 and R2) is the most frequent disagreement type overall for each annotator.

[29] On the German subcorpus, we counted a significantly lower proportion of disagreements due to misinterpretations of pitch changes than on the other three larger subcorpora (Papuan Malay, Wooi, and Yali) ($\chi^2 = 373.63$, df = 1, $p < 0.001$).

Table 11: Distribution of disagreement types over student annotators

| | Annotator | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Type of disagreement** | **R1** | | **R2** | | **R3** | | **R4** | |
| **MELODIC CUE ONLY** | **1,676** | **61.69%** | **1,423** | **55.80%** | **1,333** | **59.19%** | **841** | **53.33%** |
| - latching (false negative) | 292 | 10.75% | 439 | 17.22% | 1,123 | 49.87% | 572 | 36.27% |
| - pitch change (false positive) | 1,384 | 50.94% | 984 | 38.59% | 210 | 9.33% | 269 | 17.06% |
| **AMBIVALENCE OF PAUSES** | **831** | **30.59%** | **918** | **36.00%** | **754** | **33.48%** | **636** | **40.33%** |
| - hesitation (false positive) | 726 | 26.72% | 786 | 30.82% | 228 | 10.12% | 170 | 10.78% |
| - truncation (false negative) | 35 | 1.29% | 31 | 1.22% | 179 | 7.95% | 141 | 8.94% |
| - planned pause (false negative) | 70 | 2.58% | 101 | 3.96% | 347 | 15.41% | 325 | 20.61% |
| **UNCLEAR** | **210** | **7.73%** | **209** | **8.20%** | **165** | **7.33%** | **100** | **6.34%** |
| **Total** | **2,717** | **100.00%** | **2,550** | **100.00%** | **2,252** | **100.00%** | **1,577** | **100.00%** |
| **MECHANICAL CASES (excluded)** | 135 | (4.73%) | 133 | (4.96%) | 142 | (5.93%) | 108 | (6.41%) |

Table 11 also shows that there are two complementary strategies in dealing with difficulties relating to melodic boundary cues, R1 and R3 providing the clearest evidence for these strategies. R1 posits boundaries both at major pitch changes and at all kinds of pauses, resulting in high percentages for false positives of the pitch-change and the hesitation-type and the by far highest number of boundaries overall (cp. Table 3). R3, on the other hand, makes little use of evidence from pitch, but makes relatively good use of pauses (see also next section), in addition to relying on syntactic evidence for her native German, as noted in section 4 above. This strategy results in the lowest number of boundaries overall (cp. Table 3), with almost half of the disagreements being false negatives of the latching-type. R4 and, to a lesser extent, also R2 are more balanced in their use of the different boundary cues and hence are also somewhat more balanced in their disagreements.

To summarize this section, we have found that disagreements regarding IP boundaries can be accounted for systematically. They mostly arise in a smallish number of cue constellations where they are likely to arise, because the boundary cues are not properly synchronized. As expected, disagreements occur most commonly at IP boundaries which only involve melodic cues (55.56% of all disagreements) or where the status of pauses is ambivalent (30.94% of all disagreements). The distribution of disagreements thus indirectly supports the conclusion of the previous section that IP boundaries are robustly identifiable on the basis of the melodic and rhythmic boundary cues specified in our task design.

Furthermore, all disagreement types occur in all languages and there is no evidence that one disagreement type is specific to one language or group of languages. That is, the analysis of disagreements also supports the hypothesis that IP boundary identification is handled essentially in the same way in familiar and unfamiliar languages, with one important exception. Major pitch changes appear to be more difficult to interpret in unfamilar languages than in familiar ones and hence tend to give rise to substantially more false positives in unfamiliar languages. But note that this does not mean that melodic coherence is generally more easily identifiable in familiar languages. The substantial amount of disagreements in the German subcorpus arising from latching cases clearly rules out this interpretation.

Differences of a different kind arise with the regard to the distribution of pauses across the different subcorpora, as will be shown in the next section. This will also allow us to return to the question of why the interrater agreement results for Papuan Malay are significantly worse than for the other languages (cp. Table 6 above).

## 6. The significance of pauses

The ambivalence of pauses as IP boundary cues was noted repeatedly in the preceding sections. On the one hand, pauses are probably the easiest IP boundary cue to identify. But not every pause occurs at an IP boundary. Rather, IP-external pauses have to be distinguished from IP-internal ones. Only the former provide an important practical cue for a boundary. Furthermore, not every IP boundary is marked by a clearly identifiable pause (cp. latching). Consequently, the more IP-external pauses occur in a corpus and the easier they are identifiable as such (through their length, for example), the easier it should be consistently to segment this corpus into IPs. The present section is concerned with verifying this prediction and with showing how it affects interrater agreement results for each of our four main subcorpora.

Given the relatively easy identifiability of pauses and the fact that pauses often coincide with IP boundaries, the question naturally arises as to whether our relatively high interrater agreement results are not simply due to the fact that student annotators have made good use of pauses as boundary cues, especially in the case of unfamiliar languages. If this were the case, it would raise considerable doubt as to the validity and usefulness of other boundary cues and, more specifically, our assumption that IPs are identified (and processed) on the basis of both melodic and rhythmic cues. Therefore, in the second part of this section, we compare the interrater agreement between student annotators' segmentations and a purely pause-based segmentation with the interrater agreement between student annotators' segmentations and our consensus version, which makes use of both melodic and rhythmic cues.

To set the ground for the investigation of these issues, Table 12 contains figures about external and internal pauses in our corpus for the four main subcorpora German, Papuan Malay, Wooi, and Yali based on the consensus version. For each language, we provide the number of external and internal pauses (absolute frequency) as well as their relative frequency per IP and per word, and the average duration of external and internal pauses in milliseconds. For the relative frequencies and the lengths, we provide overall means calculated on a whole subcorpus of a particular language as well as means of means calculated per session. The latter figures are relevant for the statistical tests we perform when comparing the frequency and length of pauses between two languages. The last two rows provide overall and per-session means of the probability that a pause signals an IP boundary, calculated as the number of external pauses divided by the number of all pauses in a particular language. This measure indicates the reliability and strength of pauses as IP boundary cues.

Table 12: Some facts about pauses in our corpus

| External pauses | German | Papuan Malay | Wooi | Yali |
|---|---|---|---|---|
| **absolute frequency** | 882 | 1,631 | 777 | 429 |
| **relative frequency per IP** | 0.5046 | 0.6139 | 0.8328 | 0.7786 |
| **mean relative frequency per IP (per session)** | 0.5263 | 0.6335 | 0.8283 | 0.7862 |
| **relative frequency per word** | 0.0998 | 0.1572 | 0.2184 | 0.2138 |
| **mean relative frequency per word (per session)** | 0.1053 | 0.1642 | 0.2210 | 0.2299 |
| **mean duration (in milliseconds)** | 627 | 561 | 1,177 | 1,005 |
| **mean of mean durations (per session)** | 633 | 548 | 1,225 | 1,007 |
| **Internal pauses** | **German** | **Papuan Malay** | **Wooi** | **Yali** |
| **absolute frequency** | 162 | 102 | 16 | 8 |
| **relative frequency per IP** | 0.0927 | 0.0384 | 0.0171 | 0.0145 |
| **mean relative frequency per IP (per session)** | 0.0977 | 0.0414 | 0.0229 | 0.0164 |
| **relative frequency per word** | 0.0183 | 0.0098 | 0.0045 | 0.0040 |
| **mean relative frequency per word (per session)** | 0.0196 | 0.0109 | 0.0062 | 0.0050 |
| **mean duration (in milliseconds)** | 435 | 408 | 481 | 325 |
| **mean of mean durations (per session)** | 420 | 406 | 454 | 295 |
| **probability of IP boundary given pause** | 0.8448 | 0.9411 | 0.9798 | 0.9817 |
| **mean probability of IP boundary given pause (per session)** | 0.8504 | 0.9424 | 0.9743 | 0.9804 |

The figures in Table 12 show that pauses are more reliable and useful as IP boundary cues in Wooi and Yali (in both subcorpora 98% of all pauses occur at IP boundaries) than in Papuan Malay (94% of all pauses occur at IP boundaries) and finally German (only 84% of pauses occur at IP boundaries). These differences are statistically highly significant.[30]

Looking at frequencies of external and internal pauses in detail, we find that the German subcorpus contains fewer external pauses between IPs than the other subcorpora: The per-session mean for German is 0.53 (only about every second pair of subsequent IPs is separated by an audible external pause) compared to per-session means of 0.63 for Papuan Malay, 0.83 for Wooi, and 0.79 for Yali. German thus also contains a higher number of cases where two IPs follow each other without any audible pause (what we here call *latching*). Wilcoxon-Mann-Whitney tests yield statistically significant differences in this regard between German and all the other languages, Papuan Malay and all the other languages, but not between Wooi and Yali.[31] The same distribution is observed with regard to the measure of number of external pauses per word. Here again German and Papuan Malay significantly differ from each other and from Wooi and Yali in having less external pauses per word.[32]

[30] We use non-parametric Wilcoxon-Mann-Whitney tests to compare per-sessions means of the measures in Table 12 between languages. Pauses are significantly less reliable IP boundary cues in German compared to Papuan Malay (W = 35, *z* = -4.23, p < 0.001) and even less so compared to Wooi and Yali (German vs. Wooi: W = 6, *z* = -4.32, p < 0.001; German vs. Yali: W = 1, *z* = -3.53, p < 0.001). Pauses in Papuan Malay are also significantly less reliable IP boundary cues than in Wooi (W = 56, *z* = -2.50, p = 0.013) and in Yali (W = 19.5, *z* = -2.47, p = 0.015), while there is no statistically significant difference in reliability between Wooi and Yali (W = 34,*z* = -0.19, p = 0.887).

[31] Details for the frequency of latching: German vs. Papuan Malay: W = 88, *z* = -2.69, p = 0.006, German vs. Wooi: W = 9, *z* = -4.19, p < 0.001, German vs. Yali: W = 10, *z* = -2.93, p = 0.002; Papuan Malay vs. Wooi: W = 22, *z* = -3.81, p < 0.001, Papuan Malay vs. Yali: W = 19.5, *z* = -2.47, p = 0.015; Wooi vs. Yali: W = 43, *z* = 0.66, p = 0.553.

[32] Details for external pauses per word: German (per-session mean = 0.11) vs. Papuan Malay (per-session mean = 0.16): W = 44, *z* = -3.98, p < 0.001, German vs. Wooi (per-session mean = 0.22):W = 8, *z* = -4.23, p < 0.001,

As could be expected, for internal pauses the converse holds. Both German and Papuan Malay have significantly more internal pauses per IP as well as per word than the other two languages. Here, however, the differences between Papuan Malay and the two other Papuan languages are only marginally significant, which means that here only German differs strongly from all the other languages.[33]

Turning now to mean durations of external and internal pauses, we find that not only do the Wooi and Yali subcorpora contain more external pauses marking IP boundaries than Papuan Malay and especially German (and thus fewer cases of latching), they also exhibit a higher average duration of these external pauses than Papuan Malay and German, that is, pauses are longer and thus probably more noticeable. The average duration of external pauses (per-session means) is 633 ms in German and 548 ms in Papuan Malay compared to 1,225 ms in Wooi and 1,007 in Yali. External pauses are thus about twice as long in Wooi and Yali than in German and Papuan Malay.[34]

To sum up the first part of our discussion of the significance of pauses, (external) pauses seem to be more robust cues for IP boundaries in Wooi and Yali than in Papuan Malay and German, both in terms of frequency and duration. This, in all likelihood, is a major factor contributing to the lower agreement results for Papuan Malay compared to the other two Papuan languages Wooi and Yali already noted in section 4 above.

However, it is far from clear that this difference is one that can properly be attributed to a systematic difference on the level of linguistic structure. It is more likely that it is due to coincidental properties of the subcorpora for each language. The Papuan Malay and German subcorpora are, for example, much better gender-balanced than the Wooi and Yali subcorpora, which are heavily male-dominated. It is also well possible that the Papuan Malay and German speaker are more familiar and at ease with the task of retelling a film than the Wooi and Yali speakers, who come from backgrounds where watching films is not part of everyday culture. Note that the duration of internal hesitation pauses does not vary much between languages: German (420 ms) vs. Papuan Malay (406 ms) vs. Wooi (454 ms) vs. Yali (295

German vs. Yali (per-session mean = 0.23): $W = 3$, $z = -3.40$, $p < 0.001$. Papuan Malay vs. Wooi: $W = 42$, $z = -3.04$, $p = 0.002$, Papuan Malay vs. Yali: $W = 20$, $z = -2.43$, $p = 0.013$). Wooi vs. Yali: $W = 25$, $z = -1.03$, $p = 0.336$.

[33] Internal hesitation pauses occur more frequently in German than in Papuan Malay, both per IP (per-session means of 0.10 vs. 0.04; $W = 293$, $z = 3.30$, $p < 0.001$) and per word (per-session means of 0.02 vs. 0.01; $W = 265$, $z = 2.49$, $p = 0.012$). German also has more internal pauses than Wooi and Yali, both per IP (Wooi and Yali per-session means: 0.02) (German vs. Wooi: $W = 196.5$, $z = 3.75$, $p < 0.001$; German vs. Yali: $W = 103$, $z = 3.27$, $p < 0.001$) as well as per word (Wooi and Yali per-session means: 0.01) (German vs. Wooi: $W = 187$, $z = 3.35$, $p < 0.001$; German vs. Yali: $W = 100$, $z = 3.07$, $p < 0.001$). The Papuan Malay subcorpus contains (marginally) significantly more internal pauses than both Wooi (per IP: $W = 174$, $z = 2.11$, $p = 0.037$; per word: $W = 170$, $z = 1.95$, $p = 0.054$) and Yali (per IP: $W = 91.5$, $z = 1.92$, $p = 0.059$; per word: $W = 89$, $z = 1.77$, $p = 0.082$). In contrast, there is no difference in the relative frequency of internal pauses between Wooi and Yali, neither per IP ($W = 39$, $z = 0.28$, $p = 0.813$) nor per word ($W = 37$, $z = 0.09$, $p = 0.962$).

[34] In detail, Papuan Malay has significantly shorter external pauses than German ($W = 253$, $z = -2.13$, $p = 0.033$) and especially Wooi ($W = 5$, $z = -4.48$, $p < 0.001$) and Yali ($W = 6$, $z = -3.29$, $p < 0.001$). German external pauses are also significantly shorter on average than those in Wooi ($W = 6$, $z = -4.32$, $p < 0.001$) and Yali ($W = 7$, $z = -3.13$, $p < 0.001$), whereas there is no statistically significant difference in external pause duration between Wooi and Yali ($W = 47$, $z = 1.03$, $p = 0.336$).

ms).[35] This suggests that longer external pauses in Wooi and Yali are not simply due to slower speech rates.

If it is indeed the relatively higher robustness of pauses as boundary cues in the Yali and Wooi subcorpora that accounts for the better agreement results for these two languages when compared to Papuan Malay, the question naturally arises why the agreement results for German are not also lower than the ones for Yali and Wooi, given that pauses in the German subcorpus are even less useful boundary cues than they are in the Papuan Malay subcorpus. This is in all likelihood related to the fact observed in the preceding section that the disagreement type *pitch change* is significantly less frequently attested in the German subcorpus than in the three Papuan subcorpora. That is, while in the case of the Papuan Malay subcorpus, difficulties arose with regard to both (lack of) pauses and misinterpreted pitch changes, in the other three subcorpora, only one of these two disagreement sources occurs frequently, (lack of) pauses in the case of German, misinterpreted pitch changes in the case of Wooi and Yali.

Furthermore, if it is indeed the robustness of pauses as boundary cues that primarily accounts for the relatively high interrater agreement results on the Wooi and Yali subcorpora, it may well be questioned whether melodic cues play a significant role at all in the identification of IP boundaries in unfamiliar languages. That is, it could be the case that the segmentation of the subcorpora in unfamiliar languages is actually based primarily on rhythmic cues, in particular pauses. If this were the case, Cohen's kappa values for the pairwise interrater agreement between a purely pause-based segmentation and the segmentations provided by the student annotators should be higher than the corresponding values for a comparison of our consensus version and the student annotators' segmentations. Conversely, if the agreement values for the comparison with our consensus version are significantly higher, this would support the conclusion that melodic cues play an important role in the unfamiliar languages as well. The following figures show that the latter is the case. Note that given the distributional facts about pauses across the four subcorpora presented above, it is to be expected that the difference is largest for the German subcorpus (because pauses here are least reliable as boundary cues), somewhat smaller for Papuan Malay, and smallest for Wooi and Yali.

Table 13 provides the relevant Cohen's kappa values (mean kappa per narration) both for the corpus as a whole and with regard to the four main subcorpora. For each student annotator, the interrater agreement values with the consensus version are repeated from Table 5 and Table 8. These are given first in the lefthand column under each student annotator. The corresponding righthand column provides the agreement values with the purely pause-based segmentation.[36] In the rightmost column of Table 13,

[35] No comparison yields a statistically significant difference in the duration of internal pauses (German vs. Papuan Malay: W = 184, *z* = 0.40, p = 0.703; German vs. Wooi: W = 67.5, *z* = -0.25, p = 0.823; German vs. Yali: W = 67.5, *z* = 1.68, p = 0.100; Papuan Malay vs. Wooi: W = 74.5, *z* = -0.08, p = 0.958; Papuan Malay vs. Yali: W = 70.5, *z* = 1.64, p = 0.109; Wooi vs. Yali: W = 32, *z* = 1.79, p = 0.086).

[36] The pause-based segmentation was created from the consensus segmentation (CONS) by combining two adjacent IPs into one in case they were directly adjacent („latching") or separated by a period of silence of less than 150 milliseconds. Conversely, IPs from the consensus version were split into two at internal pauses of more than 150 msecs. Internal pauses were measured and annotated by hand as part of the original transcription process with a resolution of 100 msecs, for example as "(0.2)" = 200 msecs or "(0.5)" = 500 msecs.

we also compare our consensus version with the purely pause-based segmentation. Perhaps somewhat surprisingly, these values show that the consensus version agrees significantly better with the pause-based segmentation than all student annotators, except R3.[37] That is, at least three of the student annotators did not use pauses as boundary cues more often than the consensus version.

Table 13: Comparison of Cohen's interrater agreement coefficients for pause-based and consensus segmentation; statistically significant differences highlighted; cp. details in legend

| | **R1** | **R2** | **R3** | **R4** | **CONS** |
|---|---|---|---|---|---|
| | **vs. CONS vs. Pause** | **vs. CONS vs. Pause** | **vs. CONS vs. Pause** | **vs. CONS vs. Pause** | **vs. Pause** |
| **Overall** | **0.7422** vs. **0.5593**[a] | **0.7437** vs. **0.5959**[b] | **0.7381** vs. **0.7067**[c] | **0.8241** vs. **0.6691**[d] | 0.7060 |
| **German** | **0.7926** vs. **0.4587**[e] | **0.7752** vs. **0.5395**[f] | **0.6897** vs. **0.5523**[g] | **0.8643** vs. **0.5506**[h] | 0.5693 |
| **Papuan Malay** | **0.6814 vs. 0.5279**[i] | **0.7369** vs. **0.6047**[j] | **0.7016** vs. **0.7540**[k] | **0.7818** vs. **0.6947**[l] | 0.6881 |
| **Wooi** | **0.7335 vs. 0.6784**[m] | **0.6984** vs. **0.6354**[n] | 0.8299 vs. 0.8338 | **0.8456** vs. **0.7939**[o] | 0.8700 |
| **Yali** | **0.8148 vs. 0.7298**[p] | **0.7601 vs. 0.6526**[q] | 0.8378 vs. 0.7970 | **0.7928** vs. **0.7053**[r] | 0.8253 |

[a] V = 1,825, *z* = 6.70, p < 0.001; [b] V = 1,769, *z* = 6.72, p < 0.001; [c] V = 1,206, *z* = 2.14, p = 0.032; [d] V = 1,769, *z* = 6.55, p < 0.001; [e] V = 0, *z* = 3.72, p < 0.001; [f] V = 0, *z* = 3.72, p < 0.001; [g] V = 0, *z* = 3.72, p < 0.001; [h] V = 0, *z* = 3.72, p < 0.001; [i] V = 0, *z* = 3.92, p < 0.001; [j] V = 0, *z* = 3.92, p < 0.001; [k] V = 180, *z* = -2.80, p = 0.004; [l] V = 78, *z* = 3.40, p = 0.001; [m] V = 4, *z* = 2.75, p = 0.004; [n] V = 1, *z* = 2.87, p = 0.005; [o] V = 0, *z* = 3.02, p = 0.004; [p] V = 0, z = 2.20, p = 0.031; .[q] V = 0, z = 2.20, p = 0.031; [r] V = 0, z = 2.20, p = 0.031.

If we look at all languages taken together, averaged over 60 sessions (row OVERALL in Table 13), the data show that all student annotators agree more with our consensus segmentation than with a segmentation based on pauses. This basically also holds when the results for the individual subcorpora are compared with each other. The difference in agreement values is largest (and always significant) in the case of German, followed by Papuan Malay, and smallest (and in some instances not significant) for Wooi and Yali, as was to be expected on the basis of the distributional facts concerning pauses reviewed above. More specifically, the difference is significant for all four subcorpora in the case of R1, R2, and R4, while for annotator R3 it is not significant for Wooi and Yali. R3 also once again stands out in showing higher agreement values with the pause-based segmentation for two of the Papuan languages (Papuan Malay and Wooi), but the difference is only significant in the case of Papuan Malay. This finding points to the fact that she used primarily rhythmic cues in segmenting the unfamiliar languages. In the case of Papuan Malay, where latching is frequent, this leads to the significantly higher agreement with the pause-based segmentation.

Consequently, IPs were nearly always split at manually annotated internal pauses except for those internal pauses annotated as "(.)" (= very short pause, < 150 msecs). In addition, we also split IPs at non-linguistic events like coughing, sneezing or clearing one's throat, which had also been manually annotated during the initial transcription.

[37] Since one and the same segmentation by each student annotator is compared once to the consensus version and once to the paused-based version, we use Wilcoxon signed-rank tests for paired data here instead of Wilcoxon-Mann-Whitney tests for unpaired data. R1 vs. PAUSES compared to CONS vs. PAUSES: V = 1,828, *z* = - 6.72, p < 0.001; R2 vs. PAUSES compared to CONS vs. PAUSES: V = 1,754, *z* = - 6.18, p < 0.001; R3 vs. PAUSES compared to CONS vs. PAUSES V = 898, *z* = 0.13, p = 0.903; R4 vs. PAUSES compared to CONS vs. PAUSES: V = 1,468, *z* = - 4.07, p < 0.001.

The data presented in this section have shown that there are significant differences in the distribution of pauses in the four main subcorpora of our study which impact on the interrater agreement results. Specifically, pauses are less useful boundary cues in the German and Papuan Malay subcorpora than in Wooi and Yali. Consequently, the relatively high interrater agreement results for the latter two languages can in part be explained by the fact that here pauses coincide with IP boundaries to 98% (but the converse does not hold: approx. 20% of the IP boundaries in these subcorpora lack external pauses). Still, three out of four student annotators clearly also made use of melodic cues in segmenting the unfamiliar languages as shown by the fact that their segmentations agree better with the consensus version than with an exclusively pause-based segmentation.

Pauses thus have an important role to play in IP-boundary identification, but because of their ambivalence (IP-internal pauses) and optionality, IP-boundary identification cannot be reduced to pause identification. Rather, melodic cues, which theoretically are of prime importance, are also practically relevant despite the fact that they are less easily identifiable. This holds for familiar as well as for unfamiliar languages. In the following discussion section, we will look more closely at the theoretical status of melodic and rhythmic boundary cues and their interplay.

## 7. Discussion

The empirical results reviewed in the preceding three sections make it clear that IP boundaries are robustly identifiable by listeners with differing degrees of prosodic expertise across a substantial multilingual corpus. The inclusion of languages unfamiliar to the annotators proves that IP identification is possible on the basis of prosodic cues only. That is, annotators do not have to understand the content or the syntactic structure of the phrase chunked as one IP in order to be able to identify it as an IP. Instead, it appears to be possible to identify IP boundaries on the basis of the kind of general melodic and rhythmic boundary cues used in the instructions of our segmentation experiment (cp. section 2). This, in turn, suggests that the prosodic cues for IP boundaries and hence also IPs as prosodic units may be considered universal in a sense to be briefly explicated in this section.

Strictly speaking, the student annotators in the current experiment did not identify phonological units per se, at least not with regard to the languages unfamiliar to them. With regard to these languages, they did not know anything about the prosodic system in general and the phonological structure of IPs in particular. In this respect, the current study differs sharply from the kind of interrater agreement study briefly mentioned at the outset of section 2, where annotators are trained to identify phonological categories defined in terms of a specific framework such as ToBI. The claim made repeatedly throughout this paper and also at the beginning of this section – that IPs are robustly identifiable across familiar and unfamiliar languages – is based on the fact that there is robust interrater agreement between the student annotators' segmentation and the consensus version which identified IPs as phonological units (see section 3 for details).

Consequently, what the current study shows is that, at least for the languages under investigation, IPs can be consistently identified in spontaneous speech without being familiar with their phonological structure, simply on the basis of boundary cues which are not specific for a particular language. These cues, while usually mentioned when discussing the practicalities of segmenting spoken language into IPs, are generally not part of the hypothesized phonological structure of an IP. That is, final lengthening and pauses may be mentioned as optionally occurring at an IP boundary, but they are typically not considered to be part of the phonological structure in the same way as a final boundary tone, for example. Similarly, the pitch resets which occur between off- and onsets of consecutive IPs and are crucial melodic cues for boundaries are usually not included in phonological representations, if only for the simple reason that such representations do not pertain to chains of IPs, but to a single IP.

Still, there is a systematic relation between the units delimited by the boundary cues used in this study and language-specific phonological units of the type *intonational phrase*, at least in the languages under investigation. We propose to conceive of this relation along the lines of Gussenhoven's (2004: 49-96, inter alia) account of the relation between universal biological codes and the language-specific phonological organization of pitch variation. Specifically, we assume that the chunking of speech into IP-sized units is a universal necessity of human speech, arising from the interplay of the exigencies of the physiology of speaking (e.g. breathing) and cognitive demands on speech planning and processing (cp. Goldman-Eisler 1968, Levelt 1989, Chafe 1987, 1994, inter alia).[38] The physiology of speaking and the planning demands are also the source of universal melodic and rhythmic boundary characteristics of these IP-sized units, specifically melodic coherence and planning-related interruptions of speech delivery (planning pauses and unit-final lengthening). These boundary characteristics can be further grammaticalized into language-specific phonological categories, giving rise to a phonologically organized category *intonational phrase*. Such grammaticalizations typically involve the development of a limited set of unit-final (and, more rarely, also unit-initial) pitch movements, which usually form part of a more comprehensive system of grammaticalized pitch movements serving other functions such as marking focus (pitch accent) or distinguishing lexemes (lexical tones).

Note that this scenario specifically targets IPs and not all kinds of prosodic phrasing units such as intermediate phrases, accentual phrases or phonological words. These other levels of phrasing are possibly not universal. Inasmuch as they can be shown to be universal, we would expect that an independent motivation can be found for them in speech physiology, speech planning or another factor inherent in the processing of human speech.

As for IPs, we believe that it is very likely that an IP level is part of the prosodic system of all natural languages. This claim, however, can only be (dis-)confirmed by phonological analyses of the prosodic systems of all languages. If true, IPs would be a prime example of a universally attested *phonological* category. But note that, in principle, our scenario allows for the possibility that there is a language where spontaneous speech is produced in IP-sized chunks (delimited by the universal boundary cues), but

[38] It is likely that prosodic phrasing is not only grounded in processing constraints in language production but also has a very important role to play in language comprehension. See Frazier *et al.* (2006) for discussion and references.

where the phonological analysis of the prosodic system does not require (or support) an IP-sized unit.[39] More interestingly, perhaps, the scenario also predicts that IP units and their boundaries are grammaticalized to different degrees, i.e. that prosodic systems exist where the IP-level is but weakly grammaticalized, its structure consisting simply of a single final boundary tone, for example.

The assumptions on which the above scenario is based are of course in need of further theoretical and empirical scrutiny. In line with the empirical focus of this contribution, we briefly comment here only on some of the empirical issues.

To begin with, the segmentation task carried out here strictly speaking only shows that IP boundary cues are identifiable across a number of genetically and areally unrelated languages. It does not, of course, show that they are attested in all natural languages. However, we are not aware of any reports in the literature that claim that there are natural languages that lack the rhythmic or melodic boundary cues that have been used in this study. Hence, we believe that it is warranted to assume that these cues are indeed found in spontaneous natural speech across all spoken natural languages unless this assumption is empirically proven to be wrong.

Furthermore, we have only shown that native speakers of German are capable of identifying IP boundaries in languages unfamiliar to them. Thus, it could in principle be the case that the boundary cues in the four languages under investigation by chance happen to be similar enough to be identifiable across all of them. One would have to carry out this experiment with native speakers of languages with markedly different prosodic systems in order to be sure that the claimed identifiability in fact holds universally. In particular, one would, of course, also like to know whether native speakers of the three Papuan languages investigated here would also show robust interrater agreement across familiar and unfamiliar languages.[40]

Finally, there are of course many issues with regard to the question of whether and, if so, how the assumed universal boundary cues are actually linked to the physiology of speaking and the cognitive demands on speech processing. A particular challenge in this regard is to account for the complex interplay between the two basic cue types (melodic and rhythmic) repeatedly noted in the preceding sections. Specifically, the phenomenon of latching (IPs separated by melodic coherence only) needs to be accounted for.

## 8. Summary

The present work has provided evidence for the following claims:

[39] See Hyman (2011) for an argument along these lines with regard to syllables.

[40] A replication of the experiment using the same corpus with speakers of the three Papuan languages is not quite straightforward for quite a number of practical reasons. Importantly, the practical orthographies used for the three Papuan languages were relatively easy to process for the German annotators as the phoneme-grapheme correspondences are very regular and, for the most part, easily identifiable for them. That is, German listeners could relatively easily match the audio recording with the transcript, with the occasional problem arising from the mismatches discussed in section 5. German orthography, on the other hand, would be very difficult to process for the Papuan speakers. Consequently, the German corpus would have to be retranscribed in a practical orthography matching as closely as possible the sound values commonly associated with Indonesian graphemes, literate Papuan speakers being literate in Standard Indonesian.

1) Intonational phrases are empirically viable units according to standard measures for interrater agreement in categorization tasks. Multi-rater as well as pair-wise kappa coefficients show a substantial and statistically significantly above chance agreement on the placement of IP boundaries and thus the reliability of IP segmentation This holds for languages familiar and unfamiliar to the annotator (cp. section 4).
2) IP boundary identification can, and probably should, be based on prosodic cues only. Paying attention to non-prosodic information in the material to be segmented (syntactic boundaries, semanto-pragmatic units) leads to more disagreements.
3) Disagreements/uncertainties in IP segmentation largely arise in three instances (ignoring mistakes introduced by the experimental setup; cp. section 5):
    a) instances of latching where the speaker segments speech primarily by melodic means with little rhythmic support (and in particular, no audible pauses).
    b) instances of misinterpreted major pitch changes: IP-internal major pitch rise or fall is interpreted as a boundary tone. This misinterpretation is more common in unfamiliar languages than in familiar ones.
    c) dysfluencies where the speaker interrupts the proper delivery of an IP by a hesitation and completely aborts it (truncation). In such instances, IP boundaries are underdetermined and segmentation is arbitrary to a certain extent.
4) Melodic coherence, pauses, unit-final lengthening and unit-initial anacrusis are universal cues for IP boundaries. On the basis of these cues, it is possible to segment recordings in unknown languages with roughly the same reliability as one's native language.
5) The segmentation task becomes easier the better the fit between pauses and IP boundaries, i.e. the more frequently IPs are bounded by (external = planning) pauses and the less frequently they are interrupted by (internal = hesitation) pauses. However, melodic coherence is also an important boundary cue in practical-operational terms and segmentations based on the combination between melodic and rhythmic cues agree better with an expert segmentation (our consensus version) than with a purely pause-based segmentation (cp. section 6).
6) Annotators vary in the cues they most consistently pay attention to, some focusing on melodic evidence, while others primarily pay attention to rhythmic cues (cp. section 5). It is an interesting question whether such differences also occur in normal language processing or whether they are simply an artefact of the experimental set-up.

## References

Belo, Maurício C., John Bowden, John Hajek, Nikolaus P. Himmelman & Alex V. Tilman (2002–2006). *Dobes Waima'a documentation*. DobeS Archive MPI Nijmegen. Available at http://dobes.mpi.nl/projects/waimaa/.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. In *Glot International* **5**. 341–345.

Boersma, Paul & David Weenink (2015). Praat: doing phonetics by computer [Computer program] (version 5.4.09). Available (September 2015) at http://www.praat.org/.

Breen, Mara, Laura C. Dilley, John Kraemer & Edward Gibson (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* **8**. 277–312.

Buhmann, Jeska, Johanneke Caspers, Vincent J. van Heuven, Heleen Hoekstra, Jean-Pierre Martens & Marc Swerts (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In M. G. Rodriguez & C. P. S. Araujo (eds.) *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC).* Paris: Evaluations and Language Resources Distribution Agency. 779–785.

Chafe, Wallace L. (1977). The recall and verbalisation of past experience. In R. W. Cole (ed.) *Current Issues in Linguistic Theory*. Bloomington: Indiana University Press. 215–246.

Chafe, Wallace L. (1980). The deployment of consciousness in the production of a narrative. In Chafe (1980). 9–50.

Chafe, Wallace L. (ed.) (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Norwood, NJ: Ablex.

Chafe, Wallace L. (1987). Cognitive constraints on information flow. In Russell Tomlin (ed.) *Coherence and Grounding in Discourse*. Amsterdam: John Benjamins. 21–52.

Chafe, Wallace L. (1994). *Discourse, Consciousness, and Time*. Chicago: The University of Chicago Press.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**. 37–46.

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

Cole, Jennifer, Yoonsook Mo & Soondo Baek (2010b). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes* **25**. 1141–1177.

Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many annotators. *Psychological Bulletin* **76**. 378–382.

Fletcher, Janet (2010). The Prosody of Speech: Timing and Rhythm. In William J. Hardcastle, John Laver & Fiona E. Gibbon (eds) *The Handbook of Phonetic Sciences*. Oxford: Wiley-Blackwell Publishing. 523–602.

Frazier, Lyn, Katy Carlson & Charles Clifton Jr (2006). Prosodic phrasing is central to language comprehension. *TRENDS in Cognitive Sciences* **10**. 244-249.

Goldman-Eisler, Frieda (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Gussenhoven, Carlos (2004) *The phonology of tone*. Cambridge: Cambridge University Press.

Halliday, Michael A. K. (1967). *Intonation and grammar in British English*. The Hague: Mouton.

't Hart, J., R. Collier & A. Cohen (1990). A perceptual study of intonation: an experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press.

van Heuven, Vincent J. & Ellen van Zanten (eds.) (2007). *Prosody in Indonesian Languages*. Utrecht: LOT. Available (September 2015) at http://dspace.library.uu.nl/handle/1874/296769.

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

Holle, Henning & Robert Rein (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. In Hedda Lausberg (ed) *Understanding body movement: A guide to empirical research on nonverbal behaviour (With an introduction to the NEUROGES coding system).* Frankfurt am Main: Peter Lang Verlag. 261–275

Hyman, Larry M. (2011). Does Gokana really have no syllables? Or: what's so great about being universal? In *Phonology* **28**. 55–85.

Kirihio, Jimmi K., Volker Unterladstetter, Apriani Arilaha, Freya Morigerowsky, Alexander Loch, Yusuf Sawaki & Nikolaus P. Himmelmann (2009–2015). Dobes Wooi documentation. DobeS Archive MPI Nijmegen. Available at http://dobes.mpi.nl/projects/wooi/.

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

Jun, Sun-Ah (ed.) (2005). *Prosodic Typology. The phonology of intonation and phrasing*. Oxford: Oxford University Press.

Jun, Sun-Ah (ed.) (2014). *Prosodic Typology* II. *The phonology of intonation and phrasing*. Oxford: Oxford University Press.

Ladd, D. Robert (2008). *Intonational phonology* (2nd edition). Cambridge: Cambridge University Press.

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

Levelt, Willem J.M. (1989). *Speaking: From intention to articulation*. Cambridge, Mass.: MIT Press.

Mann, Henry & Donald Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18**. 50–60.

Mo, Yoonsook, Jennifer Cole & Eun-Kyung Lee (2008). Naive listeners' prominence and boundary perception. In P. A. Barbosa, S. Madureira & C. Reis (eds.) *Proceedings of the Fourth International Conference on Speech Prosody Campinas, Brazil, May 6–9, 2008*. 735–736. Available from ISCA Archive http://www.isca-speech.org/archive/sp2008.

Pierrehumbert, Janet (1980). *The phonology and phonetics of English Intonation*. PhD dissertation, MIT.

de Pijper, Jan-Roelof & Angelien A. Sanderman (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America* **96**. 2037–2047.

Pitrelli, John F., Mary E. Beckman & Julia Hirschberg (1994). Evaluation of Prosodic Transcription Labelling Reliability in the ToBI Framework. In *Proceedings of the 1994 International Conference on Spoken Language Processing (Yokohama, Japan)*. 123–126.

Riesberg, Sonja, Kristian Walianggen und Siegfried Zöllner (2012-2016). Dobes Yali documentation. DobeS Archive MPI Nijmegen. Available at http:// http://dobes.mpi.nl/projects/celd/.

van Rijsbergen, Cornelis Joost (1979). *Information Retrieval* (2nd edition). London: Butterworths.

Sanderman, Angelien A. (1996). *Prosodic phrasing : production, perception, acceptability and comprehension.* Eindhoven: Technische Universiteit Eindhoven. Available from http://www.tue.nl/en/publication/ep/p/d/ep-uid/142743/.

Shattuck-Hufnagel, Stefanie & Alice E. Turk (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* **25**. 193–247

Silverman, Kim, Mary E. Beckman, John F. Pitrelli, Mari Ostendorf, Colin W. Wightman, Patti Price, Janet B. Pierrehumbert, & Julia Hirschberg (1992). TOBI: a Standard for Labeling English Prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing (Banff, Canada)*. 867–70.

Stoel, Ruben B. (2005). *Focus in Manado Malay*. Leiden: CNWS Publications.

Stoel, Ruben B. (2006). The intonation of Banyumas Javanese. In *Proceedings of the Speech Prosody 2006 conference*. Dresden: TUD press. 827–830.

Streefkerk, Barbertje M. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. PhD dissertation, Amsterdam.

Wilcoxon, Frank (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**. 80–83.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).* 1556–1559.

Yoon, Tae-Jin, Sandra Chavarría, Jennifer Cole & Mark Hasegawa-Johnson (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the ISCA International Conference on spoken language processing (Interspeech 2004) Jeju Island, Korea, October 4–8, 2004*. 2729–2732. Available from ISCA Archive http://www.iscaspeech.org/archive/interspeech_2004.