Phonology **35** (2018). Supplementary materials

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

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

Supplementary materials

Appendix A: Details of the instructions given to the student annotators (§2)

Our written instructions regarding IPB cues, given to the annotators and explained once verbally, were as follows in the original German:

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

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

English translation

Your task is to segment an audio recording containing the narrative of a short film into **intonational phrases**, i.e. into sequences that are perceivable as a **distinct unit** by means of a coherent melody / a coherent pitch contour.

To keep in mind

Boundaries between two intonational phrases are typically characterised by two features:

1. an interruption of the **rhythmic** delivery by a (sometimes only very short) pause, **lengthening** of the last segment at the end of a unit and/or **increased speaking rate** at the beginning of a new unit (*anacrusis*);
2. a disruption of the pitch contour / melody line: a **pitch jump** (up or down) between the end of a unit and the beginning of the subsequent one; intonational phrases often exhibit a constant decline in fundamental frequency, which at the boundary of a unit is reset to the default pitch level of the speaker in a given context (*reset*). This is typically followed by another decline in fundamental frequency (*declination*).

Pauses, however, may sometimes also occur within an intonational phrase, e.g. if the speaker is searching for a word or corrects him-/herself = hesitation pauses. Hesitation pauses are often filled (*uhm*, etc.) but not necessarily. What is important is that the pitch levels before and after a hesitation pause fit together continuously. That is, rather than a new onset of the melody line, the original pitch contour is continued after the pause.

Along with these explanations, the annotators were presented with five audio examples of boundary cues to illustrate the following typical configurations at IPBs:

(i) Two IPs set off by a clear melodic break (a new onset is clearly audible from a downward jump in pitch after a strongly rising boundary tone), accompanied by a pause of 240 ms and greatly reduced intensity of the second IP.

(ii) Two IPs set off primarily by a clear melodic break only (new onset by a downward jump in pitch after a strongly rising boundary tone) accompanied by a very short (70 ms), but noticeable, period of silence.

(iii) Two successive IPs without any intervening silence, but with final lengthening at the end of the first IP and a clear melodic break (a falling boundary tone followed by an upward jump in pitch).

(iv) One IP with an internal hesitation pause of 690 ms, after which the pitch resumed at approximately the same level as before the hesitation.

(v) Two IPs involving minor unit-internal hesitations and no intervening pause, but a clear melodic break (a major upward jump in pitch) and increased speaking rate at the beginning of the second IP.

The examples for these configurations were taken from a short personal narrative in German that was not part of the corpus used in the segmentation task. They were played several times. Reference to boundary tones in the above descriptions has been added only to make it easier for the expert reader to identify the type of example we have used. In the actual instructions, the focus was on the auditory impression.

Note that while our instructions go into a moderate degree of technical detail, we did not make direct reference to analytical constituents of melodic contours such as boundary tones, even though all languages in our corpus use them. The concept of a boundary tone only makes sense in a theoretical model, knowledge of which we could not presuppose on the part of the participants in this study. Nor do our instructions refer to boundary cues that are difficult to perceive without specific measurements such as domain-initial strengthening (cf. Fougeron & Keating 1997, Keating *et al.* 2004).

Appendix B: Further details of data and procedure (§3)

1 Recording procedure

One person watched the pear film on a laptop screen, and then recounted it to another person, who had not seen the film before. The interlocutor was instructed to behave ‘naturally’ in the context of retelling a movie, i.e. to ask clarification questions and to provide feedback whenever and wherever appropriate. While all interlocutors engaged in appropriate (verbal and non-verbal) back channelling, only very few actually asked clarification questions, never exceeding three questions in one telling. All verbal utterances made by the interlocutor were included in the recordings and transcripts used for this study, but not in the segmentation task. Only the narrators’ speech was segmented into IPs.

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

2 Corpus compilation

The corpus used in this study was originally compiled for the AUVIS project ('*Audiovisuelles Data Mining in multimodalen Sprachdaten*/ Audiovisual data-mining in multimodal language data'; see https://tla.mpi.nl/projects_info/auvis/ for more information). The main goal of this project was to explore possibilities for automatically annotating and searching audio and video streams of unannotated or only partially annotated recordings from unrelated languages, with a particular focus on under-documented and under-resourced languages. As a case study for realistic search scenarios, the project involved an exploration of the alignment between gestural, prosodic and grammatical units. In gesture research, all annotation is standardly done by multiple annotators, which was one reason to work with multiple annotators for the prosodic annotation as well.

The version of the AUVIS corpus used in the current study differs from the version used in gesture-related studies with regard to one German retelling, which was replaced by another one at a later point, when it became apparent that the narrator of the retelling was aware of the fact that the study was concerned with gestures.

The first group of recordings in Table I consists of eighteen pear film narratives in (standard colloquial) German, one narrative in the vernacular dialect of Cologne (Kölsch) and one narrative in (American) English. Six of these narratives were recorded with analogue audio and video recorders in the 1990s, and are therefore of somewhat lower quality, especially with regard to the video (which did not play a role in the current study). The remaining narratives were recorded in 2012 with up-to-date audio/video equipment for the specific purposes of the AUVIS project. At the time of recording, the speakers involved were mostly students in their early twenties at the University of Cologne. Five recordings involve speakers aged between 30 and 50.

The second group comprises narratives in Papuan Malay, the *lingua franca* of West Papua, the western half of the island of New Guinea governed by Indonesia (see Kluge 2017 for a recent description). The pear-film narratives in Papuan Malay were recorded at the Center for Endangered Languages Documentation (CELD) in Manokwari, the capital of the province of Papua Barat (West Papua). The narrators, as well as their interlocutors, were all of approximately the same age (early to mid-twenties) and were enrolled as English students at the local university. See Riesberg & Himmelmann (2012–14) for further details.

The third group consists of three lesser-known languages of Eastern Indonesia, for which language documentation corpora have been compiled in documentation projects based in Cologne. Two of these languages, Wooi (Kirihio *et al.* 2009–15, Sawaki 2016) and Waima'a (Belo *et al.* 2002–06), are Austronesian languages spoken in coastal areas of West Papua and East Timor respectively. Both speech communities are small (fewer than 3,000 speakers each) and multilingual, and are currently shifting to regional standards (Papuan Malay and Tetum respectively). The pear-film narratives in Wooi and Waima'a were all recorded in the field, and are generally of a lower quality than the recordings made at the CELD (there are more background noises of different kinds). The age of the Wooi speakers is more mixed than in the other language groups, ranging from speakers in their early twenties to those over 50. The third language, Yali (Riesberg *et al.* 2012–16, Riesberg 2017), is a Papuan language (Trans-New-Guinea phylum) spoken in the highlands of West Papua. The number of speakers is somewhat higher (around 10,000), and only younger generations are multilingual in varieties of Malay (both Standard Indonesian and Papuan Malay, to differing degrees). The recordings were made at the CELD with young native speakers in their early twenties who were enrolled as students at the local university or (in one case) as a secondary school student.

3 Experimental procedure

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

The tier for the narrator was left blank. After they had identified a stretch of the audio stream which they assumed to form an IP, the annotators' task was to copy the respective portion of the transcript from the plain text file, and paste it into the appropriate selection on the narrator tier in ELAN. The selection was made in the waveform view of the audio file that is part of the standard annotation setup in the ELAN program.

Annotators worked on the task on their own, without any time constraints (some taking less than a week per package, others close to a month). They received the narratives in packages per group, starting with Group I (Germanic), then Group II (Papuan Malay) and finally Group III (Eastern Indonesian languages). The labels of the packages included language names, and each narrative was clearly labelled as to the language used, but no further information on the languages was provided.

The order of the narratives in a group was alphabetical, based on the abbreviated names of the narrators, except for Group II, which was arranged in such a way that male and female narrators followed each other in roughly alternating order. In the Germanic part of the corpus, alphabetic ordering resulted in a well-mixed sequence of female and male narrators. Most narrators in the Eastern Indonesian part of the corpus were men, except for Waima'a (two females). The sequence here was Wooi first, then Waima'a and finally Yali.

4 Statistical procedures

Since the task of the annotators was to provide a segmentation into IPs of a given transcription of a narrative which we provided them in a practical orthography including word boundaries, we can treat the IP-segmentation task as a binary classification: between each consecutive pair of words in the transcription, the annotators could either posit an IPB or not. For a transcription containing n words, there are $n-1$ consecutive word pairs and thus $n-1$ potential IPBs. We focus here on this binary classification, and disregard the exact location at which the annotators put an IP start or end boundary on the ELAN timeline.

In practice, annotators occasionally forgot to copy and paste a word from the transcription into the ELAN timeline, or accidentally copied one word twice. For our evaluation, we had to correct these copy-and-paste errors by occasionally adding or deleting a word. This was usually unproblematic, because the intended IPBs were still clear, due to the temporal alignment of the IP segments created by the annotator in ELAN with the audio signal. Moreover, the number of these copy-and-paste errors was relatively low: the least accurate annotator (R3) made 200 copy-and-paste errors in all, amounting to about three errors per narrative.

When evaluating interrater agreement, we cannot simply compare the raw agreement between annotators to a baseline assuming equal probabilities of 0.5 for positing or not positing an IPB between two consecutive words. Instead, we have to take into account the fact that there are many more non-boundaries between words than boundaries, i.e. a boundary is much less likely than a non-boundary (the average length of IPs in consensus segmentation (CONS) is 4.26 words; SD = 2.79). We therefore use the standard κ measures of interrater agreement that incorporate information about the relative frequency of the different categories (in our case, *boundary vs. non-boundary*). In order to assess overall agreement between all annotators, we use Fleiss' κ (Fleiss 1971). In addition, we compare the student annotators' segmentations individually to CONS using Cohen's κ (Cohen 1960) for pairwise comparisons, as well as well-known measures from information retrieval – namely the error rate, precision, recall and *f*-score (the harmonic mean of precision and recall).

Where appropriate, we evaluate differences in interrater agreement between languages, as well as the segmentation accuracy of individual annotators on different subsets of the corpus, by calculating means and variances of these measures on the basis of the 60 individual narratives in our corpus and by comparing them using non-parametric statistical tests. In most cases, we use the Wilcoxon-Mann-Whitney rank sum test (Wilcoxon 1945, Mann & Whitney 1947) for unpaired samples. We assume the conventional significance level of $p \leq 0.05$ throughout.

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

ADDITIONAL REFERENCES

Belo, Maurício C., John Bowden, John Hajek, Nikolaus P. Himmelmann & Alex V. Tilman (2002–06). *Dobes Waima'a documentation*. DobeS Archive, Max Planck Institute for Psycholinguistics, Nijmegen. Available (February 2018) at http://dobes.mpi.nl/projects/waimaa/.

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

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

Fougeron, Cécile & Patricia A. Keating (1997). Articulatory strengthening at edges of prosodic domains. *JASA* **101**. 3728–3740.

Keating, Patricia A., Taehong Cho, Cécile Fougeron & Chai-Shune Hsu (2003). Domain-initial articulatory strengthening in four languages. In John Local, Richard Ogden & Rosalind Temple (eds.) *Phonetic interpretation: papers in laboratory phonology* VI. Cambridge: Cambridge University Press. 145–163.

Kirihio, Jimmi K., Volker Unterladstetter, Apriani Arilaha, Freya Morigerowsky, Alexander Loch, Yusuf Sawaki & Nikolaus P. Himmelmann (2009–15). *Dobes Wooi documentation*. DobeS Archive, Max Planck Institute for Psycholinguistics, Nijmegen. Available (February 2018) at http://dobes.mpi.nl/projects/wooi/.

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

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

Riesberg, Sonja & Nikolaus P. Himmelmann (2012–14). *CELD Papua*. DobeS Archive, Max Planck Institute for Psycholinguistics, Nijmegen. Available (February 2018) at http://dobes.mpi.nl/projects/celd/.

Riesberg, Sonja, Kristian Walianggen & Siegfried Zöllner (2012–16). *Dobes Yali documentation*. DobeS Archive, Max Planck Institute for Psycholinguistics, Nijmegen. Available (February 2018) at http://dobes.mpi.nl/projects/celd/.

Sawaki, Yusuf (2016). *A grammar of Wooi: an Austronesian language of Yapen Island, Western New Guinea*. PhD dissertation, Australian National University.

Wilcoxon, Frank (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* **1**. 80–83.

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