Scalar implicature rates vary within and across adjectival scales

Background. In recent literature, the observation of scalar diversity has generated a lot of interest: lexical scales significantly differ from each other in how likely they are to lead to SI, e.g., *The employee is smart* → *not brilliant* vs. *The movie is funny* → *not hilarious* (Gotzner et al., 2018; Pankratz & van Tiel, 2021; Sun et al., 2018; van Tiel et al., 2016, a.o.). But while studies of scalar diversity have investigated inter-scale variation, not enough attention has been paid to potential variation introduced by the carrier sentences that scales occur in —even though different sentential contexts significantly affect the calculation of the *some but not all* SI (Degen, 2015). Though van Tiel et al. (2016) has conducted a more limited test of carrier sentences, they found no difference among them. In this paper, we carry out the first systematic investigation of the role of sentential context on scalar diversity. Focusing on scales formed by two gradable adjectives, we manipulate the comparison class (CC): whether a noun (e.g., *scientist* vs. *employee*) is likely to have the property described by the scalar adjective (e.g., smartness/brilliance). Our results show a significant effect of CC on the likelihood of SI.

Scale set. We first collected 77 adjectival scales from previous work and then normed them for cancellability and asymmetric entailment. Following de Marneffe and Tonhauser (2019), we conducted two forced-choice experiments (participant N=80): for cancellability, a statement like *X is brilliant... and even smart* had to be judged "Odd" and *X is smart... and even brilliant* "Not odd"; for asymmetric entailment, *X was brilliant... but not smart* had to be judged "Contradictory" and *X was smart... but not brilliant* "Not contradictory". Since our main interest is the effect of CC on SI calculation, the norming studies used proper nouns (*Logan is smart*) or pronouns (*It was tasty*). For a scale to pass the norming, above 60% of the responses needed to be the expected ones for both the *but not* and the *even* test. The resulting scale set that forms the basis of all subsequent experiments is 45 adjectival scales.

Elicitation. To gather CCs, we conducted an elicitation experiment: participants (N=100) saw stronger scalemates (e.g., *brilliant*, *hilarious*) and were instructed to write down a noun that was likely to have that property. For each scale, we selected two nouns: one that occurred with high frequency (henceforth "biased") and one that was very infrequent (≈ 1 count; henceforth "neutral").

Hypothesis 1 (H1): likelihood. The experiment measured the likelihood of the stronger scalar property obtaining with biased vs. neutral nouns. Participants (N=61) saw sentences like *On a 0-100 scale, how likely are {scientists/employees} to be brilliant?* and had to pick a point on a sliding scale. Results revealed significantly higher likelihood ratings with the biased noun ($p < 0.001$, Fig. 3). Gricean accounts take SI to arise via listeners' reasoning about what the speaker could have said, but did not (Grice, 1967; Horn, 1972). Based on this, we can make a prediction for our CC manipulation. With biased nouns, where the stronger adjective was especially likely to be true of the individual in the CC (e.g., *brilliant scientist*), the fact that the speaker chose not to utter it (but use a weaker term) is especially meaningful. H1 thus predicts higher rates of SI calculation for biased than for neutral CCs.

Hypothesis 2 (H2): threshold distance. A competing hypothesis is that SI calculation is modulated by the adjectival threshold distance between the two scalemates given a CC. Following the degree semantics tradition (Kresswell, 1976; Kennedy & McNally, 2005; Kennedy, 2007, a.o.), we assume that a gradable adjective A denotes a relation between a threshold value θ and the degree to which an individual x in the CC C instantiates the adjective property ($[\![A]\!]^C = \lambda\theta\lambda x.\mu_A(x) \geq \theta$). The closer the thresholds of the two scalemates on the relevant adjectival scale (e.g., smartness), the more overlap between the meanings of the two adjectives, as more individuals can be described with

Figure 1: Scales.

both the weaker *and* the stronger adjective . When there is more overlap between the two adjectives, SI calculation leads to greater strengthening of the weaker scalemate, since there are fewer degrees that fall under both *smart* and *not brilliant* —"smart (SI)" covers a smaller interval on the left than the right in Fig. 1. We argue that this situation discourages SI calculation, as the informational state of the listener (more precisely defined below), rarely warrants such dramatic information gain. We conceptualize this as the listener's counterpart of Horn's (1984) R/Q principles: 1) remain faithful to the semantic contribution of the weaker scalemate ("interpret no more than you must", R); 2) strengthen the meaning of the scalemate as much as possible ("interpret as much as you can", Q).

To reflect the fact that there is uncertainty over adjectival thresholds, we cast these ideas in probabilisic terms, and treat adjectival thresholds as probability distributions ranging over degrees of the relevant scale (Lassiter & Goodman, 2013; Qing & Franke, 2014). To obtain θ distributions, we presented participants (N=240) with sentences like *The {scientist/employee} is {smart, possibly brilliant/brilliant}*. The statement involving the weaker adjective was followed by *possibly brilliant* in order to block SI calculation. Participants were asked *On a 0-100 scale, how smart is the employee/scientist?* and responded on a sliding scale. Qualitative predictions are illustrated in Fig. 2. The SI-enriched interpretation of the weaker adjective given the negation of the stronger scalemate ($P(\theta_w|\theta_{\neg s})$) is computed through Bayesian update as shown in (1). Fig. 2A illustrates one possible situation where there is high overlap between the strong and the weak threshold distributions. High overlap between $P(\theta_{smart})$ and $P(\theta_{brilliant})$ results in an SI-enriched distribution for $P(\theta_{smart})$ that has very low overlap with its non-SI counterpart (Cf. 2B). Intuitively, high overlap between $P(\theta_{smart})$ and $P(\theta_{brilliant})$ entails that, prior to SI calculation, states where *x is smart* is highly probable are also states where *x is not brilliant* is unlikely, where $x \in$ CC. SI calculation therefore has the unwelcome consequence of making initially low probability states (i.e., smart and not brilliant) highly probable in the posterior. This results in an SI-enriched meaning that's strongly strengthened and distant from the original semantic contribution of the adjective. Given this, high overlap between $P(\theta_s)$ and P(θ_w) should discourage SI calculation. We operationalize the distance between $P(\theta_s)$ and $P(\theta_w)$ with the difference score D (equations (2-3)), which is a function of the mean (μ) and standard deviations (σ) of the relevant $P(\theta_{w/s})$ given a noun n (or CC) (Toscano & McMurray, 2010). H2 therefore predicts higher SI rates for neutral compared to biased CCs for higher values of D.

SI calculation. We used an inference task to test the likelihood of SI calculation. Participants (N=81) saw sentences such as "Mary: *The {scientist/employee} is smart.*" and were asked *Would you conclude from this that Mary thinks that the {scientist/employee} is not brilliant?* They responded by clicking "Yes" (= SI calculation) or "No" (= no SI calculation). Contra H1, we find lower rates of SI calculation in the biased condition ($p < 0.05$, Fig. 4). The by-item results are also in line with this: likelihood and SI rate are negatively correlated ($r = -0.42$, $p < 0.001$, Fig. 5). To check H2's prediction, we plot the neutral−biased SI rate against the difference scores D, and find the predicted significant positive correlation ($r = 0.36$, $p < 0.02$, Fig. 6). The SI rate data therefore supports H2.

Semantic distance. Though we have spelled out H2 in probabilistic terms, it is also compatible with the non-probabilistic semantic distance proposal of van Tiel et al. (2016) (going back to Horn 1972), who found higher SI rates *across* scales with more semantically distant scalemates. The probabilistic account proposed here explicitly articulates a mechanism (i.e., a trade-off between the listener-oriented Q and R principles) that derives the previously observed correlation between semantic distance and SI calculation within and across scales.

Conclusion. We systematically manipulate the CC for adjectival scales and find that nouns more likely to have the adjectival property lead to fewer SIs. We explain this in terms of threshold distance. Our findings also highlight the methodological importance of controlling carrier sentences. In ongoing work, we explore the implications of H2 for scales where the stronger term is a relative vs. absolute adjective (see also Gotzner et al. (2018) for the relevance of scale structure for scalar diversity).

$$P(\theta_w|\theta_{\neg s}) \propto P(\theta_{\neg s}|\theta_w)P(\theta_w) \quad (1) \qquad d_n = (\mu_{s_n} - \mu_{w_n})/\sigma_{s_n}\sigma_{w_n} \quad (2) \qquad D = d_{n_{neut.}} - d_{n_{bias.}} \quad (3)$$

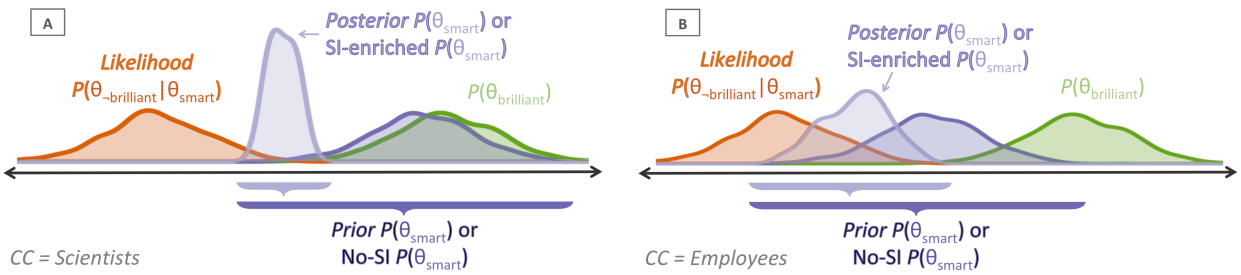

Figure 2: Simulations.

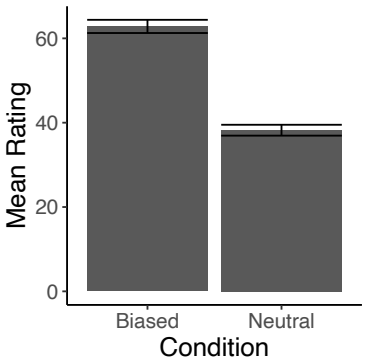

Fig. 3: Likelihood (of strong adj.)

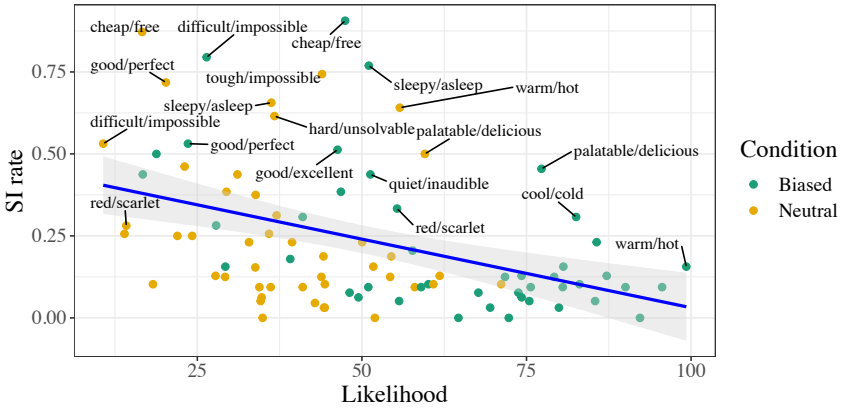

Fig. 5: By-item correlation: SI rates and likelihood

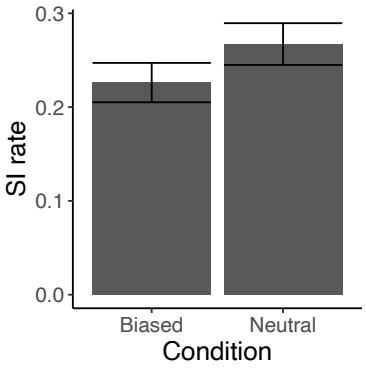

Fig. 4: SI calculation rate

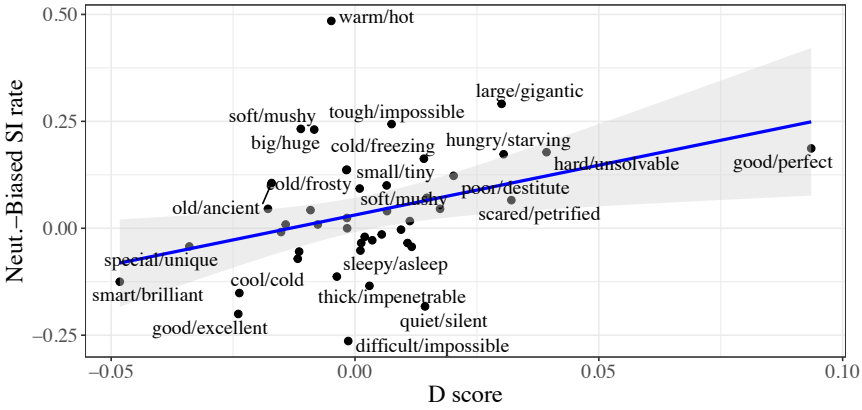

Fig. 6: By-item correlation: SI rate difference and *D* score

Selected references: Degen. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. Semantics and Pragmatics. | Gotzner et al. (2018). Scalar diversity, negative strengthening, and adjectival semantics. Frontiers in Psychology. | Horn. (1972). On the semantic properties of logical operators in English. Ph.D. thesis. | Kennedy. (2007). Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives. Linguistics and Philosophy. | Kennedy & McNally. (2005). Scale Structure and the Semantic Typology of Gradable Predicates. Language | Lassiter & Goodman (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. Proc. of SALT. | Qing & Franke (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. Proc. of SALT. | Toscano & McMurray. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science. | van Tiel et al. (2016). Scalar diversity. Journal of Semantics.