A Dual-Route RSA Model of Pragmatic Inference: Dissociating Selective Attention to Common Ground from Theory of Mind Reasoning

Gemma Seghi (gemma.seghi@uni-tuebingen.de)[1], Laura Friedrich[1], Asya Achimova[1], Michael Franke[1], Gerhard Jäger[1]

[1]Department of Linguistics, University of Tübingen

The director task is a referential communication paradigm in which a listener must identify a target object based on instructions from a speaker with limited visual access: some objects are occluded from the speaker but visible to the listener (grey cells). Although the task was designed to assess Theory of Mind (ToM) use, accuracy alone cannot determine which perspective-taking mechanism is employed. In fact, ToM is cognitively costly [2] and optimal performance can also be achieved through a simpler heuristic: selectively attending to objects in common ground (CG) (white cells) and ignoring occluded ones, without explicitly reasoning about the speaker's perspective [5].

Experiments. Building on [3, 5], we propose two variants of the director task designed to test whether previous context shapes the perspective-taking strategy adopted by listeners (Figure 1). Both experiments contain 45 trials in blocks of 5, divided into a training phase (25 trials) and a test phase (20 trials). In Experiment 1's training phase, the cell hidden from the speaker is marked in grey, so listeners only need to identify the intended referent. In Experiment 2's training phase, the hidden cell is not shown, and listeners must jointly infer the target and the speaker's blindspot from the utterances within each block. The blindspot position changes every block and items, as well as blindspot positions, are matched across experiments. We hypothesize that Experiment 1 encourages a cheaper CG-filtering strategy, whereas Experiment 2 promotes continued use of full-blown ToM. The test phase is identical across experiments: the blindspot is shown in grey, but 2 trials per block use a suspicion manipulation adapted from [5]. On these trials, the object assumed to be hidden differs from the one actually hidden, yielding overinformative utterances that use color [4] to distinguish objects that should not both be visible to the speaker. This manipulation should elicit suspicion in those listeners who actively track the speaker's perspective, rather than simply ignoring the occluded objects. Given that both strategies should yield the same target choices, they are distinguished through two additional measures: reaction times and a post-task suspiciousness questionnaire. We predict more suspicion reports and longer RTs on suspicious trials in the test phase of Experiment 2 compared to the test phase of Experiment 1.

Model. We introduce an RSA model with a dual-route architecture. Given $w \in W$ (objects), $u \in U$ (utterances), and $b \in B$ (blindspots), a perspective-taking listener is modeled as a mixture of a cheap CG-based strategy and a ToM strategy: $L^{\text{persp}}(w \mid u) = (1 - \rho_i)L^{\text{cg}}(w \mid u, b) + \rho_i L^{\text{tom}}(w \mid u)$. The cheap listener $L^{\text{cg}}(w \mid u, b) \propto [\![u]\!]_{c(b)}(w)P(w)$ is a literal listener that evaluates the utterance over the CG domain $c(b)$ determined by the blindspot, containing only mutually visible objects [1]. The ToM listener instead jointly infers the intended target and the speaker's blindspot, $L^{\text{tom}}(w, b \mid u) \propto S_L(u \mid w, b)P(w, b)$, by inverting a cooperative speaker model $S_L(u \mid w, b) \propto \exp\big(\alpha[\log L^{\text{cg}}(w \mid u, b) - \text{cost}(u)]\big)$, which reasons about the CG domain induced by the blindspot while balancing informativeness against utterance cost. The target distribution is obtained by marginalizing over blindspots, $L^{\text{tom}}(w \mid u) = \sum_b L^{\text{tom}}(w, b \mid u)$. We assume that the mixture weight ρ_i, representing participant i's propensity to use the ToM route, depends on an individual baseline and on the training context: $\rho_i = \sigma(\gamma_0 + a_i + \gamma_{\text{ctx}}\text{Context}_i)$.

References

[1] Daphna Heller, Christopher Parisien, and Suzanne Stevenson. Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, 149:104–120, 2016. doi: https://doi.org/10.1016/j.cognition.2015.12.008.

[2] Cecilia Heyes. Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, 9(2):131–143, 2014. doi: 10.1177/1745691613518076.

[3] Julian Jara-Ettinger and Paula Rubio-Fernandez. Quantitative mental state attributions in language understanding. *Science Advances*, 7:970, 11 2021. doi: 10.1126/sciadv.abj0970.

[4] Paula Rubio-Fernandez. How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7(153), 2016. doi: 10.3389/fpsyg.2016.00153.

[5] Paula Rubio-Fernandez. The director task: A test of theory-of-mind use or selective attention? *Psychonomic Bulletin & Review*, 24:1121–1128, 2016. doi: 10.3758/s13423-016-1190-7.

Figures and Tables

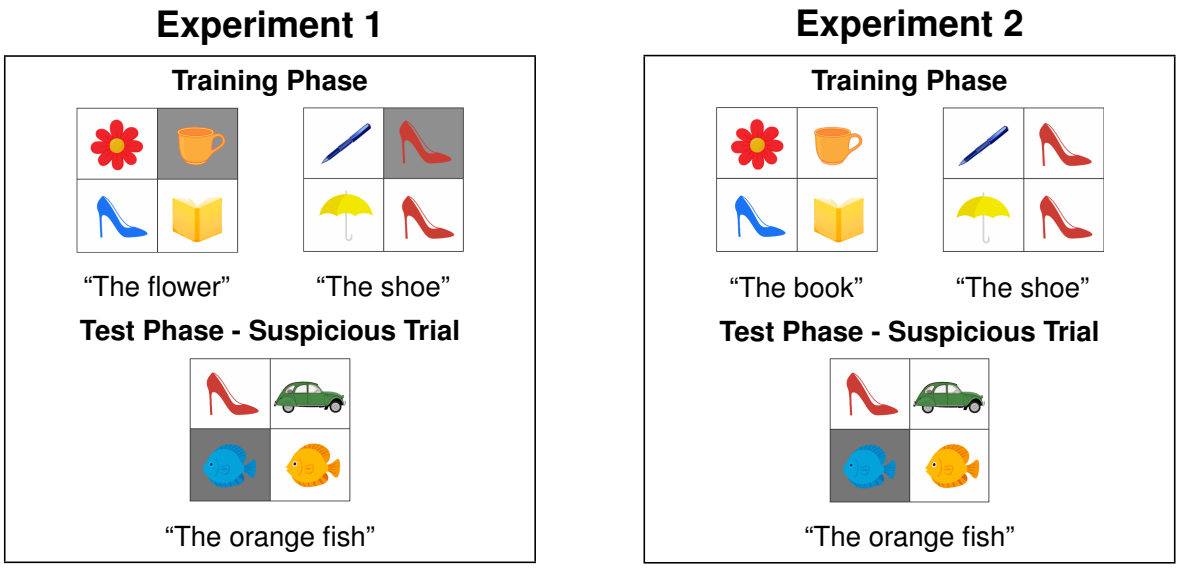

Figure 1: Experiment 1 (left): During training, the speaker's blindspot is marked in grey, and the listener only has to identify the target object. Experiment 2 (right): During training, the listener must identify the target object on each trial and infer the speaker's blindspot within each five-trial block. The test phase is identical across experiments: the blindspot is shown in grey, so no blindspot inference is required, but on suspicious trials the speaker's utterance is overinformative given their presumed blindspot.