Phonology **35** (2018) 207–245. © Cambridge University Press 2018
doi:10.1017/S0952675718000039

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

Nikolaus P. Himmelmann
Meytal Sandler
Jan Strunk
Volker Unterladstetter
University of Cologne

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.

* E-mail: SPRACHWISSENSCHAFT@UNI-KOELN.DE, MEYTAL.SANDLER@GMAIL.COM, JAN.STRUNK@UNI-KOELN.DE, VOLKER.UNTERLADSTETTER@UNI-KOELN.DE.

We are very grateful to four anonymous reviewers for *Phonology*, the associate editor, Bob Ladd, and the editors for extensive, detailed and constructive comments, questions and suggestions, which have led to major revisions. 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 analysed here.

Authors' contributions: NPH: overall design of study and paper, main author of §1, §2, §6 and §7, final revision of all other sections; MS: contributor to interrater study (including recordings), main administrator of interrater study; JS: statistical analyses, draft of §4 and §5, contributor to interrater study; VU: draft of §3, contributor to interrater study (including recordings). All authors contributed to the consensus version.

Research for this paper was funded by grant 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 from 2002 to 2016 supported the compilation of the West Papuan corpora used here. See Appendix B for details.

The appendices are available as online supplementary materials at https://doi.org/10.1017/S0952675718000039.

1 Introduction

Spoken language is produced in chunks delimited by prosodic cues such as a coherent intonation contour and pauses. These chunks are recognised in all models of prosodic analysis, albeit with different names and definitional criteria. TONE GROUP (Halliday 1967) and INTONATION UNIT (Chafe 1980, 1994) are widely known, alongside INTONATIONAL 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 'perceptually, the boundaries of an Intonational Phrase are quite clear', while 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.

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

In contrast, the current study is concerned exclusively with IPBs, and involves the segmentation of a corpus of more than three hours of spontaneous narrative speech (see Table I in §3). 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 familiarising 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 (e.g. English) and applied to data from another language (e.g. Mandarin). Results are often surprisingly good, in that boundary classifiers trained on foreign language data achieve results similar to those 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 for human annotators show important parallels to this line of work, which are discussed in §6.

This 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 a short film in four languages: German, the native language of the annotators; Papuan Malay, the lingua franca of the major centres of West Papua (Indonesia); Wooi, an Austronesian language spoken on Yapen Island in West Papua; and Yali, a Papuan highland language spoken in West Papua. Two of the authors have first-hand experience with the West Papuan languages.[1] All other annotators participating in the experiment were unfamiliar with them.

The core questions to be answered by this study are given in (1).

(1) a. Do the segmentation results for the whole corpus and for each individual language show above-chance interrater agreement according to standard kappa metrics?
b. Is there significant variation in interrater agreement for familiar *vs.* unfamiliar languages? What are possible reasons for (the lack of) such variation?

With respect to (1b), there are two ways in which familiarity with a language may influence interrater agreement in the segmentation task. First, it could be that the prosodic cues used as segmentation criteria come in language-specific forms, and are more readily recognised in familiar languages. *Prima facie*, such language-specific forms are less likely for pauses, probably the perceptually strongest cue for IPBs. But they have some plausibility for other IPB cues, such as pitch resets and 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, Mo & Baek 2010), prosodic boundary perception is not influenced only by prosodic factors, but also by non-prosodic ones, in particular syntactic structure and semantic and pragmatic coherence. IPBs have a strong tendency to overlap with clause boundaries, and there is a concomitant tendency to hear IPBs at clause boundaries. The unfamiliar-language condition completely removes the potential influence of non-prosodic factors, giving 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 introduces a different set of factors, it also increases the potential for conflict between different segmentation cues (cf. 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.

[1] Throughout this article, 'West Papua' is used as a geographic reference to the Indonesian western half of the island of New Guinea.

The paper is structured as follows. §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. §3 details task design and data, and the empirical core of the study is presented in §4 and §5. The experimental results provided in §4 demonstrate robust interrater agreement for the whole corpus, as well as for individual languages. The main question in 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 pauses rather than IPBs, especially in unfamiliar languages. §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 greater extent in unfamiliar languages than in the familiar German.

§6 discusses the theoretical import of our results for current concepts of IPs and their functions. It reviews different possible interpretations of the 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 any language. What we might call a universal PHONETIC IP needs to be distinguished from language-specific PHONOLOGICAL IPs, which can be interpreted as a language-specific grammaticisation of the universal phonetic IP.[2] Our results support a view of prosodic categories as partially universal, in as much as they are grounded in the mechanics of speaking, but partially also language-specific, in as much as they reflect the contingencies of historical developments in the grammaticisation of prosodic features.

2 Prosodic interrater agreement studies and their targets

Interrater agreement studies of prosodic phenomena can be classified into two types. One type tests 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 Tones and Break Indices (ToBI) 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, Mo & Baek 2010: 1143–1145).

This type of study is concerned with 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

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

the training for this type of study is the provision of examples illustrating typical auditory and acoustic correlates of the intended categories. Labellers are usually provided with acoustic data (minimally waveform and F0 contour), in addition to audio files.

The other type of study tests the perception of prosodic prominences and boundaries by naive 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. Mo *et al*. (2008) is a prototypical study along these lines,[3] with analytical follow-ups on phonetic factors in Cole, Mo & Hasegawa-Johnson (2010), and on syntactic (and other non-prosodic) factors in Cole, Mo & Baek (2010). 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. Mo *et al*. (2008: 736) summarise the instructions regarding prominences and boundaries 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'.

In their study, the annotators marked prominences and boundaries on print-outs of the transcripts, which included word boundaries, speech errors and disfluencies, but no punctuation or capitalisation. The relevant findings of this study are: (i) there is significant interrater agreement with regard to boundaries, with a mean Cohen's κ coefficient of 0.58 across all pairs of transcribers (the values for prominences are much lower); (ii) 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 to 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.

In some ways, Buhmann *et al*.'s (2002) study, based on Dutch corpus data, is very similar. However, their procedure is different in a number of important respects. First, working with non-expert annotators, they included an intensive training period in which the annotators, after receiving instructions and examples, worked through a 15-minute learning corpus, and were given feedback on their performance. Second, the test corpus was substantially larger than those used in most other studies, consisting of more than 8000 words (45 minutes) of read, scripted and

[3] The method originates in the perception-oriented approach to intonation developed in Eindhoven as summarised in 't Hart *et al*. (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.

unscripted speech. Third, an online working environment was used, which included the audiovisual 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*. obtained a fairly high interrater agreement. For boundaries, the Cohen's κ coefficients for interrater pairs range from 0.70 to 0.88 (Buhmann *et al*. 2002: 782).

In their instructions on detecting prosodic boundaries, Buhmann *et al*. (2002: 779) use the term 'break', a non-technical category which is presumably part of the non-expert understanding of spoken language. They distinguish strong and weak breaks, defining them as in (2) (2002: 780–781).

(2) a. Strong breaks (symbol '‖') are defined as severe interruptions of the normal flow of speech. They are typically realised as a clear pause or even an inhalation.

e.g. he was there ‖ and so was his girl-friend

b. 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.

e.g. I can tell you | this was un|be|lievable

Note that while the instructions in Mo *et al*. (2008) are concerned with what they presume to be a function of chunking (cf. '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 study in this paper belongs to the second type, in that it tests the perception of prosodic boundaries by non-expert listeners. But there are two major points of difference. The most important is that our study compares the performance of annotators across familiar and unfamiliar languages. This task design assumes 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 fact that the coherence of the melody sets 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 Appendix A for details) characterise IPs as distinct units perceivable by means of a coherent melody. They draw attention to two major types of IPB cues: (i) the interruption of the rhythmic delivery by, *inter alia*, a pause or final lengthening, and (ii) 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.

As in the Buhmann *et al*. study, our annotators were thus also clearly instructed to follow prosodic cues for boundaries only, but unlike that study, a distinction was made between melodic and rhythmic cues. Importantly, the instructions also reflected 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 consistent 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 of 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 it is 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. By '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 between and within IPs. There is thus a need to distinguish between IP-external and IP-internal pauses. External pauses are pauses that occur between two adjacent IPs. On one 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 they are referred to as 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, 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 (cf. §3). 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, if correctly identified, are therefore an important practical cue for IPBs. A large number 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.

On the other hand, it is much more difficult to perceive melodic coherence consistently when conscious attention is paid to it in a segmentation task. In our instructions, we drew attention to jumps in pitch between offsets and onsets of IPs as indicators of interrupted coherence. However, such pitch jumps are often not larger than the micro-perturbations caused by obstruents; the correlation with rhythmic interruptions provides 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 also made use of these additional cues, even though they were not mentioned in our instructions. However, this aspect will not be further discussed in this paper.

To summarise, our study focuses on prosodic boundary cues and, in the case of languages unfamiliar to the annotators, actually forced them to pay attention to them exclusively. Annotators were advised to pay attention to both melodic and rhythmic cues in their identification of IPBs. These cues reinforce each other when they are temporally aligned (cf. Pijper & Sanderman 1994, Krivokapić & Byrd 2012), but may lead to disagreements when not synchronised. 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 sounds associated with the actions depicted (such as a bicycle accident).

The sixty pear stories were told in different languages, primarily German and three languages from Eastern Indonesia, the major fieldwork site of the first author. Table I provides details of the corpus, which is partitioned into

number of narratives			length		total number of words
			total	mean	
Group 1: Germanic	German	18	53m 28s	02m 58s	8836
	Kölsch	1	02m 31s	02m 31s	286
	English	1	10m 06s	10m 06s	1418
	subtotal	20	1h 06m 05s	05m 12s	10540
Group 2: Papuan Malay	Papuan Malay	20	1h 04m 00s	03m 12s	10373
Group 3: Eastern Indonesian	Wooi	12	34m 53s	02m 54s	3557
	Waima'a	2	08m 15s	04m 08s	1406
	Yali	6	17m 42s	02m 57s	2007
	subtotal	20	1h 00m 50s	03m 20s	6970
	total	60	3h 10m 55s	03m 55s	27883

Table I
Composition of the corpus.

three groups for processing and presentation purposes, each comprising twenty stories. For practical and explorative purposes, the corpus also includes one pear story in English, one in Kölsch (the German dialect of Cologne) and two in Waima'a, an Austronesian language from East Timor. Segmentation results for these varieties do not differ from those 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 modelling. Appendix B provides further details of recording procedures and corpus compilation, as well as our statistical procedures.

The three languages from Eastern Indonesia that this study mainly focuses 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, has a fairly strict SVO pattern and has bare nouns as the most frequent type of noun phrase. Wooi has an elaborate subject-marking paradigm, as well as a complex set of NP 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, and has a moderate amount of (postpositional) case

marking and complex verbal morphology, with hundreds of forms in a paradigm (cf. Riesberg 2017).

Prosodically, these three languages illustrate systems very different from German, but found in many other parts of the world. As is typical of 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 is 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 Manado Malay (Stoel 2007) and Waima'a (Himmelmann 2010). See §6.1 for further discussion and exemplification.

Wooi is similar to Papuan Malay in delimiting IPs by the combination of a phrase accent and a boundary tone, but differs in having both lexical stress and lexical pitch accents, similar to Papiamento (Remijsen & van Heuven 2005). The small group of Austronesian West Guinea languages it belongs to are 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 a similar prosodic system in neighbouring Yale (Kosarek).

Prior to the study, all sixty pear-story narratives had been transcribed by native speakers of the respective languages using ELAN, a multimedia annotation tool for multimodal research.[4] 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 were given a .wav 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 the studies mentioned in §2, disfluencies were not marked as such, but the transcript did contain a representation of unclear segments which could not be transcribed (indicated roughly by one 'x' for each unclear syllable). Further details on experimental procedure are given in Appendix B.

Four linguistics students, all native speakers of German, were recruited for this task, and paid a fixed rate for participating. They were students in different linguistics programmes at the University of Cologne, with varying degrees of familiarity with prosodic analyses. R1 was a female

[4] ELAN is a multimedia annotation tool for multi-modal research (see http://tla.mpi.nl/tools/tla-tools/elan/).

We thank Sonja Riesberg for help with the Yali data. See references in Appendix B for further information on the data sources and full acknowledgements for transcriptions and translations.

Bachelor student in Linguistics, R2 and R3 were Master students in Linguistics, male and female respectively. R1–R3 all had a basic introduction to prosody as part of the introductory courses of their BA programme. R4 was a female Master student in Linguistics, specialising in phonetics and writing an MA thesis on a prosodic topic at the time of her involvement in the project.

In addition, the authors of the paper[5] produced a consensus version, which, importantly, involved native speaker input in the creation phase, and was 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 by 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 in Appendix A. Most of the narratives had been transcribed before the current study was designed. Second, the transcriptions were independently checked by two of MS, JS and VU. Third, these three authors compared their changes to the original transcripts, and produced a first consensus version, after resolving any disagreements through re-listening and discussion. As a final step, this version was checked by NPH, who focused on problematic cases and overall consistency in instances where the exact placement of the boundary was arguably arbitrary (due to noise in the recording, for example, or disfluencies, as discussed below). In contrast to the four student annotators, the authors made regular use of instrumental evidence in the form of F0 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, first-hand knowledge of the languages and their prosodic systems, we decided to treat the consensus version (henceforth 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 fewer 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 the German example in (3), latching occurs in three successive IPs. The main cues for IPBs here are pitch jumps interrupting the melodic contour, downward after *gelegt* and *bereitstanden*, and upward after *heraus*, as shown in Fig. 1. All

[5] All authors are native speakers of German, except for MS, who is a native speaker of Hebrew but speaks German fluently.

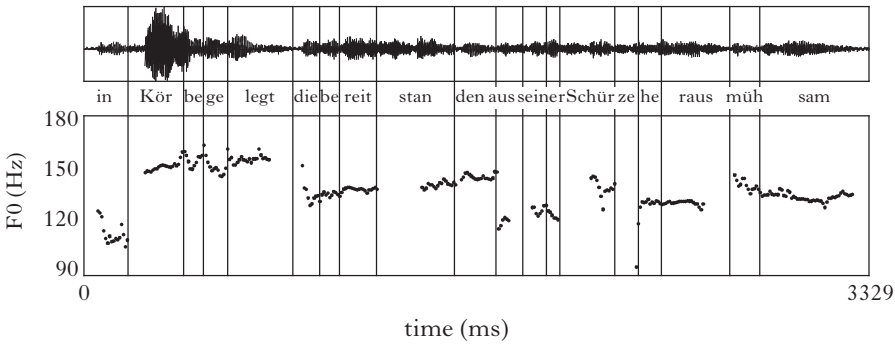

Figure 1
Waveform and F0 track for the German example in (3).

student annotators agreed with the boundary after *mühsam*, but only two had boundaries after *gelegt* und *bereitstanden*, and only one after *heraus*.[6]

(3) in Körbe gelegt =
in baskets put.PRTC
die bereitstanden =
that stand.by.3PL
aus seiner Schürze heraus =
out.of his apron out
mühsam (700 ms)
painstakingly
'into baskets, that stood there, from out of his apron, painstakingly' (DEU_pear_Flor)

The other factor giving rise to disagreements involves 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, either interrupting or abandoning it. Consequently, disfluencies could be handled by a convention stipulating that all instances of disfluency either always or never induce a boundary. While we drew attention to the problem of IP-internal disfluencies in the instructions, we did not propose conventions for handling these, as this would have required major training efforts to be useful.

In the CONS version, we tried to distinguish consistently between hesitations (IP-internal disfluencies) and truncations, i.e. the abandonment of a unit currently underway. This distinction is primarily based on pitch evidence, but also on the length of the interruption. Interruptions lasting

[6] The following conventions are used in the examples: each line is one IP; '=' indicates latching; pause length is given between parentheses; < > surround false starts (< > on the morpheme interlinearisation 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, APPL = applicative, DAT = dative, DET = determiner, HES = hesitation particle, NSG = non-singular, PL = plural, PRTC = participle, REL = relative marker, SG = singular, TOP = topic marker, VEN = venitive.

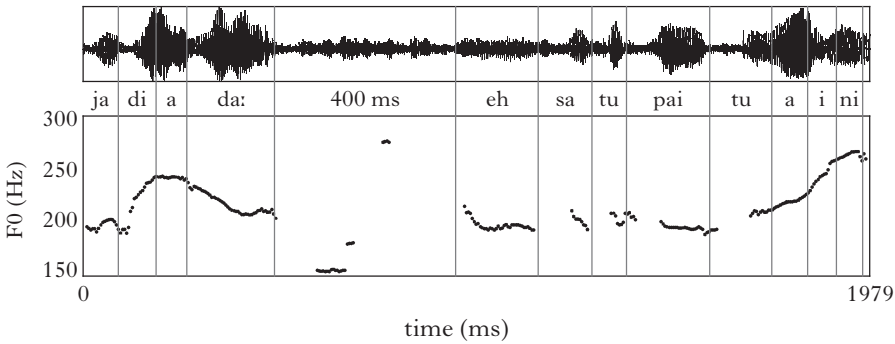

Figure 2

Waveform and F0 track for the Papuan Malay example in (4). The F0 traces seen during the 400 ms pause are caused by background noises.

more than one second were generally considered to be truncations. Otherwise, a disfluency was considered to be IP-internal only if speech was resumed after the disfluency on the same pitch level as before. In this case, it is likely that the speaker will continue with the delivery of an IP begun before the disfluency. This is illustrated by (4), from Papuan Malay, where the F0 extraction in Fig. 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*).

(4) jadi ada: (400 ms) eh satu paitua ini =
so there.is HES one adult this
'so uhm there was this man' (PMY_pear_Lala)

In truncations, on the other hand, there is clear evidence for the start of a new IP, for example in (5), from Wooi. As shown in Fig. 3, the speaker

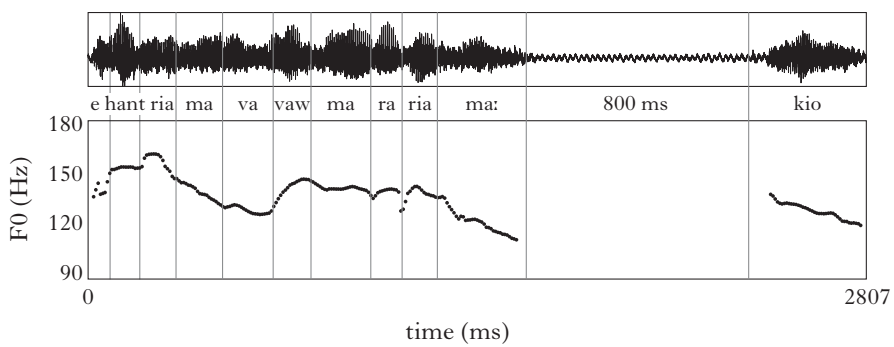

Figure 3

Waveform and F0 track for the Wooi example in (5).

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 F0 between *ma:* and *kio* is almost four semitones, so that it is safe to assume that the speaker has no intention of connecting back to the previous pitch contour.

(5) ehanti ria ma vavaw mara ria ma: (800 ms)
someone <3SG>go VEN DET.NSG TOP <3SG>go VEN
kio (2000 ms) 'there was someone coming, he came …
<3SG>take he took …' (DEU_pear_Alex)

While there are many instances in which the distinction between a hesitation and a truncation is reasonably clear, it 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.

4 Interrater agreement results

In this section, we first look at interrater 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 behaviour 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 entire corpus comprises 27883 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, leaving 27823 potential IPBs in all (one fewer than the number of words for each of the sixty narratives). Table II provides an

annotator	IPs	mean IP length (words)	SD of IP length
R1	8441	3.29	2.05
R2	7898	3.51	2.20
R3	5159	5.35	3.84
R4	5864	4.72	2.95
CONS	6499	4.26	2.79
mean	6772	4.09	2.82

Table II
Overview of IP segmentation by annotator.

overview of the segmentations created by the five annotators (four students and CONS), and shows that the corpus was divided into 6772 IPs on average, resulting in a mean IP length of about four words. For the entire corpus, as shown in Fig. 4, we obtain a raw agreement of 77.54% across all five annotators, and a statistically significant Fleiss' κ score of 0.71, which represents substantial agreement (see Landis & Koch 1977). If we consider only the four student annotators, we find a raw agreement of 78.21% and a statistically robust and substantial interrater agreement ($n = 27823$, $\kappa = 0.68$, $z = 277$, $p < 0.001$). Figure 4 provides the number and percentage of cases in which a particular subset of the five annotators posited an IPB, ranging from zero for cases where no annotator posited an IPB to five for places where all annotators assumed an IPB. The rightmost column shows the total of all 'unanimous' decisions, i.e. cases where all annotators agreed that there was or was not a boundary. These results show that recordings of spontaneous speech in different languages can be reliably segmented into IPs even by non-expert annotators without special training.

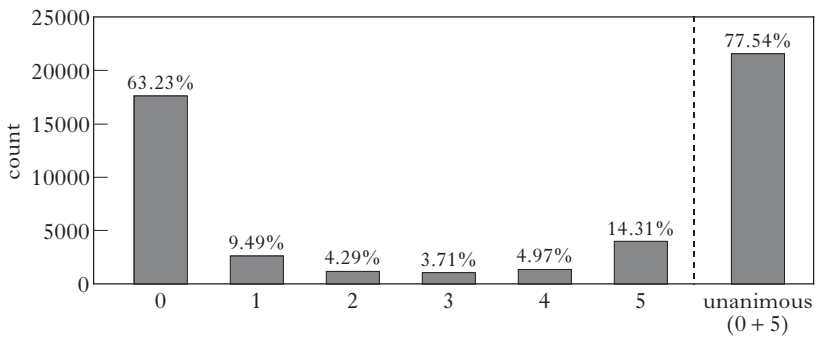

Figure 4

Overall agreement on the IP segmentation of the whole corpus ($n = 27823$, Fleiss' $\kappa = 0.71$, $z = 375$, $p < 0.001$).

If we take our consensus segmentation as reference and compare it with individual student annotators' segmentations, we obtain the results in Table III. Individual student annotators' segmentations agree quite well with the CONS segmentation, with Cohen's κ (overall) ranging from 0.74 to 0.82, in all cases statistically significantly above chance (R1: $\kappa = 0.74$, $n = 27823$, $z = 125$, $p < 0.001$; R2: $\kappa = 0.75$, $n = 27823$, $z = 126$, $p < 0.001$; R3: $\kappa = 0.74$, $n = 27823$, $z = 125$, $p < 0.001$; R4: $\kappa = 0.82$, $n = 27823$, $z = 138$, $p < 0.001$). All four student annotators were thus able to provide a reliable IP segmentation that agreed to a large extent with the authors' expert segmentation.

	annotator			
measure	R1	R2	R3	R4
true positives	5984	5797	4572	5279
false positives	2397	2041	527	525
true negatives	18987	19343	20857	20859
false negatives	455	642	1867	1160
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)	0.74	0.75	0.74	0.82
mean κ per narrative	0.74	0.74	0.74	0.82
SD of κ per narrative	0.09	0.07	0.09	0.06

Table III

Comparison of annotators to reference segmentation on the whole corpus.

Student annotators nonetheless differed amongst themselves in their tendency to either assume more or fewer IPBs than CONS: R1 and R2 posited a relatively large number of IPBs (cf. Table II), and segmented the narratives into relatively short IPs, resulting 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 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, assumed fewer IPBs and therefore longer IPs (cf. Table II), resulting in high precision values of about 90%, as well as lower recall values of around 71% and 82% respectively (cf. Table III). R4 had the lowest standard deviation. R4 thus displayed most agreement with CONS across all 60 narratives. This is probably related to the fact that R4 was the only student annotator who had had in-depth training in prosodic analysis, albeit not specifically for the present study.

Nevertheless, 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 (cf. §2) can be applied reliably and consistently in familiar and unfamiliar languages. To further scrutinise this finding, we now turn our focus to individual languages in our corpus, and to 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, using the format of Fig. 4. Interrater agreement is remarkably similar across the four languages, the value for three of the languages being close to the overall Fleiss' κ of 0.71 (German: $\kappa = 0.72$; Wooi: $\kappa = 0.74$; Yali: $\kappa = 0.75$). Papuan Malay is somewhat lower ($\kappa = 0.68$). The test statistics thus display substantial agreement between the five annotators' segmentations of each of these four subcorpora.[7] These results suggest that the familiar *vs.* unfamiliar language distinction is not the most important factor in 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.

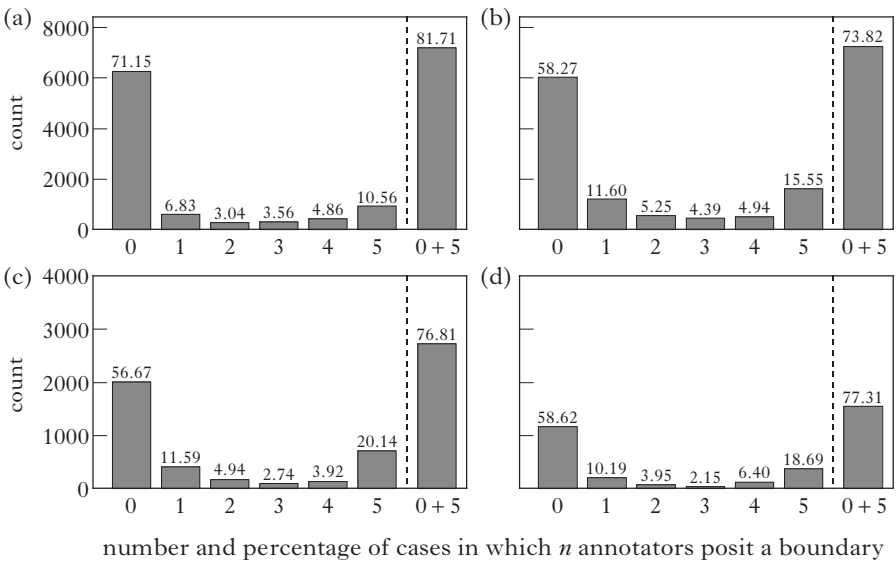

Figure 5

Interrater agreement on the IP segmentation for the individual languages in the corpus: (a) German ($n = 8818$, Fleiss' $\kappa = 0.72$, $z = 214$, $p < 0.001$); (b) Papuan Malay ($n = 10353$, Fleiss' $\kappa = 0.68$, $z = 219$, $p < 0.001$); (c) Wooi ($n = 3545$, Fleiss' $\kappa = 0.74$, $z = 139$, $p < 0.001$); (d) Yali ($n = 2001$, Fleiss' $\kappa = 0.75$, $z = 106$, $p < 0.001$).

To conclude, let us see whether the statistics for the individual annotators agree with this overall pattern. Table IV gives an overview of the number and average length of IPs in the segmentations by annotator and language. At first sight, the table appears to reveal one conspicuous difference between German and the West Papuan languages: German IPs appear to be substantially longer, both overall and for individual

[7] The results for the three minor subcorpora in our corpus, Kölsch, English and Waima'a, are fully in line with the results for the larger subcorpora.

(a)

annotator	IPs	mean (words)	SD (words)
R1	2238	3.93	2.71
R2	1887	4.65	2.92
R3	1085	8.03	4.72
R4	1583	5.53	3.48
CONS	1748	5.02	3.27
mean	1708	5.13	3.55

(b)

annotator	IPs	mean (words)	SD (words)
R1	3502	2.95	1.49
R2	3214	3.21	1.68
R3	2157	4.78	3.08
R4	2315	4.45	2.67
CONS	2657	3.88	2.36
mean	2769	3.73	2.33

(c)

annotator	IPs	mean (words)	SD (words)
R1	1213	2.92	1.50
R2	1289	2.74	1.45
R3	914	3.86	2.47
R4	889	3.96	2.29
CONS	933	3.78	2.37
mean	1048	3.37	2.07

(d)

annotator	IPs	mean (words)	SD (words)
R1	612	3.26	2.08
R2	711	2.81	1.61
R3	531	3.75	2.51
R4	498	4.00	2.56
CONS	551	3.62	2.48
mean	581	3.44	2.27

Table IV
Number and mean length of IPs per annotator and language: (a) German; (b) Papuan Malay; (c) Wooi; (d) Yali.

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. Rather, the units identified in the West Papuan languages might instantiate another, smaller kind of prosodic phrase (e.g. a 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 IV 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=INSTR) '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-

organising functions). As seen in Table V, the sample reflects the imbalance in the average number of words per IP across the four languages found in Table IV. However, no comparable imbalance is found with regard to the average number of *content* words per IP. Consequently, the higher average number of words per IP in German must be due to the higher number of orthographically independent function words.[8]

The data in Table V 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 (cf. §6.1). Table V should, however, give some plausibility to the claim that we are dealing with units of a comparable size (i.e. comparable informational content), and allow us to continue to speak of IPs in the further discussion of our results.

	IPs	words	mean length of IPs (words)	content words	content words per IP
German	270	1408	5.20	487	1.8
Papuan Malay	300	1223	4.08	530	1.8
Wooi	180	654	3.63	288	1.6
Yali	90	303	3.37	162	1.8

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

Apart from the difference in the mean length of IPs, the statistical trends in Table IV are surprisingly similar to those in Table II for the whole corpus. CONS and R4 again posited a similar number of IPs, resulting in similar mean IP length for the four individual subcorpora in Table IV. Compared to the other annotators, R1 and R2 again segmented the narratives into shorter units. There are thus individual differences in annotator behaviour that hold across the different subcorpora. This may indicate that segmentation strategies are similar across the four languages.

That this is not necessarily the case, however, is shown by R3, who segmented the German narratives, which she is able to understand, into IPs with an average length of more than eight words.[9] Boundaries here were preferably placed at clause boundaries, ignoring the fact that clauses in

[8] Recall from Table II 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.

[9] R3 also has the greatest 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.

spontaneous speech are often chunked into several IPs.[10] (6) is a typical case, where R3 considered a longish clause with several PPs to be a single IP. All other annotators, including CONS, chunked this clause into five IPs.

(6) dann kam ihm <ein-> (200 ms)
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 (900 ms)
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 behaved more like the other annotators with regard to the three unfamiliar West Papuan languages. This suggests that R3 used different segmentation strategies in familiar and unfamiliar languages. Segmentation in the familiar languages takes non-prosodic factors more strongly into account, while segmentation in the unfamiliar languages relies exclusively on prosodic cues. The inclusion of non-prosodic factors in IP segmentation may thus increase the potential for disagreements (cf. §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.

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 simply due to the fact that student annotators 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 inter-rater-agreement results reported in the previous section. First, pauses may

[10] 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 focusing on, rather than the end of declination units.

happen to be better boundary predictors in the West Papuan languages, thereby offsetting 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, so that the fact that interrater agreement is similar across the four languages is 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 the 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 §3. Non-silent interruptions such as coughing and sneezing were not included in the statistical model.

Table VI provides, for each language, the absolute frequency of external and internal pauses, as well as their relative frequency per IP and their average duration. 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, and the last row shows that pauses are more reliable as IPB cues in Wooi and Yali than in German and Papuan Malay. 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: both German and 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

		German	Papuan Malay	Wooi	Yali
external pauses	absolute frequency	882	1631	777	429
	relative frequency per IP	0.5046	0.6139	0.8328	0.7786
	mean duration (ms)	627	561	1,177	1,005
internal pauses	absolute frequency	162	102	16	8
	relative frequency per IP	0.0927	0.0384	0.0171	0.0145
	mean duration (ms)	435	408	481	325
probability of IPB given pause		0.8448	0.9411	0.9798	0.9817

Table VI

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

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 German and Papuan Malay, 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. For example, the German and Papuan Malay subcorpora are better gender-balanced than the Wooi and Yali subcorpora, which are heavily male-dominated. The German and Papuan Malay speakers were 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. This suggests that the 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 VI probably contribute to the high 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: did annotators base 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 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 the five levels shown in Fig. 6, 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, all pauses longer than 600 ms were put into one category.

We fitted our logistic regression model using the glm (generalised linear model) function 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 is as in (7).

(7) IPB ~ Pause length + Annotator + Language + Pause length:Annotator
+ Pause length:Language + Annotator:Language
+ Pause length:Annotator:Language

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

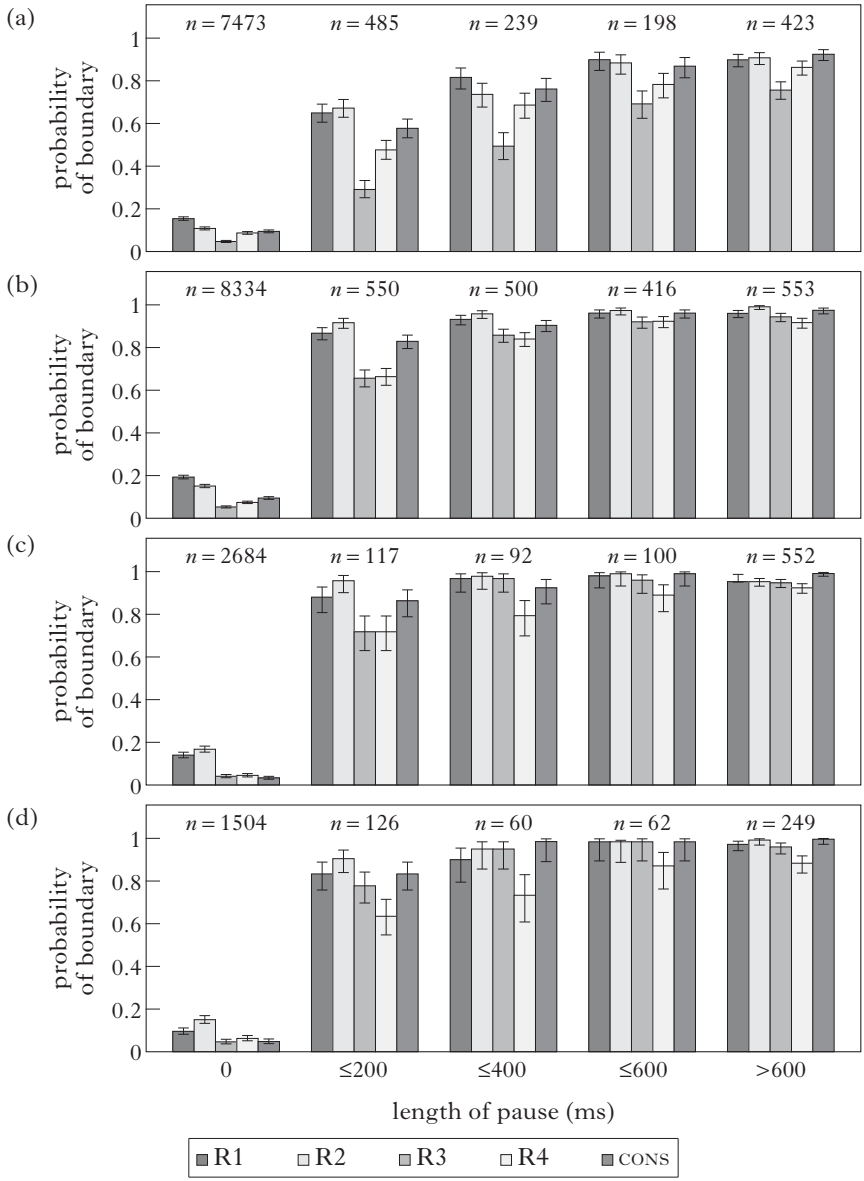

Figure 6

Effect display of logistic regression model predicting the probability that each annotator will assume an IPB: (a) German; (b) Papuan Malay; (c) Wooi; (d) Yali.

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 modelling results in

Fig. 6 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.

The overall trends 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 Fig. 6 are lowest 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 (cf. Table VI), not to the fact that annotators made more use of pauses in the unfamiliar West Papuan languages than in the familiar German.

Second, annotators did not posit IPBs in unfamiliar languages solely on the basis of pauses. Otherwise, one would expect zero probabilities in the case of no pause (the leftmost group of bars) and a probability of 1 in the case of a longer pause (≥ 400 ms). Instead, the student annotators assumed a comparable, though of course relatively low, likelihood of latching cases across all four subcorpora, and were also quite constant in their relative propensity to allow for latching. R1 and R2 were more likely to posit IPBs without a pause than R3 and R4 both in German and in all three unfamiliar languages. Conversely, while the predicted probabilities of the student annotators assuming a substantial IPB rise (to > 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 VI).

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

To compare the four student annotators more directly to the reference segmentation, we fitted an additional logistic regression model to our data. This time, however, the dependent variable is agreement with CONS: 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- and three-way interactions between them. The model formula in short form is given in (8).

(8) Agreement with CONS ~ (Pause length + Annotator + Language)^3

A likelihood-ratio test indicated that the three-way interaction is required for a good model fit ($\chi^2 = 110.49$, 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 both to cases without any pause (0 ms) and to 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 *vs*. unfamiliar languages. This is further evidence that the segmentation behaviour of student annotators for unfamiliar languages was not completely different as a result of their relying exclusively on pauses.

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 pauses in Wooi and Yali coincide with IPBs in 98% of all cases (but the converse does not hold; approximately 20% of the IPBs in these subcorpora lack external pauses). While the predictive power of pauses for IPBs thus varies across the languages, there are no clear trends separating familiar from unfamiliar languages. Specifically, there is no evidence that student annotators relied 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 (cf. e.g. Soto *et al*. 2013: Table 6).

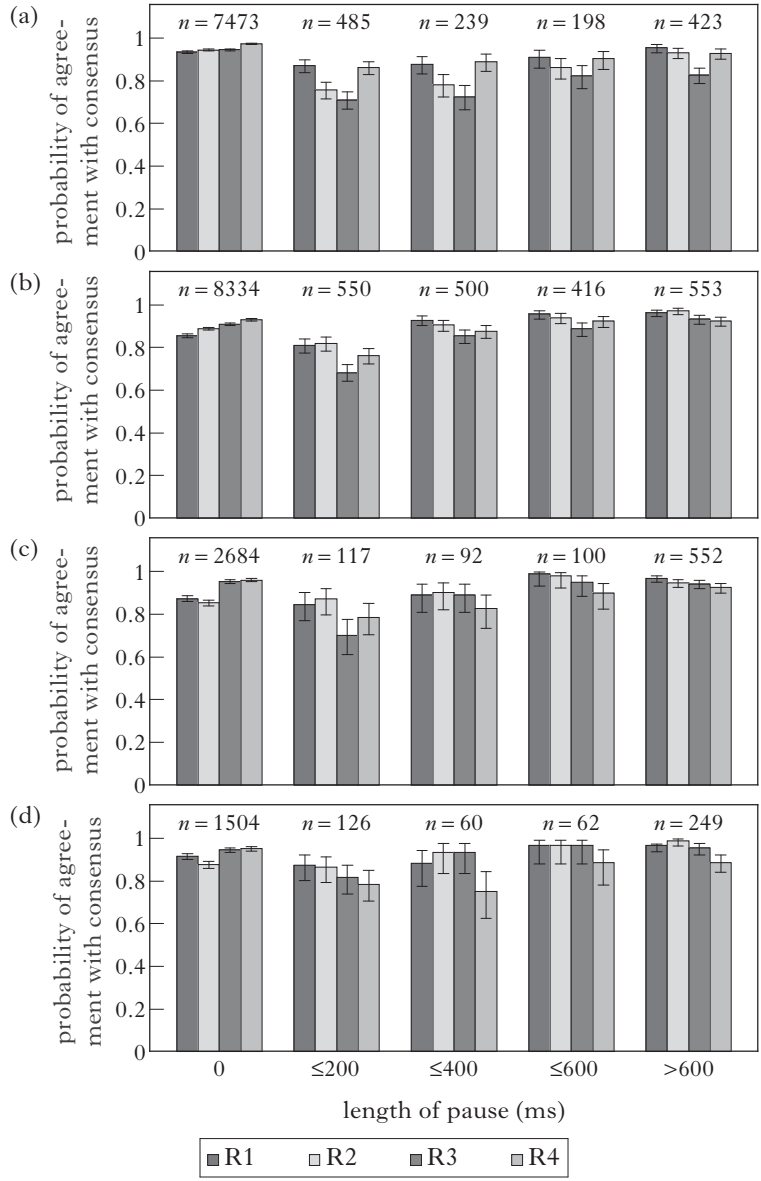

Figure 7

Effect display of model predicting agreement by each annotator with CONS: (a) German; (b) Papuan Malay; (c) Wooi; (d) Yali.

6 Discussion

The empirical results reviewed in the preceding two sections make it clear that the cues for IPBs provided in our instructions (cf. §2) were 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. What our data show is that German listeners are able to identify the kinds of prosodic cues they are familiar with from their native language 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 it would appear to be largely devoid of theoretical importance. 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 the reviewers of this paper) on two interrelated grounds. First, the same kind of cues might identify different kinds of units in unrelated languages, an issue taken up in §6.1. Second, it might be the case that native speakers of other languages hear completely different things, and that the units identified are therefore essentially *German* perceptual IPs, and irrelevant to the native speakers of the unfamiliar languages. We address this issue in §6.2.

If we can counter 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. §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 typological work, 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 of the cross-linguistic comparability of prosodic units here, but will try to justify the plausibility of 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 IP level and the next lower level, widely known as phonological phrase (PhP, the term we will use) or intermediate phrase.[11] 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 level. 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 side with the arguments against a proliferation of prosodic levels and the requirement that each level be defined by specific properties that distinguish its boundary from boundaries at other levels (cf. 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 the languages. In Papuan Malay and Wooi, IPs are optionally 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 a pause or pitch reset. The overall melodic and rhythmic coherence is thus not interrupted, as illustrated by the Papuan Malay example in (9).

(9) 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)

As seen in Fig. 8, the high PhP boundary tone on *pi* (the final syllable of *topi*, which functions as head noun for the following relative clause) is immediately followed by a fall that continues across the next word, the relative pronoun *yang*. PhP-final syllables may be slightly lengthened, but this is the exception rather than the rule, and is not found in Fig. 8. IPBs, on the other hand, are generally followed by a new pitch onset, and often by a pause, i.e. they involve an interruption of rhythmic and melodic coherence. Additionally, IPBs in both Papuan Malaya and Wooi involve *two* tonal targets, a phrase accent and a final boundary which occurs in a two-syllable window at the end of the unit (cf. §3). This is illustrated in Fig. 8 by the combination of a high phrase accent and a falling boundary tone on the final verb *jatu*. Both penultimate and final syllables tend to be considerably lengthened.

Crucially, the boundary strength *within* each unit type may vary, and such differences may be perceived by listeners (see e.g. Ladd 2008: 293–297, Wagner & Watson 2010, Krivokapić & Byrd 2012). As noted in §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 indicate a major prosodic

[11] We do not discuss the next lower level, the 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.

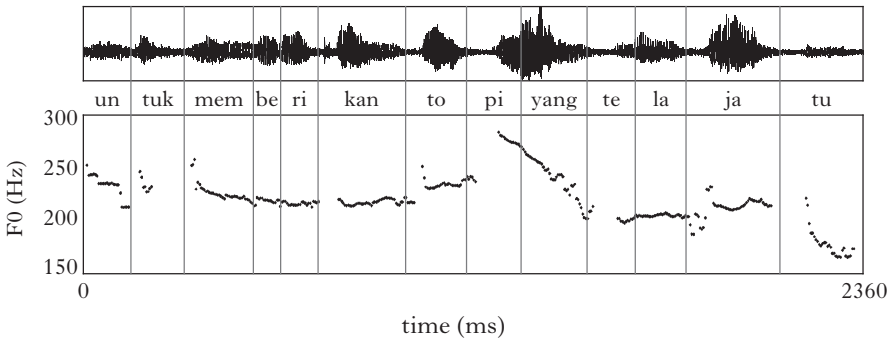

Figure 8
Waveform and F0 track for the Papuan Malay example in (9).

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 (cf. Table II) – had a tendency to mark PhP boundaries occurring within larger IPs.

Turning to our segmentation data, our expert segmentation (CONS) distinguished PhPs from IPs in three of the four languages, 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.[12] 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, 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 the same two major phrasing levels above the phonological word were used in three of the four languages. In as much as the student annotators' segmentations match the expert annotation across the four subcorpora (cf. Fig. 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 cross-linguistic collections such as Jun (2005,

[12] For German, we follow the GToBI analysis 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.

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 with the same name are at least roughly comparable.

Still, use of the same analytical framework and procedure may not be sufficient to support cross-linguistic sameness. How can we be sure that levels with 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 proposed one such parameter in §4, with respect to differences in mean IP length across the four main subcorpora. The data in Table V show that the units are of a comparable size with regard to their information 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 for how the informational content of IPs should be specified; one such is 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 (cf. 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 variability in size. The units identified in our segmentation data are highly variable in size, ranging from discourse particles and short phrases without content words, to NPs or PPs, to clausal and multiclausal units. This is typical of 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 none of the above proves that the units identified in our experiment are relevant and perceptible for 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 West Papuan speakers are sensitive to substantially different kinds of segmentation cues and possibly also arrive at substantially different segmentations. The next section will address this objection.

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

To fully counter the objection that our findings only show that German speakers hear German IPs, we 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 it would be difficult to find enough Wooi and Yali speakers who could engage in tasks requiring the processing of written language.

However, 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 (cf. §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*. (2018), which investigates both prominence and boundary perception using the RPT method. 22 speakers of Papuan Malay annotated transcripts of 56 excerpts of spontaneous narrative and conversational speech produced by 28 different speakers. While interrater agreement for prominence was negligible (Fleiss' $\kappa = 0.10$), interrater agreement for boundaries (Fleiss' $\kappa = 0.41$) was within the range found in comparable studies for English.

38 of the 56 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 VII provides agreement statistics within the group of Papuan Malay native speakers and within the group of German student annotators. In addition, it also shows mean κ values for agreement between members of each of these two groups and with

		Papuan Malay speakers	German students
agreement within groups	raters	22	4
	Fleiss' κ	0.40	0.57
	pairs of raters	231	6
	mean of Cohen's κ	0.41	0.57
	SD of Cohen's κ	0.20	0.08
agreement with CONS	raters	22 + 1 CONS	4 + 1 CONS
	pairs of raters	22	4
	mean of Cohen's κ	0.48	0.60
	SD of Cohen's κ	0.18	0.05
		Papuan Malay speakers *vs.* German students	
agreement across groups	raters	22 + 4	
	pairs of raters	88	
	mean of Cohen's κ	0.40	
	SD of Cohen's κ	0.15	

Table VII
Interrater agreement within and between different groups of annotators on 38 excerpts of Papuan Malay pear stories (480 boundary decisions).

our consensus version. 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.

As noted above, Fleiss' κ for agreement within the group of Papuan Malay native speakers (0.40) is comparable to results obtained in similar studies for English. Interrater agreement among the four German student annotators is clearly higher (Fleiss' $\kappa = 0.57$). This difference in agreement values is likely due to the different experimental methods and the stricter time constraints which the native speakers were subjected to in the RPT approach. In addition, it may have to do with 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 0.48, 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 VII also supports this conclusion. The mean agreement between all different pairs of one Papuan Malay native speaker

annotator and one German student annotator (mean $\kappa = 0.40$) is quite close to the mean agreement among Papuan Malay native speakers themselves (mean $\kappa = 0.41$). This indicates that native and non-native speakers segment the Papuan Malay speech into comparable units. It also supports the conjecture that the greater 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 again suggest substantial interrater agreement between the segmentations produced by German and Papuan Malay speakers. We therefore believe that it is plausible to assume not only that the units identified in our experiment are 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 requires 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 markedly from the kind of interrater agreement study briefly mentioned in §2, in which annotators were 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 (cf. §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 familiarity with their phonological structure, simply on the basis of phonetic boundary cues which appear not to be specific to a particular language. This finding can be interpreted in a number of ways. In the two preceding subsections, we argued against the view that it shows only that German speakers are able to identify German IPs everywhere. Instead, we propose that it supports what we call the UNIVERSAL PHONETIC IP HYPOTHESIS (UPIPH), which claims that all natural languages make use of the same kinds of phonetic cues for IPs, and that these cues can be perceived by speaker-hearers even in unfamiliar languages. The main cues are the interruption of melodic coherence, as manifested in pitch resets between IPs and major rhythmic breaks, particularly pauses. Both types of cues are considerably more complex than just stated, and involve language-specific and probably also speaker-specific further features.

In addition, IPs may be – and usually are – phonologically organised units, with the phonological organisation being manifested in particular in tonal events. The prototypical example of this are the edge tones found in all prosodic systems described so far. They are the clearest *phonological* markers for prosodic boundaries, and tend to be intricately interlinked with segmental articulatory gestures (e.g. Krivokapić & Byrd 2012). In this view, IP boundary tones are regularised (grammaticised) descendants 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) account of the relation between universal biological codes and the language-specific phonological organisation 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 from cognitive demands on speech planning and processing (cf. §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).[13] These boundary characteristics can be further grammaticised into language-specific phonological categories, giving rise to a phonologically organised category, the intonational phrase. Such grammaticisations 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 grammaticised pitch movements, serving other functions such as marking information status (postlexical pitch accents) and distinguishing lexemes (lexical tones).

Note that this scenario specifically targets IPs, and 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 (cf. §6.1).

We believe that it is quite likely that phonological IPs are part of the prosodic system of all natural languages. If this is the case, 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 (cf. §6.1). In principle, however, the UPIPH allows for the possibility that there are languages in which spontaneous speech is produced in IP-sized chunks (delimited by universal phonetic

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

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 grammaticised to different degrees, i.e. that prosodic systems exist where the IP level is only weakly grammaticised, its structure consisting simply of a single final boundary tone, for example.

The UPIPH, of course, needs further conceptual and empirical scrutiny. Empirically, it predicts 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 the languages represented in a sample should ideally also be included among the annotators.

While the current sample covers a range of prosodic systems (cf. §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 use of this. 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.

Empirical testing 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 §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 be applicable 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',[14] then we would expect this signature to occur across a worldwide 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 (cf. §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.

(i) Intonational phrases are empirically viable units, according to standard measures for interrater agreement. Multirater as well as pairwise κ coefficients show a substantial and statistically significantly above-chance

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

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 (cf. §4).

(ii) 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, semantico-pragmatic units) leads to more disagreements.

(iii) 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.

(iv) 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 grammaticised into phonological IPs, which are language-specific structural units arising from, and continually undergoing, processes of diachronic change.

REFERENCES

Boersma, Paul & David Weenink (2015). *Praat: doing phonetics by computer* (version 5.4.09). 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 Manuel González Rodríguez & Carmen Paz Suarez Arauj (eds.) *Proceedings of the 3rd International Conference on Language Resources 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: University of Chicago Press.

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

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

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. Available (February 2018) 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*. 2nd edn. Malden, Mass.: Wiley-Blackwell. 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.
Goldman Eisler, F. (1968). *Psycholinguistics: experiments in spontaneous speech*. London & New York: Academic Press.
Grice, Martine, Stefan Baumann & Ralf Benzmuller (2005). German intonation in autosegmental-metrical phonology. In Jun (2005). 55–83.
Gussenhoven, Carlos (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
Halliday, M. A. K. (1967). *Intonation and grammar in British English*. The Hague & Paris: Mouton.
Hart, Johan 't, René Collier & Antonie 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. *Lg* **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.
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.
Kamholz, David C. (2014). *Austronesians in Papua: diversification and change in South Halmahera-West New Guinea*. PhD dissertation, University of California, Berkeley.
Katsika, Argyro, Jelena Krivokapić, Christine Mooshammer, Mark Tiede & Louis Goldstein (2014). The coordination of boundary tones and its interaction with prominence. *JPh* **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**. http://dx.doi.org/10.1098/rstb.2013.0397.
Krivokapić, Jelena & Dani Byrd (2012). Prosodic boundary strength: an articulatory and perceptual study. *JPh* **40**. 430–442.
Ladd, D. Robert (1986). Intonational phrasing: the case for recursive prosodic structure. *Phonology Yearbook* **3**. 311–340.
Ladd, D. Robert (2008). *Intonational phonology*. 2nd edn. 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 Augusto Hoffmann 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, Enhua Shen & Fanji Gu (eds.) *Advances in cognitive neurodynamics: Proceedings of the International Conference on Cognitive Neurodynamics 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). Naïve listeners' prominence and boundary perception. *Proceedings of the 4th Conference on Speech Prosody*. Campinas, Brazil. 735–738. Available (February 2018) at http://www.isca-speech.org/archive/sp2008/.
Pijper, Jan Roelof de & Angelien A. Sanderman (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *JASA* **96**. 2037–2047.
Pitrelli, John F., Mary E. Beckman & Julia Hirschberg (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proceedings of the 3rd International Conference on Spoken Language Processing* (*ICSLP 94*). Yokohama: Acoustical Society of Japan. 123–126.
R Core Team (2017). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org.
Remijsen, Bert (2001). *Word-prosodic systems of Raja Ampat languages*. PhD dissertation, University of Leiden.
Remijsen, Bert & Vincent J. van Heuven (2005). Stress, tone and discourse prominence in the Curaçao dialect of Papiamentu. *Phonology* **22**. 205–235.
Riesberg, Sonja (2017). An introduction to the Yali-German dictionary with a short grammatical sketch. In Siegfried Zöllner, Ilse Zöllner & Sonja Riesberg (eds.) *A Yali* (*Angguruk*) – *German dictionary*. Canberra: Asia-Pacific Linguistics. 1–43. Available at http://hdl.handle.net/1885/127381.
Riesberg, Sonja, Janina Kalbertodt, Stefan Baumann & Nikolaus P. Himmelmann (2018). On the perception of prosodic prominences and boundaries in Papuan Malay. In Sonja Riesberg, Asako Shiohara & Atsuko Utsumi (eds.) *Perspectives on information structure in Austronesian languages*. Berlin: Language Science Press.
Sanderman, Angelien A. (1996). *Prosodic phrasing: production, perception, acceptability and comprehension*. PhD dissertation, University of Eindhoven.
Sanderman, Angelien A. & René Collier (1997). Prosodic phrasing and comprehension. *Language and Speech* **40**. 391–409.
Shattuck-Hufnagel, Stefanie & Alice Turk (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* **25**. 193–247.
Silverman, Kim, Mary E. Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet B. Pierrehumbert & Julia Hirschberg (1992). ToBI: a standard for labeling English prosody. In John J. Ohala, Terrance M. Nearey, Bruce L. Derwing, Megan M. Hodge & Grace E. Wiebe (eds.) *Proceedings of the 1992 International Conference on Spoken Language Processing*. Edmonton: University of Alberta. 867–870.
Soto, Victor, Erica Cooper, Andrew Rosenberg & Julia Hirschberg (2013). Cross-language phrase boundary detection. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. http://dx.doi.org/10.1109/ICASSP.2013.6639316.
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, University of Amsterdam.

Tokizaki, Hisao (2002). Prosodic hierarchy and prosodic boundary. *Bunka-to Gengo* **56**. 81–99.

Wagner, Michael (2010). Prosody and recursion in coordinate structures and beyond. *NLLT* **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. *Interspeech 2004: Proceedings of the 8th European Conference on Spoken Language Processing, Jeju Island, Korea*. 2729–2732. Available (February 2018) at http://www.isca-speech.org/archive/interspeech_2004/.