Prediction and adaptation in intonational processing
Timo B. Roettger1, Michael Franke2, & Jennifer Cole3
1Northwestern University, 2University of Osnabrück

Introduction: Intonation plays an integral role in comprehending spoken utterances. However, intonation exhibits a large degree of variability, raising the question as to how a listener reliably uses intonation to infer what a speaker intends to communicate. Despite this uncertainty, listeners rapidly integrate intonational information to predictively map a given intonational cue onto respective speaker intentions [e.g. 1,2,3] and adapt intonation-based inferences by what appears to be quick adjustment of beliefs about speaker production likelihoods [4,5]. Although less focused on in the literature, the absence of a pitch accent can also be predictive [6]. Indeed, rational interpreters should not make a categorical difference between presence and absence of a cue but show behaviour merely as a function of differential production likelihoods.

Against this background, we present mouse-tracking data pertaining to the incremental interpretation of present or absent intonational cues prior to lexically disambiguating material, with a between-subject manipulation of stochastic reliability of these cues.

Method: In a preregistered study, sixty German listeners were exposed to two different intonational contexts in a forced choice mouse tracking experiment [7]. As an answer to the question ‘Has the wuggy collected the violin then?’ (English translation of original German), they were exposed to either (1) ‘The wuggy then has collected the violin.’ or (2) ‘The wuggy then has collected the piano.’ In (1), a pitch accent on the verb indicates that the object has already been mentioned in the question. In (2), a pitch accent on the object indicates that the speaker refers to a contrastive object. Absence of a pitch accent on the auxiliary in (2) can be utilized as an informative cue to the contrastive nature of the referent. As a control condition, listeners were exposed to intonationally uninformative utterances, in which listeners had to wait for lexical disambiguation. Subjects were randomly assigned to two groups. In the reliable speaker group (RS), the mapping of pitch accent position and referent resolution was always reliable (verb ➔ given; object ➔ contrastive); in the unreliable speaker group (US), the mapping was crossed in 1/3 of trials, making the function mapping stochastically unreliable. Acoustic stimuli were resynthesized such that listeners could not utilize any prosodic information preceding the critical pitch accent (or absence thereof). Listeners had to click on visually presented pictures to indicate the intended referent during the unfolding auditory stimulus. The computer mouse trajectories (x,y coordinates) of their responses were recorded with a 100 Hz sampling rate. We present results for matching trials (i.e. discourse status and intonation match) and the turn-towards-target measurement (TTT), operationalized as the last point in time at which the trajectory horizontally turns towards the target (see Figure 1).

Results and discussion: Raw data and scripts can be assessed from our osf.io repository (osf.io/dnbuk). The results are shown in Figure 2. There was a significant effect of intonational context, such that verb elicited significantly earlier TTTs than object (p<.001), which elicited earlier TTTs than lexical (p<.001, see Figure 1). There was also a significant interaction of group and verb, indicating substantially later TTTs in US (p<.001). OBJECT was not significantly affected by US. These results conceptually replicate earlier findings, demonstrating the rapid integration of intonational cues during intention recognition [e.g. 1-3]. The acoustically early cue associated with verb allows listeners to infer the intended given referent before the lexical material becomes available. Listeners also use the absence of this cue (no accent on <has>) to anticipate reference resolution. We propose a formal model of rational rapid cue integration by linking the TTT measure to the listeners posterior odds over interpretation alternatives, given an incrementally unfolding utterance. Based on a few defensible qualitative constraints on beliefs about speaker production likelihoods (e.g., a pitch accent on the auxiliary is more likely when referring to the given referent than referring to the competitor) we show that observed differences between verb and object are compatible with rational incremental interpretation.

Beyond the present findings, we will present a series of follow up experiments, investigating the interplay between perceptual salience of the intonational cues (presence vs. absence of pitch accent) and their position within the utterance (prenuclear vs. nuclear).

The present line of research contributes to our understanding of how listeners deal with the ubiquitous uncertainty in processing intonation. We suggest that listeners infer speaker intentions based on both bottom-up acoustic cues as well as on dynamically adaptable probabilistic expectations about likely intonational contours in a given context.
Figure 1. Horizontal cursor position (A) and spatial resolution (B) of time- and space-normalized averaged trajectories.

![Diagram A: Mean horizontal cursor position over time](image)

![Diagram B: Mean trajectories](image)

Figure 2. Estimates (solid points) and 95%-Credible intervals (whiskers) for turn-towards-target measurements across focus conditions (color) and speaker-groups (shape). Small semi-transparent dots indicate mean values of subjects.

Unreliable trials mainly affected the predictive value of verb focus, leading to a substantial increase in the TTT measure.

References


