Two-dimensional parsing, the iambic-trochaic law, and the typology of rhythm

Michael Wagner, Alvaro Iturralde Zurita, and Sijia Zhang (McGill University)

Humans appear to be wired to perceive acoustic events rhythmically. English speakers tend to perceive alternating short and long sounds as sequences of binary groups with a final beat (iambs), but alternating soft and loud sounds as trochees (Bolton, 1894; Woodrow, 1909). This generalization (often called the 'Iambic-trochaic Law' (ITL), following Hayes 1995), has been hypothesized to be a universal of auditory processing (Hay and Diehl, 2007). Kusumoto and Moreton (1997); Iversen et al. (2008), Bhatara et al. (2013), and Crowhurst and Teodocio Olivares (2014), however, found that the duration-side of the ITL fails to apply in Japanese, French, and Spanish, suggesting that rhythm perception is shaped by language experience. This has been attributed to cross-linguistic differences in word order (Iversen et al., 2008) or stress-systems (Bhatara et al., 2013). This prior work has an important limitation: If one parses a sequence of sounds (e.g. iterations of the sylllables 'ba' and 'ga') into binary groups, there are not 2 but (at least) 4 potential percepts (e.g., BAga, baGA, GAba, gaBA). Prior research usually asked in some way or other about the perceived foot (*Did you hear [X x] or [x X]?)*. This task only narrows things down to 2 out of 4 possibilities (e.g. both BAga and GAba are trochees). Crowhurst and Teodocio Olivares (2014) used a speech segmentation task which also only narrows narrows things down to two possibilities (BAga/baGA for a 'baga' response; GAba/gaBA for a 'gaba' response).

Wagner (under review) argues that the ITL is a simple consequence of the cue distribution for the perceptual dimensions of grouping and prominence. Production data show that in words and phrases, prominent syllables are both louder and longer, but these two cues anti-correlate when encoding grouping: initial syllables are louder, final syllables longer. Using two tasks to fully determine the percept (*Which syllable is initial?*, *Which syllable is prominent?*), one can see that the ITL is simply a consequence of this cue distribution. The perception data can be predicted from the cue distribution seen in production, including the ITL effect: If a sound is sufficiently long, it will be perceived as final and stressed, and if it is sufficiently loud, as initial and stressed.

Our **first** contribution is to **replicate** the perception findings in Wagner (2020) (perception of syllable sequences e.g, *..bagaba...*, results in Fig. 1): Listeners make consistent prominence choices when intensity and duration correlate (consistent cues for prominence), and are closer to chance when they anti-correlate. They make more consistent grouping decisions when the cues anti-correlate (consistent cues for grouping since louder=initial; longer=final), and are closer to chance when they correlate. The foot decision, which can be reconstructed from the grouping and prominence decisions, shows the ITL pattern when only one cue is manipulated in an extreme way (trochees for an intensity difference; iambs for a duration difference), but shows little systematicity when both cues are manipulated. The choice between iamb and trochee is epiphenomenal, the choice of prominence and grouping highly systematic.

Our **second and central** contribution is to establish the beginnings of a **parsing typology** based on experiments in 4 more languages. The coefficients (log odds) for intensity and duration in logistic models of the prominence and grouping decisions (Fig. 2) show that cue interpretation is not universal. In Mandarin and Japanese duration does not reliably cue grouping, and all languages differ significantly from each other in at least one coefficient. However, the differences between languages are small and non-significant for duration when looking at prominence, and for intensity when looking at grouping. So language learners could use the more invariable cue (duration for prominence; intensity for grouping) to bootstrap into the signal, even if the respective other cue is language specific. Children have been shown to show ITL effects early on, while duration ITL effects only emerge later (and only for some languages). This literature mostly used speech segmentation tasks. Given the findings here, maybe early on, children use duration mostly to parse for prominence. New work will be needed to look both at prominence and grouping to tease apart the acquisiotn path. Existing quantitative measures of rhythm in the acquisition literature have been criticized as tapping phonotactics or phonemic differences rather than rhythm (see Arvaniti 2012 i.a.). The typology based on the cues for grouping and prominence perception jointly may provide a better quantitative map of cross-linguistic differences in what we intuitively call 'rhythm.'

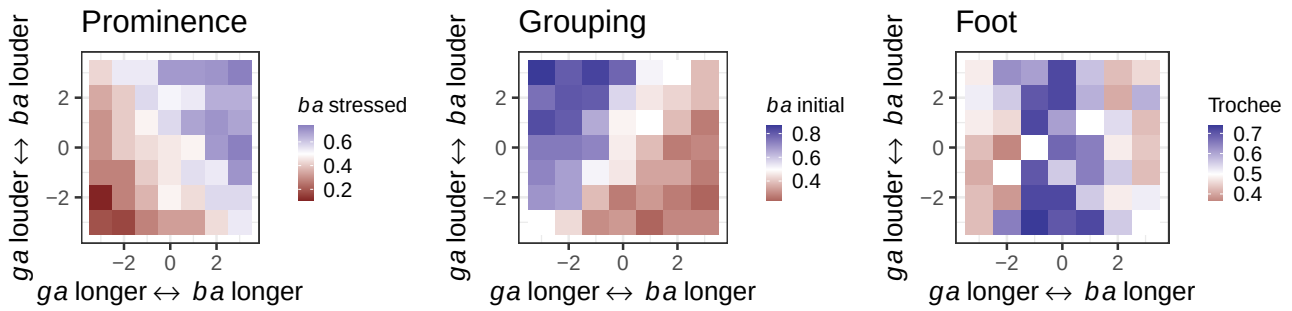

Figure 1: Listeners (50 adult native speakers of North American English) heard sequences of syllables *bagabaga...* or *gabagaba...*. The heatmaps show proportions of responses from the prominence task (*Which syllable was stressed (ba or ga)?*) and the grouping task (*Did you hear baga or gaba?*), plotted by relative duration (7 steps on x-axis) and intensity step (7 steps on y-axis); the foot decision (right) was reconstructed from these two responses.

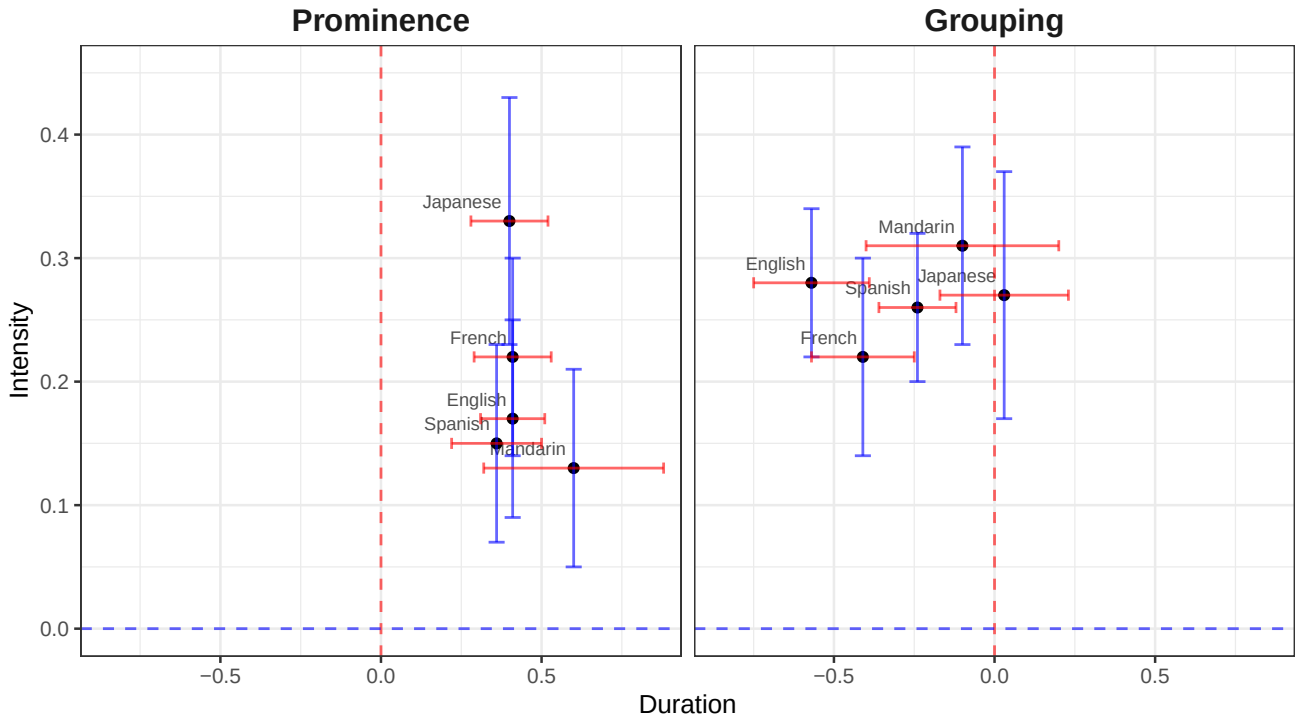

Figure 2: The parsing typology based on a small sample of 5 languages (NA English, French, Japanese, Mandarin, Mexican Spanish), plotting the coefficients from logistic MER models for the individual languages for each decision (50 listeners per language). Coefficients corresponds to the predicted change in log odds given a unit change in intensity/duration. Duration (x-axis) and intensity (y-axis) coefficients are shown for the prominence decision (left) and the grouping decision (right). Error bars show 2*se estimated by the logistic models. All error bars that don't cross the dashed zero line (red for duration, blue for intensity) came out significant (only two were not significant: duration in the grouping decision models in Mandarin and Japanese). Note that there is little variation on the horizontal (duration) for the prominence dimension (small differences, non-significant interaction duration*Language in all-language MER model), and little variation on the vertical (intensity) for the grouping decision (small differences, non-significant interaction duration*Language)