Semantic adaptation supports pragmatic reasoning

Introduction. The language comprehension system is highly adaptable, rapidly recalibrating linguistic representations in response to input. It has been observed that some adaptation effects even generalize to related categories. For instance, [1] provide evidence of phonetic recalibration spreading to phonemes sharing key features (e.g., updates in /d/ generalize to other voiced stops like /b/), but it remains unclear whether comparable effects occur in the domain of meaning. This question is critical, as adapting to speaker input has been claimed to support pragmatic coordination by enabling finer-grained inferences about speaker meaning [2-5], yet direct evidence of a link between adaptation and pragmatic reasoning is lacking. We investigate whether similar generalization effects occur in semantic adaptation, and whether listeners use these generalized updates to guide pragmatic inference. Our results suggest a positive answer to both questions: we show that listeners track a speaker's usage preferences for vague quantifiers (e.g., *many*), and generalize these preferences to other vague predicates, such as relative adjectives (e.g., *smart*). Importantly, these cascading updates modulate the likelihood that these adjectives trigger scalar implicatures (SIs, e.g., *smart* $\rightsquigarrow$ *not brilliant*), in a manner consistent with the *semantic distance hypothesis* of SI [6,7]: the greater the distance between the updated threshold of the weaker scalemate and its stronger counterpart, the more likely an SI is to arise. These results provide the first evidence of generalization in semantic adaptation and highlight it as a central mechanism underlying pragmatic reasoning.

Experiment (n=60) The experiment followed an exposure-and-test paradigm [3,9]. During exposure, participants believed they were conversing with a remote partner viewing the same displays of 50 circles containing varying percentages (20–80$\%$) of the target color (blue or yellow). In reality, they interacted with a chatbot programmed with predetermined usage preferences for the quantifier *many*. *Speaker 1* used an acceptance threshold (θ) of 40$\%$, whereas *Speaker 2* applied a stricter criterion, using the quantifier only for proportions of 60$\%$ or higher (Fig. 1, left). Participants were adapted through metalinguistic disagreements [4,5,7,8]. They first described each display to the 'other participant' using the quantifiers *many* or *few*, and the bot then agreed or disagreed with their description based on the participant's predication and the speaker's programmed preferences (Fig. 1, middle). Speaker type was manipulated between participants. Next, participants moved to the test phase, which consisted of three blocks: 1) to test whether participants generalized quantifier thresholds to relative adjectives, we collected degree estimates for the adjectives later used in the SI task. The stimuli consisted of 38 scales where the weak member was a relative adjective and the strong member an absolute or extreme adjective (a subset of the scales used in [11]). Participants were told that the *other participant* had produced an utterance such as *the employee is smart* and were asked, *On a scale from 0 to 100, how smart do you think the other participant believes the employee is?* Responses were provided by selecting a point on a slider. 2) Next, participants completed an inference task to measure the SI rates elicited by the adjectives from the previous block. They were asked, *Would you conclude that the other participant thinks the employee is not brilliant?* and responded by clicking 'Yes' (= SI calculation) or 'No' (= no SI calculation). 3) Finally, we assessed adaptation to Speaker 1/2's quantifier preferences. Participants indicated with a binary 'Yes/No' response whether the speaker from the exposure phase would use the quantifier to describe a given display (See Fig. 1, right, for full design).

Results. Participants showed higher acceptance rates of the quantifier *many* in the Speaker 1 group (40$\%$ bot) than in the Speaker 2 group (40$\%$ bot; Fig. 2, left). An area under the curve (AUC) analysis confirmed this difference ($p < 0.01$; Fig. 2, right), indicating that participants adapted, on average, to the different thresholds exhibited by the two speakers during the exposure phase. Degree estimates are shown in Fig. 3. As seen in the plot, the correlation line generally falls below the unity line (dashed diagonal), indicating that estimates were higher on average in the Speaker 2 group than in the Speaker 1 group. This difference was confirmed by a two-sample t-test ($p < 0.05$). Finally, Fig. 4 (left panel) shows the SI results. A mixed-effects model confirmed that SI rates were higher on average in the Speaker 1 group than in the Speaker 2 group ($p < 0.01$). As shown in Fig. 4 (right panel), this difference was consistent across items, as indicated by the correlation line running parallel to the unity line. **Discussion&Conclusion**. Our results replicate previous findings that listeners track speakers' usage preferences for vague quantifiers [4,10] and extend them by showing generalization to related expressions such us vague adjectives: participants exposed to Speaker 1, who displayed a lower threshold for *many*, gave lower degree estimates for relative adjectives on average. Our results further show that participants used these generalized updates for pragmatic rea-

soning: those in the Speaker 1 group—with lower adjectival thresholds and thus greater semantic distance between scalemates—exhibited higher SI rates than those in the Speaker 2 group, as predicted by the *semantic distance hypothesis*. These findings provide preliminary evidence for generalization in semantic adaptation and highlight it as a central mechanism supporting pragmatic reasoning. Time permitting, we will also present results from a second study showing parallel effects, where listeners generalize speaker preferences for vague quantifiers to the thresholds of maximum standard absolute adjectives and use these generalizations to guide pragmatic inferences about imprecision.

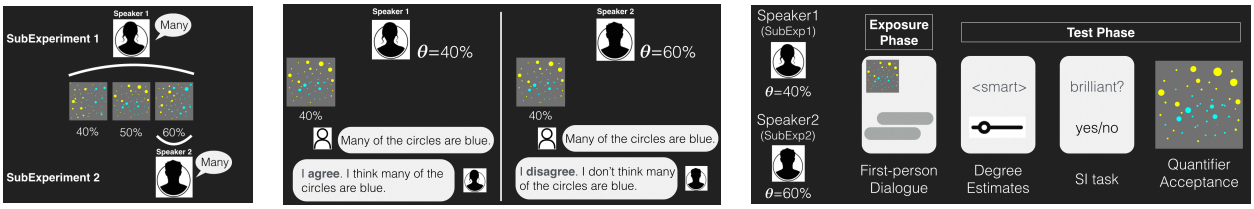

Figure 1: **Left:** Speaker 1/2 preferences for the quantifier *many* during the exposure phase; **Middle:** (Dis)agreement dialogues used in the exposure phase; **Right:** Overview of the experimental design.

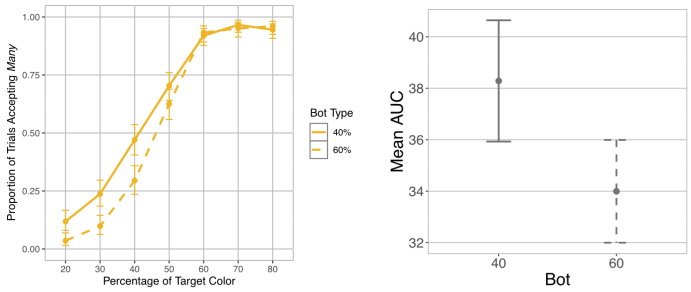

Figure 2: **Left:** Acceptance rates for the quantifier *many* in the test phase. Error bars indicate the 95% confidence intervals.; **Left:** Area under the curve (AUC) analysis. Error bars indicate the 95% confidence intervals.

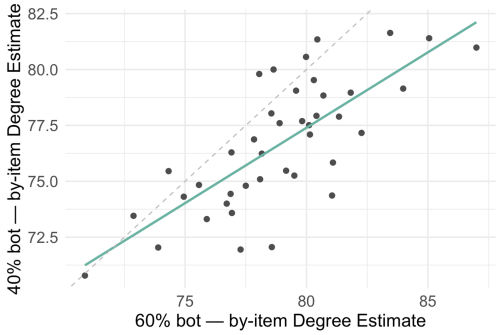

Figure 3: By-item correlation of degree estimates in Speaker 1/2 (40%/60% bot) groups. Dashed line represents unity line.

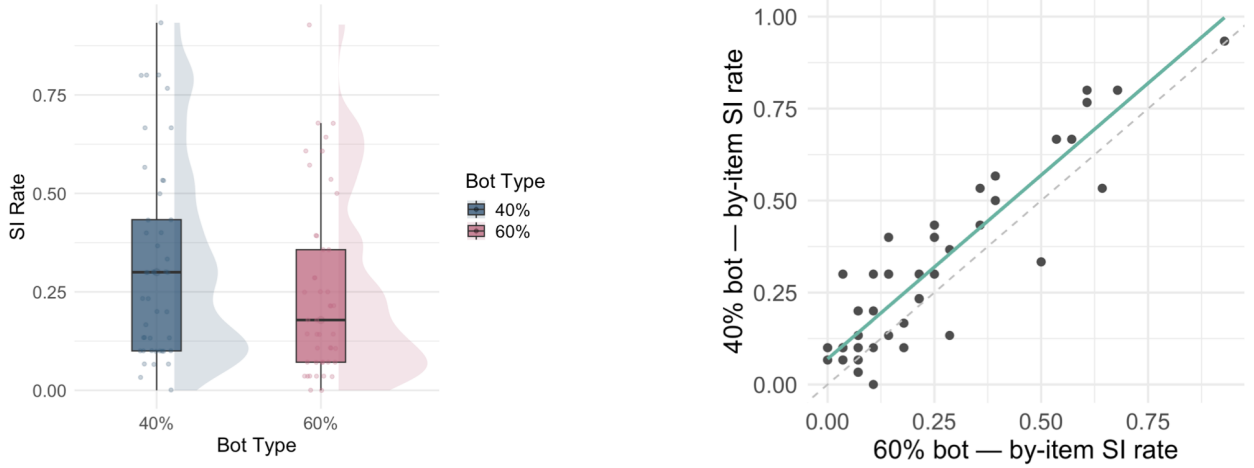

Figure 4: **Left:** SI rates by bot-type group. Floating dots correspond to item means. The adjacent shaded shape represents the smoothed density of responses. **Right:** By-item correlation of SI rates in Speaker 1 (40% bot) and Speaker 2 (60% bot) groups. Dashed line represents unity line.

References: [1] Kraljic & Samuel (2006); [2] Xiang et al. (2020); [3] Schuster & Degen (2019); [4] Pecsock & Aparicio (2024); [5] Wu & Aparicio (*to appear*); [6] Horn (1972); van Tiel et al. (2016); [7] Wu & Aparicio (2024); [8] Wu & Aparicio (2025); [9] Yildirim et al. (2016); [10] Heim et al. (2020); [11] Aparicio & Ronai (2025).