**Title: Sequences of Intonation Units Form a 1 Hz Rhythm**

Authors: Maya Inbar[1,2], Eitan Grossman[1] and Ayelet N. Landau[2]*

Affiliation:

[1] Department of Linguistics, the Hebrew University of Jerusalem, Mt. Scopus, Jerusalem, 91905, Israel

[2] Departments of Psychology and of Cognitive Science, the Hebrew University of Jerusalem, Jerusalem, Mt. Scopus, 91905, Israel

* Corresponding author: ayelet.landau@mail.huji.ac.il

Word count: 3000

**Abstract**

Studies of speech processing investigate the relationship between temporal structure in speech stimuli and neural activity (Giraud & Poeppel, 2012; Gross et al., 2013; Park, Ince, Schyns, Thut, & Gross, 2015). The speech envelope, a representation of speech that captures amplitude fluctuations in the acoustic speech signal, has been used to probe this temporal relationship. The envelope is most commonly understood to reflect the succession of syllables at ~5 Hz (Ding et al., 2017; Greenberg, Carvey, Hitchcock, & Chang, 2003; Räsänen, Doyle, & Frank, 2018; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017). Despite clear evidence for auditory tracking of the envelope at lower frequencies (Gross et al., 2013; Park et al., 2015), it is not well understood what linguistic information is captured in the speech envelope beyond the syllable level. Here, we harness linguistic theory to focus on Intonation Units (IUs), while analyzing their temporal structure as captured in the speech envelope.

IUs are prosodic units (Chafe, 1994; Du Bois, Cumming, Schuetze-Coburn, & Paolino, 1992; Himmelmann, Sandler, Strunk, & Unterladstetter, 2018) which are defined by a specific pattern of syllable delivery, together with resets in pitch and articulatory force. Linguistic studies of spontaneous speech indicate that this prosodic segmentation paces new information in language use across numerous and diverse languages. Therefore, IUs provide an important structural cue for the cognitive dynamics producing and comprehending speech universally.

We study the relation between IUs and the periodic components of the speech envelope in six languages. We apply analysis methods from investigations of neural synchronization (Vinck, van Wingerden, Womelsdorf, Fries, & Pennartz, 2010) to study recordings from every-day speech contexts of over 100 speakers. We find that sequences of IUs form a consistent 1 Hz rhythm. Our results demonstrate that IUs form a significant periodic cue within the speech envelope that could be utilized by the cognitive and neural systems when tracking speech.

Keywords: intonation units, prosody, low-frequency neural activity, information flow, phase-synchronization

## 1. Introduction

Speech processing is commonly investigated by the measurement of brain activity as it relates to the acoustic speech stimulus. Such research has revealed that neural activity tracks amplitude modulations present in speech. It is generally agreed that a dominant element in neural tracking of speech is a 5 Hz rhythmic component, which corresponds to the rate of syllables in speech. The speech stimulus is also tracked at lower frequencies (<4 Hz), but the functional role of these fluctuations is not fully understood. They are assumed to relate to the "musical" elements of speech which are above the word level – called prosody. However, prosody in the neuroscience literature is rarely investigated for its structure and function in cognition.

In linguistics, on the other hand, prosody is investigated for its structural and functional roles in speech and cognition. This line of research has identified prosodic segmentation cues that are common to all languages, and that characterize what is termed Intonation Units (IUs; Figure 1a). Importantly, in addition to providing a systematic segmentation to ongoing naturalistic speech, IUs capture the pacing of information, parceling a maximum of one new idea per IU. Thus, IUs provide a valuable construct for quantifying how ongoing speech serves cognition in individual and interpersonal contexts. The first goal of our study is to introduce this understanding of prosodic segmentation from linguistic theory to the neuroscientific community. The second goal is to put forth a temporal characterization of IUs, and hence offer a precise, theoretically driven interpretation of the low-frequency auditory tracking and its relevance to cognition.

Regardless of the phonological, morphosyntactic and so-called rhythmic structure of a language, when speakers talk, they produce their utterances in chunks with a specific prosodic profile. The prosodic profile intersects rhythmic, melodic, and articulatory characteristics (Chafe, 1994; Du Bois et al., 1992; Himmelmann et al., 2018). Rhythmically, chunks may be delimited by pauses, but more importantly, by a fast-slow dynamic of syllables (Figure 1b). Melodically, chunks have a continuous pitch contour, which is typically sharply reset at the onset of a new unit. In terms of articulation, the degree of contact between articulators is strongest at the onset of an IU (Keating, Cho, Fougeron, & Hsu, 2003).

Across languages, these auditorily-defined units are also functionally comparable in that they pace the flow of information in the course of speech (Chafe, 1987, 1994; Du Bois, 1987; Pawley & Syder, 2000). For example, when speakers develop a narrative, they do so gradually, introducing the setting, participants and the course of events in sequences of IUs, where no more than one new piece of information relative to the preceding discourse is added per IU (Box 1). This has been demonstrated both by means of qualitative discourse analysis (e.g., Chafe, 1987, 1994; Ono & Thompson, 1995), and by quantifying the average amount of content items per IU. Specifically, the amount of content items per IU was found to be very similar across languages albeit strikingly different grammatical profiles (Himmelmann et al., 2018). Another example for the shared role of IUs in different languages pertains to the way speakers plan their (speech)

actions. When a speaker coordinates a transition during a turn-taking sequence they rely on prosodic segmentation (i.e., IUs): semantic/syntactic phrasal completion is not a sufficient cue for predicting when a transition will take place, and IU design is found to serve a crucial role in timing the next turn-taking transition (Bögels & Torreira, 2015; Ford & Thompson, 1996; Gravano & Hirschberg, 2011).

Here we use recordings of spontaneous speech in natural settings to characterize the temporal structure of sequences of IUs in six languages. The sample includes well-studied languages from the Eurasian macro area - languages that are spoken by large or relatively large speech communities (English, Russian and Hebrew), as well much lesser-known and -studied languages, spoken in the Indonesian-governed part of Papua by smaller speech communities. Importantly, our results generalize across this linguistic diversity, despite the substantial differences in socio-cultural settings and all aspects of grammar, including other prosodic characteristics (Himmelmann et al., 2018). In contrast to previous research, we estimate the temporal structure of IUs using direct time measurements rather than estimations based on word count or syllable length (Box 2). In addition, we quantify the temporal structure of IUs in relation to the speech envelope, which is an acoustic representation relevant to neural processing of speech. We find that sequences of IUs form a consistent 1 Hz rhythm in the six sample languages, and relate this finding to recent neuroscientific accounts of the roles of slow rhythms in speech processing.

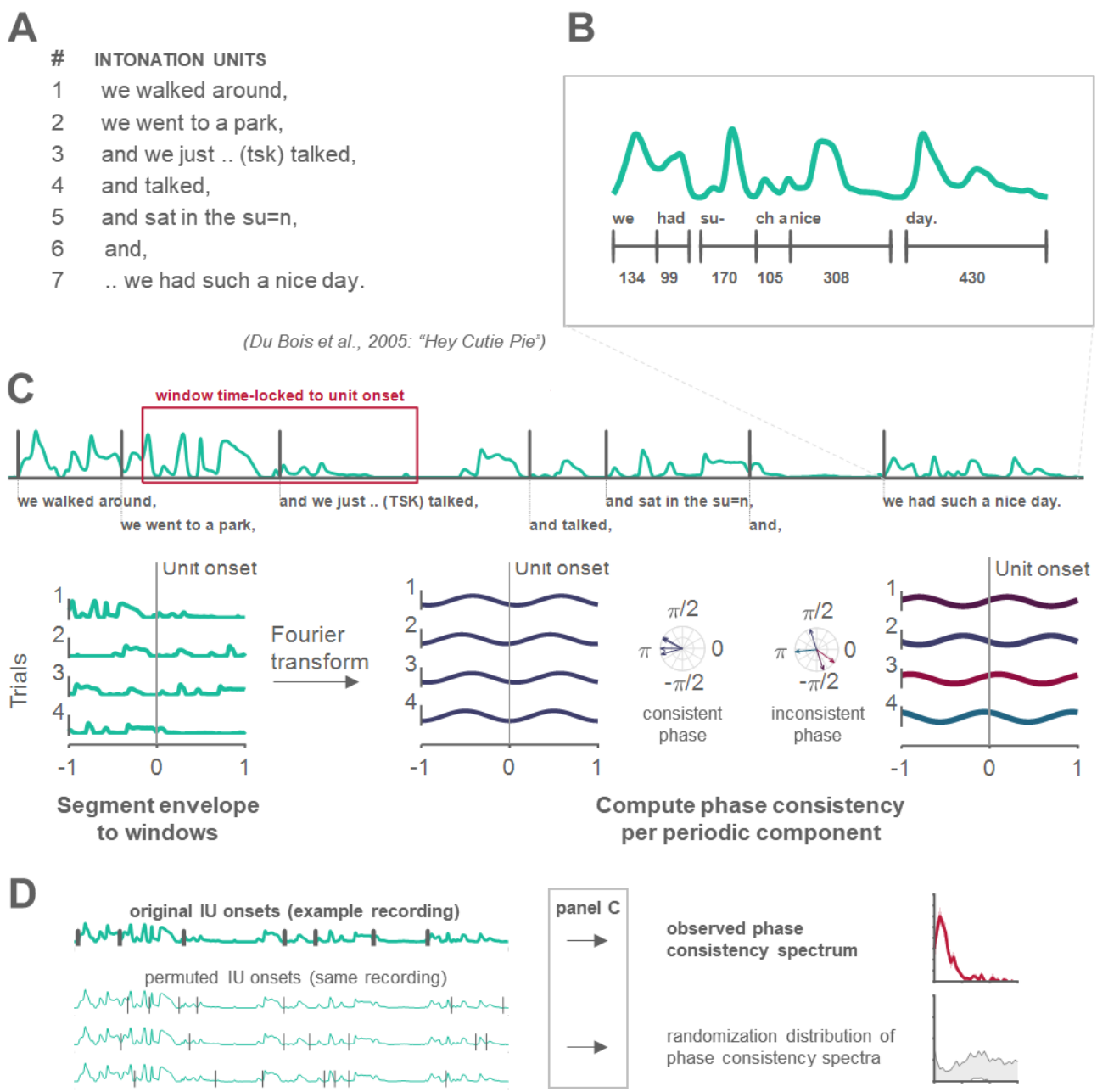


**Figure 1. Analysis pipeline.** (a) An example intonation unit sequence from a conversation in Du Bois et al. (2005). (b) Illustration of one of the characteristics contributing to the delimitation of IUs: the fast-slow dynamic of syllables. A succession of short syllables is followed by comparatively longer ones; new units are cued by the resumed rush in syllable rate following the lengthening (syllable duration measured in ms). (c) Illustration of the phase-consistency analysis: 2-second windows of the speech envelope (green) were extracted around each IU onset (gray vertical line), decomposed and compared for consistency of phase angle within each frequency. (d) Illustration of the IU-onset permutation in time, which was used to compute the randomization distribution of phase consistency spectra (see Materials and Methods).

**Box 1. Information is temporally structured in social interaction.**

Time is crucial for organizing information in interaction, not only via prosodic cues but also through body conduct, such as gaze-direction, head and hand gestures, leg movements and body torques (Mondada, 2018). This is especially evident in task-oriented interaction, for example, in direction-giving sequences during navigation, or instruction sequences more broadly, where the many deictic words (e.g., this, there) can only be interpreted correctly when accompanied by a timely gesture. Following is such a fragment from a judo-instruction class (Du Bois et al., 2005). The prosodic-based segmentation to IUs is represented by a break in line. To facilitate reading and considering the current focus on function rather than form, transcription conventions were simplified (the original transcription can be retrieved with the sound file from the linked corpus). Speaker overlap is marked by square brackets ( [ ] ), minimally audible to medium pauses (up to 0.6 seconds) are marked by sequences of dots ( … ) and punctuation marks represent different classes of pitch movements, indicating roughly the degree of continuity between one unit and the next (comma – continuing; period – final; double dash – cut short).

**"Throw me" (SBC057: 24:19– 25:19)** (Du Bois et al., 2005)

| | | | | |
|---|---|---|---|---|
| 1 | Jed: | ... The more you reach that arm through, | 22 | ... and you roll like that. |
| 2 | | ... Let me add one -- | 23 | Okay, |
| 3 | | Let me add one thing to it, | 24 | and if you do this, |
| 4 | | to make it -- | 25 | in demo, |
| 5 | | cause it is a hard fall if you're, | 26 | .. uh .. in demo fashion, |
| 6 | | throwing it contest style, | 27 | you'll be so low to the ground, |
| 7 | | if you're throwing it demo style, | 28 | by the time that your opponent hits the ground, |
| 8 | | somebody once told me, | 29 | that you'll be real nice to him. |
| 9 | | that that, | 30 | And that way I mean, |
| 10 | | the way to practice, | 31 | you can intensify, |
| 11 | | all the Makikomi-waza, | 32 | you can always intensify throwing. |
| 12 | | ... was .. to go down ... like this. | 33 | .. throw it harder. |
| 13 | | To take this foot if you're throwing right-hand throws, | 34 | But the mark of a ... good .. judo throw, |
| 14 | | and to take this foot, | 35 | is to be able to, |
| 15 | | and stick it all the way through l[ike thi]s. | 36 | do you see the people throw Seoi so nice, |
| 16 | Nick: | [Yep], | 37 | they can do -- |
| 17 | Jed: | ... And just stick it. | 38 | they could throw their grandmother, |
| 18 | | ... All the way, | 39 | and it doesn't even hurt. |
| 19 | | behind you, | 40 | She goes ah. |
| 20 | | .. like this. | 41 | That felt good. |
| 21 | | And as you go down, | | |

**Box 2. What is a word?**

The notion of a word has been argued to be untenable for both language-specific and cross-linguistic analyses (e.g., Haspelmath, 2011). We demonstrate why this is so for cross-linguistic comparison with the following example in Seneca, a member of the Northern Iroquoian branch of the Iroquoian language family spoken in Northeast America (Chafe, 2015, p. 191).

***wáőwö'géosäh***

| | | | | |
|---|---|---|---|---|
| *wa'-* | *höwa-* | *a'kër-* | *osrah-* | *Ø* |
| FACTUAL- | 3RD PERSON.AGENT/MASCULINE.SINGULAR.PATIENT- | ashes- | sprinkle- | PUNCTUAL |

'they sprinkled ashes on him'

The first line includes a word in the language, that is, a unit of meaning whose unit-ness is defined by morphosyntactic processes in the language. The second line includes a breakdown to meaning components (separated by hyphens and tabs), obtained through comparative evidence from related languages. Due to extensive sound changes over the years, Seneca shows a high degree of fusion between meaning components, that is, the boundaries between them are obscured and not necessarily available to speakers. Note also that these meaning components cannot normally appear as independent words, that is, without the neighboring meaning components. The third line includes a gloss per meaning component, differentiating between those with grammatical meaning (part of a grammatical paradigm; in small caps) and content items. The fourth line includes the corresponding English translation. As evident from this example, a noteworthy distinction between Seneca and some better-known languages of the world such as English is that Seneca regularly packages an event, its participants and other meaning components within a single morphosyntactic word. Consequently, for one and the same message, Seneca IUs would contain fewer words compared to English IUs.

## 2. Materials and Methods

### 2.1 Data

We studied the temporal structure of IUs using six corpora of conversations and narratives that were transcribed and segmented into IUs according to the unified criteria devised by Chafe, Du Bois and colleagues (Chafe, 1994; Du Bois et al., 1992). Three of the corpora were segmented by specialist teams working on their native language: The Santa Barbara Corpus of Spoken American English (Du Bois et al., 2005), The Haifa Corpus of Spoken Hebrew (Maschler et al., 2017), and The Prosodically Annotated Corpus of Spoken Russian (Kibrik & Podlesskaya, 2014). The other three corpora were segmented by teams with varying degrees of familiarity with the languages, as part of a project studying the ability to identify IUs in unfamiliar languages: The DoBeS Summits-PAGE Collection of Papuan Malay (Himmelmann & Riesberg, 2016), The DoBeS Wooi Documentation (Kirihio et al., 2015), and The DoBes Yali Documentation (Riesberg, Walianggen, & Zöllner, 2016). Further information regarding the sample may be found in Table 1. Appendix A includes additional information regarding the construction of the sample and the coding and processing of IUs. We also analyzed in the same manner a corpus of an additional language that was transcribed semi-automatically to prosodic units following slightly different rhythmic and melodic criteria: Rhapsodie: A Prosodic-Syntactic Treebank for Spoken French (Lacheret et al., 2014). Further information regarding this additional analysis can be found in Appendix B. From all language samples, we extracted IU onset times, noting which speaker produced a given IU. Additionally, we computed the speech envelope for each sound file following standard procedure (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009; Gross et al., 2013; Figure 1c and Appendix A).

| Source of recordings and transcriptions | Number of recordings | Audio duration (minutes) | Number of IUs | Number of speakers with > 5 IUs | % IUs following inter-IU interval < 1 s |
|---|---|---|---|---|---|
| The Santa Barbara Corpus of Spoken American English | 10 | 9:58 | 460 | 19 | 78.3% |
| The Haifa Corpus of Spoken Hebrew | 10 | 6:30 | 507 | 24 | 80.9% |
| The Prosodically Annotated Corpus of Spoken Russian | 20 | 13:22 | 412 | 20 | 83% |
| The DoBeS Summits-PAGE Collection of Papuan Malay | 20 | 64:07 | 2995 | 33 | 89.1% |
| The DoBeS Wooi Documentation | 12 | 34:59 | 1033 | 18 | 64.6% |
| The DoBes Yali Documentation | 5 | 13:25 | 561 | 10 | 80.2% |

*Table 1*. Summary information on the sample of speech segments used in the study.

## 2.2 Phase-consistency analysis

We analyzed the relation between IU onsets and the speech envelope using a point-field synchronization measure, adopted from the study of rhythmic synchronization of neural spiking activity and Local Field Potentials (Vinck et al., 2010). In this analysis, the rhythmicity of IU sequences is measured through the phase consistency of IU onsets with the periodic components of the speech envelope (Figure 1c). The speech envelope is typically modulated at a rate of 2-7 Hz (Chandrasekaran et al., 2009; Ding et al., 2017; Varnet et al., 2017), and is understood to reflect predominantly the sequence of syllables in speech. The vocal nuclei of syllables are the main source of envelope peaks, while syllable boundaries are the main source of envelope troughs (Figure 1b). IU onsets can thus be expected to coincide with troughs in the envelope, since each IU onset is necessarily also a syllable boundary. Therefore, one can expect a high phase consistency between IU onsets and the frequency component of the speech envelope corresponding to the rhythm of syllables, at ~5 Hz.

A less trivial finding would be a high phase consistency between IU onsets and other periodic components in the speech envelope. Specifically, since IUs typically include more than one syllable, such an effect would pertain to frequency components below ~5 Hz. In this analysis we hypothesized that the syllable organization within IUs gives rise to slow periodic components in the speech envelope. If low-frequency components are negligible in the speech envelope, estimating the phase of the low-frequency components at the time of IU onsets would lead to random phase angles, a result that would translate to low phase consistency. In another scenario, if the speech envelope captures slow rhythmicity in language other than that arising from IUs, different IUs would occur in different phases of the lower frequency components, translating again to low phase consistency. Contrary to these scenarios, finding phase consistency at a degree higher than expected under the null hypothesis would indicate both that the speech envelope captures the rhythmic characteristics of IUs and would characterize the period of this rhythmicity.

We extracted 2-second windows of the speech envelope centered on each IU onset, and decomposed them using FFT with a single Hann window, no padding, and after demeaning. This yielded phase estimations for frequency components at a resolution of 0.5 Hz. Then, we measured the consistency in phase of each FFT frequency component across speech segments using the pairwise phase consistency metric (PPC; Vinck et al., 2010), yielding a consistency spectrum. We calculated consistency spectra separately for each speaker that produced > 5 IUs and averaged the spectra within each language. Note, that the PPC measure is unbiased by the number of 2-second envelope windows entering the analysis (Vinck et al., 2010), and likewise that in a turn-taking sequence, it is inevitable that part of the 2-second envelope windows capture speech by more than one participant. The results remain unchanged when creating a single spectrum per conversation without separating speakers, or when including all speakers that produced more than two IUs (not shown).

### 2.3 Statistical assessment

We assessed the statistical significance of peaks in the average consistency spectra using a randomization procedure (Figure 1d). Per language, we created a randomization distribution of consistency estimates with 1000 sets of average surrogate spectra. These surrogate spectra were calculated using the speech envelope as before, but with temporally permuted IU onsets that maintained the association with envelope troughs. Troughs are defined by a minimum magnitude of 0.01 (on a scale of 0-1), and with a minimal duration between troughs of 200 ms, as would be expected from syllables, on average. By constraining the temporal permutation of IU onsets, we address the fact that each IU onset is necessarily a syllable onset, and therefore is expected to align with a trough in the envelope. We then calculated, for each frequency, the proportion of consistency estimates (in the 1000 surrogate spectra) that were greater than the consistency estimate obtained

for the observed IU sequences. We corrected p-values for multiple comparisons across frequency bins ensuring that on average, FDR will not exceed 1% (Genovese, Lazar, & Nichols, 2002).

## Results

We studied the temporal structure of IU sequences through their alignment with the periodic components of the speech envelope, using a phase-consistency analysis. This analysis builds on a cognitively-oriented linguistic theory which is supported by empirical speech analysis with wide cross-linguistic validity. We hypothesized that one of the characteristics of IUs – the fast-slow dynamic of syllables – would give rise to slow periodic modulations in the speech envelope. Figure 2 displays the observed phase consistency spectra in the six sample languages. IU onsets appear at significantly consistent phases of the low-frequency components of the speech envelope, indicating that their rhythm is captured in the speech envelope, hierarchically above the syllabic rhythm at 5 Hz. We find the consistent phase relation in all six languages (English, Hebrew, Russian, Papuan Malay, Wooi and Yali) at approximately 1 Hz.

We sought to confirm that this effect was not a result of an amplitude transient at the beginning of IU sequences. To this end, we repeated the analysis, submitting only IUs that followed an inter-IU interval below 1 s, which corresponded to 65% of the IUs in Wooi, and ~80% of the IUs in the other five languages (Table 1). The consistency estimates at 1 Hz were still larger than expected under the null hypothesis that IUs lack a definite rhythmic structure (Figure S2).

Our results are consistent with preliminary characterizations of the temporal structure of IUs (Chafe, 1987, 2018; Jun, 2005). The direct time measurements we used obviate the pitfalls of length measurements in word count or syllable count (e.g., Chafe, 1994; Himmelmann et al., 2018; Jun, 2005; Pawley & Syder, 2000). The temporal structure of IUs cannot be inferred from reported word counts, because what constitutes a word varies greatly across languages (Box 2). Syllable count per IU may provide an indirect estimation of IU length, especially if variation in syllable duration is taken into account (e.g., Silber-Varod & Levy, 2014), but it does not capture information about the temporal structure of IU *sequences*.

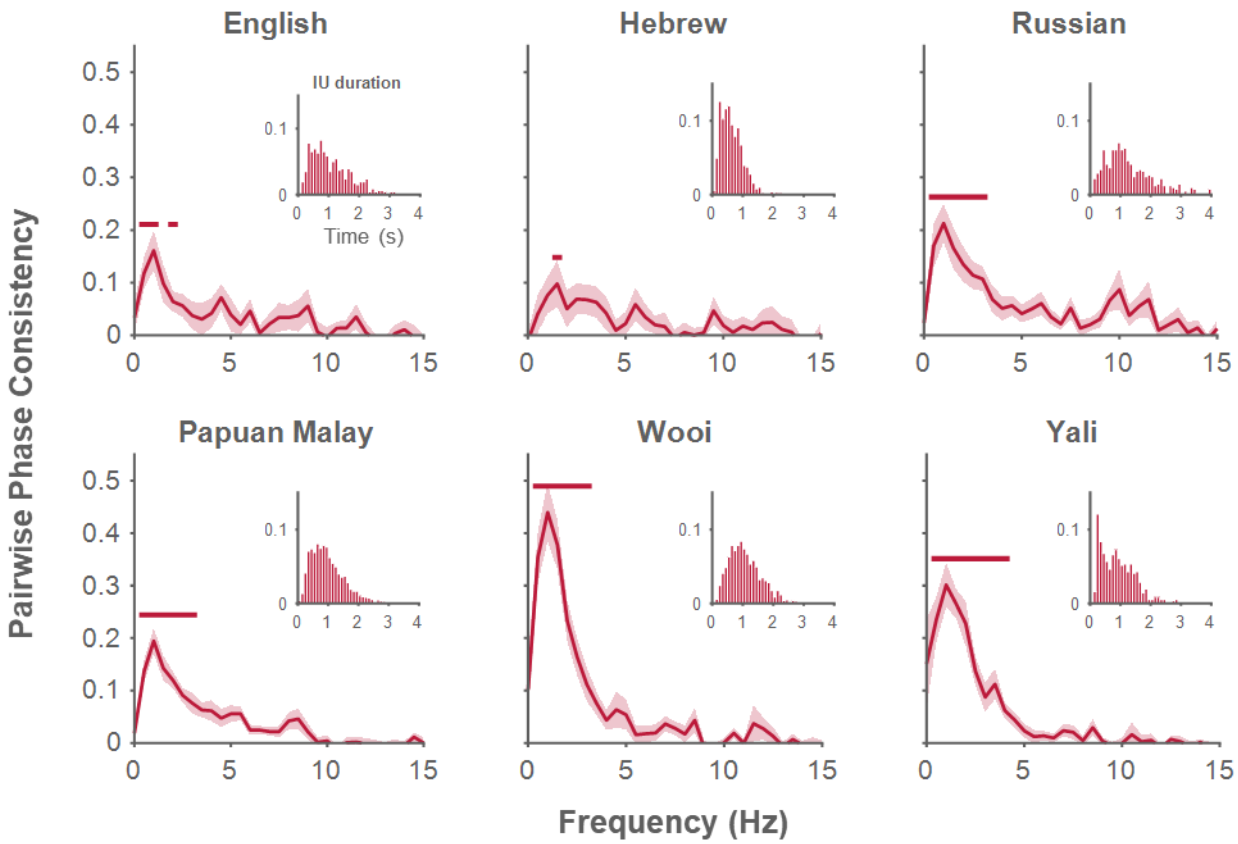


**Figure 2. Characterization of the temporal structure of Intonation Units.** Phase-consistency analysis results. Shaded regions denote ±SEM across speakers. Significance is denoted by a horizontal line above the spectra, after correction for multiple comparisons across neighboring frequency bins using an FDR procedure. Inset: Probability distribution of IU durations within each language corpus, calculated for 50 ms bins and pooled across speakers.

## Discussion

Neuroscientific studies suggest that neural oscillations participate in segmenting the auditory signal and encoding linguistic units during speech perception. Many studies focus on the role of oscillations in the theta range (3-5 Hz) and in the gamma range (>40- Hz). The levels of segmentation attributed to these ranges are the syllable level and the fine-grain encoding of phonetic detail, respectively (Giraud & Poeppel, 2012). Studies identify also slower oscillations, in the delta range (<2- Hz), and have attributed them to segmentation at the level of phrases, both prosodic and semantic/syntactic (Bonhage, Meyer, Gruber, Friederici, & Mueller, 2017; Bourguignon et al., 2013; Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Gross et al., 2013; Keitel, Ince, Gross, & Kayser, 2017; Meyer, Henry, Gaston, Schmuck, & Friederici, 2017; Park et al., 2015). Consistently, other findings demonstrated a decrease in high-frequency neural activity at points of semantic/syntactic completion (Nelson et al., 2017) or natural pauses between phrases (Hamilton, Edwards, & Chang, 2018). This pattern of results yields a slow modulation aligned to phrase structure. We harness linguistic theory to offer a conceptual framework for such slow modulations. We quantify the phase consistency between the slow modulations and IU onsets, and for the first time demonstrate that prosodic units with established functions in cognition give rise to a low-frequency rhythm in the auditory signal available to listeners.

Previous research has proposed to dissociate delta activity that represents acoustically-driven segmentation following prosodic phrases from delta activity that represents knowledge-based segmentation of semantic/syntactic phrases (Meyer, 2017). From the perspective of studying the temporal structure of spontaneous speech, we suggest that the distinction maintained between semantic/syntactic and prosodic phrasing might be superficial. That is because the semantic/syntactic building blocks always appear within prosodic phrases in natural language use (Auer, Couper-Kuhlen, & Müller, 1999; Hopper, 1987; Kreiner & Eviatar, 2014; Mithun, 2009). Studies investigating semantic/syntactic building blocks often compare the temporal dynamics of intact grammatical structure to word lists or grammatical structure in an unfamiliar language (e.g., Bonhage et al., 2017; Ding et al., 2016; Nelson et al., 2017). We argue that such studies need to incorporate a possibility that ongoing processing dynamics might reflect perceptual chunking, owing to the ubiquity of prosodic segmentation cues in natural language experience. This possibility is further supported by the fact that theoretically-defined semantic/syntactic boundaries are known to enhance the perception of prosodic boundaries. In a study that investigated the role of syntactic structure in guiding the perception of prosody in naturalistic speech (Cole, Mo, & Baek, 2010), syntactic structure was found to make an independent contribution to the perception of prosodic grouping. Another study equated prosodic boundary strength experimentally (controlling in a parametric fashion word duration, pitch contour, and following-pause duration), and found the same result: semantic/syntactic completion contributed to boundary perception (Buxó-Lugo & Watson, 2016). Even studies that use visual serial word presentation paradigms

rather than auditory stimuli are not immune to an interpretation of prosodically-guided perceptual chunking, which is known to affect silent reading (for a review, see Breen, 2014; Fodor, 1998).

Independent of whether delta activity in the brain of the listener represents acoustic landmarks, abstract knowledge or the prosodically-mediated embodiment of abstract knowledge (Kreiner & Eviatar, 2014), our results point to another role for slow rhythmic brain activity (1 Hz). We find that orthogonal to divergent grammatical systems, speakers and speech mode, speakers express their developing ideas at a rate of approximately 1 Hz. Previous studies have shown that in the brains of listeners, a wide network interacts with low-frequency auditory-tracking activity, suggesting an interface of prediction and attention-related processes, memory and the language system (Kayser, Ince, Gross, & Kayser, 2015; Keitel et al., 2017; Park et al., 2015; Piai et al., 2016; Schroeder & Lakatos, 2009). We expect that via such low-frequency interactions, this same network constraints spontaneous speech production, orchestrating the management and communication of conceptual foci (Chafe, 1994).

**Acknowledgement**

MI is supported by the Humanities Fund PhD program in Linguistics and the Jack, Joseph and Morton Mandel School for Advanced Studies in the Humanities. Our work would not have been possible without the tremendous efforts carried by the creators of the corpora and their teams. We deeply thank them and the many people they recorded all over the world.

## Appendix A: Supplementary Material

### Data

We analyzed data from six corpora of spontaneous speech, that fall into two groups.

The first group consists of corpora in English, Hebrew and Russian, transcribed and segmented by professional teams, each working on their native language: The Santa Barbara Corpus of Spoken American English (Du Bois et al., 2005), The Haifa Corpus of Spoken Hebrew (Maschler et al., 2017), and The Prosodically Annotated Corpus of Spoken Russian (Kibrik & Podlesskaya, 2014). We constructed a subsample from each of the three corpora according to the following protocol: 10 recordings of conversations were sampled randomly from the English and Hebrew corpora. The Russian corpus includes narratives rather than conversations, so 20 recordings were sampled randomly to cover a comparably wide number of speakers. The English and Hebrew corpora include transcriptions segmented into intonation units (IUs) that are not timestamped to the audio file. Author MI manually measured the IU- onset and -offset times in the first 30 seconds of each recording, continuing until the next speaker change in each recording. Pauses between IUs were not considered as part of the neighboring IUs' duration, in deviation from the transcription guidelines of Du Bois et al. (1992). Spectrogram analyses in Praat 6.0.23 (Boersma & Weenink, 2016) accompanied this procedure, including pitch and intensity contours produced by Praat's default settings. Measurements in milliseconds were entered into ELAN 4.9.4 (Wittenburg, Brugman, Russel, Klassman, & Sloetjes, 2006), on separate tiers for different speakers. The Russian corpus includes the onset and offset times of IUs, measured in a similar fashion. The 20 sampled recordings were analyzed in full. Finally, in English and Hebrew, additional information was recorded for each IU, but was not further analyzed: the intonation contour attributed by the corpus constructors, the position within the turn, and a categorization to type of IU – Fragmentary, Regulatory or Substantive (Chafe, 1994).

The second group consists of corpora in Papuan Malay, Wooi and Yali: The DoBeS Summits-PAGE Collection of Papuan Malay (Himmelmann & Riesberg, 2016), The DoBeS Wooi Documentation (Kirihio et al., 2015), and The DoBes Yali Documentation (Riesberg et al., 2016). Papuan Malay (ISO 639-3 code: pmy) is the *lingua franca* of West Papua, the western half of the island of New Guinea governed by Indonesia. Wooi (ISO 639-3 code: wbw) and Yali (ISO 639-3 code: yac) belong to different language families, have very different grammatical profiles, and are spoken in different regions in West Papua – on the coast and in the highlands, respectively. All three corpora include retellings of the plot of a silent film by a narrator to an interlocuter. More data on these languages and corpora can be found in Himmelmann et al., (2018), a study that was dedicated to the ability to identify IUs in unfamiliar languages. Researchers with varying degrees of familiarity with the languages transcribed them, and the agreement between them was quantified and found to be equally high to the degree of agreement in their segmentation of familiar

languages. Nikolaus Himmelmann, kindly provided us with the consensus transcriptions that the team constructed, and that were already entered into ELAN (Wittenburg et al., 2006) and time-aligned to the recording. We retrieved the respective audiofiles from the DoBeS archive. The Eastern Indonesia data in Himmelmann et al. (2018) included an additional language that we did not attempt to analyze because it included only two recordings. Additionally, we could not retrieve the audiofile of one of the original 6 narratives from the Yali consensus set, so only 5 Yali recordings were analyzed.

Data were extracted from all the ELAN files with the aid of custom-written scripts in R 3.3.3 (R Core Team, 2017) to data frames including IU onset and offset times, the computed duration, and the speaker of each IU. The remaining analyses were performed in MATLAB (MATLAB version 9.2.0.5, R2017a), using custom-written scripts and the FieldTrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011).

**Speech envelope computation**

The amplitude envelope was computed for the audio file of each speech segment following the methods used in several studies demonstrating neural entrainment to the envelope (e.g., Gross et al., 2013), methods that partly follow Chandrasekaran et al. (2009).

Speech segments that were recorded in stereo were converted to mono by averaging the channels. Recordings with a sampling rate above 20 kHz were downsampled to 20 kHz. Speech segments were band-pass filtered into 10 bands between 200 Hz and half the audio file's sampling frequency, with cut-off points designed to be equidistant on the human cochlear map. Amplitude envelopes for each band (the narrowband envelopes) were computed as absolute values of the Hilbert transform. These narrowband envelopes were downsampled to 1000 Hz and subsequently averaged, yielding the wideband envelope. The wideband envelope was smoothed using a 50 ms sliding Gaussian filter and divided by its maximal value to be on a scale of 0-1.

## Appendix B: Additional Analyses

### Characterization of the temporal structure of another prosodic unit

We conducted an identical analysis for prosodic units in an additional corpus, Rhapsodie: A Prosodic-Syntactic Treebank for Spoken French (Lacheret et al., 2014). This corpus included a semi-automatic & timestamped segmentation into prosodic units. The segmentation used a different combination of rhythmic and melodic cues, most notably relying on the notions of prosodic prominence and disfluency and assuming that these define different degrees of prosodic cohesion. The different degrees of prosodic cohesion yield a hierarchy of prosodic units, from which we extracted the level of rhythmic groups, delimited by weak or strong prominences on phonological word-final syllables that are non-disfluent. Below rhythmic groups in the hierarchy of prosodic units are metrical feet (delimited by all non-disfluent prominences, whether or not

they are word-final, and hence likely do not correspond to IU final boundaries). Above rhythmic groups in the hierarchy of prosodic units are intonation packages (delimited by strong prominences). The gradient prominence annotation is based on unquantified auditory inspection. It is likely based on pitch, volume, and duration variations, the latter of which might be compatible with the notion of word-final lengthening used in the segmentation of IUs. We selected the level of rhythmic groups to impose the minimal requirement that the prominence be phonological word-final, noting that the prominence annotation does not entail that all phonological words end in prominence. Information regarding the French sample may be found in Table 2.

| **Source of recordings and transcriptions** | **Number of recordings** | **Audio duration (minutes)** | **Number of IUs** | **Number of speakers with > 5 IUs** | **% IUs following inter-IU interval < 1 s** |
|---|---|---|---|---|---|
| Rhapsodie: A Prosodic-Syntactic Treebank for Spoken French | 14 | 13:09 | 900 | 14 | 95.2% |

*Table S2.* Summary information on the sample of speech segments used in the analysis of the French corpus.

The analysis of the French corpus provided the unique opportunity to test the notion that prosodic boundaries arise, among other criteria, from the fast-slow dynamic of syllables. Unlike the six IU-segmented corpora used in the main analysis, the French corpus includes a time-aligned annotation of syllables. Therefore, we could use actual syllable onset times for the permutation of IU time-courses instead of indirectly estimating syllables onsets based on troughs in the envelope. The results of this analysis are presented in Figure S1.

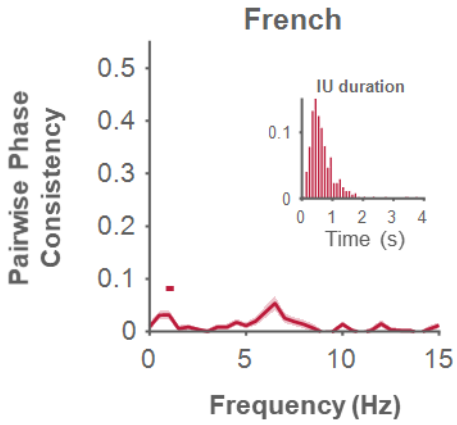


**Figure S1. Phase-consistency in the French corpus.** Phase-consistency analysis results. Shaded regions (red) denote ±SEM across speakers. Significance was assessed for the 1 Hz frequency component and is denoted by a horizontal line above the peak (uncorrected). Inset: Probability distribution of IU durations within the language corpus, calculated for 50 ms bins and pooled across speakers.

**Addressing a question of transient effects**

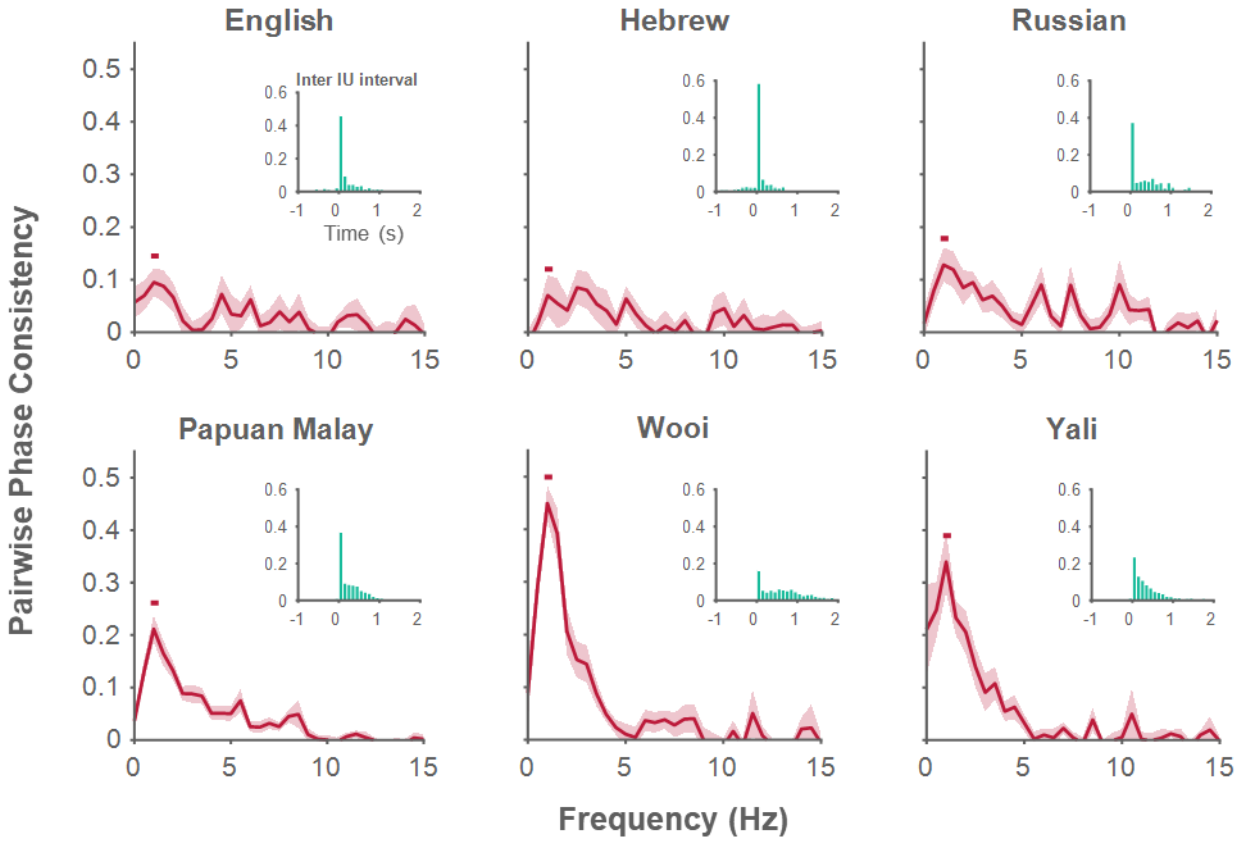


**Figure S2. Removal of IUs following long pauses does not qualitatively change the characterization of the temporal structure of Intonation Units** (item corresponding to Figure 2). Phase-consistency analysis results. Shaded regions denote ±SEM across speakers. Significance was assessed for the 1 Hz frequency component of each spectrum and is denoted by a horizontal line above the peak (uncorrected). Inset: Probability distribution of all inter-IU-interval durations within each language corpus, calculated for 50 ms bins and pooled across speakers.

## References


Auer, P., Couper-Kuhlen, E., & Müller, F. (1999). The Study of Rhythm: Retemporalizing the Detemporalized Object of Linguistic Research. In *Language in Time: The Rhythm and Tempo of Spoken Interaction* (pp. 3–34). New York: Oxford University Press.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004

Bonhage, C. E., Meyer, L., Gruber, T., Friederici, A. D., & Mueller, J. L. (2017). Oscillatory EEG dynamics underlying automatic chunking during sentence processing. *NeuroImage*, *152*(September 2016), 647–657. https://doi.org/10.1016/j.neuroimage.2017.03.018

Bourguignon, M., De Tiège, X., De Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., … Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping*, *34*(2), 314–326. https://doi.org/10.1002/hbm.21442

Breen, M. (2014). Empirical investigations of the role of implicit prosody in sentence processing. *Linguistics and Language Compass*, *8*(2), 37–50. https://doi.org/10.1111/lnc3.12061

Buxó-Lugo, A., & Watson, D. G. (2016). Evidence for the influence of syntax on prosodic parsing. *Journal of Memory and Language*, *90*, 1–13. https://doi.org/10.1016/j.jml.2016.03.001

Chafe, W. (1987). Cognitive constraints on information Flow. *Coherence and Grounding in Discourse*, (June 1984), 21–51. https://doi.org/10.1075/tsl.11.03cha

Chafe, W. (1994). *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Chafe, W. (2015). *A Grammar of the Seneca Language* (Vol. 149). Oakland: University of California Press.

Chafe, W. (2018). *Thought-based Linguistics*. Cambridge University Press. https://doi.org/10.1017/9781108367493

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7). https://doi.org/10.1371/journal.pcbi.1000436

Cole, J., Mo, Y., & Baek, S. (2010). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, *25*(7–9), 1141–1177. https://doi.org/10.1080/01690960903525507

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2017.02.011

Du Bois, J. W. (1987). The Discourse Basis of Ergativity. *Language*, *63*(4), 805–855.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2005). Santa Barbara corpus of spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., Cumming, S., Schuetze-Coburn, S., & Paolino, D. (1992). *Discourse Transcription*. *Santa Barbara Papers in Linguistics* (Vol. 4).

Fodor, J. D. (1998). Leaning to Parse? *Journal of Psycholinguistic Research*, *27*(2), 285–319. https://doi.org/10.1353/cjl.2004.0023

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and grammar* (pp. 134–184). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.003

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, *15*(4), 870–878. https://doi.org/10.1006/nimg.2001.1037

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, *25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech - A syllable-centric perspective. *Journal of Phonetics*, *31*(3–4), 465–485. https://doi.org/10.1016/j.wocn.2003.09.005

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12). https://doi.org/10.1371/journal.pbio.1001752

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, *28*(12), 1860-1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax, *45*(1), 31–80. https://doi.org/10.1515/flin.2011.002

Himmelmann, N. P., & Riesberg, S. (2016). The DoBeS Summits-PAGE Collection of Papuan Malay 2012-2016. The Language Archive MPI Nijmegen. https://doi.org/https://hdl.handle.net/1839/00-0000-0000-0019-FF78-5

Himmelmann, N. P., Sandler, M., Strunk, J., & Unterladstetter, V. (2018). On the universality of intonational phrases – a cross-linguistic interrater study. *Phonology*, *35*(2).

Hopper, P. J. (1987). Emergent Grammar. *Annual Meeting of the Berkeley Linguistics Society*, *13*, 139–157.

Jun, S. (2005). Prosodic Typology. In *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 430–458). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199249633.001.0001

Kayser, S. J., Ince, R. A. A., Gross, J., & Kayser, C. (2015). Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(44), 14691–14701. https://doi.org/10.1523/JNEUROSCI.2243-15.2015

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2003). Domain-initial articulatory strengthening in four languages. *Papers in Laboratory …*, *235*(January), 1–21. https://doi.org/10.1177/009614429902500406

Keitel, A., Ince, R. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage*, *147*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.11.062

Kibrik, A. A., & Podlesskaya, V. I. (2014). Spokencorpora. Rasskazy o snovidenijax i drugie korpusa zvučaščej reči (Prosodically annotated corpus of spoken Russian). Retrieved from http://spokencorpora.ru/

Kirihio, J. K., Unterladstetter, V., Arilaha, A., Morigerowsky, F., Loch, A., Sawaki, Y., & Himmelmann, N. P. (2015). The DobeS Wooi Documentation 2009-2015. The Language Archive MPI Nijmegen. https://doi.org/https://hdl.handle.net/1839/00-0000-0000-0014-C76C-1

Kreiner, H., & Eviatar, Z. (2014). The missing link in the embodiment of syntax: Prosody. *Brain and Language*, *137*, 91–102. https://doi.org/10.1016/j.bandl.2014.08.004

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., … (2014)., et al. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken french. In *Language Resources and Evaluation Conference*. Retrieved from http://www.projet-rhapsodie.fr

Maschler, Y., Polak-Yitzhaki, H., Fishman, S., Miller Shapiro, C., Goretsky, N., Aghion, G., & Fofliger, O. (2017). The Haifa Corpus of Spoken Hebrew. Retrieved from http://weblx2.haifa.ac.il/~corpus/corpus_website

MATLAB version 9.2.0.5. (2017). MATLAB. Natick, Massachusetts: The MathWorks, Inc.

Meyer, L. (2017). The neural oscillations of speech processing and language comprehension: State of the art and emerging mechanisms. *European Journal of Neuroscience*, 1–13. https://doi.org/10.1111/ejn.13748

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, *27*(9), 4293–4302. https://doi.org/10.1093/cercor/bhw228

Mithun, M. (2009). Re(e)volving complexity: Adding intonation. In T. Givón & M. Shibatani (Eds.), *Syntactic Complexity: Diachrony, acquisition, neuro-cognition, evolution* (pp. 53–80). Amsterdam: John Benjamins Publishing Company.

Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Social Interaction*, *51*(1), 85–106. https://doi.org/10.1080/08351813.2018.1413878

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114

Ono, T., & Thompson, S. A. (1995). What can conversation tell us about syntax? In P. W. Davis (Ed.), *Alternative Linguistics: Descriptive and Theoretical Modes* (pp. 213–272). Amsterdam: John Benjamins Publishing Company.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*. https://doi.org/10.1155/2011/156869

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Current Biology*, *25*(12), 1649–1653. https://doi.org/10.1016/j.cub.2015.04.049

Pawley, A., & Syder, F. H. (2000). The One-clause-at-a-time Hypothesis. In H. Riggenbach (Ed.), *Perspectives on Fluency*. Ann Arbor: University of Michigan Press.

Piai, V., Anderson, K. L., Lin, J. J., Dewar, C., Parvizi, J., & Dronkers, N. F. (2016). Direct brain recordings reveal hippocampal rhythm underpinnings of language processing, *113*(40).

https://doi.org/10.1073/pnas.1603312113

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*(November 2017), 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Riesberg, S., Walianggen, K., & Zöllner, S. (2016). The Dobes Yali Documentation 2012-2016. The Language Archive MPI Nijmegen. https://doi.org/https://hdl.handle.net/1839/00-0000-0000-0017-EA2D-D

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Silber-Varod, V., & Levy, T. (2014). Intonation Unit Size in Spontaneous Hebrew : Gender and Channel Differences Department of Electrical and Computer Engineering , Rutgers University , USA. *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, 658–662.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179

Vinck, M., van Wingerden, M., Womelsdorf, T., Fries, P., & Pennartz, C. M. A. (2010). The pairwise phase consistency: A bias-free measure of rhythmic neuronal synchronization. *NeuroImage*, *51*(1), 112–122. https://doi.org/10.1016/j.neuroimage.2010.01.073

Wittenburg, P., Brugman, H., Russel, A., Klassman, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from http://tla.mpi.nl/tools/tla-tools/elan/