Explaining scalar diversity with predictability of alternatives

Eszter Ronai & Ming Xiang (The University of Chicago)

Previous research has revealed that different scalar expressions give rise to scalar inferences (SIs) at different rates. This variation has been termed *scalar diversity*. In this study, we investigate the role of the predictability of scalar alternatives in explaining this variation in SI rates. We first collect 60 different pairs of scalar expressions, which represent a better balance across grammatical categories than previous work on scalar diversity. Investigating these 60 different scales, we first successfully replicate the basic finding of scalar diversity. We then propose a new factor to capture observed variation in SI rates: a language production-based metric of how predictable the stronger alternative (*all*, or *brilliant*) is. We operationalize this factor as cloze predictability and experimentally show that it can indeed explain some of the observed variance.

Background. In calculating SI, hearers infer messages beyond what is literally, explicitly said by the speaker. In doing so, they also reason about what is not said: the stronger alternative—*all* in (1) and *brilliant* in (2).

(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.

Recent work has found considerable variation across different scales in the rates of SI calculation; for instance, the SI in (1) arises much more robustly than the one in (2) (van Tiel et al. 2016; see also Doran et al. 2012; Beltrama & Xiang 2013). The question arises, then, how to capture this observed variation: can we identify some properties of different scales that influence how robustly they lead to SI calculation? Van Tiel et al. (2016) found the distinctness of the stronger scalar term, while Sun et al. (2018) found local enrichability to be such a property. However, there is still a lot of variance unaccounted for in the empirical results.

Hypothesis. Our hypothesis is that how likely an SI is to arise from a given scale can (in part) be explained by how predictable a stronger alternative is, given the weaker scalar. We operationalize predictability as cloze probability, commonly used to measure the predictions the parser makes in language comprehension. Cloze probability can be defined as the probability of a target word (here, the specific stronger scalar) completing a particular sentence frame —a measure that indexes how expected a word is in a context (Taylor, 1953; see also i.a. Kutas & Hillyard, 1984). The intuition is that there may be differences across scales in how strongly the weaker scalar evokes a specific stronger alternative. For instance, for the weaker scalar *some*, it might be the case that the stronger alternative *all* is always evoked (in a sentential context); for a weaker scalar such as *intelligent*, a number of competing alternatives may be activated, such as *brilliant, hardworking, kind, crafty*, etc. Our measure is related to alternative availability (van Tiel et al.; see Discussion below for details), but is operationalized in a novel way, using a language production task in a discourse context.

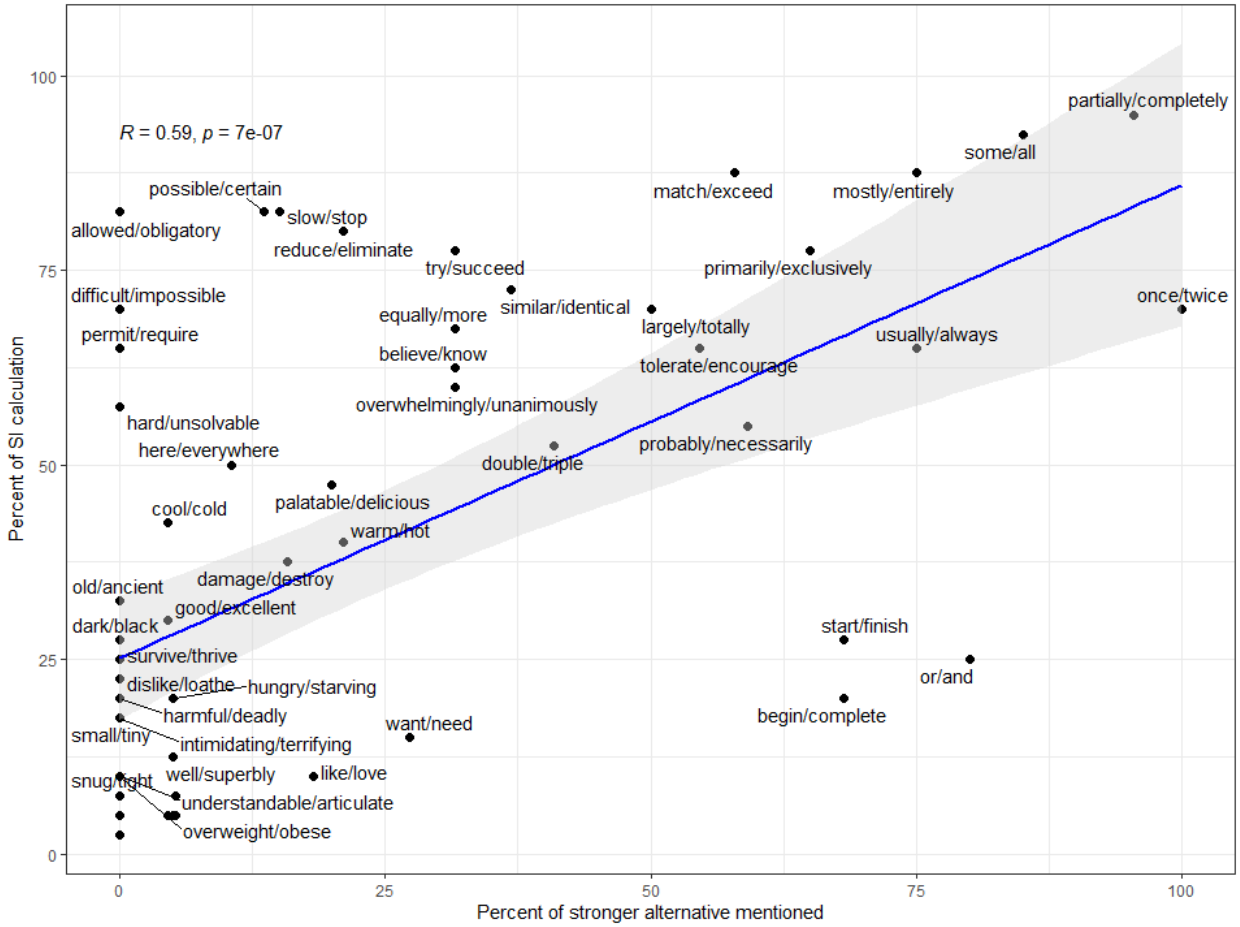
Corpus study. In previous work on scalar diversity, the set of scales studied included mostly (70%, e.g. van Tiel et al.) or entirely (e.g. Gotzner et al.) adjectival scales. If our goal is to identify properties of SI that hold generally, across all scales, then we should devote equal attention to scales from other grammatical classes. For this reason, we modified existing scale sets (from van Tiel et al. 2016; de Marneffe & Tonhauser, 2019) and supplemented them with corpus work. Specifically, we conducted the following corpus searches in the Corpus of Contemporary American English: *X or even Y*; *not just X but Y*; *X but not Y*. These searches were conducted for adjectives, verbs, and adverbs. To ensure that X and Y formed a lexical scale (in the sense of Horn, 1972), semantic tests for asymmetric entailment and cancellability were used to filter the results from the corpus. The resulting final set consists of 60 lexical scales, which were tested in both Experiment 1 and 2.

Experiment 1. 40 native speakers of American English participated in a replication of van Tiel et al. (2016), which used an inference task to investigate the likelihood of SI calculation. Participants were presented with a sentence 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" (indicating that the SI was calculated) or "No" (indicating that the SI was not calculated).

Experiment 2 was a modified cloze task (participant N=61). Participants were presented with a dialogue context, and instructed to complete the answer in that dialogue with the first word that comes to mind (3):

(3) Sue: The student is intelligent.
Mary: So you mean she's not _____.

The experiment included conditions that addressed different research questions and are not discussed here—due to counterbalancing reasons, we ended up collecting 19-22 completions per scale. Our critical prediction is that if alternative predictability captures scalar diversity, then the more frequently the stronger alternative is mentioned in the cloze task (Exp. 2), the higher the SI rate should be for that scale (Exp. 1).



Results and discussion. Exp. 1 successfully replicated the scalar diversity effect —see the y axis of the figure above. The x axis of the figure above shows the results of Exp. 2; in the coding of the results, synonyms of the stronger scalar were also counted. As can be seen in the figure, the predictability of the stronger alternative, i.e. the % likelihood of the stronger alternative being mentioned in the cloze task, is significantly correlated with SI rate ($p < 0.001$). The results of the cloze task also remain a significant predictor of SI rate in further statistical analyses, which use mixed effects logistic regression models and take into account other known factors that explain scalar diversity, e.g. boundedness. That is, confirming our hypothesis, we find that scalar diversity is predicted by how strongly a particular weaker scalar evokes the relevant stronger alternative. Our proposal of alternative predictability is closely related to van Tiel et al.'s hypothesis that the availability of the stronger alternative should predict scalar diversity. This is because for SI to arise, it has to be the case that the speaker could have actually considered using the stronger scalar term instead of the weaker one they uttered. Van Tiel et al. operationalized availability via four different metrics; however, none of them were found to predict scalar diversity.

Conclusion. We successfully replicated scalar diversity on a set of lexical scales drawn from a diverse

range of grammatical categories. Crucially, we showed that a new, production-based measure of alternative predictability can capture scalar diversity —also in line with van Tiel et al.'s proposal of alternative availability.
Ongoing and future work. In the presentation, we will discuss further work on uncovering factors that predict some of the observed variance in SI rates. In particular, we will show that empirically-collected posterior degree estimates also play a role in explaining scalar diversity. For instance, the more different the world states that a weak and a strong scalar term are taken to describe (i.e. the more different *The student is intelligent* from *The student is brilliant* on an underlying degree scale), the higher the SI rate from that scale —also lending support to van Tiel et al.'s proposal of semantic distance. Lastly, we also show that the more similar the world states that the weak term and the negated strong term are taken to describe (i.e. the more similar *The student is intelligent* and *The student is not brilliant*), the higher the SI rate.

Selected references. Beltrama & Xiang (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. Proceedings of SuB 17. | Doran et al. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. Language. | Gotzner et al. (2018). Scalar diversity, negative strengthening, and adjectival semantics. Frontiers in Psychology. | Sun et al. (2018). A link between local enrichment and scalar diversity. Frontiers in Psychology. | van Tiel et al. (2016). Scalar diversity. Journal of Semantics.