PREPUBLICATION DRAFT

On the universality of intonational phrases – 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 boundaries across familiar and unfamiliar languages. Four annotators segmented a corpus of more than three hours of spontaneous speech into Intonational Phrases. The corpus included narratives in their native German, but also in three languages of Indonesia unknown to them. The results show significant agreement across the whole corpus, as well as for each subcorpus. We discuss the interpretation of these results, including the hypothesis that it makes sense to distinguish between phonetic and phonological Intonational Phrases and that the former are a universal characteristic of speech, allowing listeners to segment speech into Intonational Phrase-sized units even in unknown languages.

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 different 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 in the type of discourse and conversation analysis inspired by Chafe (1994).

Intonational phrases (IPs) are widely held not to pose particular problems of identification. Thus, 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, …

[1] We are very grateful to the four anonymous reviewers for PHONOLOGY, Associate Editor Bob Ladd and Editor Ellen Kaisse for extensive, detailed and constructive comments, questions and suggestions which have led to major revisions of the version first submitted. We also thank the members of the Cologne Phonetics colloquium for helpful discussion of the first draft of this paper. We owe a very big thanks to the many students and colleagues who participated in the transcription and segmentation of the recordings analyzed here.
Authors' contributions: 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: statistical analyses, draft of sections 4 and 5, contributor to interrater study and consensus version. VU: draft of section 3, contributor to interrater study (including recordings) and consensus version.
Research for this paper was funded by grant no. 01UG1240A from the German Federal Ministry for Education and Research (*Bundesministerium für Bildung und Forschung*) to Nikolaus P. Himmelmann. We are also grateful for funding from the Volkswagen Foundation which supported the compilation of the West Papuan corpora used here from 2002-2016. See FN 5 for details.

To date this assumption has not been subject to scrutiny in ways standard to research concerned with segmentation tasks, i.e. by evaluating interrater agreement. As reviewed in section 2, previous interrater studies on IP boundaries (IPBs) are typically limited in that 1) they involve short (< 30 seconds) examples specifically recorded for the task or excerpted from longer recordings; and 2) they usually combine several tasks, i.e. labelling prosodic boundaries and prominences (e.g. pitch accents).

The current study, in contrast, is exclusively concerned with IPBs and involves the segmentation of a corpus of more than three hours of spontaneous narrative speech (cp. Table 1). Most importantly, it is primarily concerned with the question of whether IPs are cross-linguistically identifiable across unrelated languages, which, as far as we know, has not been addressed in the literature. Specifically, we ask whether non-native listeners are able to identify IPBs in unfamiliar languages without being able to understand the utterances to be segmented and without familiarizing themselves with the prosodic system of the language in question.

Experiments from machine learning suggest that at least some cues for IPBs are applicable across unrelated languages. In such experiments, models for IPB detection are trained on data from one language (English, for example) and applied to data from another language (Mandarin, for example). Results are often surprisingly good in that boundary classifiers trained on foreign language data achieve results within the range of classifiers trained on data from the same language. Soto *et al.* (2013) provide an instructive example comparing classifiers trained on English, German, Mandarin and Italian. Our findings with human annotators show important parallels to this line of work, as discussed in 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 comparable between languages, as it consists of retellings of the Pear Film (Chafe 1980) in German, the native language of the annotators; Papuan Malay, the lingua franca 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 experience with the West Papuan languages.[2] All other annotators participating in the experiment were unfamiliar with them.

The core questions to be answered by this study are:

Q1. Do the segmentation results for the whole corpus and for each individual language show above-chance interrater agreement according to standard kappa metrics?

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

As for the second question, there are two ways in which familiarity with a language may become relevant in the segmentation task and influence interrater agreement. First, it could be that the prosodic cues used as segmentation criteria come in language-specific forms and are more readily recognized in familiar languages. Prima facie, such language-specific forms are less likely for pauses, probably the

[2] Throughout this article, “West Papuan” is used as a geographic reference to the Indonesian western half of the island of New Guinea.

perceptually strongest cue for IPBs. But they have some plausibility for other IPB cues like pitch resets or unit-final lengthening. 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 other factors (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. There is a strong tendency for IPBs to overlap with clause boundaries and a concomitant tendency to hear IPBs at clause boundaries. The unfamiliar-language condition completely removes the potential influence of non-prosodic factors, with two possible outcomes. On the one hand, interrater agreement could be significantly less strong for unfamiliar languages because of the missing non-prosodic information. However, as non-prosodic information brings in a different layer of factors, it also increases the potential for conflict between different segmentation cues (cp. Ladd 2008: 288–290). Consequently, interrater agreement in familiar languages could be worse than in unfamiliar ones, as, in the latter, annotators are forced to focus exclusively on prosody.

The paper is structured as follows. Section 2 reviews previous interrater studies concerning IPBs and highlights the points where our study diverges from these. It also provides details on the boundary cues focused on here and their complex interrelationship. Section 3 details task design and data. The empirical core of this study is presented in sections 4–5. The experimental results provided in section 4 demonstrate robust interrater agreement for the whole corpus, as well as for individual languages. The main question for evaluating this result is whether the robust interrater agreement is due to the fact that pauses play a major role in detecting IPBs. It could be the case that annotators identify pause units rather than IPs, especially in unfamiliar languages. Section 5, therefore, takes a closer look at the experimental results and the distribution of pauses in the corpus and shows that annotators do not rely on pauses to a higher extent in unfamiliar languages than in the familiar German.

Section 6 discusses the theoretical import of our results for current concepts of IPs and their functions. It reviews different possible interpretations of the interrater results, including the view that they only show that German hearers can identify German-like IPs in other languages. The main alternative interpretation is the hypothesis that an IP-sized unit is found across all languages and that the phonetic cues delimiting its boundaries can be perceived by speakers of all languages. What we might call a universal *phonetic* IP needs to be distinguished from language-specific *phonological* IPs,[3] which can be interpreted as a language-specific grammaticization of the universal phonetic IP. Our results support a view of prosodic categories as 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 grammaticization of prosodic features.

[3] Special thanks to Associate Editor Bob Ladd for suggesting this terminology and for a great many further suggestions for improving the exposition.

2. Prosodic interrater agreement studies and their targets

Interrater agreement studies of prosodic phenomena can be 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 annotation schemes, the Rhythm and Pitch (RaP) system (Dilley & Brown 2005) and the To(nes and) B(reak) I(ndices) system (Silverman *et al.* 1992, Pitrelli *et al.* 1994). They also present a useful survey of previous interrater studies of this type and their methodological challenges (see also Cole *et al.* 2010b: 1143–1145).

This type of study targets language-specific phonological categories, i.e. tonal targets and different prosodic boundaries. The annotation schemes tested differ in the consistency and directness of the auditory and acoustic evidence used, but the decisions are clearly about (abstract) phonological categories and not about phonetic events. Part of the training for this type of study is the provision of examples illustrating typical auditory and acoustic correlates of the intended categories. Labelers are usually provided with acoustic data (minimally wave-form and f_0 contour) in addition to audio files.

The other type of study targets the perception of prosodic prominences and boundaries by naïve listeners without expertise in prosodic theory and annotation, and 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 syntactic, semantic or pragmatic information. A prototypical study along these lines is Mo *et al.* (2008)[4], with analytical follow-ups in Cole *et al.* (2010a) on phonetic factors, and Cole *et al.* (2010b) on syntactic (and other non-prosodic) factors. In this study, more than 70 undergraduate students of linguistics marked prosodic prominences and boundaries in 18 short excerpts of spontaneous American English, based solely on their auditory impressions. The instructions regarding prominences and boundaries are 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, which included word boundaries, speech errors and disfluencies, but no punctuation or capitalization. The relevant findings of this study are: a) there is significant interrater agreement with regard to boundaries, with a mean Cohen's κ coefficient of 0.582 across all pairs of transcribers (the values for prominences are much lower); b) there is significant variation with regard both to speakers, where Fleiss' κ coefficients (measuring agreement between all listeners at the same time) range from 0.35–0.95, and to listeners, with some pairs only reaching a Cohen's κ as low as 0.24, while others agree to a large extent, as reflected in a Cohen's κ coefficient of 0.85.

[4] The method originates 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 of work on prominence perception in this tradition.

In some ways, Buhmann *et al.* (2002), based on Dutch corpus data, is a very similar study. However, their procedure is different in a number of important regards. First, while working with non-expert annotators, they include an intensive training period in which, after having received instructions and examples, 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 8,000 words (45 minutes) of read, scripted and unscripted speech. Third, an on-line working environment was used, which included the audio-visual display of waveforms as well as time-aligned text. Finally, the test corpus was pre-segmented into pause-bounded phrases of roughly ten seconds, using automatically detected pauses (> 0.5 seconds) as indicators for strong prosodic boundaries. Given the intensive training and the pre-segmentation, it is not surprising that Buhmann *et al.* obtain a fairly high interrater agreement. For boundaries, the Cohen's κ coefficients for interrater pairs range from 0.695 to 0.884 (Buhmann *et al.* 2002: 782).

Regarding instructions on detecting prosodic boundaries, Buhmann *et al.* (2002: 779) speak of "breaks", thus targeting a non-technical category which presumably is part of the non-expert understanding of spoken language. They distinguish strong and weak breaks, defining them 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), Buhmann *et al.* focus on auditory impressions, with an emphasis on pauses and no explicit appeal to coherent melody contours.

The current study 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 (see the following section). But there are also two major points of difference which warrant attention here. The most important difference is that our study compares the performance of annotators across familiar and unfamiliar languages. This task design presupposes that the chunking of speech can be auditorily identified across languages, which in turn presupposes that some relevant cues occur cross-linguistically. In the latter regard, note 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 adjacent ones. Furthermore, Fletcher (2010) provides a wealth of references for pauses (2010: 573–575) and tempo changes (2010: 540–547) as cross-linguistically attested boundary cues.

The cross-linguistic identifiability of boundary cues, however, has not been explored systematically and is the topic of this investigation. Hence, it is important which cues we used and how we explained them to the annotators. This is the second point where the present study diverges from Mo *et al.* (2008) and Buhmann *et al.* (2002). Our written instructions (see Supplement 1 for details) characterize IPs as distinct units perceivable by means of a coherent melody. They draw attention to two major types of IPB cues: 1. the interruption of the rhythmic delivery by, inter alia, a pause or final lengthening; and 2. the disruption of the pitch contour by a jump in pitch (up or down) between the end of one unit and the beginning of the next.

Like the Buhmann *et al.* study, our annotators were thus also clearly instructed to follow prosodic cues for boundaries only, but unlike Buhmann *et al.*, a distinction was made between melodic and rhythmic cues. Importantly, the instructions also reflect the complex interdependence between melodic and rhythmic cues, and the fact that both are ambivalent as boundary cues. 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 emphasis on a particular syllable). Similarly, pauses are heard as boundaries only when the melodic contour appears to have reached its projected endpoint.

However, the reverse also holds: The identification of a coherent contour partly depends on its interplay with rhythmic cues. The clearest example for this is the fact that there are limits to the length of a silence 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 silences 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 silence.

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 refer primarily to the segmentation task at hand. However, it is not very speculative to assume that this also holds for speaker-hearers engaged in the actual production and comprehension of speech.

The ambivalence of pauses as boundary indicators arises from the fact that they 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 adjacent IPs. According to a widespread view (e.g. Goldman-Eisler 1968, Levelt 1989, Chafe 1994, Krivokapić 2014), they usually arise because speakers need time to plan the next IP (hence *planning pauses*), but may sometimes also be used deliberately as an IPB signal. Also, external pauses often give the speaker the opportunity to breathe. Internal pauses, in contrast, are pauses that occur during the production of an IP. They mostly result from production difficulties, such as problems with lexical access, self-corrections, etc., and are also called hesitation pauses (cp. next section). Evidence from gestural coordination in articulation suggests that these two pause types can be distinguished by the position of the articulators during the resting period (Krivokapić 2014:4f, see also Katsika *et al.* 2014:75f). This research also suggests that external pauses are themselves planned.

In practical-operational terms, pauses are probably the easiest IPB cue to identify. External pauses, when correctly identified, are therefore an important practical cue for IPBs. Lots of internal pauses, in contrast, may render identification of IPBs more difficult as they can be misinterpreted as IPB cues, especially when the hearer does not understand the content of a given segment.

Melodic coherence, on the other hand, is much more difficult to perceive consistently when paying conscious attention to it in a segmentation task. In our instructions, we highlight jumps in pitch between off- and onsets of IPs as indicators of interrupted coherence. However, such pitch jumps often are not larger than the micro-perturbations caused by obstruents, the correlation with rhythmic interruptions providing the best diagnostic for distinguishing between these two types of pitch jumps.

There are many further phonetic cues that occur at IPBs such as fading intensity, creaky voice, the absence of coarticulation, unit-initial glottal stops, etc. (Shattuck-Hufnagel & Turk 1996, Ladd 2008, Wagner & Watson 2010). These cues, however, tend to be less frequent and systematic. When they occur, they contribute to the two overarching perceptual constructs, melodic and rhythmic coherence. Fading intensity and creaky voice, for example, contribute to the interruption of melodic coherence. It is likely that our annotators have also made use of these additional cues, even though they are not mentioned in our instructions. This aspect, however, will not be further discussed in this paper.

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. Both melodic and rhythmic cues are to be used in identifying IPBs. They reinforce each other when occurring in temporal alignment (cp. Pijper & Sanderman 1994, Krivokapić & Byrd 2012), but may lead to disagreements when not synchronized. Pauses have a special status because they can be identified relatively easily and consistently, but they are not unequivocal boundary cues because of the occurrence of IP-internal pauses.

3. Data and procedure

The corpus used in this study consists of sixty retellings of the Pear Film, a six-minute film made in 1975 for the cross-linguistic study of cognitive, cultural and linguistic aspects of narrative production (Chafe 1980). The soundtrack does not contain speech, consisting only of the sounds associated with the depicted actions (such as a bike accident).

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 of the corpus, which is partitioned into three groups for processing and presentation purposes, each comprising twenty stories. For practical and explorative purposes, the corpus also includes smallish samples from additional varieties: Kölsch (the German dialect of 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 concerned with cross-linguistic comparison, because they are too small for valid statistical modeling. Supplement 2 provides further details on recording procedures and corpus compilation.

Table 1: Composition of the corpus

No. of narratives		**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 three languages from Eastern Indonesia that this study mainly focusses on are typologically and genetically very diverse and show very different prosodic characteristics. While both Papuan Malay and Wooi are Austronesian languages, they belong to two different major branches of this family (Western-Malayo Polynesian and South Halmahera-West New Guinea, respectively) and have very different grammatical profiles. Papuan Malay has little morphology, adheres to a fairly strict SVO pattern and has bare nouns as the most frequent type of noun phrase. Wooi has a complex subject marking paradigm, as well as a complex set of noun phrase markers, makes frequent use of serial verb constructions and, while also following a basic SVO pattern, places negation and other particles at the end of the clause (rather than before or after the verb as in Papuan Malay). Yali belongs to a different language family altogether (Trans-New Guinea), is an SOV language, has a moderate amount of (post-positional) case marking and complex verbal morphology, with hundreds of forms in a paradigm (cp. Riesberg 2017).

Prosodically, these three languages illustrate systems very different from German, but found in many other parts of the world. As typical for Malayic and other western Indonesian languages, Papuan Malay has neither tone nor stress, but two major levels of prosodic phrasing. The IP is marked by the combination of a phrase accent and a boundary tone occurring within a two-syllable window at the end of the phrase, similar to what has been described by Maskikit-Essed & Gussenhoven (2016) for Ambon Malay and Stoel (2007) for Manado Malay. The smaller *Phonological Phrase* is marked by a high tone on the final syllable, similar to what has been described by Stoel (2007) for Manado Malay and by Himmelmann (2010) for Waima'a. See section 6.1 for further discussion and exemplification.

Wooi is similar to Papuan Malay in delimiting IPs with the combination of a phrase accent and a boundary tone, but differs in having lexical stress *and* lexical pitch accents, similar to Papiamento (Remijsen & van Heuven 2006). Note that the small group of Austronesian West Guinea languages it belongs to are well-known for their unusual prosodic systems. Remijsen (2001) and Kamholz (2014) provide details. Finally, Yali is a typical Papuan lexical pitch accent language where each content word is marked with a final high tone, with more complex regularities holding for the (clause-final) verbal complex. See Heeschen (1992:13f) for a description of the similar prosodic system in the neighboring Yale (Kosarek) language.

Prior to the current study, all sixty pear story narratives had been transcribed by native speakers of the respective languages using ELAN.[5] 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 to segment the narratives into IPs on the basis of the audio stream and the plain text script. For each narrative, 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, unlike in other studies mentioned in section 2, disfluencies were not marked as such, but the transcript did contain a representation of unclear segments which could not be transcribed (indicated by roughly one x per unclear syllable). Further details on experimental procedure are given in Supplement 2.

Four linguistics students, all native speakers of German, 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, cp. Table 2. R1-3 had a basic introduction to prosody as part of the introductory courses of their BA program.

Table 2: Student annotators and the authors' consensus version

R1	Bachelor student (female) in Linguistics
R2	Master student (male) in Linguistics
R3	Master student (female) in Linguistics
R4	Master student (female) in Linguistics, specializing in phonetics, writing MA thesis on prosodic topic at the time of involvement in the project
CONS/Authors	each narrative originally transcribed in IPs with native speaker input, transcriptions independently checked for consistency by 2 of the authors, final check by first author; all authors are native speakers of German except for MS, who is a native speaker of Hebrew but speaks German fluently

[5] We thank Sonja Riesberg for 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 the documentation projects. ELAN is a multimedia annotation tool for multi-modal research, see http://tla.mpi.nl/tools/tla-tools/elan/.

In addition, the authors produced a consensus version which, importantly, involved native-speaker input in the creation phase and 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 narrative was transcribed by a native speaker, or a language specialist working together with a native speaker. The primary segmentation unit of the transcription was the IP, defined in the same way as for the participants in the current study. Most of the transcriptions were done before the current study was designed. Second, the transcriptions were independently checked by two of the three last-named authors. Third, the three last-named authors compared their changes to the original transcripts and produced a first consensus version by resolving disagreements through relistening and discussion. As a final step, this version was checked by the first author, who focused on problematic cases and overall consistency in instances where the exact placement of the boundary is arguably arbitrary (due to noise in the recording, for example, or due to disfluencies, as further discussed shortly). In contrast to the four student annotators, the authors made regular use of instrumental evidence in the form of f_0 plots and waveforms produced by PRAAT (Boersma & Weenink 2015) in order to decide especially difficult cases. Given that the consensus version is based on phonological hypotheses regarding the structure of IPs in each language and was created by annotators with expert training in prosody and, in the case of NPH and VU, with first-hand knowledge of the languages and their prosodic systems, we decided to treat the consensus version (CONS) as the reference segmentation in the analysis, against which the performance of the other annotators can be evaluated.

Instances of disagreement in the creation of the CONS version never exceeded 20% of the boundaries in a given narrative and involved less than 10% of all boundaries in the corpus. Most disagreements pertained to two types of well-known problematic cases. First, boundary decisions tend to be difficult when the speaker produces a sequence of IPs in rapid succession without intervening pauses, known as *latching* in the discourse- and conversation-analytic literature. In example (1) from German, latching occurs in three IPs in a row. The main cues for IPBs here are pitch jumps interrupting the melodic contour, downward after *gelegt* and *bereitstanden*, upward after *heraus* (cp. Figure 1). All student annotators agree with the boundary after *mühsam*, but only two have boundaries after *gelegt* und *bereitstanden*, and only one after *heraus*.[6]

[6] Conventions in the examples: each line is one IP; = indicates latching; pause length is given in ( ); < > surround false starts (< > on morpheme interlinearization tier indicates infixes in Wooi). Pauses and false starts were not marked as such in the transcripts given to the student annotators. Glosses for grammatical categories: ACT – actor voice, DAT – dative, DET – determiner, NSG – non-singular, PL – plural, PRTC – participle, REL – relative marker, SG – singular, TOP – topic marker, and VEN – venitive.

(1) *in* *Körbe* *gelegt* =
in baskets put:PRTC

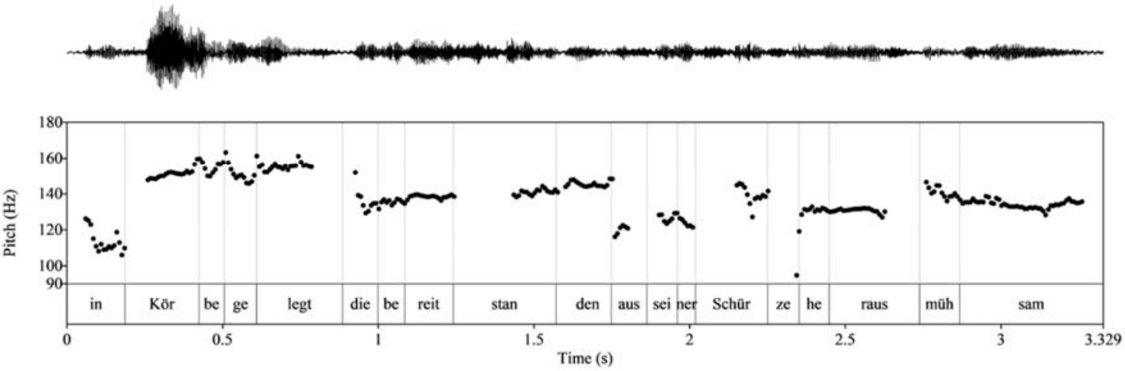
die *bereitstanden* =
that stand.by:3PL

aus *seiner* *Schürze* *heraus* =
out.of his apron out

mühsam (0.7)
painstakingly

'into baskets, that stood there, from out of his apron, painstakingly.' (DEU_pear_Flor)

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

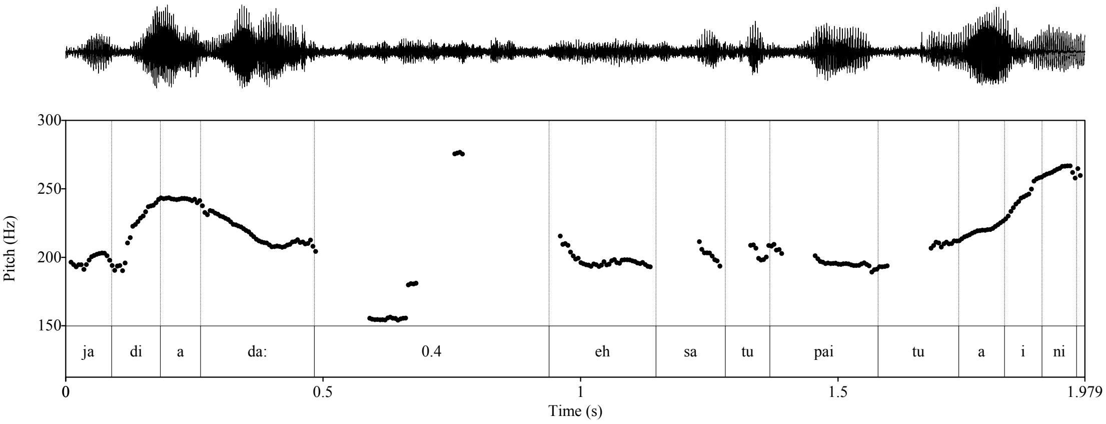


The other factor giving rise to disagreements relates to disfluencies. Disfluencies are a special case, because they are inherently ambiguous with regard to the boundary issue, as the speaker does not properly deliver an IP already in production, and either interrupts or abandons it. Consequently, disfluencies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

In the consensus version, we tried to distinguish consistently between hesitations (IP-internal disfluencies) and truncations, i.e. the abandonment of a unit currently under way. This distinction is primarily based on pitch evidence, but also on the length of the interruption. Interruptions lasting more than one second were generally considered truncations. Otherwise, a disfluency was considered to be IP-internal only if speech delivery was resumed after the disfluency on the same pitch level that was reached before. The idea here is that if one were to cut out the disfluency, the IP would display an overall coherent intonation contour, making it likely that the speaker continues with the delivery of an IP begun before the disfluency. This is illustrated by example (2) from Papuan Malay, where the f_0 extraction in Figure 2 clearly shows that the pitch on *satu* continues on almost exactly the same level as it was on *ada:* right before the hesitation break (the IP-internal pause is partially filled by the hesitation marker *eh*).

(2) *jadi* *ada:* (0.4) *eh* *satu* *paitua* *ini* =
so there.is uh one adult this
‘so uhm there was this man’ (PMY_pear_Lala)

Figure 2: Waveform and f_0 extraction of example (2)[7]

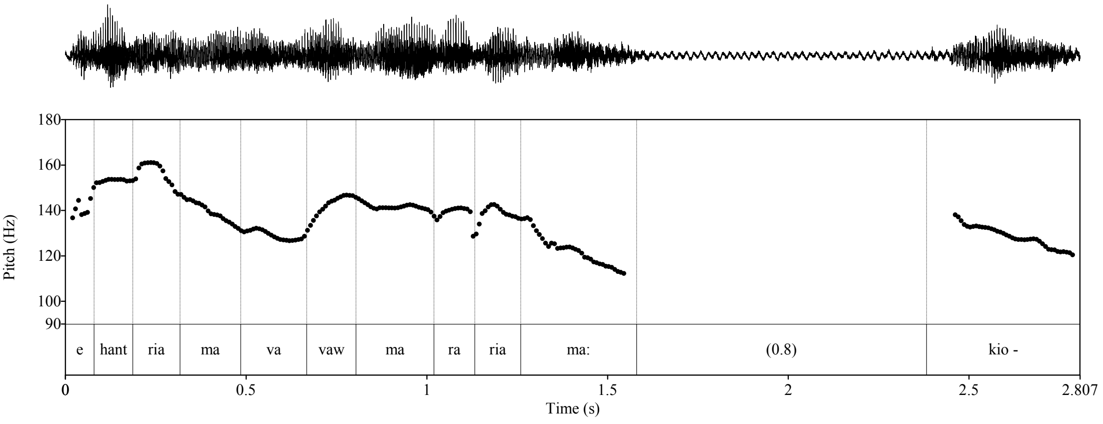


In truncations, on the other hand, there is clear evidence for the start of a new IP, for example in (3) taken from Wooi. Here the speaker aborts the utterance at the end of *ria ma:* and after a short break starts a new one instead of repairing or resuming the old one. The truncation is clearly cued by a pitch reset (falling pitch on *ma:* followed by a new onset on *kio*) and considerable lengthening of the last syllable. The difference in f_0 between *ma:* and *kio* is almost four semitones, so that it is safe to assume that there is no intention of the speaker to connect back to the previous pitch contour.

(3) *ehanti* *ria* *ma* *vavaw* *mara* *ria* *ma* (0.8)
someone <3SG>go VEN DET:NSG TOP <3SG>go VEN

kio (2.0)
<3SG>take
‘there was someone coming, he came… he took –’ (DEU_pear_Alex)

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



[7] The f_0 traces seen during the pause of 0.4 ms are caused by background noises.

While there are many instances in which the distinction between a hesitation and a truncation is reasonably clear, the distinction is also to some degree arbitrary in that 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.

Our statistical procedures are described in Supplement 2.

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

In this section, we first look at overall agreement on the entire corpus to assess the validity and reliability of the IP as a cross-linguistically identifiable unit. Second, we compare the segmentations of individual annotators to our consensus (CONS) segmentation to look for differences in the behavior of individual annotators. Third, we compare interrater agreement on individual languages to determine whether annotators agree equally on the segmentation of IPs across different languages.

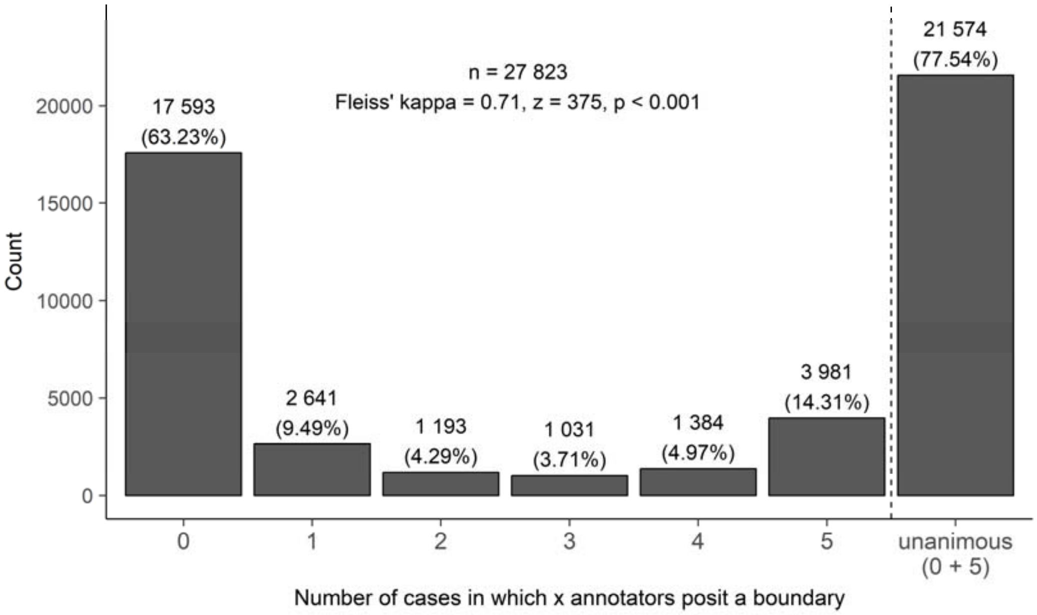
The whole corpus comprises 27,883 words. Since the start of the first IP and the end of the last IP in a narrative always coincide with the first and last words and are thus given by definition, we excluded them from the evaluation and therefore have to consider 27,823 potential IPBs in all (one less than the number of words for each of the sixty narratives). Table 3 provides an overview of the segmentations created by the five annotators (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	6,499	4.26	2.79
(grand) mean	6,772	4.09	2.82

On the entire corpus, we obtain a raw agreement of 77.54% for all five annotators (R1–R4 and CONS) and a statistically significant Fleiss' κ score of 0.71 (cp. Figure 4), which represents substantial agreement according to Landis & Koch (1977). When we only consider the four student annotators, we obtain a raw agreement of 78.21% and a statistically robust and substantial interrater agreement ($\kappa = 0.68$, $n = 27{,}823$, $z = 277$, $p < 0.001$). Figure 4 provides the number and percentage of cases in which *x* of the five annotators posit an IPB, ranging from zero for places where no annotator has posited a boundary, to five for places where all annotators have assumed an IPB. The rightmost column shows the total of all unanimous decisions, i.e. cases where all annotators agreed that there is no boundary and cases where all agreed that there is a boundary. These results show that recordings of spontaneous speech in different languages can be segmented into IPs reliably even by non-expert annotators without special training.

Figure 4: Overall agreement on the IP segmentation of the whole corpus



If we take our consensus segmentation (CONS) as reference and compare individual student annotators' segmentations to it, we obtain the results presented in Table 4. Individual student annotators' segmentations agree quite well with the authors' consensus segmentation, with Cohen's κ statistics (overall) ranging from 0.74 for R3 to 0.82 for R4, all of which are highly statistically significantly above chance.[8] All four student annotators are thus able to provide a reliable IP segmentation that agrees to a large extent with the authors' expert segmentation.

Table 4: Comparison of annotators to reference segmentation on the whole 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 κ (overall)	0.7393	0.7481	0.7392	0.8237
Mean κ per narrative	0.7422	0.7437	0.7381	0.8241
Std. dev. of κ per narrative	0.0903	0.0711	0.0920	0.0630

Student annotators nonetheless differ amongst each other in their tendency to either assume more or fewer IPBs than CONS: R1 and R2 posit relatively many IPBs (cp. Table 3) and segment the narratives into relatively short IPs, which results in high recall values above 90% (i.e. more than 90% of the IPBs marked in CONS are also found in these segmentations) but lower precision values of

[8] 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).

slightly above 70% (i.e. only about 70% of the boundaries marked by these student annotators are also found in CONS). R3 and R4, in contrast, assume fewer IPBs and therefore longer IPs (cp. Table 3), resulting in high precision values of about 90%, as well as lower recall values of approx. 71% and 82%, respectively (cp. Table 4).

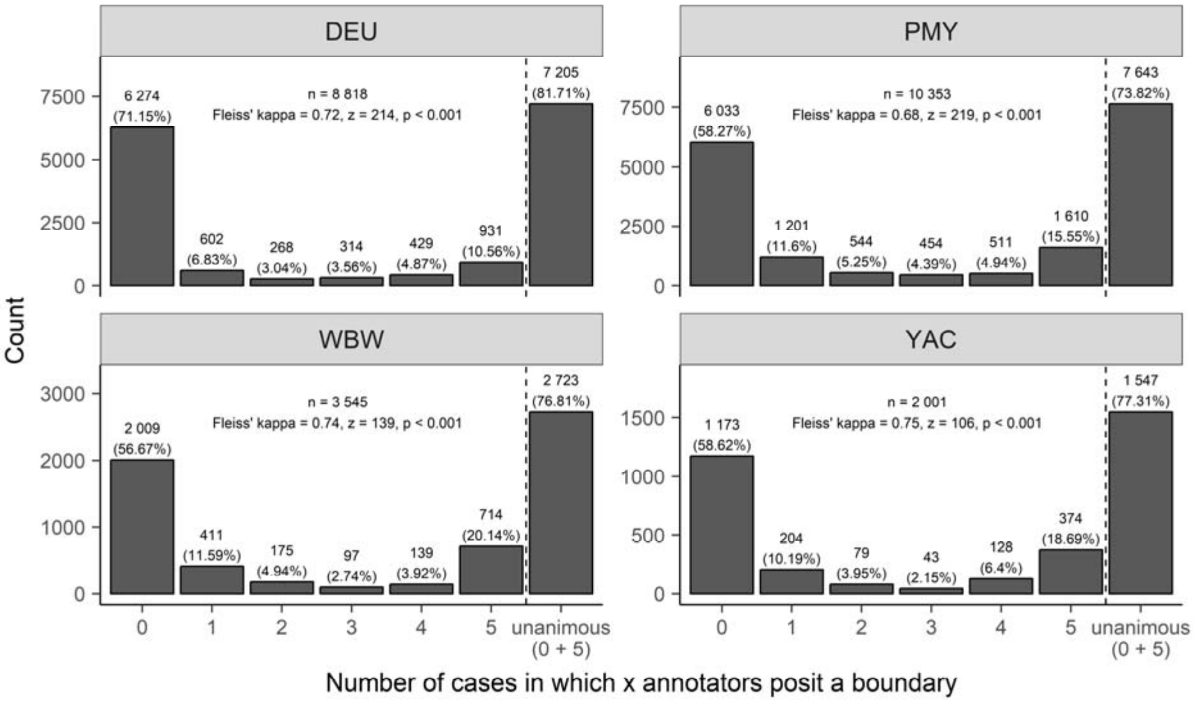
R4 has the lowest standard deviation (cp. Table 4), namely, 0.06 vs. 0.09 for R1, 0.07 for R2, and 0.09 for R3. R4 is thus also the most consistent of the four annotators with regard to agreement with CONS across all 60 narratives. This is probably related to the fact that R4 is the only student annotator who has had in-depth training in prosodic analysis, albeit not specifically for the present study.

Still, the overall results demonstrate a well-above-chance agreement between annotators of different levels of expertise in determining IPBs in an extensive corpus of spontaneous narrative speech in both familiar and unfamiliar languages. This suggests that phonetic boundary cues for IPs (cp. section 2) can be applied reliably and consistently in familiar and unfamiliar languages. To further scrutinize this finding, we now turn our focus to individual languages in our corpus and possible differences with regard to the interrater-reliability of IP segmentation on these subcorpora.

Figure 5 show interrater-agreement values for the four larger subcorpora in the format of Figure 4. Interrater agreement is remarkably similar across these four languages, the value for each language being similar to the overall Fleiss' κ of 0.71. The highest Fleiss' κ value is attained for Yali (κ = 0.75), followed by Wooi (κ = 0.74), and then German (κ = 0.72). Interrater agreement on Papuan Malay is somewhat lower (κ = 0.68). The test statistics thus confirm substantial agreement between the five annotators' segmentations of each of these four subcorpora.[9] These results suggest that the familiar vs. unfamiliar language distinction is not the most important factor determining interrater agreement. That is, it does not seem to be necessary to understand spontaneous speech in order to be able to consistently segment it into IPs.

[9] The results on the three minor subcorpora in our 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).

Figure 5: Interrater agreement for individual languages in the corpus.



To conclude, let us see whether statistical patterns for the individual annotators agree with this overall pattern or diverge from it. Table 5 gives an overview of the number and average length of IPs in the segmentations by annotator and language.

Table 5: Number and mean length 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

On first sight, Table 5 would appear to reveal one conspicuous difference between German and the West Papuan languages: German IPs appear to be substantially longer, both overall and for individual annotators, including CONS. This may raise doubts as to the claim that the units identified in all four corpora are of the same granularity, i.e. that they are all IP-sized. Alternatively, the units identified in the West Papuan languages might instantiate another, smaller kind of prosodic phrase (e.g. the so-called phonological or intermediate phrase), which happens to be delimited by the same boundary cues as IPs in German.

However, the difference in mean IP length in words in Table 5 is largely due to differences in grammatical structure and orthographic conventions, i.e. the frequency and the orthographic representation of function words. In German, articles, prepositions and particles such as *ja* and *also*, for example, are very frequent and written as separate orthographic words. Yali enclitic postpositions, on the other hand, form an orthographic unit with their morphosyntactic hosts (e.g. orthographic <inggiken> is morphological *inggik=en* (hand=INSTRUMENTAL) 'with (his) hands'). More generally, the West Papuan languages have fewer function words than German, and many are not written separately.

To lend support to this explanation, we arbitrarily selected 15 IPs from each narrative in the four languages and counted the number of content words per IP. Content words include nouns, verbs (but not auxiliaries), adjectives and lexical adverbs such as *tomorrow* or *boldly* (but not *again, thereafter* and the like which primarily have grammatical or discourse organizing functions). As seen in the fourth column of Table 6, the sample reflects the imbalance in the average number of words per IP across the four languages found in Table 5. However, no comparable imbalance is found with regard to the average number of *content* words per IP (cp. the rightmost column of Table 6). Consequently, the higher average number of words per IP in German must be due to the higher number of orthographically independent function words.

Table 6: Average number of content words per IP per language (based on sample from CONS version)

	No of IPs in sample[10]	No of words in sample	Mean length of IPs (in words)	No of Content Words in sample	Content Words per IP
German	270	1,408	5.20	487	**1.8**
Papuan Malay	300	1,223	4.08	530	**1.8**
Wooi	180	654	3.63	288	**1.6**
Yali	90	303	3.37	162	**1.8**

The data in Table 6 suggest that with regard to content words — and thus informational content — the units delimited in each of the four languages are roughly equivalent. Clearly, this evidence does not settle all questions concerning the cross-linguistic comparability of the units identified by the annotators (cp. section 6.1). Table 6 should, however, give some plausibility to the claim that we are dealing with units of a comparable size (comparable informational content), and allow us to continue to speak of IPs in the further discussion of our results.

[10] Recall from Table 3 that each of the four languages is represented by a different number of narratives in the corpus. As this sample is based on 15 IPs per narrative, the numbers of IPs per language differ quite significantly from each other.

Apart from the difference in the mean length of IPs, statistical trends in Table 5 are surprisingly similar to those in Table 3 for the whole corpus. CONS and R4 again posit a similar number of IPs, resulting in similar mean lengths of IPs also for the four individual subcorpora in Table 5. R1 and R2 again segment the narratives into shorter units compared to the other annotators. There are thus individual differences in annotator behavior that hold across the different subcorpora. This may indicate that segmentation strategies are similar across the four languages.

That this is not necessarily so, however, is shown by R3, who segments the German narratives, which she is able to understand, into IPs with an average length of more than eight words.[11] Boundaries here are preferably placed at clause boundaries,[12] ignoring the fact that clauses in spontaneous speech often are chunked into several IPs. Example (4) illustrates a typical case where R3 fails to identify four IPBs in succession before she posits a boundary in agreement with the other annotators at the end of the whole clause. All other annotators, including CONS, chunk this clause into five IPs.

(4)	*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*	=
on	a	other	bicycle	

<auf	*der* ->	*<auf*	*einer* ->	=
on	the	on	a	

auf	*der*	*staubigen*	*Landstraße*	*entgegen*	(0.9)
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_Flor)

In contrast, R3 behaves more like the other annotators with regard to the three unfamiliar West Papuan languages. This suggests that R3 used different segmentation strategies in familiar vs. unfamiliar languages. Segmentation in the familiar languages more strongly takes into account non-prosodic factors, while segmentation in the unfamiliar languages has to rely exclusively on prosodic cues. The inclusion of non-prosodic factors in IP segmentation may thus increase the potential for disagreements (cp. section 1). While sentence boundaries, for example, are typically also IPBs, the reverse does not hold. This is especially clear in narrative speech, where long strings of syntactically coordinated constructions (*and then … and … and …*) may occur.

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

[12] 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 suggests that it is indeed clause and sentence boundaries that R3 is orienting to rather than the end of declination units.

The data presented in this section show robust interrater agreement for IPB identification across the whole corpus, as well as for individual subcorpora. However, IPBs often coincide with pauses and in the computational literature it has been noted that, among all possible predictors for IPBs, pauses are usually the strongest (e.g. Soto *et al.* 2013). Hence the question arises whether the high interrater agreement is not simply due to the fact that student annotators have made good use of pauses as boundary cues, especially in unfamiliar languages.

5. The significance of pauses

There are different ways in which pauses could have influenced the interrater agreement results reported in the previous section. First, pauses may happen to be better boundary predictors in the West Papuan languages, thereby off-setting the advantages resulting from familiarity with German. Second, annotators may have based their decisions exclusively on pauses in the unfamiliar languages, but on a complex mix of prosodic, syntactic, semantic and pragmatic factors in the familiar German, the fact that interrater agreement is similar across the four languages being due to chance.

In this section, we first describe how we determined pauses and their length in our recordings, then present some raw figures on pause frequencies in our corpus, and finally discuss two logistic regression models incorporating information on pauses.

Pause extraction was based on the CONS version. As the recordings were done under field conditions, they contain substantial noise which made it unfeasible to do this automatically. Instead, pauses were annotated manually during the transcription stage detailed in section 3. Non-silent interruptions such as coughing and sneezing were not included in the statistical model.

Table 7 provides, for each language, the absolute frequency of external and internal pauses, as well as their relative frequency per IP, and their average duration in milliseconds. The last row gives the probability that a pause signals an IPB, calculated as the number of IP-external pauses divided by the number of all pauses in a particular language. This measure is an indication of the reliability of pauses as IPB cues.

Table 7: Frequency of internal and external pauses in the four main subcorpora.

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 duration (in milliseconds)	627	561	1,177	1,005
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 duration (in milliseconds)	435	408	481	325
probability of IPB given pause	0.8448	0.9411	0.9798	0.9817

The last row in Table 7 shows that pauses are more reliable as IPB cues in Wooi and Yali than in Papuan Malay and German. Moreover, the German subcorpus contains fewer external pauses between

IPs than the other subcorpora, with Papuan Malay being somewhat closer to German than to Wooi and Yali. German thus also contains more instances of latching. For internal pauses, the converse holds: German and also Papuan Malay have more internal pauses per IP than the other two languages. Finally, external pauses are on average only about 50% longer than internal pauses in German and Papuan Malay, but more than twice as long in Wooi and Yali, and thus probably more noticeable.

Pauses are thus more robust cues for IPBs in Wooi and Yali than in Papuan Malay and German, both in terms of frequency and duration. However, it is not clear that this difference can be attributed to a systematic difference in linguistic structure. It is more likely due to coincidental properties of the different subcorpora. The Papuan Malay and German subcorpora are, for example, better gender-balanced than the Wooi and Yali subcorpora, which are heavily male-dominated. The Papuan Malay and German speakers are probably also more at ease with the task of retelling a film than the Wooi and Yali speakers, for whom watching films is not part of everyday culture. Note that the duration of internal hesitation pauses does not vary much between languages (cp. Table 7). This suggests that longer external pauses in Wooi and Yali are not simply due to slower speech rates.

The differences in the frequency and length of pauses documented in Table 7 have likely contributed to the good interrater agreement scores in two of the three West Papuan languages, Wooi and Yali. Hence, the core question of this section becomes even more pressing: Have annotators based their boundary decisions in the unfamiliar languages on pauses to a significantly larger degree than in the familiar German, perhaps even exclusively so? Figure 6 shows that this is not the case.

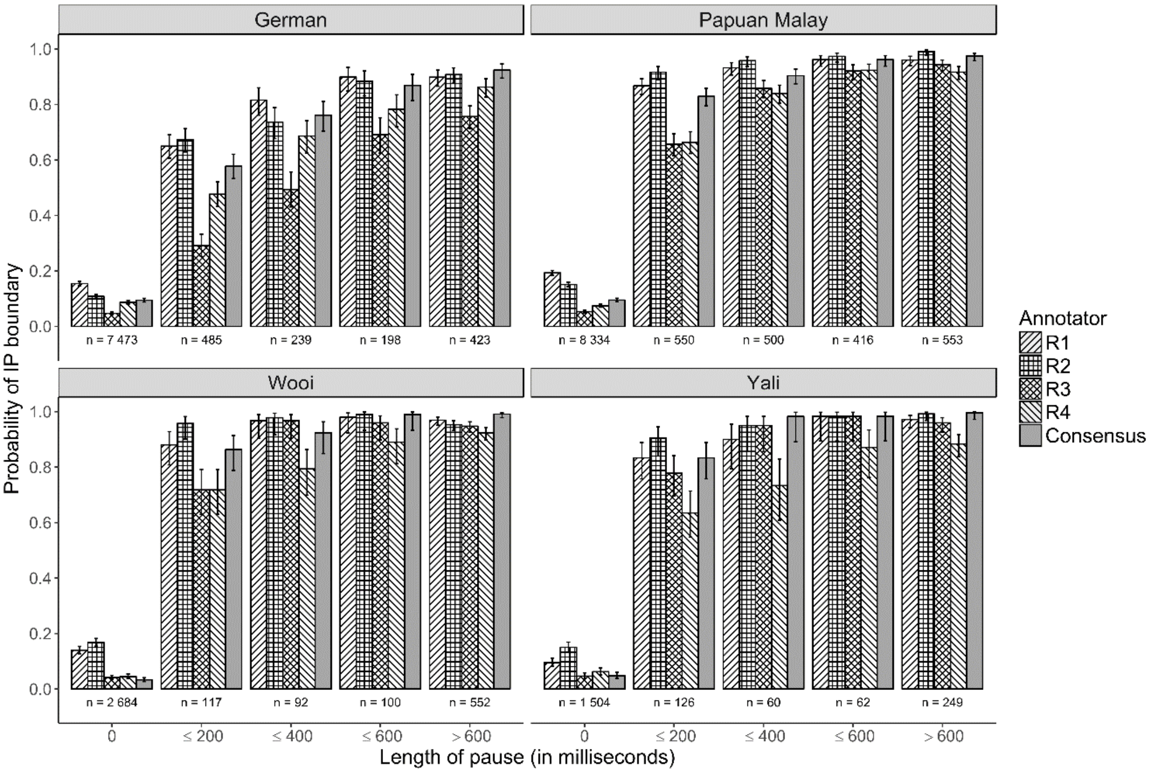
Figure 6 is based on a logistic regression model of our experimental data that predicts the probability of assuming an IP boundary between two words depending on the particular language, the annotator who is making the decision and the length of a possible pause between the two words in question. We decided to code pause length as an ordinal variable with five levels (zero: 0 ms, very short: ≤ 200 ms, short: ≤ 400 ms, medium: ≤ 600 ms and long: > 600 ms) to make it easier to relate the probability of an IPB at a certain pause length category to the actual number of cases in our experimental results that this probability is derived from. Since there are very few cases of long pauses, we put all pauses longer than 600 ms into one category.

We fit our logistic regression model using the function *glm* (generalized linear model) in the R statistical environment (R Core Team 2017), starting with the maximal model including all two- and three-way interactions in addition to the three simple factors. The model formula in expanded form looks as follows:

$$\begin{aligned} IPB \sim\ & pause\ length + annotator + language + pause\ length{:}annotator \\ & + pause\ length{:}language + annotator{:}language \\ & + pause\ length{:}annotator{:}language \end{aligned}$$

We then tested whether the interactions were necessary for a good model fit. The likelihood-ratio test of the three-way interaction indicated that it is required in the model ($\chi^2 = 149.77$, df = 48, $p < 0.001$), which accordingly cannot be further simplified. The high number of factor levels (five levels of pause length, four languages and five annotators) and the inclusion of two- and three-way interactions mean that our model comprises 100 coefficients, making it very hard to discuss it in the usual tabular format. For this reason, we present the modeling results as an effect display (Fox 2003), which, for each language, shows the predicted probability of an IPB for each pause length category and each annotator as a bar graph with confidence intervals based on the model; cp. Figure 6.

Figure 6: Effect display of logistic regression model predicting the assumption of IPBs



The overall trends in Figure 6 are not surprising: Lack of a pause correlates strongly with no IPB, while pauses of 600 ms or longer are associated with a very high likelihood of an IPB. Note that the number of decisions varies substantially across the pause length categories, with the leftmost group of bars representing between 56% and 71% of all decisions made with regard to a given subcorpus.

Three more specific observations are relevant in the current context. First, the correlation between pauses and IPBs indeed varies according to the distribution of pauses in the four subcorpora. It is weakest in German and strongest in Wooi and Yali, with Papuan Malay clustering more strongly with the latter two. Accordingly, the predicted probabilities of an IPB in Figure 6 are lower overall for German and increase more slowly with a higher pause length than in the other three languages for all annotators. The weaker association of IPs with pause length in German, however, is due to the distribution of pauses in the respective corpora (cp. Table 7) and not to the fact that annotators make more use of pauses in the unfamiliar West Papuan languages than in the familiar German.

Second, annotators do not posit IPBs in unfamiliar languages solely on the basis of pauses. Otherwise, one would expect zero probabilities in the case of "no pause" (leftmost group of bars) and a probability of 1 in the case of longer pauses (≥400ms). Instead, the student annotators assume a comparable, though of course relatively low, likelihood of latching cases across all four subcorpora and are even quite constant in their relative propensity to allow for latching. R1 and R2 are more likely to posit IPBs without a pause than R3 and R4 in the familiar German, as well as in all three unfamiliar languages. Conversely, while the predicted probabilities of the student annotators assuming an IPB rise substantially (to above 0.9) for longer pauses in the unfamiliar languages, they are fully in line with, and often even lower than, the respective probabilities predicted for CONS. This suggests that the high probability of assuming an IPB for longer pauses, especially for Wooi and Yali, results from the high reliability of long pauses as IPB cues in these languages (Table 7).

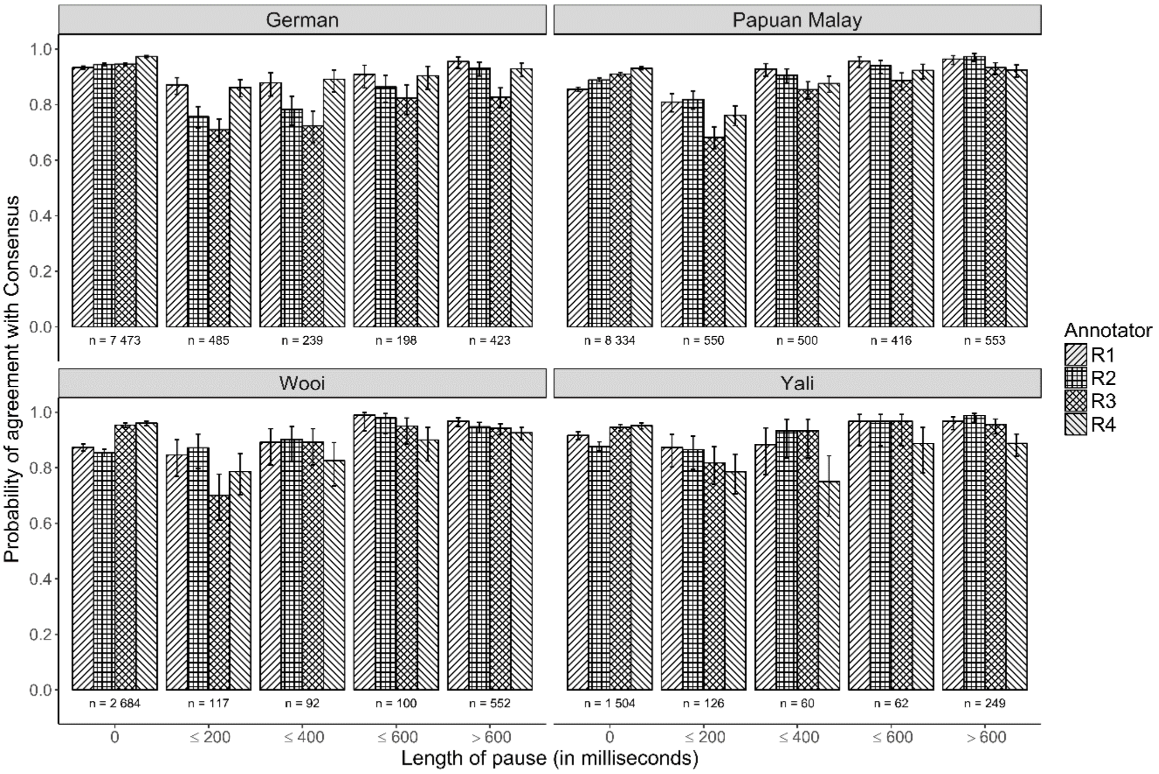
Third, according to the model, the four student annotators in general show a stable tendency to assume more or fewer IPBs compared to CONS across all four languages and, crucially, also across the different pause conditions: R1 and R2 are more likely to posit an IPB than CONS in all languages and for all pause lengths (except for the longest pauses, where CONS sometimes has a higher predicted probability of an IPB and thus seems to be more sensitive to pauses than the student annotators), while R3 and R4 are predicted to be less likely to assume an IPB than CONS in all four languages and for all pause lengths. The observation that R3 segments the familiar German subcorpus according to syntactic and semantic criteria rather than on the basis of prosodic cues alone (cp. section 4) is also reflected in the low sensitivity of R3 to pauses in German (cp. Figure 6).

To compare the four student annotators more directly to the reference segmentation, we fit an additional logistic regression model to our data. This time, however, the dependent variable is agreement with CONS, that is, for each boundary decision, the dependent variable was set to "true" if the student annotator agreed with CONS in that particular case and to "false" if he or she did not. As independent variables, we again included *pause length*, *language* and *annotator* as well as all possible two-way and three-way interactions between them. The model formula in short form is thus:

$$agreement\ with\ CONS \sim (pause\ length + annotator + language)^3$$

A likelihood-ratio test indicated that the three-way interaction is required for a good model fit (χ^2 =110.49, df = 36, p < 0.001) and that the model should not be further simplified. Figure *7* displays the effects of *pause length*, *language* and *annotator* according to the final model. Despite the significance of the three-way interaction, it shows a largely uniform probability of agreeing with CONS across languages and pause lengths. Unsurprisingly perhaps, the probability of agreeing with CONS for individual annotators within one language is reduced somewhat for cases with short pauses (≤ 400 ms) compared to cases without any pause (0 ms) and cases with longer, more noticeable pauses (> 400 ms). This effect is apparent for all student annotators in all four languages. Crucially, however, there is no clear contrast in the pattern of agreement with CONS in familiar versus unfamiliar languages. This is further evidence that student annotators do not exhibit a completely different segmentation behavior in unfamiliar languages by exclusively relying on pauses.

Figure 7: Effect display of model predicting agreement with CONS



This section has shown differences in the distribution of pauses in the four main subcorpora of our study. Specifically, pauses are less useful boundary cues in German and Papuan Malay than in Wooi and Yali. Consequently, the relatively high interrater agreement for the latter two can partly be explained by the fact that here pauses coincide with IPBs to 98% (but the converse does not hold: approx. 20% of the IPBs in these subcorpora lack external pauses). While the predictive power of pauses for IPBs thus varies across languages, there are no clear trends separating familiar from unfamiliar languages. Specifically, there is no evidence that student annotators rely more heavily on pauses in the unfamiliar languages than in their native German. Instead, other boundary cues (pitch, final lengthening, etc.) also play a role in boundary identification and contribute to the overall high interrater agreement across our corpus. In this regard, our results match findings from the automatic boundary detection literature which also find that non-silence features add extra predictive power to boundary classifiers (cp., for example, Soto *et al.* 2013, Table 6).

6. Discussion

The empirical results reviewed in the preceding two sections make it clear that the cues for IPBs provided in our instructions (cp. section 2) 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 identification of these cues is possible even when annotators do not understand the content of the audio signal and are not familiar with the prosodic system of the language in question.

This section discusses how this finding may be explained and what it implies for our understanding of prosodic phrasing. Staying strictly on the level of (phonetic) boundary cues, one could argue that there is not much to explain here. What our data show is that German listeners are able to identify the kinds of prosodic cues they are familiar with from their native German across a range of diverse and unrelated languages. This may be mildly interesting when compared to the ability of German speakers to identify other kinds of phonetic phenomena across unfamiliar languages (e.g. a specific consonant or vowel), but otherwise it would appear to be largely devoid of theoretical interest. The findings become theoretically relevant on the assumption that our annotators identify prosodic units of the same basic type, i.e. IPs, across unrelated languages. This assumption of 'sameness' can be challenged (and has been challenged by almost all reviewers for PHONOLOGY) on two interrelated grounds. First, the same kind of cues might identify different kinds of units in unrelated languages, an issue taken up in section 6.1. Second, it might be the case that native speakers of other languages hear completely different things and that, therefore, the units identified are essentially *German* perceptual IPs and irrelevant to the native speakers of the unfamiliar languages. We address this issue in section 6.2.

Inasmuch as we succeed in countering these challenges, our findings suggest the hypothesis that there is a universal phonetic basis to IP chunking that allows speakers to identify IPs across familiar and unfamiliar languages. Section 6.3 briefly expounds this hypothesis, pointing out some of the empirical and theoretical issues that need to be resolved to further substantiate it.

6.1 On the cross-linguistic comparability of prosodic units

The challenges in comparing grammatical categories across languages are well-known in typology and have recently again become a major concern in the field (e.g. Lazard 2002, Haspelmath 2010). With regard to prosodic units, Hyman's (2015) examination of the evidence for syllables in Gokana is an instructive example. We cannot provide a comprehensive discussion on the cross-linguistic comparability of prosodic units here, but will try to give plausibility to the claim that the units identified by our annotators are the 'same' across the languages of the sample.

The core issue with regard to our data pertains to a specific region of the prosodic hierarchy, i.e. the level of IPs and the next lower level, widely known as *Phonological Phrase* (the term we will use and abbreviate as PhP) or *intermediate phrase*.[13] We thus assume that the units delimited by all annotators are larger than syllables and phonological words but smaller than utterances, paragraphs or other kinds of macro units proposed above the IP. It is a matter of controversy how many levels need to be assumed between phonological words and IPs and whether such levels are actually found in all languages. In this regard, we clearly side with the arguments against a proliferation of prosodic levels

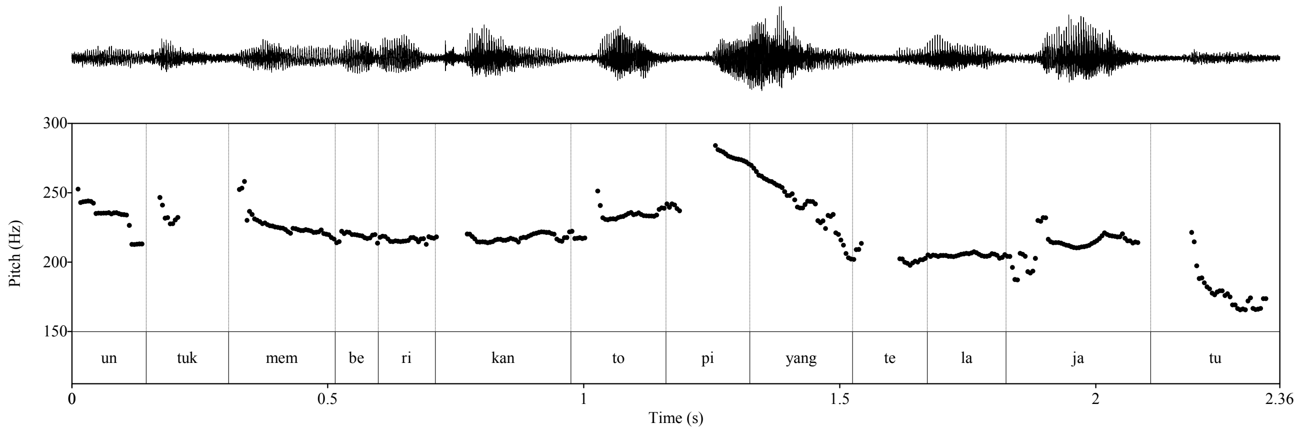
[13] We do not discuss the next lower level known as *minor* or *accentual phrase*, as our units tend to be longer than the one or two phonological words usually constituting an accentual phrase in the prototypical exemplar languages Korean and Japanese.

and the requirement that each level be defined by specific properties that distinguish the boundary of this level from boundaries at other levels (cp. Ladd 1986, 2008:288-299; Frota 2000; Tokizaki 2002; Wagner 2010; Krivokapić & Byrd 2012:438).

A case in point are the highly conspicuous and systematic PhP boundaries occurring in two of our languages. In Papuan Malay and Wooi, IPs are optionally further segmented into PhPs, which are marked by a high tone on the final syllable of the phrase. Importantly, PhP boundaries in these two languages do *not* involve pause and pitch reset. The overall melodic and rhythmic coherence is thus not interrupted, as illustrated by the Papuan Malay example (5). Here the high PhP boundary tone on *pi* (the final syllable of *topi* 'hat' which functions as head noun for the following relative clause) is immediately followed by a fall that continues across the next word, i.e. the relative pronoun *yang*. PhP-final syllables may be slightly lengthened, but this is the exception rather than the rule and not found in example (5). IPBs, on the other hand, are generally followed by a new onset in pitch and often by a pause, i.e. they involve an interruption of rhythmic and melodic coherence. Additionally, in both Wooi and Papuan Malay, IPBs involve *two* tonal targets, a phrase accent and a final boundary which occurs in a two syllable window from the end of the unit (cp. section 3). Example (5) illustrates this with the combination of a high phrase accent and a falling boundary tone on the final verb *jatu* 'fall'. Both penultimate and ultimate syllables tend to be considerably lengthened.

(5) *untuk memberikan topi yang tela jatu*
for ACT:give:APPL hat REL already fall
'to give back the hat that had fallen down' (pear_Virgin2)

Figure 8: Waveform and f_0 extraction of example (5)



Crucially, the boundary strength *within* each unit type may vary and such differences may be perceived by listeners (Ladd 2008:293-297; Wagner & Watson 2010; and Krivokapić & Byrd 2012, inter alia). As noted in section 2, IPBs without pauses are more difficult to perceive than ones where pause, pitch reset, final syllable lengthening and possibly other features such as creaky voice and fading intensity all point to a major prosodic boundary. This is clearly reflected in our interrater agreement data where disagreements rarely arise at such clearly marked boundaries.

Given the variability in boundary strength and the fact that many boundary cues are highly language-specific (such as the edge-tone combination just illustrated for Papuan Malay), it is not surprising that some annotators occasionally interpreted PhP boundaries as IPBs. In fact, R1 – the student annotator with the shortest IPs on average (cp. Table 3) – had a tendency to mark PhP boundaries occurring within larger IPs.

Turning to our segmentation data, our expert segmentation (CONS) distinguishes PhPs from IPs in three of the four languages, i.e. German, Papuan Malay and Wooi. Importantly, PhP boundaries in these languages do not involve the interruption of melodic coherence (pitch jumps) and are thus clearly distinguished from IPs.[14] Insofar as our analyses of these languages are correct, it follows that the units identified as IPs are larger than PhPs in all three languages and, moreover, also comparable with regard to the phonetic boundary cues used in our segmentation instructions. Thus, our first argument for the claim that the units identified in the different subcorpora are of the same type is that the expert annotation followed standard procedures in prosodic analysis, using standard criteria for distinguishing prosodic phrasing levels, and that for three of the four languages the same two major phrasing levels above the phonological word were used. Inasmuch as the student annotators' segmentations match the expert annotation across the four subcorpora (cp. Figure 7), they also target the same phrasing level, i.e. IPs. This argument may be less forceful for Yali, where we do not assume an additional phrasing level between phonological word and IP.

This type of argument implicitly underlies all cross-linguistic work on prosody and particularly typological collections such as Jun (2005, 2014). In these collections, the prosodic descriptions of *all* languages assume an IP level without explicitly arguing for the cross-linguistic comparability of the language-specific IP constructs. The tacit assumption appears to be that if the same procedures are followed in the analysis of two or more languages then the postulated units of the same name are at least roughly comparable.

Still, use of the same analytical framework and procedure may not be deemed to be sufficient to support cross-linguistic sameness. How can one be sure that levels of the same name really have the same status and function in two different prosodic systems? Phonetic similarities and analytic consistency may be suggestive, but they hardly constitute full proof. Other, preferably non-prosodic parameters for assessing similarity are needed to further substantiate claims of cross-linguistic similarity.

We already proposed one such parameter in section 4, addressing differing mean lengths of IPs across the four main subcorpora. The data in Table 6 show that the units are of a comparable size with regard to their informational content, i.e. they contain on average 1.6-1.8 content words. This informational measure is relevant on the widely shared assumption that IPs are major processing units in speech production and comprehension. There are very few proposals how to specify the

[14] For German, we follow the GToBI analysis as described in Grice *et al.* 2005. See also http://www.gtobi.uni-koeln.de/index.html). Note also that PhPs in all three languages are delimited by a single edge tone, while IPs involve a combination of two edge tones.

informational content of IPs, among them Chafe's (1994:108–119) proposal that IPs present exactly one 'new idea'. But there is wide agreement that IPs represent informational 'chunks' that the speaker processes as one unit and presents to the hearer as such (cp. Sanderman & Collier 1997, Frazier *et al.* 2006, Krivokapić 2007, Wagner & Watson 2010). It is unclear to what extent this also holds for lower level prosodic constituents such as PhPs.

A second non-prosodic parameter for cross-linguistic comparison is size variability. The units identified in our segmentation data are highly variable as to their size, ranging from discourse particles and short phrases without content words, to NPs or PPs, to clausal and multi-clausal units. This is typical for IPs, whereas lower-level prosodic units are more regularly associated with syntactic constituents of a narrowly delimited size. Langus *et al.* (2012: 286) explicitly contrast the PhP and the IP in this regard and note that the IP is "a more variable constituent as to its domain".

In sum, there are good reasons to assume that the units identified in our segmentation experiment are essentially of the same kind across familiar and unfamiliar languages. A fundamental challenge to the line of argument presented in this section, however, is that all of the above does not prove that the units identified in our experiment are relevant and perceptible for the native speakers of the West Papuan languages. It might well be that we are consistently identifying IP-sized units across the four languages but that these units are constructs of an analytical framework based on European languages and that the West Papuan speakers are sensitive to substantially different kinds of segmentation cues and thus possibly also arrive at substantially different segmentations. The next section will address this objection.

6.2 What do the native speakers of the West Papuan languages hear?

To fully counter the objection that our findings only show that German speakers hear German IPs, one would have to replicate the experiment with native speakers of the West Papuan languages. A replication using the same corpus with speakers of the three West Papuan languages, however, is not straightforward for a number of practical reasons, including the substantial size of the corpus (> 3 hours). Crucially, the practical orthographies used for the West Papuan languages were relatively easy to process for the German annotators as the phoneme-grapheme correspondences are very regular and easily identifiable for them. German listeners could relatively easily match the audio recording with the transcript. German orthography, on the other hand, is not so easy to process for the West Papuan speakers. Furthermore, levels of literacy, and in particular the computer literacy needed to handle the ELAN program, vary dramatically among the West Papuan speakers and, for both Wooi and Yali, it would be difficult to find enough speakers who could engage in tasks requiring the processing of written language.

Still, we have been conducting pilot experiments with speakers of Papuan Malay to determine ways to collect comparable interrater agreement data for this population. These pilot experiments more closely follow the procedures of Mo *et al.* (2008), using smallish sets of excerpts of spontaneous speech and having speakers mark boundaries on printouts of the transcripts of these excerpts (cp.

section 2). Most importantly, following the *Rapid Prosody Transcription* (RPT) method (Cole & Shattuck-Hufnagel 2016), speakers were allowed to hear each excerpt only twice. Consequently, the results of these pilot projects are not directly comparable to the current results. However, they provide at least some support for the claim that Papuan Malay speakers can make use of the same boundary cues as those used in the segmentation task reported here and arrive at roughly the same kinds of units as the German annotators.

One of these pilot experiments is reported in Riesberg *et al.* (in press) which investigates both prominence and boundary perception with the RPT method. 22 speakers of Papuan Malay annotated transcripts of 56 excerpts of spontaneous narrative and conversational speech produced by 28 different speakers of Papuan Malay. While interrater agreement for prominence was negligible (Fleiss' κ of 0.103), interrater agreement for boundaries (Fleiss' κ of 0.407) was within the range found in comparable studies for English.

Thirty eight of the fifty six excerpts used in this experiment come from the Papuan Malay pear stories also used here. Hence, we can compare the units identified by the Papuan Malay speakers with those in our CONS version as well as with the units identified by our student annotators. This allows us to calculate interrater agreement statistics within and across the different groups of annotators. Table 8 provides agreement statistics within the group of Papuan Malay native speakers and within the group of German student annotators, respectively. In addition, it also shows mean κ values for agreement between members of each of these two groups and our consensus version, respectively. Finally, we also computed agreement across groups by comparing the boundary decisions of each of the Papuan Malay native speakers with those of each of our German student annotators.

Table 8: Interrater agreement within and between different groups of annotators on 56 excerpts of Papuan Malay pear stories (480 boundary decisions)

Agreement within groups	Papuan Malay speakers	German students
raters	22	4
Fleiss' κ	0.399	0.569
pairs of raters	231	6
mean of Cohen's κ (Light's κ)	0.405	0.571
std. dev. of Cohen's κ	0.197	0.083

Agreement with consensus	Papuan Malay speakers	German students
raters	22 + 1 (CONS)	4 + 1 (CONS)
pairs of raters	22	4
mean of Cohen's κ	0.478	0.601
std. dev. of Cohen's κ	0.183	0.050

Agreement across groups	Papuan Malay speakers vs. German students
raters	22 + 4
pairs of raters	88
mean of Cohen's κ	0.396
std. dev. of Cohen's κ	0.148

As noted above, the Fleiss' κ statistic for agreement within the group of Papuan Malay native speakers (0.399) is comparable to results obtained in similar studies for English. Interrater agreement among the four German student annotators is clearly higher (Fleiss' κ 0.569). This difference in agreement values is likely due to the differing experimental methods and the stricter time constraints the native speakers were subjected to in the RPT approach. In addition, it may be due to the fact that German annotators based their decision exclusively on phonetic cues for IPBs, while the Papuan Malay speakers probably also made use of syntax, semantics and pragmatics.

A direct comparison of the IP segmentations created by Papuan Malay native speakers with our consensus segmentation results in a mean κ value of 0.478, representing moderate agreement according to Landis & Koch (1977). This suggests that our consensus segmentation does agree to a large extent with intuitions of native speakers and does not constitute a completely irrelevant German-based IP segmentation of the data.

The comparison across the native and non-native groups of annotators in the bottom of Table 9 also supports this conclusion. The mean agreement between all different pairs of one Papuan Malay native speaker annotator and one German student annotator (mean κ = 0.396) is quite close to the mean agreement among Papuan Malay native speakers themselves (mean κ = 0.405). This indicates that the native and non-native speakers segment the Papuan Malay speech into comparable units. It also supports the conjecture that the higher agreement among German annotators is due to differences in the experimental methods.

We are currently also running an experiment where Papuan Malay speakers identify IPBs in excerpts of the German pear stories used here. Preliminary results suggest that we again find substantial interrater agreement between the segmentations produced by German and by Papuan Malay speakers. We therefore believe that it is plausible to assume that the units identified in our experiment are not only the same prosodic analytical constructs (i.e. IPs), but that speakers from different populations would arrive at similar segmentations, given the same instructions. Obviously, this hypothesis is in need of further empirical scrutiny. We nonetheless conclude our study with a brief exploration of the theoretical implications that arise if it can be shown to be empirically well supported.

6.3 The Universal Phonetic IP Hypothesis

Strictly speaking, the student annotators in our experiment did not identify phonological units, at least not in 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. The current study thus differs sharply from the kind of interrater agreement study briefly mentioned in section 2, where annotators are trained to identify phonological categories defined within a specific framework such as ToBI. The claim made repeatedly throughout this paper – 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 (cp. sections 3 and 6.1).

At least for the languages under investigation, the current study therefore shows that IPs can be consistently identified in spontaneous speech without being familiar with their phonological structure, simply on the basis of phonetic boundary cues which appear to be not specific to a particular language. This finding can be interpreted in a number of ways. In the two preceding subsections, we have argued against the view that it only shows that German speakers are able to identify German IPs everywhere. Instead, we propose that it points to what could be called the *Universal Phonetic IP Hypothesis* (UPIPH). According to this hypothesis, all natural languages make use of the same kinds of phonetic cues for IPs and these cues can be perceived by speaker-hearers even in unfamiliar languages. The main cues are the interruption of melodic coherence as manifest in pitch resets between IPs and major rhythmic breaks, in particular pauses. Both types of cues are considerably more complex than just stated and involve language- and probably also speaker-specific further features.

In addition, IPs may be – and usually are – phonologically organized units, with the phonological organization targeting in particular tonal events. The prototypical example of this are specific edge tones which are the clearest *phonological* markers for prosodic boundaries and tend to be intricately interlinked with segmental articulatory gestures (e.g. Krivokapić & Byrd 2012). The phonological organization may include further grammaticized (regularized) variants of the universal phonetic IPB cues. In this view, IP boundary tones, for example, are regularized (grammaticized) variants of the universal pitch resets associated with the interruption of melodic coherence.

We propose to conceive of the relation between universal phonetic IPs and language-specific phonological IPs 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 physiology of speaking (e.g. breathing) as well as cognitive demands on speech planning and processing (cp. section 6.1). The physiology of speaking and processing demands are also the source of the universal melodic and rhythmic boundary characteristics of the universal phonetic IP, specifically melodic coherence and processing-related interruptions of speech delivery (planning pauses and unit-final lengthening).[15] These boundary characteristics can be further grammaticized into language-specific phonological categories, giving rise to a phonologically organized category *intonational phrase*. Such grammaticizations 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 grammaticized pitch movements serving other functions such as marking information status (postlexical pitch accents) or distinguishing lexemes (lexical tones).

[15] It is therefore highly likely that these boundary cues are also instances of the kind of language-general cues required for language acquisition.

Note that this scenario specifically targets IPs. It does not necessarily apply to other levels of prosodic phrasing. Thus, for example, to support the claim that there are also universal phonetic PhPs, it would be necessary to identify a distinct set of phonetic cues for PhP boundaries, which should likewise be derivable from aspects of speech physiology or processing (cp. section 6.1).

As for IPs, we believe that it is quite likely that phonological IPs are part of the prosodic system of all natural languages. If true, IPs would be a prime example of a universally attested *phonological* category (in addition to being a universally attested phonetic category). Such a claim, however, presupposes not only the analysis of the prosodic systems of all languages, but also that the units labelled IPs in these analyses are cross-linguistically comparable with regard to independent parameters such as informational content and size variability (cp. section 6.1). In principle, however, the UPIPH allows for the possibility that languages exist where spontaneous speech is produced in IP-sized chunks (delimited by the universal phonetic boundary cues), but where the phonological analysis of the prosodic system does not require (or support) an IP level. More importantly, perhaps, the hypothesis predicts that IP units and their boundaries are grammaticized to different degrees, i.e. that prosodic systems exist where the IP level is but weakly grammaticized, its structure consisting simply of a single final boundary tone, for example.

The UPIPH is, of course, in need of further conceptual and empirical scrutiny. Empirically, it makes the prediction that segmentation tasks of the type employed in this study will result in substantial interrater agreement across every combination of languages, speaker populations and speaking styles (an obvious limitation of this study is its restriction to narrative speech). Unlike in the current study, native speakers of all languages represented in the sample should ideally also be included among the annotators.

While the current sample covers a range of prosodic systems (cp. section 3), crucial test cases are still to be investigated. Syllable tone languages such as Mandarin or Thai, for example, may provide particular challenges. In such languages, tonal sandhi may provide a conspicuous cue to melodic coherence, and it remains to be seen whether non-native annotators can make good use of that. Conversely, it may turn out that (monolingual) Mandarin or Thai native speakers encounter difficulties in segmentation tasks involving German or Wooi data where these tonal sandhi cues are absent.

The empirical examination of the UPIPH is not restricted to the exact task design used in this study, which would not be feasible in many speech communities for the reasons noted in section 6.2. In fact, it is not restricted to segmentation tasks targeting IPBs and referring to the universal phonetic cues of melodic and rhythmic coherence. In principle, it should apply to any kind of evidence associated with phonetic IPs. Thus, for example, if the (auditory) processing of IPs indeed involves a brain signature of the type proposed by Steinhauer *et al.* (1999), who claim that IPBs are associated with a *Closure*

Positive Shift,[16] then one would expect this signature to occur across a world-wide sample of speakers and languages.

Conceptually, it needs to be further clarified and empirically tested whether and how the presumed universal phonetic boundary cues are linked to the physiology of speaking and the cognitive demands on speech processing (cp. section 6.1). A fully explicit account of this link should also cover the complex interplay between the two basic phonetic cue types for IPBs (melodic and rhythmic) that has been discussed throughout the preceding sections.

7. Summary

The present work has provided evidence for the following claims:

1) Intonational phrases are empirically viable units according to standard measures for interrater agreement. Multi-rater as well as pair-wise κ coefficients show a substantial and statistically significantly above chance agreement on the placement of IPBs and thus demonstrate the reliability of IP segmentation. This holds both for languages familiar and unfamiliar to the annotator (cp. section 4).
2) IPB 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) Melodic coherence, pauses, unit-final lengthening and increased unit-initial speaking rate are universal cues for IPBs. On the basis of these cues, it is possible to segment narratives in unknown languages with roughly the same reliability as in one's native language.
4) The empirical findings support the hypothesis of universal phonetic IP chunking linked to the physiology of speaking and the cognitive demands on speech processing. Languages differ as to whether and to what degree phonetic IPs are further grammaticized into phonological IPs, which are language-specific structural units arising from, and continually undergoing, processes of diachronic change.

References

Beckman, Mary, Julia Hirschberg & Stefanie Shattuck-Hufnagel (2005). The Original ToBI System and the Evolution of the ToBI Framework. In Jun (2005). 9–54.

Boersma, Paul & David Weenink (2015). *Praat: doing phonetics by computer* (version 5.4.09). Available 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*

[16] Li *et al.* (2008) claim that this signature occurs with both PhPs and IPs in Chinese, though with different onset and peak latencies.

and Evaluation (LREC). Paris: Evaluations and Language Resources Distribution Agency. 779–785.

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

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

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.

Cole, Jennifer & Stefanie Shattuck-Hufnagel (2016). New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology* **7**. 1–29.

Dilley, Laura C. & Meredith Brown (2005). The RaP (Rhythm and Pitch) Labeling System. Version 1.0. Available at https://pdfs.semanticscholar.org/5f73/1dbcafb2b64da6eb15daa67718866bc74cc9.pdf.

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.

Fox, John (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software* **8**. 1–27. Available at http://www.jstatsoft.org/v08/i15/.

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

Frota, Sónia (2000). *Prosody and focus in European Portuguese. Phonological phrasing and intonation*. New York: Garland Publishing.

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

Grice, Martine, Stefan Baumann & Ralf Benzmüller (2005). German Intonation within the Framework of Autosegmental-Metrical Phonology. In Jun (2005). 55–83.

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.

Haspelmath, Martin (2010). Comparative concepts and descriptive categories in cross-linguistic studies. *Language* **86**. 663 –687.

Heeschen, Volker (1992). *A dictionary of the Yale (Kosarek) language (with sketch of grammar and English index)*. Berlin: Reimer.

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.

Hyman, Larry M. (2015). Does Gokana really have syllables? A postscript. *Phonology* **32**. 303–306.

Kamholz, David C. (2014). *Austronesians in Papua: Diversification and change in South Halmahera-West New Guinea*. PhD dissertation, UC Berkeley.

Katsika, Argyro, Jelena Krivokapić, Christine Mooshammer, Mark Tiede & Louis Goldstein (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics* **44**. 62–82.

Krivokapić, Jelena (2007). *The planning, production, and perception of prosodic structure*. PhD dissertation, University of Southern California.

Krivokapić, Jelena (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B* **369**. 20130397. Available at http://rstb.royalsocietypublishing.org/content/369/1658/20130397.

Krivokapić, Jelena & Dani Byrd (2012). Prosodic boundary strength: An articulatory and perceptual study. *Journal of Phonetics* **40**. 430–442.

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 (1986). Intonational phrasing: The case for recursive prosodic structure. *Phonology* **3**. 311 –340.

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.

Langus, Alan, Erika Marchetto, Ricardo A.H. Bion & Marina Nespor (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language* **66**. 285–306.

Lazard, Gilbert (2002). Transitivity revisited as an example of a more strict approach in typological research. *Folia Linguistica* **36**. 141–190.

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

Li, Weijun, Lin Wang, Xiaqing Li & Yufang Yang (2008). Closure Positive Shifts Evoked by Different Prosodic Boundaries in Chinese Sentences. In Rubin Wang *et al.* (eds.) *Advances in Cognitive Neurodynamics ICCN 2007*. Dordrecht: Springer. 505–509.

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

Mo, Yoonsook, Jennifer Cole & Eun-Kyung Lee (2008). Naive listeners' prominence and boundary perception. In P. A. Barbosa *et al.* (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/.

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. *Proceedings of the 1994 International Conference on Spoken Language Processing (Yokohama, Japan)*. 123–126.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Remijsen, Bert (2001). *Word-prosodic systems of Raja Ampat languages*. PhD dissertation, Leiden University: LOT Dissertation Series **49**.

Remijsen, Bert & Vincent J. van Heuven (2005). Stress, tone, and discourse prominence in the Curacao dialect of Papiamentu. *Phonology* **22**. 205–235.

Riesberg, Sonja (2017). An Introduction to the Yali-German Dictionary with a Short Grammatical Sketch. In Sonja Riesberg (ed.) *Wörterbuch Yali (Angguruk) – Deutsch*. Canberra: Pacific Linguistics. http://hdl.handle.net/1885/127381.

Riesberg, Sonja, Janina Kalbertodt, Stefan Baumann & Nikolaus P. Himmelmann (in press). On the perception of prosodic prominences and boundaries in Papuan Malay. In S. Riesberg, A. Shiohara & A. Utsumi (eds.) *A cross-linguistic perspective on information structure in Austronesian languages*. Berlin: Language Science Press.

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/.

Sanderman, Angelien A. & René Collier (1997). Prosodic phrasing and comprehension. *Language and Speech* **40:4**. 391–409.

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. *Proceedings of the 1992 International Conference on Spoken Language Processing (Banff, Canada)*. 867–70.

Soto, Victor, Erica Cooper, Andrew Rosenberg & Julia Hirschberg (2013) *Cross-Language Phrase Boundary Detection*. ICASSP Vancouver, Canada. http://makino.linguist.jussieu.fr/idp09/actes_fr.html/.

Steinhauer, Karsten, Kai Alter & Angela D. Friederici (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience* **2**. 191–196.

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

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

Tokizaki, Hisao (2002). Prosodic Hierarchy and Prosodic Boundary. *Bunka-to Gengo* (Sapporo University) **56**. 81–99.

Wagner, Michael (2010). Prosody and recursion in coordinate structures and beyond. *Natural Language & Linguistic Theory* **28**. 183–237.

Wagner, Michael & Duane G. Watson. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes* **25**. 905–945.

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

PREPUBLICATION DRAFT

SUPPLEMENTARY MATERIALS to

On the universality of intonational phrases – a cross-linguistic interrater study

Nikolaus P. Himmelmann, Meytal Sandler, Jan Strunk & Volker Unterladstetter
Universität zu Köln

Supplement 1. Details of the instructions given to the student annotators (section 2)

Our written instructions regarding IPB cues, given to the annotators and explained once verbally, read as follows 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>Wissenswertes</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.

English translation:

> Your task is to segment an audio recording containing the narrative 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 characterized 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 speaking 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 (*uhm*, etc.) but not necessarily so. What is important is that the pitch levels before and after a 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 to illustrate the following typical configurations at IPBs:

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 increased speaking rate at the beginning of the second IP.

The examples for these configurations 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. Reference to boundary tones in the above descriptions 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.

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, even though all languages in our corpus use them. The concept of a boundary tone only makes sense in a theoretical model, knowledge of which we could not presuppose on the part of the participants in this study. Nor do our instructions refer to boundary cues that are difficult to perceive without specific measurements such as domain-initial strengthening (cp. Fougeron & Keating 1997, Keating et al. 2004).

Fougeron, Cécile & Keating, Patricia A. (1997). Articulatory strengthening at edges of prosodic domains. *JASA* **106**. 3728-3740.

Keating, Patricia A., Taehong Cho, Cécile Fougeron & Chai-Shune Hsu (2004). Domain-initial articulatory strengthening in four languages. In J.K. Local, R. Ogden and R.A.M. Temple (eds.) *Papers in Laboratory Phonology VI: Phonetic interpretation*. Cambridge: CUP. 145–163.

Supplement 2. Further details on data and procedure (section 3)

Recording procedure: One person watched the pear film on a laptop screen and then recounted 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) back channeling, only very few actually asked clarification questions, never exceeding three questions in one telling. All verbal utterances made by 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.

With the exception of a few German recordings mentioned below, all recent recordings were done with a Sony digital video recorder (e.g. HDR-CX730E or similar) mounted on a tripod and an external microphone (in most instances, a stereo on-camera condenser microphone).

Corpus compilation: The corpus used in this study was originally compiled for the AUVIS project (*Audiovisual data-mining using event segmentation in multimodal language data as an example*, cp. https://tla.mpi.nl/projects_info/auvis/ for more information). The main goal of this project was to explore possibilities for automatically annotating and searching audio and video streams of unannotated or only partially annotated recordings from unrelated languages, with a particular focus on under-documented and under-resourced languages. 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.

The version of the *AUVIS Corpus* used in the current study 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.

The first group of recordings in Table 1 consists of eighteen pear film narratives in (Standard colloquial) German, one narrative in the vernacular dialect of Cologne (Kölsch) and one narrative in (American) English. Six of these narratives were recorded with analog audio and video recorders in the 1990s and are therefore of somewhat lower quality, especially with regard to the video (which did not play a role in the current study). The remaining narratives were recorded in 2012 with up-to-date audio/video equipment 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 narratives in Papuan Malay, the *lingua franca* of West Papua, the western half of the island of New Guinea governed by Indonesia (see Kluge 2017 for a recent description). The pear film narratives in Papuan Malay were recorded at the Center for Endangered 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. Cp. Riesberg & Himmelmann (2012–2014).

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, Sawaki 2016) and Waima'a (Belo *et al.* 2002–2006), are Austronesian languages 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 narratives in Wooi and Waima'a were all recorded in the field sites and are generally of a lower quality than the recordings done at the CELD (there are more background noises of different kinds). The age of the Wooi speakers is more mixed than in the other language groups, ranging from speakers in their early twenties to mature speakers of 50 years and older. The third language, Yali (Riesberg *et al.* 2012–2016, Riesberg 2017), 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.

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/.

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

Sawaki, Yusuf (2016). *A Grammar of Wooi: An Austronesian Languages of Yapen Island, Western New Guinea*. PhD dissertation, Australian National University.

Riesberg, Sonja (2017). An Introduction to the Yali-German Dictionary with a Short Grammatical Sketch. In Sonja Riesberg (ed.) *Wörterbuch Yali (Angguruk) – Deutsch*. Canberra: Pacific Linguistics. Available at http://hdl.handle.net/1885/127381.

Riesberg, Sonja & Nikolaus P. Himmelmann (2012-2014). *Papuan Malay. Summits-Page Collection.* DoBeS Archive MPI Nijmegen. Available at http://www.mpi.nl/DOBES/.

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

Experimental Procedure: The ELAN file given to the annotators contained two annotation tiers, one for the narrator, and one for the interlocutor. To facilitate orientation within the recording, we left the utterances of the interlocutors in place and included them on separate lines in the plain text transcription file. Note that interlocutor utterances were few and far between, in particular in the West Papuan narratives. More than half of the latter do not include any interlocutor interventions and such interventions rarely exceed half a dozen per retelling. Thus, even if such interventions may have influenced annotator decisions by triggering boundary decisions at intervention points, the overall influence of interlocutor utterances on the task is negligible.

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 selection was made in the waveform view of the audio file that is part of the standard annotation setup in the ELAN program.

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 narratives in packages per group, starting with Group I (Germanic), then Group II (Papuan Malay), and finally Group III (Eastern Indonesian languages). The labels of the packages included language names, and each narrative was clearly labeled as to the language used, but no further information on the languages was provided.

The order of the narratives in a group was alphabetical based on 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 in roughly 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.

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

Statistical procedures: Since the task of the annotators was to segment into IPs a given transcription of a narrative which we provided to them in a practical orthography including word boundaries, we can treat the IP segmentation task as a binary classification: Between each consecutive pair of words in the transcription, the annotators can either posit an IPB or not. For a transcription containing *n* words, there are (*n* - 1) consecutive word pairs and thus (*n* - 1) potential IPBs. We 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.

In practice, annotators occasionally forgot to copy and paste a word from the transcription into the ELAN time line or accidentally 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 IPBs 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 least accurate annotator (R3) made 200 copy-and-paste errors in all, amounting to about 3 errors per narrative.

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 IPB 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 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 case, *boundary* vs. *non-boundary*). In order to assess overall agreement between all annotators, we use Fleiss' κ (Fleiss 1971). In addition, we compare the student annotators' segmentations individually to our consensus segmentation (CONS) using Cohen's κ (Cohen 1960) for pairwise comparisons, 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).

Where appropriate, we 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 narratives in our corpus and by comparing them using non-parametric statistical tests. In most cases, we use the so-called Wilcoxon-Mann-Whitney rank sum test (Wilcoxon 1945; Mann & Whitney 1947) for unpaired samples. We assume the conventional significance level of $p \leq 0.05$ throughout.

In section 5, we additionally use multivariate logistic regression to investigate the student annotators' reliance on pauses (of different lengths) in familiar versus unfamiliar languages.

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

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

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.

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