

Prediction of turn-taking in naturalistic conversations: the role of semantic, syntactic and prosodic cues

Corresponding Author (andromachi.tsoukala@uni-oldenburg.de), Andromachi Tsoukala¹,

Esther Ruigendijk¹, ¹University of Oldenburg and Cluster of Excellence Hearing4all

Background: Smooth conversational turn-taking is thought to rely on prediction, namely turn-end anticipation, supported by various (non)speech cues; yet, the relative contributions of semantic, syntactic, and prosodic cues remain unclear. Prior paradigms (button-press tasks [1, 2]) require actively monitoring for turn-ends, which may not reflect real-life processing. To overcome this limitation, we used eye-tracking while participants watched audiovisually (AV) recorded dyadic conversations in German, and measured spontaneous proactive gaze shifts to speakers prior to their turn's onset. By comparing the original stimuli to conditions with distorted *semantic* (500Hz low-pass filtering (LPF) of content words [2]), *syntactic* (500Hz LPF of function words [2]), and *prosodic* cues (flat pitch [1]), we assessed effects on anticipatory gaze timing. We hypothesised participants will increasingly shift their gaze to the next speaker over time (H1), and that cue distortion will lead to more gradual increases in gaze shifts to the next speaker compared to the original condition (H2a), with effects being larger when semantic cues are distorted and smaller when prosody is distorted (H2b).

Methods: As data collection is ongoing, we report pilot results: 10 native German participants (Mean_{AGE} = 24.7) viewed $\approx$ 2-minute-long AV stimuli (N = 20) showing two people conversing, then provided subjective effort ratings and answered comprehension questions. We extracted gaze samples around non-overlapping “gap” events (N = 103), i.e. silent intervals during turn transitions (floor transfers). Our “prediction” time window for analyses includes -500ms from gap onset to +200ms, to account for saccade planning [3]. We measured the proportion of looks to post-gap vs pre-gap speaker in 20ms bins, and modelled the empirical logs-odds ratio in GAMMs [4] with a smooth for Time varying by Condition, factor-smooths for participant, and an AR(1) correction for autocorrelation.

Results and Discussion: In line with H1, we found an effect of Time involving a steady increase in looks to the next speaker within our prediction time window, from $\approx$ -200ms (Figure 1). Differences in the smooths for Time by Condition (Figure 2) suggest that shortly after 0ms to 200ms of gap onset, there was a steeper increase in looks to the next speaker in the original condition compared to all cue-distorted ones, consistent with H2a. Differences among the cue-distorted conditions were mostly observed initially (between $\approx$ -500ms to -200ms) but attenuated soon after, thus not providing clear support for H2b. These novel, albeit preliminary, results suggest that semantic, syntactic and prosodic cues can make non-trivial contributions to turn-end anticipation. Analyses of further data will help clarify to what extent the distortion of different cues leads to distinct anticipatory gaze patterns, and to address questions about perceived effort (ongoing analyses of pupil dilation & effort ratings).

References

- [1] De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3).
- [2] Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6, 89.
- [3] Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: optimizing reciprocity. *Frontiers in psychology*, 6, 98.
- [4] Ito, A., & Knoeferle, P. (2023). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods*, 55(7).

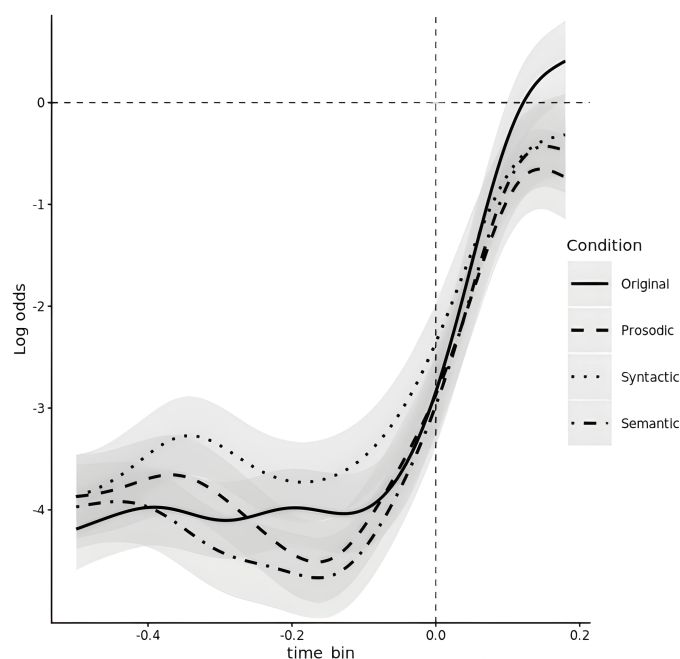


Fig. 1: Estimated log-odds of looks to next speaker over Time in seconds (shaded 95% CI).

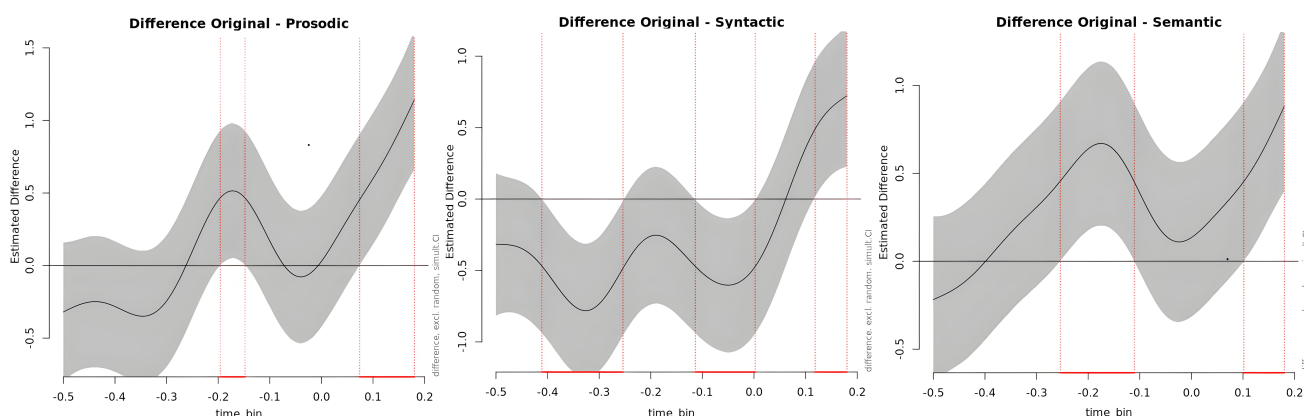


Fig. 2: Population-level difference slopes (Original – Cue-distorted conditions) showing odds of looks to next speaker over Time (seconds), with shaded 95% CI and highlighted intervals indicating significant differences (CI excludes zero).