Quantifying scalar diversity: a first look

Eszter Ronai (The University of Chicago) & Ming Xiang (The University of Chicago)
ronai@uchicago.edu

Background. Previous research has revealed that different scalar expressions give rise to scalar inferences (SIs) at different rates; for instance, the SI in (1) arises much more robustly than the one in (2) [1,2,4,6,7]. This variation has been termed *scalar diversity*. However, in previous work, the observation of scalar diversity was based on descriptive statistics, e.g. that SI rates range from 4% to 100% across scales. In this study, we take a first step towards providing a more rigorous measure to quantify scalar diversity, using relative entropy. Our testing ground is 60 scales that represent a better balance across grammatical categories than previous work. In two experiments, we find that while overt exhaustification with *only* and a biasing Question Under Discussion (QUD) both increase SI rates, only the former substantially reduces scalar diversity.

Corpus study. Previous work has focused mostly (70%, e.g. [7]) or entirely (e.g. [4]) on adjectival scales. If our goal is to identify properties of SI that hold across all scales, then we should devote equal attention to scales from other grammatical classes. We thus supplemented existing scale sets [3, 7] with corpus work, conducting the following COCA searches: *X or even Y*; *not just X but Y*; *X but not Y* (for adjectives, verbs, adverbs). Semantic tests for asymmetric entailment and cancellability were used to filter the corpus results. The resulting final set consists of 60 lexical scales.

Exp. 1 (participant N=80) used an inference task to investigate the likelihood of SI calculation [7]. Participants saw sentences such as "Mary: *The student is intelligent.*" and were asked the question "Would you conclude from this that Mary thinks the student is not brilliant?". They responded by clicking "Yes" (= SI calculation) or "No" (= no SI calculation). In a between-participants manipulation, we also tested sentences that contained the focus particle *only*: *The student is only intelligent.*

Exp. 2 (participant N=40) added a within-participants two-condition QUD manipulation to the inference task. Mary's statement was preceded by a question that contained either the stronger or the weaker scalar: "Sue: *Is the student brilliant/intelligent?*"; "Mary: *She is intelligent*".

Entropy measure. To quantify scalar diversity, we used relative entropy —see the equation in (3). Specifically, we treated the normalized % of "Yes" responses (i.e. the SI rates) across different scales as a probability distribution. We tested whether a given SI rate provides enough information to identify the scale that it came from. In our calculations, we compared each set of SI rates (two conditions from Exp. 1 and 2 each) to the uniform distribution. The uniform distribution represents a scenario where each scale leads to the same SI rate —here, the % of "Yes" responses gives 0 information about the identity of the scale that it came from, and consequently scales cannot be identified by their associated SI rates. Using the measure of relative entropy (in our case, the entropy of the uniform distribution minus the entropy of the given SI rates), we are able to quantify how "diverse" the SI rates are that we obtained in our experiments.

Results and discussion. In Exp. 1, we found that sentences with *only* resulted in significantly higher rates of "Yes" responses than sentences without *only* (bare SI) ($p < 0.001$) —see left facet in Fig. 1. In Exp. 2, we found that significantly more SIs were derived when the preceding question contained the stronger scalar term than when it contained the weaker one ($p < 0.001$) (replicating [5]) —see right facet in Fig. 1. Thus, overt exhaustification with *only* and a pragmatic QUD manipulation both increase inference calculation rates. The relative entropy measures (as compared to the uniform distribution) are as follows: bare SI (Exp. 1)=0.466; sentences with *only* (Exp. 1)=0.046; weak-scalar QUD condition (Exp. 2)=0.404; strong-scalar QUD condition (Exp. 2)=0.137. This suggests that the bare SI condition, which was in previous literature impressionistically taken to show scalar diversity, does indeed substantially differ from the uniform distribution, and so does the weak-scalar QUD condition. When sentences include the focus particle *only*, however, scalar diversity is greatly lessened; the entropy of the *only* results barely differed from the uniform distribution. Lastly, SI rates under the strong-scalar QUD condition appear to fall somewhere in the middle. Altogether, it seems that encoding the implicature meaning in the semantics (with *only*) substantially reduces scalar diversity, but a pragmatic QUD manipulation (in either condition) does not.

Conclusion. We replicate scalar diversity on 60 scales. Looking at these SI rates, as well as overt exhaustification (*only*) and a QUD manipulation, we offer a first attempt at quantifying how "diverse" the inference rates are. We find that only over exhaustification reduces scalar diversity.

(1) Mary ate some of the cookies. → SI: Mary ate some, but not all, of the cookies.

(2) The student is intelligent. → SI: The student is intelligent, but not brilliant.

(3) Let $p(x)$ and $q(x)$ be probability mass functions over the same set $\mathcal{X}$. The relative entropy of $p(x)$ with respect to $q(x)$ is given by

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

$p(x)$ is the observed % of "Yes" responses across scales. $q(x) = 1/60$ is the uniform probability mass function over the 60 scales.

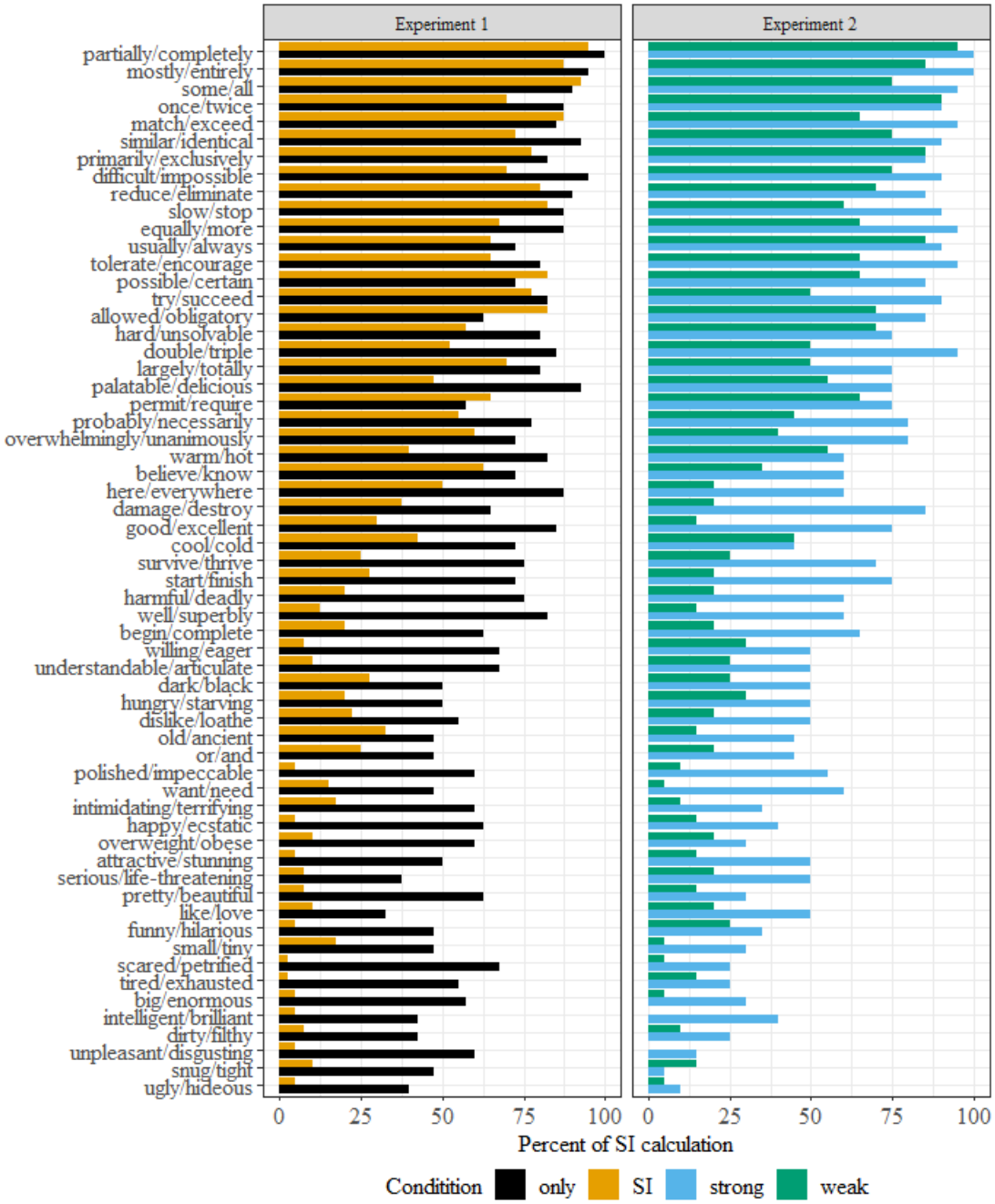

Figure 1: Results of Experiment 1 (left facet) and Experiment 2 (right facet)

References. [1] Beltrama & Xiang (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. Proc. of SuB 17. | [2] Doran et al. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. Language. | [3] Marneffe & Tonhauser (2019). Inferring meaning from indirect answers to polar questions. Questions in Discourse. | [4] Gotzner et al. (2018). Scalar Diversity, Negative Strengthening, and Adjectival Semantics. Frontiers. | [5] Ronai & Xiang (2021). Exploring the connection between Question Under Discussion and scalar diversity. Proc. of the LSA. | [6] Sun et al. (2018). A link between local enrichment and scalar diversity. Frontiers. | [7] van Tiel et al. (2016). Scalar diversity. J. of Semantics.